Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Yaz* 2021
Summer 2021
Cilt: 12- Sayı: 2
Volume: 12- Issue: 2

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: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

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

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

Hakem Kurulu / Referee Board

Hakem Kurulu / Referee Board

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2021, Sayı: 12-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2021, Issue: 12-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 71-96

Multilevel Effects of Student Qualifications and In-Classroom Variables on Science Achievement *

Sıdıka AKYÜZ ARU ** Mustafa KALE ***

Abstract

This research aims to determine the effects of student qualifications and some in-classroom variables related to the school teaching process on the TIMSS science achievement of 4th-grade students in Turkey. It was also aimed to determine the variables that contributed the most to explaining the achievement differences between schools at the student and classroom levels in this study, which was conducted with a causal comparison pattern. The sample of the study consists of 6378 students and classroom teachers of these students. The data of this group was analyzed using the Two-Level Hierarchical Linear Model (HLM). The effects of absenteeism, not having breakfast, use of technology in school, use of technology outside school and home on science achievement scores were found to be statistically significant as a result of HLM analysis. Teachers' perceptions of the inadequacy of the school's facilities and resources, giving feedback on homework, discussing homework in the classroom, and explaining the answers given by the students in the classroom have significant effects on science achievement at the classroom level. These results are related to students in improving the academic performance of primary school students and reveal the importance of a number of psychological and physical characteristics that may affect the teaching process positively or negatively.

Key Words: in-classroom variables, student qualifications, science achievement, TIMSS, HLM.

INTRODUCTION

Keeping the school alive for its purposes in the 21st century, ensuring the happiness and satisfaction of the parties stand before us as an equation with many variables that seems very difficult to achieve (Özdemir, 2013). Therefore, there is a need to re-address education systems at a level that can meet these functions. Accordingly, all the elements that make up the education system are subjected to the evaluation process in order to describe the current situation, to reveal deficiencies and needs, and to determine the activities to be carried out in the future (Bilican-Demir, 2014).

At this point, the information to be provided by the evaluation activity to be carried out is important. Reddy (2005) states that the most efficient way to evaluate countries' education systems is to evaluate the outcomes, which is one of the elements of the system, and emphasizes that the most realistic approach to this issue is international comparisons.

Turkey has been participating in the International Mathematics and Science Study (TIMSS), which is one of the international large-scale test applications since 1999 in this context. TIMSS provides the opportunity to make the necessary changes in light of the scientific data obtained by comparing the Turkish education system with the other education systems at the international level. In addition, TIMSS provides an opportunity to classify the students participating in the application according to their competence levels in line with their achievement scores, to determine their abilities, and to evaluate them.

* This research was produced from the doctoral dissertation study titled "Investigation of Variables Affecting Science and Mathematics Achievement of 4th Grade Students: Multilevel Analysis of TIMSS Data".

** PhD., Ministry of Education, Ankara-Turkey, SIDIKAARU@gmail.com, ORCID ID: 0000-0002-5991-2288
*** Assoc. Prof., Gazi University, Faculty of Education, Ankara-Turkey, mkale@gazi.edu.tr, ORCID ID: 0000-0002-3727-1475

To cite this article:
Akyüz-Aru, S., & Kale, M. (2021). Multilevel effects of student qualifications and in-classroom variables on science achievement. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 71-96. doi: 10.21031/epod.756083

Received: 22.06.2020
Accepted: 30.03.2021

It is seen that the percentage of students who remain below the low level in the field of science (18%) is approximately 3.5 times the median value of TIMSS when the proficiency levels of Turkish students are examined in the result reports (1999-2015) (Karip, 2017). This result shows that 225 thousand 4th grade students pass from primary school to secondary school with a performance below the low level in the field of science, that is, without basic skills (Karip, 2017). The fact that almost half of Turkish students (n = 3250) are at low and lower levels of competence shows that they have difficulties in implementing the basic information they have learned, adapting this information to the problems they encounter, and even remembering (Yücel & Karadağ, 2016), in other words. Such problems directly affect academic achievement (Mullis, Martin, Gonzalez, & Kennedy, 2003). The results of the studies conducted on TIMSS data in different years (Büyüköztürk, Çakan, Tan, & Atar, 2014; Martin et al., 2000; Olson, Martin, & Mullis, 2008; Özden, 2007; Uzun, Bütüner, & Yiğit, 2010) show that the science achievement of Turkish students is lower than the mean of overall achievement.

Effects of Student Qualifications and In-Classroom Variables on Achievement

Meta-analysis studies examining the relationships between student qualifications and achievement (Hattie, 2009; Marzano, 2003) show that factors directly related to the student have a high effect on academic achievement. It was found that the majority of the variance in student achievement was explained by features at the student level in a study by Mohammadpour, Shekarchizadeh, and Kalantarrashidi (2015) examining the characteristics of students, schools, and countries affecting the TIMSS 2007 science scores of 8th-grade students from 29 countries.

It is seen that many student qualities are discussed in the literature. Kaya (2008) discussed TIMSS 2003, with the variables of gender, self-confidence, and home resources at the student level in relation to student achievements. Aydın (2015) defined student-level data as student affective characteristics and student characteristics. İpekçioğlu-Önal (2015) identified student-level factors as gender, educational resources used at home, family participation, homework, and bullying. Sarı, Arıkan, and Yıldızlı (2017) examined affective characteristics (self-efficacy, attitude, and learning value), resources at home, belonging to the school, bullying, and teaching activities at the student level.

However, it is stated in different studies that student characteristics such as absenteeism, nutrition, and technology use are not adequately examined in research even though the power to explain the variability in achievement is high (Asigbee, Whitney, & Peterson, 2018; Garcia & Weiss, 2018; İsmail & Awang, 2008; Khalid, 2017; Kolasa, Díaz, & Duffrin, 2018; Liouaeddine, Bijou, &, Naji, 2017).

Klem and Connell (2004) and Ackerman (2013) emphasized that absenteeism, students' participation in school activities, and regular attendance are factors that directly and positively affect students' motivation; students who continuously participate in school processes develop positive attitudes towards lessons and their homework performance increases. In addition, it is another factor emphasized in studies where students who are oppressed under challenging school conditions, adverse climate, and heavy education programs tend not to go to school and even to leave school (Akey, 2006; Doğan, 2014). Likewise, another factor that closely affects students' academic achievement is nutrition. The Organisation for Economic Co-operation and Development (OECD) reports show that many students struggle with hunger and insomnia to pay attention to the lesson even in the most developed countries (OECD, 2015). Clinton (2013) found that children who are advantageous in nutrition and physical activity are more open to effective learning compared to those who are disadvantaged in this regard. However, Clinton, Rensford, and Willing (2007) state that there is no direct research result in the literature that nutrition increases academic achievement to very high levels. Therefore, nutrition is only one of the factors that are thought to affect achievement.

Another factor ignored at the student level in the studies is students' use of technology. It is known that especially children begin to spend most of their time interacting with new digital media, such as e-books, tablets, and smartphones, with the development of technology (Lieberman, Bates, & So, 2009). The way that most children who spend part of their daily lives in school are affected by technology may be reflected positively or negatively on their academic achievement. This issue is worth investigating even though parents think that these can be effective and complementary in the education of early age children if educational applications that can be used, for example, on phones and other digital media tools, are used in the right direction (Chiong & Shuler; as cited in Hooper, Mullis, & Martin, 2013). Therefore, nutrition, absenteeism, and technology use at the student level of the current study were included in order to eliminate the deficiency in the literature and to see its effects on explaining the variance in student achievements.

Hox (1995) states that the qualities of the students, which are the basis of the measurement in the field of education, are affected by the characteristics of the school and classroom in which the student is studying. It is inevitable that useful learning is affected by teaching activities that are among the classroom environment and in-classroom variables since most of the teaching and learning activities take place in the classroom (Hooper et al., 2013; Nilsen, Gustafsson, & Blömöke, 2016). Hattie (2009) states in the meta-analysis study that in-classroom variables (in-classroom instructional interactions, school resources in classroom materials and lessons, in-classroom evaluation methods, etc.) significantly affect student achievement but that not enough work is done at the classroom level.

It is seen that there are studies on lesson tools, student, and parent characteristics (Aydın, 2015; Çavdar, 2015; Erşan, 2016; Korkmaz, 2012) and teaching method techniques and learning environments used in and outside the classroom, national teacher training policies and teacher education, teachers' experience, teacher qualifications, attitudes towards teacher training, the structure of school management, and leadership understanding when the studies conducted on this subject in Turkey are examined (Aktaş, 2011; Atar, 2014; Sezer, 2016). In addition, there are also studies examining the effect of different teaching methods and techniques related to student achievement at the 4th grade level of primary school (Ayvaz, 2010; Güngör, 2014; Kılıç-Özün, 2010; Selçuk, 2015), in which the results of TIMSS are comparatively examined between countries (Akkuş, 2014), the relationship between international exams and educational policies and equal opportunities in education is examined (Çelebi, Güner, Taşçı-Kaya, & Korumaz, 2014).

One of the critical factors in the quality of education that directly affects achievement is the comprehensiveness and quality of the resources available at school and for the lessons (Lee & Barro, 2001; Lee & Zuze, 2011). The results of TIMSS studies conducted to date show that the students of the teachers who do not have resource shortage in the lessons are generally more successful (Hooper et al., 2013). The concept of *in-classroom variables* was used by Blömeke, Olsen and Suhl (2016) in their study on the relationship between teacher quality and teaching quality with student achievement. Teachers' qualifications, knowledge-skilling levels, and perceptions are important factors in increasing student achievement. Discussions in classroom processes, explanations of the answers given, and homework are elements that should be planned in advance. The subject of homework is an important factor researched by many researchers to observe its impact on achievement. Trautwein (2007) explains the existence of studies on the frequency of homework and the time spent on homework and states that there is a need for studies on the effect of the effective use of homework in classroom processes on academic achievement.

Many of the studies investigating the relationships of student achievements with determinants at different levels such as students, schools, classrooms, and teachers in the literature (Acar, 2013; Aktaş, 2011; Akyüz, 2006; Atar, 2014; Ekinci-Vural, 2012; Fullarton, Lokan, Lamb, & Ainley, 2003; İpekcioglu-Önal, 2015; Mohammadpour & Abdul Ghafar, 2014; Sezer, 2016; Stemler, 2001; Taştekinoğlu, 2014; Yaman, 2004) ignore the effect of classroom-level characteristics on achievement.

In addition, it is seen that achievement is examined only in the context of student-derived factors (Kunuk, 2015), student-derived and teacher-derived factors (Akyüz, 2006; İpekcioglu-Önal, 2015; Kaya, 2008), and teacher-derived factors (Aktaş, 2011; Atar, 2014; Sezer, 2016; Yaman, 2004), only school-derived factors (Stemler, 2001) or school and student-derived factors (Acar, 2013; Aydın, 2015; Fullarton et al., 2003; Lamb & Fullarton, 2002; van den Broeck, Opdenakker, & van Damme, 2005; Yatağan, 2014) when the studies conducted at home and abroad and examining student, teacher, and school characteristics on the results of large-scale tests such as TIMSS (Akkuş, 2014; Aktaş, 2011; Akyüz, 2006; Atar, 2014; Aydın, 2015; Çavdar, 2015; Erşan, 2016; İpekcioglu-Önal, 2015; Kaya, 2008; Korkmaz, 2012; Sevgi, 2009; Stemler, 2001; Yatağan, 2014;) are examined. The relationships of achievement only with the interaction of variables at student and school levels were investigated in some studies conducted abroad (Blömeke, Suhl, & Kaiser, 2011; Hooper et al., 2013; Kyriakides, 2006; Martin, Mullis, Foy et al., 2016; OECD, 2013).

However, it is necessary to investigate in-depth and in relation to each other all the variables that may affect achievement and take place at different levels in future studies. The fact that achievement is a goal reached at the end of the processes where mental activities are effective is proof of this necessity (Nilsen et al., 2016). The versatile structure of the mind suggests that achievement is too complex a concept to be measured only by standard tests, and therefore needs to be investigated in depth by examining in different contexts. For example, Creemers and Kyriakides (2006) state that this stratified sampling structure of large-scale tests, in which students are clustered within classrooms, classrooms within schools and schools within countries, may change in relation to student, classroom, and school characteristics and will be affected by the conditions contained in these contexts. It is therefore important to conduct multi-faceted research in which achievement is assessed in the circumstances in which these conditions arise.

It is understood that there is a gap in this regard considering that *nutrition, absenteeism, and technology use* from student qualifications and *sources, classroom discussions, and homework* from in-classroom variables are not handled together in the studies conducted in Turkey (Akkuş, 2014; Aktaş, 2011; Atar, 2014; Aydın, 2015; Ayvaz, 2010; Çavdar, 2015; Çelebi et al., 2014; Erşan, 2016; Güngör, 2014; Kılıç-Özün, 2010; Korkmaz, 2012; Selçuk, 2015; Sezer, 2016). Therefore, this study focuses on the effect of these variables selected at student and classroom levels together on students' science achievement.

Objective of Research

This research aims to estimate the extent to which these variables affect the TIMSS science achievement of 4th-grade students by combining different variables at student and classroom levels with the help of the Hierarchical Linear Model. In addition, the study also aims to determine the student qualifications and in-classroom variables that most explain the inter-school achievement variables. Answers to the following questions were sought for these purposes.

Research questions:

1. Are there significant differences between the classrooms in terms of students' science achievement?
2. Do the science achievement scores of the students differ according to the in-classroom variables discussed at the grade level? What are the classroom level variables that explain this difference if there is a difference? How much of the variance in science achievement scores are explained by variables with significant effects?
3. Do the science achievement scores of the students differ according to the student qualifications discussed at the student level? What are the student-level variables that explain this difference if there is a difference? How much of the variance in science achievement scores are explained by student-level variables with significant effects?

METHOD

A causal comparison pattern, one of the quantitative research methods, is used since the study aims to determine and compare the variables affecting the science achievement measured in TIMSS 2015 among the various student and classroom characteristics of the 4th-grade students discussed within the scope of TIMSS 2015 application in this study. Causal comparison studies aim to determine the causes of a situation or differences between groups, what is effective in the formation of this situation, in other words, the causal variables affecting the variable related to the result or the results of the effect without any intervention on the participants and conditions (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2011; Fraenkel, Wallen, & Hyun, 2012).

Sample

A total of 6456 4th grade students from 242 schools and classroom teachers of students from Turkey (*n* = 249) participated in TIMSS 2015 application (LaRoche & Foy, 2016). A two-stage stratified sampling method was used in TIMSS 2015 sample selection. Schools in the first stage and random classrooms in the second stage were selected from these schools.

This study was carried out with data on all 4th-grade students participating in TIMSS 2015 and classroom teachers of these students (*n* = 249). However, it was observed that there were deficiencies in the data obtained from the sample in the data file. Therefore, Missing Value Analysis was performed for missing values in order to finalize the sampling.

It was found as a result of this analysis that Little's MCAR test was significant (< .05) and the existing lost data in the data file showed a systematic distribution. The listwise deletion method can be applied if the lost data is below 5% in such cases (Garson, 2008). Accordingly, the loss rate in each variable was examined for each student in the data set, and the listwise deletion method was not preferred because it was more than 5%.

One of the alternatives to addressing loss values that are over 5% and distributed systematically is to make predictions of lost values/assign an approximate value, also known as *imputation*. "This process can only be performed for quantitative data" (Çokluk, Şekercioğlu, & Büyüköztürk, 2018, p. 11). The three most common methods of performing these operations are using historical information, assigning mean values, and regression (Çokluk et al., 2018; Mertler & Vannatta, 2005; Tabachnick & Fidell, 2001). "Assigning mean value in these cases is the best prediction method if the researcher has not been working on research for a long time and has no other information" (Çokluk et al., 2018, p. 11). Therefore, this method was preferred, the mean was calculated by using the obtained data, and these means were assigned to the variables containing lost values.

In the last case, 6348 students and 241 classroom teachers constituted the sample of the present study. In addition, the weighting values of the students and teachers in the TIMSS 2015 data file were used in order to ensure an equal representation of all students and teachers in the selected sample in the study.

Data Collection Tools

Science achievement test

The science achievement test consists of items half scored as multiple choice and the other half scored as multi-category. Multiple-choice items have four options and one correct answer. The correct answer for each multiple-choice item is 1 point. Incorrect answers do not affect the correct answers. Students create their own answers in the items scored as multi-category. Students make explanations, verbally or numerically support their answers, draw shapes, or use data in this type of question. Items scored as multi-category are evaluated with scoring guidelines developed for each item (Martin, Mullis, & Foy, 2013). These scoring guidelines contain the basic characteristics of an appropriate and complete answer for each item. The guidelines focus on the evidence of the type of behavior that the item assesses. Student responses that are partially or completely correct are clearly defined in the manual and scored as 0-1-2 or 3. In addition, the possible different student responses in the guide also direct the experts. Only the skills required by the evaluated subject are focused on, not the writing skills of the students, while scoring items scored as multi-category (Martin et al., 2013).

Items are distributed into learning areas as 45% life sciences, 35% physical sciences, and 20% earth sciences. In addition, the items show cognitive field distribution as 40% knowing, 40% as applying, and 20% as reasoning (Martin et al., 2013).

Plausible values: The science questions in TIMSS consist of 14 different test booklets. An item pattern has been developed so that a common question can be found in both test booklets. Therefore, it is not possible to test each student on the same items. The science achievement score of each student is predicted as if the student answered all items. It is the range or distribution of possible values predicted for the competence of each student rather than each observed score at this point. Five possible values (plausible value, PVSCI1-5) were reported for each student's science achievement score in the TIMSS data file (Martin, Mullis, Hooper et al., 2016). These possible values obtained as a result of the science achievement test application were used as indicators of TIMSS science achievement, which is the result variable of the current research. The HLM program simultaneously incorporates these five possible values into the analysis by assigning multiple data and assigns a mean value. Therefore, analyses were performed by averaging the possible values in the PV1SCI-PV5SCI range in relation to each student's science test within the scope of the research.

Teacher questionnaire

All variables in the classroom level of the research were obtained by teacher questionnaire. Ten out of 21 items in this questionnaire were related to science and were filled in by the classroom teacher. The items in the teacher questionnaire data file were determined as the variables Perception of the importance of school in academic achievement, perception of safe and regular school structure, problems related to school facilities and resources, difficulties encountered, teaching limited to student needs, giving feedback to students' science homework, discussing science homework in the classroom, checking science homework, the importance given to research, lesson-day life connection, explaining the answers in the classroom, using interesting materials, completing challenging activities, classroom discussions, new content-present content connection, deciding on problem-solving periods, and explaining their thoughts in line with the relevant literature.

Student questionnaire

The variables at the student level were obtained by student questionnaire. This questionnaire, which was filled in by the students, consists of 10 items regarding students' home and school lives, their perceptions about themselves, attitudes towards mathematics and science lessons, homework and extracurricular activities, computer use, resources related to home learning, and general personal information (Hooper et al., 2013). The items in the TIMSS 2015 student questionnaire data file were determined as gender, absenteeism, nutrition, use of technology at home, use of technology in school, and use of technology in other places variables in line with the relevant literature.

Reliability of measurements: TIMSS 2015 student and teacher surveys include items in which Likert-type grading is used to measure the characteristics thought to be related to science achievement. Analyses based on Item Response Theory were performed by TIMSS experts using ConQuest 2.0 software, and measurements of the structure to be measured were obtained based on the responses given to these questionnaire items. These measurements were determined as a mean of 10 and a standard deviation of two for each structure (Martin, Mullis, Hooper et al., 2016). The scale items determined within the scope of the current research were examined by the TIMSS technical team with the Rasch Partial Credit Model within the framework of Item Response Theory.

Reliability coefficients for each scale were calculated for each country, and principal components analysis of the scale items was performed as evidence that the scales provide comparable measurements between countries. The reports presented as a result of the analyses conducted in this direction showed that the TIMSS 2015 scales were generally at an acceptable level, and Cronbach's Alpha values were higher than .70. The values related to the reliability of the scales used in the current research are given in Table 1 (Martin, Mullis, Hooper et al., 2016).

Data Analysis

The two-level hierarchical linear modeling (HLM) method was used in the analysis of the data of the study. The first stage of HLM is a preliminary analysis. Accordingly, the student-level variables obtained from the student questionnaire and the classroom level variables obtained from the teacher questionnaire were arranged in accordance with the purpose of the study by applying the following procedures. The assumptions of HLM were tested in the second stage. Data were analyzed by establishing HLM models at the last stage.

Preliminary analyses

Data editing: The original codes determined for the variables in the data set were re-coded as *X* for the variables of this research, *W* for the student level variables, and to express the classroom level variables. Some items in the student and teacher questionnaire were removed from the data in line with the purpose of the research and the relevant literature. The items used in the research are the items determined at student and classroom levels in Table 1. Items of variables with index scores were deleted from the data.

Correlation between variables: Independent variables that are not related to the dependent variable were checked. Correlation values ranged from -.08 to .29 (at .01 significance level). Therefore, the variable was not deleted from the data set.

Multicollinearity: The correlations of the independent variables in the student level were examined, and it was checked whether there was multicollinearity. The fact that this relationship level is .80 and above indicates that this problem may occur, while the fact that it is .90 and above is important evidence for the multicollinearity problem (Tabachnick & Fidell, 2001). Correlation values obtained in this direction ranged from -.07 to .34 (at .01 significance level). No variables were deleted from the file according to the results.

Missing value analysis: This section is described in the “sample” section.

Detecting and removing outliers: The differentiation of any subject from the rest of the sample is the basis for the outlier in scientific research. The values of the continuous variables were converted to Z scores, and it was checked whether there was a value excluded by ± 4 points in the present study (Mertler & Vannatta, 2005; Tabachnick & Fidell, 2001). The sample consisted of 6378 students and 241 teachers as a result of removing the outliers from the data set.

Exploratory analysis: Exploratory analysis was performed for classroom-level variables. Exploratory analysis is one of the options of the HLM program; it is a basis for deciding which variables are appropriate to include in the model. If the absolute value of t obtained in the analysis is greater than 1, the relevant variable can be included in the analysis (Raudenbush & Bryk, 2002). Fifteen variables related to science achievement were examined simultaneously, and the t values of 11 variables were found to be significant (ranging from -1.23 to 10.26) in this analysis. All variables with significant *t* values were included in the model. Variables with an insignificant *t* value [checking homework in the classroom (0.57), lesson-day life connection (0.96), completing challenging activities (-0.31), explaining thoughts during the lesson (-0.84)] were excluded from the analysis. There were six variables at the student level and 11 variables at the classroom level as a result of the preliminary analysis.

HLM analysis

HLM is a multi-level regression technique that performs the necessary analyses in accordance with the structure of hierarchical (gradual) data obtained especially in the field of education and includes intertwined random effects (Raudenbush & Bryk, 2002). The vast majority of the data obtained in social sciences are hierarchical due to the sampling structure or sampling techniques. Students exhibit a hierarchical structure to be clustered in classrooms, classrooms in schools, schools in regions, and regions in countries in the TIMSS application, which is the subject of the current study. The hierarchical Linear Model investigates the relationships between the hierarchical levels of simultaneously grouped data and thus makes it more efficient in calculating the difference between variables at levels unlike single-level analysis methods (Raudenbush & Bryk, 2002). It is recommended to use multi-level models for data analysis in studies where data are obtained from different levels, such as TIMSS, in the literature on this subject (Heck & Thomas, 2009; Hox, 2002;Raudenbush & Bryk, 2002). Hox (2002) states that the application of single-level models for data analysis in such studies will cause statistical and conceptual problems.

Single-level analysis methods require the assumptions of independence of observations and homoscedasticity to be met. These assumptions may be violated in the data obtained from large samples. Ozborne (2002) states that the data obtained for different groups in a hierarchical pattern tend to be more similar to each other at the level they are at. For example, students in a certain classroom are more similar to each other because they share the same opportunities compared to students in different classrooms. It is impossible in this case for the observations obtained from the students in the same unit to be completely independent of each other. Therefore, it would be more accurate to use multi-level models in the analysis of data in an intertwined structure in order not to violate the assumption of independence of observations. Another important issue is the violation of the assumption of homoscedasticity (Hox, 2010). The other classroom may be heterogeneous while one classroom shows a homogeneous structure in large samples. Multi-level models allow the intra-group and inter-group variance of the dependent variable to be calculated; therefore, it is possible to understand the effects of the levels.

In addition, the use of single-level analysis methods in the analysis of hierarchical data may cause the standard errors of regression coefficient predictions to be calculated smaller than they should be. This leads to an overestimation of the significance levels of predicted regression coefficients (Raudenbush & Bryk, 2002). This situation can be eliminated by including a random effect coefficient (u_{qj}) at each level in multi-level models. Thus, standard errors can be accurately predicted considering the variability in random effects. This is another advantage of the multi-level model.

"Hierarchical Linear Model was preferred in the analysis of data due to the advantages it provides compared to single-level models on different subjects explained" (Raudenbush & Bryk, 2002, pp. 3-6), and it is in accordance with the data structure of the current research as a result (Hox, 1995).

Determination of levels in HLM: Determining the level to be addressed in the examination of the relationships between variables is a stage that is considered important and needs attention when working with data showing a hierarchical structure. The number of hierarchical categories used in HLM analysis is used to name the analysis. The variables belonging to the students may be in the student-level; the variables belonging to the classrooms may be in the classroom or school level in the analyses to be made in cases where students are clustered in the classrooms (Nilsen et al., 2016). This is also called Two-Level Hierarchical Linear Modeling due to its two-level data structure (Toraman, Akay, Özdemir, & Karadağ, 2011). Independent variables withdrawn from the TIMSS 2015 dataset were defined in two main categories as student level and grade level to be included in the HLM analysis in the present study.

Determination of variables of research: It is seen that the student-level and classroom-level variables related to achievement are based on different school learning models in the literature in some of the studies conducted at the international level based on TIMSS (Kyriakides, 2006; Lamb & Fullarton, 2002; Nilsen et al., 2016; Webster & Fisher, 2000). There are studies that theoretically benefit from different models in determining the variables discussed at student and classroom levels among the studies based on TIMSS in Turkey (Akyüz, 2006; Aydın, 2015; İpekçioğlu-Önal, 2015). For example, Akyüz (2006) prepared the variables determined from the teacher and student questionnaire in relation to students' mathematics achievements based on the theoretical structure that constitutes the theoretical framework of the TIMSS 1995 study and examined them based on the Survey of Mathematics and Science Opportunities (SMSO). Similarly, Aydın (2015) examined different theoretical models and proposed a new model regarding the effects of student and school-level factors on student achievement. This model was established based on the theoretical assessment framework of TIMSS 2011 implementation. İpekçioğlu-Önal (2015) developed a new model based on previous studies on this subject in order to examine student- and teacher-level factors affecting students' science achievement and attitudes towards science. Gender, time allocated to homework, peer bullying, family participation, and educational resources at home were brought together at the student level and self-confidence in science teaching, commitment to the profession, cooperation with colleagues, emphasis on science experiments, experience, and professional development factors were brought together at the teacher level regarding students' science achievements and attitudes towards science in this model.

Different models have been developed in studies on what factors affect achievement in science, addressing various factors at student and classroom levels in relation to student achievement as a result (Lamb & Fullarton, 2002). It can be said that only one model cannot fully explain the relationship between student outcomes and many different variables based on this information. It may be statistically and theoretically more useful to choose a framework created based on variables tried and proven with different models in a study to be carried out on this subject (Hox, 2010).

Therefore, different learning models in the literature, the TIMSS assessment framework, and the Conceptional framework of determinants of students' outcomes developed by Nilsen et al. (2016) were examined to decide which variables to include in the student and classroom levels in the current research and the variables of the research were shaped as a result of these reviews. A total of 21 independent variables, 6 at the student level, and 17 at the classroom level were determined based on this. Table 1 contains detailed information about the variables of the study.

Table 1. Variables Determined Before Preliminary Analysis of the Research

Conceptual Group	Classroom Level (Level-2) Variables	Conversion	TIMSS Code	Research Code	Reliability*	Variance Explained
School Environment	Perception of the importance of school in academic achievement	Yes	ATBG06A-R	W12	.90	45
	Perception of safe and regular school structure	Yes	ATBG07A-H	W13	.89	57
Working Conditions	Problems with school facilities and resources	Yes	ATBG08A-G	W14	.89	60
	Difficulties encountered		ATBG11A-H	W15	.76	35
Teaching Practices	Teaching limited to student needs	Yes	ATBG15A-G	W16	.73	43
	Homework Given (3)	No	ATBS06CA-CB	W18A-C**	-	-
	The importance given to research	Yes	ATBS03A-K	W19	.88	47
	Participation in teaching/Quality of teaching (8)	No	ATBG14A-H	W20A-H**	-	-
	Student Level (Level-1) Variables					
Student Qualifications	Gender	No	ITSEX	X8***		
	Absenteeism	No	ASBG08	X9**		
	Not having breakfast	No	ASBG09	X10**		
	Use of technology (3)	No	ASBG10A-C	X11A-C**		

*Cronbach's Alpha, https://timssandpirls.bc.edu/publications/timss/2015methods/pdf/T15_MP_Chap15_Appendix_A.pdf, **Discontinuous, ***Categorical

Models were established with a total of 17 variables, 6 at the student level, and 11 at the classroom level after the preliminary analysis with the variables in Table 1. The dependent variable of the study is TIMSS 2015 science achievement scores. Among the independent variables, those with similar characteristics were divided into conceptual groups based on the literature and in terms of interpreting the results. The student level consists of 1, and the classroom level consists of 3 conceptual groups as shown in Table 1.

The scales used in TIMSS applications related to the variables of the research were scored as 1-3 as a result of questionnaires or index data consisting of 1-4 scored response categories [*strongly agree* (1), *slightly agree* (2), *disagree* (3), *strongly disagree* (4)] [for example *less* (3), *moderately* (2), *more* (1)]. However, these scores were converted by taking the cut-off points into account by calculating the actual ranges in order for the data to be used in the analysis (Aydın, 2015). Analyses were conducted with continuous values obtained from converted scales in this study. This approach can prevent a number of statistical difficulties in identifying the measured psychological trait with numerical data and in particular, its interpretation. For example,

> 9.6 and 8 cut-off points were determined for the peer bullying scale, which was constantly converted into a form by calculating the actual ranges. The levels of peer bullying of students were defined as *almost never* (< 8), *once a month* (< 9.6 and > 8) and *once a week* (> 9.6). (Martin, Mullis, Hooper et al., 2016, p. 15.89)

The use of these defined values (for example, peer bullying variable) provides convenience in terms of statistical identification and interpretation of the psychological structures such as attitude, value, etc. targeted to be measured.

The variable related to gender is categorical. Each item is considered as a separate variable in the measurements related to some variables that are not continuously converted into forms and have implicit characteristics (homework given, use of technology, and participation in teaching/quality of teaching). Measurements of some variables are discontinuous. Descriptive statistics regarding the variables included in the HLM analysis are given in Table 2.

Table 2. Descriptive Statistics on Variables Included in HLM Analysis

Variables	N	Mean	SD	Minimum	Maximum
X8	6378	1.51	0.50	1.00	2.00
X9	6378	3.38	0.96	1.00	4.00
X10	6378	1.74	1.04	1.00	4.00
X11A	6378	2.34	1.21	1.00	4.00
X11B	6378	2.77	1.29	1.00	4.00
X11C	6378	2.59	1.20	1.00	4.00
W12	241	9.29	1.97	2.82	15.83
W13	241	9.67	2.16	3.75	13.41
W14	241	8.90	2.23	3.19	13.57
W15	241	1.05	0.93	0.00	3.00
W16	241	8.78	1.72	3.80	14.51
W18A	241	1.25	0.45	1.00	3.00
W18B	241	1.52	0.54	1.00	3.00
W18C	241	1.18	0.39	1.00	3.00
W19	241	11.11	2.06	7.30	15.55
W20B	241	1.344	0.60	1.00	3.00
W20C	241	2.49	0.75	1.00	4.00
W20E	241	1.34	0.60	1.00	4.00
W20F	241	1.48	0.75	1.00	4.00
W20G	241	1.60	0.80	1.00	4.00
PV1(ASSSCI01)	6378	484.33	91.13	142.40	754.79
PV2(ASSSCI02)	6378	482.38	92.36	150.67	745.23
PV3(ASSSCI03)	6378	482.26	92.69	136.74	746.52
PV4(ASSSCI04)	6378	481.30	93.80	69.90	846.24
PV5(ASSSCI05)	6378	484.62	93.39	103.62	781.64

Assumptions of HLM: The assumptions for HLM analysis were checked after the completion of the preliminary analyses. The first assumption for Level-1 in HLM is about the normality of residuals. Kolmogorov-Smirnov test results of the residual files created in the SPSS software were found to be significant at this stage ($p < .001$). It is, in this case, interpreted as that the data differ from the normal distribution (Mertler & Vannatta, 2005). Afterward, skewness and kurtosis coefficients were examined. These values are calculated as zero in normally distributed data. However, the fact that these values are between ±1 is interpreted as that the distribution does not deviate excessively from normal (Çokluk et al., 2018; Kim, 2013). Skewness and kurtosis values of the data were calculated as (-0.225) and (0.250) respectively. It was observed in this case that the data showed normal distribution. Another statistical process related to the normality of 1st level errors is the examination of the histogram of the data by drawing it on the normal curve. The histogram indicates the distribution of in-school errors is approximately normal, in which case the normality count can be advocated. The Q-Q graph of the data can also be examined, especially in 100 and larger samples, to check the normality count. The points in Q-Q Plot were seen as a line in the focus of diagonals in the process. This figure is interpreted as an image of the normal distribution (Mertler & Vannatta, 2005). Finally, the Level-1 residual homoscedasticity was checked. The *ellipse shape* showed that the residual variances of Level-1 were homogeneous when the Scatter Plot Diagram obtained as a result of the SPSS procedure was examined. Test of Homogeneity was also applied in the HLM software regarding homoscedasticity. It is χ^2 = 274.93, *df* = 239, *p* > .001 according to the results of this test. The fact that the test for homoscedasticity of Level-1 variances was not found to be significant indicates that variances were distributed homogeneously between Level-2 units. Therefore, it was observed that this assumption was made regarding the data of the current research. It was observed as a result that the residuals for Level-1 showed a close to normal and homogeneous distribution; the variables were independent of the error term r_{ij} and random effects at other levels.

Scatter graphs were obtained using residual files in SPSS in the first stage of the multivariate normal distribution of errors, which is the first assumption for Level-2. The MDIST (Mahalanobis Distance) graph for each school gave the deviation of residuals from normality. Q-Q and P-P graphs were examined for residuals of intersection and slope models. The MDIST vs. CHIPT graph is expected to resemble a 45-degree line in the figures obtained for the normality assumption of the slope coefficients of the cut-off point and variables at this level. The obtained graph resembles a 45-degree line. Furthermore, the Q-Q graph for the intersection model was found to be approximately linear. Thus, it is seen that subordination is defendable. In addition, it was seen that the residual values of the second-level cut-off point coefficients were normally distributed with multivariates when the Q-Q Plots of the slope coefficients of the classroom level variables were examined. In addition, the fact that the Kolmogorov-Smirnov and Shapiro-Wilk test results regarding the slope coefficients of the variables were found to be significant explained that the hypothesis was acceptable for the relevant coefficients and that the data showed normal distribution.

As a result, it was found that the slope coefficients of the cut-off point and variables at this level showed a normal distribution for Level-2. The variables are independent of the error term u_{0j}. In addition, Level-2 errors show multiple normalities with a mean of zero.

Models for responding to the research problems were analyzed after providing the assumptions. The HLM models tested within the scope of this research are given below:

1) One-Way Analysis of Variance Random Effects Model (ANOVA Model): It is checked whether HLM is appropriate in the analysis of the data while answering the question *How much of the differences in students' science achievement arise from the difference between classrooms?* with this model at the same time. There are no explanatory variables for student or classroom levels in the model (Hox, 2002). The variance of the dependent variable is divided into two as inter-group and intra-group variance in the one-way analysis of variance. The Level-1 equation for this model is:

$$Y_{ij}=\beta_{0j}+r_{ij}$$

The science achievement of the student *i* in the classroom *j* (Y_{ij}) is predicted in this equation. β_{0j} refers to the mean science achievement score of the class *j*, r_{ij} refers to the error score of the student *i* in the class *j*, that is, the difference of the student *i* from the mean science score in the class *j*. It is assumed that each error score at the student level is normally distributed with the 0 mean and constant Level-1 (σ^2) variance. The Level-2 equation for this model is:

$$\beta_{0j}=\gamma_{00}+u_{0j}$$

The mean overall science achievement score of the classrooms γ_{00} indicates the error score of the u_{0j}, classroom *j* in the equation where the intersection coefficient (β_{0j}) at the first level of the model is taken as the dependent variable; that is, it is interpreted as the difference of the mean science achievement of the classroom *j* from the mean overall science achievement. It is assumed that u_{0j} shows a normal distribution with the mean 0 and the variance τ_{00}. u_{0j} getting closer to zero means that the variability among the classrooms is very low.

Unified model:

$$Y_{ij}= \gamma_{00}+ u_{0j}+r_{ij}$$

Inter-school and intra-school correlation coefficient (Intra-Classroom Correlation, [ICC]) *p* is calculated to determine how much of the variance in the dependent variable originates from the first level and how much originates from the second level using the following parameters.

$$\rho \text{ (inter-class)} = \tau^{00} / (\tau^{00} + \sigma^{0})$$

$$\rho \text{ (intra-classroom)} = \sigma^{2} / (\tau^{00} + \sigma^{2})$$

2) *The Model with Means as Dependent Variables* was established to answer the second question of the research. The Level-1 equation of this model is the same as the ANOVA model, and there are no student level variables. Grade variables were added to Level-2 to show the extent to which classroom characteristics predict student achievement.

Centering was performed to eliminate the bias caused by the multicollinearity problem in the installation of models (Raudenbush & Bryk, 2002). Group-mean centering was performed for the continuous variables at the student level, and grand-mean centering was performed for the continuous variables at the classroom level while centering was not performed for the categorical variables at both levels. Accordingly, the equations for the model in which the means are dependent variables are as follows:

$$Y_{ij}=\beta_{0j}+r_{ij}$$

$$\beta_{0j}= \gamma_{00}+ \gamma_{01} *(W20B_{j}) + \gamma_{02} *(W18A_{j}) + \gamma_{03} *(W18B_{j}) + \gamma_{04} *(W14_{j}) + u_{0j}$$

*General mean centering was performed for these variables

3) *Random Coefficients Regression Model*: Student variables related to science scores are assigned to the first level, and it is determined which student-level variable affects science achievement score in this model established to answer the third research question. There are no classroom-level variables in the model (Raudenbush & Bryk, 2002). The Level-2 equation is the same as in the ANOVA model. The coefficients of the student variables are interpreted as the change in the mean school achievement scores caused by one-unit variability in the independent variable at the 1st level of the model. The equations for the model are as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j* }(X9_{ij}) + \beta_{2j* }(X10_{ij}) + \beta_{3j* }(X11B_{ij}) + \beta_{4j* }(X11C_{ij}) + r_{ij}$$

$$\beta_{0j}= \gamma_{00}+ u_{0j}$$

$$\beta_{1j}= \gamma_{10}+ u_{1j}$$

$$\beta_{2j}= \gamma_{20}+ u_{2j}$$

$$\beta_{3j}= \gamma_{30}+ u_{3j}$$

$$\beta_{4j}= \gamma_{40}+ u_{4j}$$

*Group-mean centering was performed for these variables.

RESULTS

One-Way Analysis of Variance Results Related to the Random Effects Model

The results of the analysis are given in Table 3. The weighted least squares prediction for the overall science achievement mean is 481.91, and the standard error of the prediction is 3.98 when Table 3 is examined. It can be said that the actual value of the overall science achievement mean is 95% probability in the range of 479.69-483.86 points when the 95% confidence interval for the student overall science achievement mean is calculated [%95CI (γ_{oo}) = γ_{00} ± (1.96) (SE)].

γ_{oo} ± (1.96)* (SE) = 481.91 ± (1.96)* (3.91) = 479.69 – 483.86

Table 3. One-Way Analysis of Variance Random Effects Model Analysis Results

Fixed Effect	Coefficient	*t*	Standard Error (SE)	Approximate *df*
Overall science achievement mean, *γ00*	481.91*	120.92	3.98	158
Random Effect	Variance		Standard Deviation	
Level-2 Error Term, *u0j*	2983.99		54.63*	240
Level-1 Error Term, *rij*	5498.73		74.15	

**p* < .001

In Table 3, the intra-classroom variability (r_{ij}) was 5498.73, and the inter-classroom variability (u_{0j}) was 2983.99 regarding the mean science achievement score. The fact that the predicted value of inter-classroom variability (u_{0j}) was found to be significantly greater than zero (p < .001) indicates significant differences between the mean science achievements of the classrooms.

It can also be calculated through this model how much of the difference in science scores can be explained by student level, and how much can be explained by classroom-level variables. The intra-group correlation coefficient for science achievement scores was calculated by dividing the inter-school variance in Table 3 by total variance (Raudenbush & Bryk, 2002).

$$\rho = \sigma^2 / (\tau_{00} + \sigma^2) = 5498.73 / (2983.99+5498.73)=0.65$$

This result shows that 65% of the difference in science scores is due to the difference between students, and 35% is due to the difference in mean science achievement between classrooms. Most of the variability in students' science achievement scores is due to differences between students.

Results on the Model with Means as Dependent Variables

This model included [perception of the importance of school in academic achievement (W12), perception of safe and regular school structure (W13), problems related to school facilities and resources (W14), difficulties encountered (W15), teaching limited to student needs (W16), giving feedback to science homework (W18A), discussing science homework in the classroom (W18B), the importance given to research (W19), explaining the answers (W20B), using interesting materials (W20C), new content-present content connection (W20E), classroom discussions (W20F), and deciding on problem-solving processes (W20G)]. Variables whose effects were not significant were removed from the model as a result of the first analysis. Accordingly, their relationship with W12, W16, W20C, W20F, and W20G science achievement scores is positive but not significant. Similarly, variables W13, W15, W20E were found to have non-significant negative correlations with science achievement scores. Variables W20B, W18A, W18B, and W14 were included in the final analysis. The equation for the final model is given below:

$$\beta_{0j} = \gamma_{00} + \gamma_{01*} (W20B_j) + \gamma_{02*} (W18A_j) + \gamma_{03*} (W18B_j) + \gamma_{04*} (W14_j) + u_{0j}$$

*general mean centering was performed.

γ_{00} is the corrected mean overall science achievement of the classrooms in this equation. γ_{01} is interpreted as the effect of explaining the answers given in the classroom; γ_{02} is interpreted as the effect of giving feedback to the homework; γ_{03} is interpreted as the effect of discussing the homework in the classroom on the corrected mean science achievement. u_{0j} (Level-2 error term) is expressed as the difference of the mean science achievement score of the classroom j from the mean overall science achievement score when the variables in the model are taken under control.

Table 4. Model Analysis Results with Means as Dependent Variables

Fixed Effects	Coefficients	Standard Error	*t*	*sd*	Effect Size
Mean classroom mean, γ_{00} Intersection	481.88	3.66	131.67	112	
W20B, γ_{01}*	-14.68	5.75	-2.57	236	-.30
W18A, γ_{02}*	19.36	8.26	2.34	236	.39
W18B, γ_{03}*	-14.74	6.90	-2.13	236	-.30
W14, γ_{04}*	9.91	1.50	6.57	236	.20
Random Effects	Standard Deviation	Variance components	*sd*	χ^2	
Level-2 Error Term, u_{0j}	48.816	2383.01	236	2982.37	
Level-1 Error Term, r_{ij}	74.154	5498.94			

*$p < .001$

It is seen that when the other variables in the model are taken under control, the effect of the explanation of the answers given by the students (W20B) on the science achievement scores is predicted to be negatively significant (γ_{01} = -14.68, *t* = -2.57, *p* < .001). Accordingly, the frequent explanation of the answers in the teaching process in the classroom negatively affects the science achievement scores of the schools. The mean science achievement score of these schools is about 15 units less compared to the schools where the answers are explained less. Similarly, the effect of discussing science homework in the classroom (W18B) on science achievement scores was predicted to be negatively significant when the other variables in the model were taken under control (γ_{03} = -14.74, *t* = -2.13, *p* < .001). Accordingly, it is seen that the science achievement scores of the schools where the students who frequently discuss science homework in the classroom are 14.749 units lower compared to the science achievement scores of the schools where the homework is not discussed much.

The effect of giving feedback to science homework (W18A) on the science achievement scores of the students is predicted to be positively significant when the other variables in the model are taken under control, unlike these results (γ_{02} = 19.36, *t* = 2.34, *p* < .001). Accordingly, it is seen that the feedback given to the science homework in the teaching process in the classroom positively affects the science achievement scores of the schools. Another variable with a positive effect on science achievement scores is problems with school facilities and resources (W14) (γ_{04} = 9.91, *t* = 6.57, *p* < .001). Accordingly, it is seen that the science achievement of the schools is positively affected by this situation as the perception levels of the teachers about the lack of opportunities and resources of the schools they work affect the teaching when the other variables in the model are taken under control.

Effect size calculation was made in order to give an idea about whether the interpretations made in line with the variance rates and correlational relationships obtained as a result of the analysis indicate significance for daily life. Accordingly, the effect size was calculated by dividing the constant effect coefficients obtained by the analysis performed at each level by the standard deviation of the residual value at the relevant level. The effect size coefficient .41 indicates the minimum effect (Ferguson, 2009). It is seen that these values are less than .41 when the calculated effect sizes are examined. However, it can be said that these variables may cause a change that can be felt in practice on science achievement considering that even the effect sizes calculated at the .1 level in the studies conducted with large samples may contribute to the developments in the field of education (Glass, McGaw, & Smith, 1981). For example, a standard deviation increase in giving feedback to science homework is expected to create an increase of 0.39 standard deviation in the science achievement mean when the other variables in the model are taken under control. It can be said that the science achievement scores of the schools whose homework is discussed in the classroom are less than 0.30 standard deviations compared to the schools whose homework is not discussed.

Finally, the inter-classroom variance component for science achievement scores was predicted at 2983.99 in the ANOVA model. It was predicted as 2383.011 as the inter-classroom variance component with the addition of class-level variables to the model. Therefore, in-classroom variables explained 20% of the observed variance in achievement scores [(2983.99-2383.011)/2983.99].

Random-Coefficients Results Related to the Regression Model

This model included variables such as gender (X8), absenteeism (X9), not having breakfast (X10), use of technology at home (X11A), use of technology in school (X11B), and use of technology in other places (X11C). Variables whose effects were not significant were removed from the model as a result of the first analysis. Accordingly, the strong positive relationship of X8 with science achievement and the negative relationship of X11A with science achievement score was not found to be statistically significant. Variables X9, X10, W11B, and X11C were included in the final analysis. The equation for the final model is given below:

$$Y_{ij}=\beta_{0j}+\beta_{1j*}X9_{ij}+ \beta_{2j*}X10_{ij} + \beta_{3j*}X11B_{ij} + \beta_{4j*}X11C_{ij}+r_{ij}$$

*group mean centering was performed.

This equation shows the science achievement score of the student *i* in the Y_{ij}, *j* classroom and the mean science achievement score of the β_{0j}, classroom *j*. β_{1j} is expressed as a unit change in absenteeism in the classroom *j* (when the other variables in the model are taken under control); β_{2j}, is expressed as a unit change in nutrition in the classroom *j* (when the other variables in the model are taken under control); β_{3j}, is expressed as a unit change in technology use in the classroom *j* (when the other variables in the model are taken under control); β_{4j}, is expressed as a unit change in technology use in the classroom *j* (when the other variables in the model are taken under control); and these are expressed as a change in the classroom mean science achievement scores.

Table 5. Random-Coefficients Results of Regression Model Analysis

Fixed Effects	Coefficients	Standard Error	*t*	Effect Size
Cut-off point 1, β_0 Overall science achievement mean, γ_{00}*	481.89	3.98	120.84	---
X9 Slope, β_1 Cut-off point 2, γ_{10}*	18.26	1.44	12.60	0.26
X10 Slope, β_2 Cut-off point 2, γ_{20}*	-2.90	1.13	-2.55	-0.00
X11B Slope, β_3 Cut-off point 2, γ_{30}*	8.90	1.05	8.46	0.01
X11C Slope, β_4 Cut-off point 2, γ_{40}*	-2.67	1.12	-2.37	-0.00
Random Effects	Standard Error	Variance Components	*sd*	X^2
Level-2 Error Term, u_{0j}	54.83	3007.17	223	4003.17*
X9 Slope, u_1	8.32	69.25	223	278.18
X10 Slope, u_2	6.86	47.14	223	255.73
X11B Slope, u_3	6.80	46.35	223	295.48
X11C Slope, u_4	6.06	36.83	223	254.55
Level-1 Error Term, r_{ij}	69.93	4891.49		

*$p < .001$

Table 5 it is seen that the variable with the highest effect on science scores is absenteeism (X9) when examined. The effect of absenteeism on science achievement scores was predicted to be positive and significant when the other variables in the model were taken under control (γ_{10} = 18.26, SE = 1.44, $p < .001$). Accordingly, the score of a student who has almost no absenteeism can be interpreted as 18.27 units more than the science score of a student who has frequent absenteeism. Use of technology in school (X11B) is another variable with a high impact on science achievement scores. The coefficient (β_3) was statistically significant ($p < .001$). It can be said that a unit increase in the use of technology in the school (high scores taken from the scale mean low use) will create an increase of 8.90 units in the science achievement scores of the students when the other variables in the model are taken under control. The effect of not eating breakfast (X10) on science achievement scores was predicted to be negative and significant (γ_{20} = -2.90, SE = 1.13, $p < .001$). Accordingly, it can be interpreted that the mean science achievement of the students who have almost no breakfast on the days they go to school is 2 units less compared to the students who have breakfast every day and go to school. Finally, the effect of technology use in other places (X11C) on science scores was predicted to be negative and significant when the other variables in the model were taken under control. It can be said that there will be a decrease of 2.67 units in the science achievement scores of the students as the use of technology in other places (high scores taken from the scale indicate low use) increases based on this when the other variables in the model are taken under control.

It is said that an increase in a standard deviation in absenteeism will create an increase of 0.26 standard deviation in the science achievement mean, and an increase in a standard deviation in using technology in school will create an increase of 0.01 standard deviation in the science achievement mean when the calculated effect sizes are examined. A standard deviation increase in non-breakfast will result in a 0.00 standard deviation decrease in science achievement score. Similarly, a standard deviation increase in technology use elsewhere will create a 0.00 standard deviation decrease in science achievement.

It was determined that the random effect of variance was significant in terms of classroom level when the variance components related to the model were examined (X^2 = 3007.17, sd = 223, *p* < .001). Differentiation between classrooms in terms of mean science scores is indiscriminate when student-level variables are added. The indiscriminate effects of slopes of absenteeism, not having breakfast, use of technology in school, and use of technology in other places were found to be significant according to Table 5 (*p* < .05). This situation reveals that the relationship between the mean science scores of the classrooms and the variables of non-attendance and use of technology in school varies statistically significantly between the classrooms.

The student-level residual variance (4891.49) is smaller compared to the variance (5498.73) obtained in the ANOVA model. This result shows that the difference between students in science achievement scores decreases with the addition of student-level variables. Student-level variables explained 11% of the observed variance in achievement scores.

Reliability values for Level-1 coefficients were calculated to determine whether the values obtained from the sample were a reliable predictor of the actual value. The results of the HLM analysis are as follows:

Table 6. Reliability Values Regarding Level-1 Random Coefficients

Level-1 Random Coefficients	Reliability Predictions
Mean Science Achievement, γ_{00}	.93
X9, γ_{10}	.19
X10, γ_{20}	.16
X11B, γ_{30}	.19
X11C, γ_{40}	.15

Reliability predictions provide information on whether Level-1 coefficients change randomly, constantly, or incidentally. These coefficients do not change randomly or may be constant if the reliability of Level-1 coefficients is below .05 (Acar, 2013; Raudenbush & Bryk, 2002). The high reliability of the constant (.93) indicates that the science mean obtained from the sample is a reliable predictor of the actual school mean, considering Table 6. The reliability of these variables was not found to be very high when the predictions of absenteeism, not having breakfast, use of technology in school, and use of technology in other places were examined. However, it can be said that these variables, albeit at a low level, are reliable predictors of science achievement. In addition, the reliability predictions of these variables being greater than .05 indicates that these coefficients change randomly between schools.

DISCUSSION and CONCLUSION

The factors originating from students and classrooms, which are thought to affect the science achievements of 4th-grade students in a large-scale application, were discussed together in this study. Therefore, the research is important in terms of contributing to the large-scale evaluation literature, which has an important place in terms of education systems. It was seen as a result of the HLM analysis that teachers' perception had a statistically significant effect on science achievement means at the classroom level in terms of explaining the answers given by the students, giving feedback to science homework, discussing science homework in the classroom, and problems related to school facilities and resources. On the other hand, teachers' perception of the importance of school in academic achievement, teaching limited to student needs, using interesting materials, classroom discussions, deciding on problem-solving processes, teachers' perception of safe and regular school structure, difficulties faced by teachers, effects of the new content-present connection on science achievement means were not found to be significant. The effects of gender of students and their use of technology at home on science achievement means were not statistically significant at the student level. However, not having breakfast, absenteeism, use of technology in school, and use of technology in places other than school and home were found to have a significant effect on science achievement mean. Figure 1 was arranged regarding the determinants of 4th-grade TIMSS science achievement at student and classroom levels based on the results of the research.

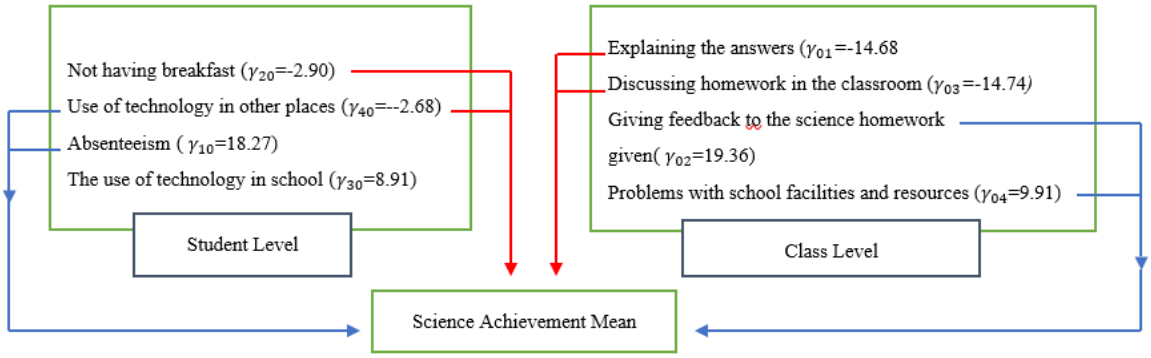

Not having breakfast (γ_{20}=-2.90)
Use of technology in other places (γ_{40}=--2.68)
Absenteeism (γ_{10}=18.27)
The use of technology in school (γ_{30}=8.91)
Student Level
Explaining the answers (γ_{01}=-14.68
Discussing homework in the classroom (γ_{03}=-14.74)
Giving feedback to the science homework given(γ_{02}=19.36)
Problems with school facilities and resources (γ_{04}=9.91)
Class Level
Science Achievement Mean

Figure 1. Determinants of 4th Grade TIMSS Science Achievement at Student and Classroom Levels

It was found in the study that 65% of the differences in science achievement scores of students are due to the difference between students, and 35% is due to the difference between classrooms. Most of the variability in achievement is due to differences between students. Meta-analysis studies on this subject (Hattie, 2009; Sirin, 2005) show that predictions of differences between schools in student achievement are divided into two. It is stated in some studies that most of the differences in student achievements are explained by school level (İpekçioğlu-Önal, 2015; Mohammadpour & Abdul Ghafar, 2014; Özgen, 2009; Yılmaz & Aztekin, 2012). The results of some other studies support the results that the source of the differences in student achievements is explained by the student level as in the results of the current study (Akyüz & Berberoğlu, 2010; Akyüz-Aru & Kale, 2019; Atar & Atar, 2012; Aydın, 2015; İpekçioğlu-Önal, 2015; Ryoo, 2001; Sevgi, 2009).

The contribution of class-level variables was found to be important in explaining the inter-school variance related to science achievements. The explanation of the answers given by the students in the lessons explained approximately 20% of the observed variance in teachers' perception and achievement scores in terms of giving feedback to science homework, discussing science homework in the classroom, and problems related to school facilities, and resources. Wenglinsky (2000) states that it is important to investigate the determinacy of variables related to in-classroom processes on student achievement, and results showing significant effects such as the effect of family or student characteristics on achievement can be obtained but the effects of in-classroom processes on achievement are generally ignored in the literature. The results confirm Wenglinsky's (2000) claim considering the magnitude of the explanation rate of the variance at the relevant level.

The variable with the greatest effect on the mean science achievement of the schools at the classroom level is giving feedback to the science homework given considering the effect sizes among the variables used in the study. The most prominent features of the classrooms in which a supportive atmosphere is created by meeting the psychological needs of the students in the classroom are the continuous increase in achievement measurements. It is generally known in these classrooms that teachers are willing to give positive feedback and empathy to their students and create environments that will enable students to study autonomously by guiding them (Ryan & Deci, 2000). Therefore, giving feedback to homework in the classroom may be a factor that reassures students mentally and psychologically. The student can see the deficiencies and mistakes, learn different solutions, and relax knowing that they can meet with the teacher with the feedback on the homework at the end of the process. It can be said that being appreciated by the teacher in return for their efforts will positively affect achievement if it is considered psychological support. Zhu and Leung (2012) stated that the time allocated to homework showed significant relationships with achievement in a study on homework practices in the classroom. It is important how this allocated time is spent. Akyüz (2006) revealed that emphasizing homework and checking homework have a positive effect on student achievement. Therefore, the importance of feedback emerges in teachers' planning for homework.

It is interesting to see that discussing homework in the classroom has a significant negative effect on achievement. The science achievement scores of the schools where the students who frequently discuss science homework in the classroom were found to be lower compared to the science achievement scores of the schools where the homework is not discussed much according to the results. Discussing homework in the classroom can be considered together with giving feedback on the homework. Today, it is known that homework is still a frequently used learning tool by teachers regardless of age and achievement difference within and between classrooms (Trautwein, 2007; Won & Han, 2010). However, the use of homework in the classroom can have a negative impact on achievement in some cases. This issue can be evaluated in terms of the psychology of the student. The discussions in the classroom about the given homework necessitate the student to be more active when the issue of giving feedback to the student about the homework is considered as a process in which the teacher is more active, and the student is passive. Discussion can be on the way the student does the homework, the way they think, the approach to the homework, and especially on mistakes. The teacher's reactions and the way they direct the process are very important. Ryan and Deci (2000) state that giving positive feedback and empathy to the students of the teachers is important in creating a supportive classroom climate by meeting the psychological needs of the students in the classroom in this case. The psychology of a 9-10-year-old student in the 4th grade may be prone to perceive the discussion of mistakes differently in the classroom; therefore, the negative interaction of achievement with discussing the homework can be evaluated within this framework.

Science lessons are processes that by their natures require the student to be curious, to question, to search for answers, to take responsibility for learning with confidence. Therefore, it would be appropriate to start these lessons with problems, dilemmas, and questions for students (Hiebert et al., 1997). van de Valle, Karp, and Bay-Williams (2010) state that *justification* should be a fundamental part of science lessons. Teachers have important duties at this point. The language and expression that teachers will use will be an important determinant of students' willingness to express their ideas. The results of the research showed that the more students were directed to explain their answers in the classroom, the more they failed the science lesson. Therefore, these results can be evaluated in terms of the language used by the teachers.

Deci and Ryan (1985) emphasize that the teacher plays a major role in meeting a number of psychological needs of students, such as interaction within the classroom. Students' perceptions of these efforts are more important even though teachers report that students make efforts to meet these psychological needs (Daniels & Perry, 2003). The class-level results of the research emphasize the importance of these perceptions. It can also be said that the attitude of teachers in supporting the explanation of the answers may cause students to be psychologically worried about explaining the answers they give, especially, in the classroom or to experience morale disorder. Therefore, performance is adversely affected by this process.

The result of the research about the school resources is that the science achievement of the schools is positively affected by this situation as the perception levels of the teachers about the lack of opportunities and resources of the schools they work affect the teaching. It can be evaluated psychologically even though this result is interesting. It can be said that this situation indirectly affects student achievement due to the negativities caused by teachers' perceptions, even if not directly based on the fact that teachers feel happier in environments where working conditions are good (Hooper et al., 2013; World Bank, 2011) when the resource shortage is evaluated under conditions. It can be seen that teachers see and enlarge some problems arising from the management and internalize them in a way that adversely affects classroom achievement whereas there are also studies showing that teachers are not affected by this (Bénabou, Kramarz, & Prost, 2009; Wößmann, 2003). However, it can be interpreted that they do not reflect it to the teaching process and solve it with their special efforts even though teachers see resource shortage as a problem according to the results of the study. It can be seen in such cases that teachers overcome resource shortages with the available facilities in cooperation with their students.

Absenteeism was the variable that contributed the highest level to explaining the variance seen in achievement scores considering the results of the study at the student level. There are studies on absenteeism in parallel with the results of the current study. Alexander and Hicks (2015) found a significant and positive relationship between absenteeism and achievement scores of 383 students. The academic performance of the students who attend the lesson regularly is higher compared to that of the absent students according to the results of the study. Similarly, absenteeism negatively affects students' grades according to the results of Khalid's (2017) study with 119 participants. Absenteeism is an important issue consisting of different components, especially the psychological background of students. Studies generally focus on *chronic absenteeism*, and the results show that achievement decreases significantly as students' absenteeism rates increase. However, Cattan, Kamhöfer, Karlsson and Nilsson (2017, p. 47) state that "the studies in the literature on this subject are insufficient".

Another result is that this situation of students coming to school without breakfast negatively affects their academic performance. There are studies in the literature mentioning a positive but low level of relationship between nutrition and academic achievement (Dwyer, Sallis, Blizzard, Lazarus, & Dean, 2001; Keeley & Fox, 2009). Shaw, Gomes, Polotskaia and Jankowska (2015) state that the controllable aspects of student health are nutrition, healthy weight preservation, and physical equivalence with their peers. It is shown that students with poor health are more likely to fail at school, fail the class, and drop out of school. Kolasa et al. (2018) state that many studies on nutrition and academic achievement abroad are carried out on the basis of nutrition programs given in schools. The common conclusion of the studies conducted to provide evidence that integrating food/nutrition education into the 4th-grade curriculum can support academic knowledge acquisitions is that science and mathematics knowledge overlaps with nutrition knowledge in a holistic way to improve academic knowledge and that nutrition knowledge can also be developed simultaneously among 4th-grade students using science and mathematics curriculum (Kolasa et al., 2018).

There is no special lesson for nutrition in the primary school curriculum in Turkey. Life science lessons include texts for regular and balanced nutrition. In addition, the contents of nutrients and balanced nutrition are briefly emphasized in a unit of the primary school science curriculum. A compulsory or optional *nutrition lesson* can be included in the primary school curriculum at this point. The content of this lesson can be determined by including families after the needs of students are identified or *nutrition workshops* can be organized. Lessons can be practically conducted with the participation of families and students in this workshop.

The use of technology in school and elsewhere is another variable with a high impact on science achievement scores. Low scores obtained from the technology use scale (computer, tablet, etc.) in the TIMSS student survey mean that technology is used more frequently. Therefore, students with higher science achievement stated that they use technology less frequently in school but more frequently outside the school for homework purposes. Students' access to technology at school and access to technology in different places other than school and home may differ in terms of duration. Longer time can be spent online for homework, etc. outside school. In addition, it can be interpreted based on the results that students use technology efficiently for homework purposes outside school. 33% of students are allowed to use computers in science lessons according to the results of TIMSS 2015 (Martin, Mullis, Foy et al., 2016). It is known that the use of the Internet during school is limited except for informatics lessons. Therefore, students can turn to technology sources outside the school for their homework. The 8th-grade students who participated in the TIMSS application were asked what the Internet was used for outside the school. The majority of students (75%) stated that they mostly use the Internet outside the school to prepare project homework with their friends (Martin, Mullis, Foy et al., 2016). Students can benefit from technology for homework and research purposes if they are given opportunities at school, as a result.

Work in this direction has accelerated in recent years. It is now frequently observed that programs containing many interesting visuals and narratives are used personally by teachers in lessons and that students interactively participate in these programs with the integration of technology into educational environments. It can be easily predicted that the use of computers under the supervision of teachers in schools will positively affect students' achievement considering the nature of the science lesson. It is seen that the issue of ensuring the integration of technology into education and teaching is emphasized within the framework of Turkey's 2023 Education Vision. Turkey shows a successful example of the integration of technology into education in the interaction of student-teacher-parent-school resources with each other in the current pandemic period. Many teachers and students have been struggling with technology in the distance education process. Perhaps the greatest gain when the process is over is the discovery that technology is a useful and inevitable element for education.

Recommendations

Student and classroom-level variables of this research in relation to student achievement are limited to the items in the student questionnaire and teacher questionnaire in the TIMSS 2015 study and measuring the characteristics at this level. It is also a causal comparison study. The result variable is science achievement scores as measured in the TIMSS application. The sample consists of only a certain number of students and teacher groups who participated in TIMSS 2015. The preference of the experimental method will allow for more detailed discussions about the results in future research. In addition, studies can be conducted with different variables and different groups. The cross-level interactions of the intersection and slope model and the relationships of the variables at different levels with each other can be examined in HLM.

The results of the research on the use of technology showed that the use of technology in other places outside the school negatively affects the science achievement scores of the students while the use of technology in school positively affects the science performance of the students. It may be suggested in light of these results to organize training on the conscious use of technology for students, parents, and teachers, especially primary school age students, regarding their use of technology at home and elsewhere, by the Ministry of National Education and similar organizations interested in education. Practical activities aiming to use the Internet in a beneficial way without damaging the social and academic development of students can be carried out. These activities can be started especially by following the daily usage periods of young primary school students such as tablets, computers, phones, etc. by classroom teachers. Afterward, it should be determined what kind of contents are preferred. The process can continue with useful content on academic and personal development and guidance on channeling technology in the right direction. Activities to be carried out on this subject should be managed by classroom teachers.

Absenteeism was the variable that had the highest effect on science achievement scores at the student level according to the results of the study. Absenteeism negatively affects science achievement scores. In addition, Turkish students are absent more frequently compared to students from other participating countries according to PISA and TIMSS 2015 data. An *absenteeism research project* can be proposed to be carried out especially in schools on habitual absenteeism (chronic absenteeism), in line with these results. The *absenteeism researcher* of each school can be appointed by the relevant District Directorate of National Education from among the teachers working in the school within the scope of the project. Teachers who will be assigned as researchers should receive seminars on absenteeism, where the results of the latest academic studies in the literature are also discussed. These teachers submit reports to the District Directorates at periods to be determined in the following process. The content of the report may be absent students on a school basis, reasons for absenteeism, interviews, etc. In addition, the researcher also detects and observes students and schools that do not have problems in attendance. The results of this observation can be used for an approach to increase attendance in schools with absenteeism.

Another result of the study is related to nutrition. Coming to school without breakfast negatively and significantly affected science scores. A compulsory or optional *nutrition lesson* can be included in the primary school curriculum in light of these results. The content of this lesson should be determined after identifying the needs of students. Families can also be included in the nutrition lesson. An elective lesson can be performed by adding a lesson hour to the school exit. Or *nutrition workshops* can be organized. Lessons can be practically conducted with the participation of families and students in this workshop. *Nutrition lesson* can be improved with activities such as nutrition problems, relationship between health and nutrition relationship, academic achievement and nutrition, regular and balanced nutrition as well as good examples from daily life.

Recommendations on nutrition can also be improved on students' nutritional breaks. As a matter of fact, students eat what they have brought in their lunch-box during these breaks. How these breaks are spent, how long they last, what students consume, and how nutrition lists are prepared, if any, should be investigated. Accordingly, food engineers can be assigned within the District Directorates of National Education. A school-based supervisor can be designated, and *interview days with the food engineer* can be arranged periodically. Useful dialogs on students' eating habits should be established in these interviews.

The results of the study at the classroom level show that the variable with the greatest effect on the mean science achievement of the schools is giving feedback on the science homework given. In addition, *discussing homework in class* has a negative significant effect on achievement. The science achievement scores of the schools where the students who frequently discuss science homework in the classroom were found to be lower compared to the science achievement scores of the schools where the homework is not discussed much according to the results. Classroom level results highlighted teachers' attitudes towards students regarding homework. It can be said that homework also has a psychological structure that affects the academic performance of primary school students when the attitude is considered as an affective factor. The sensitivity of the subject of homework emerges considering these results. Accordingly, the use of homework in the teaching process and the studies examining the attitudes of teachers during this process may contribute to the process. Field studies can be carried out by the relevant institutions on how teachers manage the homework process. Teachers should be provided with training, and their development should be supported according to the results.

One of the problems experienced by the parents of the families who have primary school students in Turkey and the parents of the students who will start the first grade is homework, and parents may have prejudices in this regard. Therefore, this bias can grow and spread to other years of primary school, and this may negatively affect the student’s perception of the homework if this bias, which is carried to school with the family at the beginning of the year, is unbreakable. Educators and especially primary school teachers have great duties in this regard. Homework to be given in first grade should be considered meticulously and in accordance with the developmental processes, psychologies, and needs of the students. Homework should not be torture for the child and the family and should be given in line with its purpose. Therefore, studies to be carried out with parents are also important. Accordingly, it may be suggested to create *interactive homework portals* in schools. Children’s interest in digital media may be a good opportunity in this regard. Thus, the reactions of families to homework can also be evaluated instantly. Teachers can comfort families with constructive feedback on homework as an audience and a participant in this process.

REFERENCES

Acar, M. (2013). *Öğrenci başarılarının belirlenmesi sınavında Türkçe dersi başarısının öğrenci ve okul özellikleri ile ilişkisinin hiyerarşik lineer model ile analizi* (Doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Ackerman, P. L. (2013). Engagement and opportunity to learn. In J. Hattie, & E. M. Anderman (Eds.), *The international guide to student achievement* (pp. 39-41). New York: Taylor & Francis.

Akey, T., M. (2006). *School context, student attitudes and behavior, and academic achievement: An exploratory analysis*. New York: MDRC. Retrieved from https://files.eric.ed.gov/fulltext/ED489760.pdf

Akkuş, M. (2014). *PISA, TIMSS ve PIRLS sonuçlarının değerlendirilmesi* (Yüksek lisans tezi). İstanbul Aydın Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Aktaş, I. (2011). *TIMSS 2007 verilerine göre öğrenci fen başarısı ile öğretmenlerinin özellikleri arasındaki ilişkinin incelenmesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Akyüz-Aru, S., & Kale, M. (2019). Effects of school-related factors and early learning experiences on mathematics achievement: A multilevel analysis to analyze the TIMSS data. *Journal of Education and Training Studies, 7*(4), 259-272. doi: 10.11114/jets.v7i4.3949

Akyüz, G. (2006). Türkiye ve Avrupa Birliği ülkelerinde öğretmen ve sınıf niteliklerinin matematik başarısına etkisinin incelenmesi. *İlköğretim Online, 5*(2), 75-86. http://ilkogretim-online.org//?mno=121113 adresinden erişilmiştir.

Akyüz, G., ve Berberoğlu, G. (2010). Teacher and classroom characteristics and their relations to mathematics achievement of the students in the TIMSS. *New Horizons in Education, 58*(1),77-95. Retrieved from https://www.researchgate.net/publication/293232168_Teacher_and_classroom_characteristics_and_their_relations_to_mathematics_achievement_of_the_students_in_the_TIMSS

Alexander, V., & Hicks, R. E. (2015). Does class attendance predict academic performance in first year psychology tutorials? *International Journal of Psychological Studies, 8*(1), 28-32. doi: 10.5539/ijps.v8n1p28

Asigbee, F. M., Whitney, S. D., & Peterson, C. E. (2018). The link between nutrition and physical activity in increasing academic achievement. *Journal of School Health, 88*(6), 407-415. doi: 10.1111/josh.12625

Atar, H. Y. (2014). Öğretmen niteliklerinin TIMSS 2011 fen başarısına çok düzeyli etkileri. *Eğitim ve Bilim Dergisi, 39*(172), 121-137. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/2894/620 adresinden erişilmiştir.

Atar, H. Y., ve Atar, B. (2012). Türk eğitim reformunun öğrencilerin TIMSS 2007 fen başarılarına etkisinin incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(4), 2621-2636.

Aydın, M. (2015). *Öğrenci ve okul kaynaklı faktörlerin TIMSS matematik başarısına etkisi* (Doktora tezi). Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya.

Ayvaz, A. (2010). *4. sınıf matematik dersi bölme işlemi alt öğrenme alanının edebi ürünlerle işlenmesinin öğrenci başarı ve tutumuna etkisi* (Yüksek lisans tezi). Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, Sakarya.

Bénabou, R., Kramarz, F., & Prost, C. (2009). The French zones d'éducation prioritaire: Much ado about nothing? *Economics of Education Review, 28*(3), 345-356.

Bilican-Demir, S. (2014). Değerlendirme. S. Tekindal (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 225-257). Ankara: Pegem Akademi.

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Teacher quality, insturactional quality and student outcomes, relationshps accross countries, cohorts and time. In T. Nilsen & J. E. Gustafsson (Eds.), *Relation of student achievement to the quality of their teachers and ınstructional quality* (pp. 21-51). Switzerland: IEA Publishing.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education, 62*, 154-171.

Büyüköztürk, Ş., Çakan, M., Tan, Ş. ve Atar, H. Y. (2014). *TIMSS 2011 Ulusal Matematik ve Fen Raporu*. Ankara: İşkur Matbaacılık.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., ve Demirel, F. (2011). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Akademi.

Cattan, S., Kamhöfer, D. A., Karlsson, M., & Nilsson, T. (2017). *The short-and long-term effects of student absence: Evidence from Sweden* (IZA DP No. 10995). Retrieved from http://ftp.iza.org/dp10995.pdf

Clinton, J. (2013). Physical activity. In: J. Hattie, and E. M. Anderman (Eds.), *The international guide to student achievement* (pp. 33-35). New York, NY: Taylor & Francis.

Clinton, J., Rensford, A., & Willing, E. (2007). *Literature review of the relationship between physical activity, nutrition and academic achievement.* New Zeland: Auckland Centre for Health Services, Research and Policy, University of Auckland.

Creemers, B. P., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347-366.

Çavdar, D. (2015). *TIMSS 2011 matematik başarısının öğrenci ve öğretmen özellikleri ile ilişkisi* (Yüksek lisans tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Çelebi, N., Güner, H., Taşçı-Kaya, G. ve Korumaz, M. (2014). Neoliberal eğitim politikaları ve eğitimde fırsat eşitliği bağlamında uluslararası sınavların (PISA, TIMSS ve PIRLS) analizi. *Tarih Kültür ve Sanat Araştırmaları Dergisi,* 3(3), 33-75. doi: 10.7596/taksad.v3i3.329

Çokluk, Ö., Şekercioğlu, G., ve Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi.

Daniels, D. H., & Perry, K. E. (2003). "Learner-centered" according to children. *Theory into Practice, 42*(2), 102-109. doi: 10.1207/s15430421tip4202_3

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Doğan, U. (2014). Validity and reliability of student engagement scale. *Bartın Üniversitesi Eğitim Fakültesi Dergisi*, *3*(2), 390-403. Retrieved from https://dergipark.org.tr/en/download/article-file/43660

Dwyer, T., Sallis, J. F., Blizzard, L., Lazarus, R., & Dean, K. (2001). Relation of academic performance to physical activity and fitness in children. *Pediatric Exercise Science*, *13*, 225-237.

Ekinci-Vural, D. (2012). *Okul öncesi eğitimin ilköğretime etkisinin aile katılımı ve çeşitli değişkenler açısından incelenmesi* (Doktora tezi). Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü, İzmir.

Erşan, Ö. (2016). *TIMSS 2011 8. Sınıf öğrencilerinin matematik başarılarını etkileyen faktörlerin yapısal eşitlik modeliyle incelemesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538. doi: 10.1037/a0015808

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

Fullarton, S., Lokan, J., Lamb, S. & Ainley, J. (2003). *Lessons from the third international mathematics and science study* (TIMSS Australia Monograph No. 4). Melbourne: Australian Council for Educational Research.

Garcia, E., & Weiss, E. (2018). *Student absenteeism-Who misses school and how missingschool matters for performance* (Economic Policy Institute Report). Retrieved from https://files.eric.ed.gov/fulltext/ED593361.pdf

Garson, D. (2008). *Data imputation for missing values*. Retrieved from https://faculty.chass.ncsu.edu/garson/PA765/index.htm

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* London: Sage.

Güngör, H. (2014). *İlkokul 4. sınıf matematik dersi "kesirler" konusunun öğretiminde yardımcı kitap kullanımının öğrenci başarısı üzerindeki etkisi.* (Yüksek lisans tezi). Yüzüncü Yıl Üniversitesi, Eğitim Bilimleri Enstitüsü, Van.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Newyork: Routledge.

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multlievel modeling techniques.* New York: Routledge.

Hiebert, J., Carpenter, C. P., Fennema, E., Fuson, K., Wearne, D., Murray, H., Olivier, A., & Human, P. (1997). *Making sense: Teaching and learning mathematics with understanding.* Portsmouth, NH: Heinemann.

Hooper, M., Mullis, I. V. S., & Martin, O. M. (2013). TIMSS 2015 Context questionnaire framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 61-80). United States Boston College: TIMSS and PIRLS International Study Center.

Hox, J. J. (1995). *Applied multilevel analysis.* Amsterdam, Netherlands: TT- publicities.

Hox, J. J. (2002). *Applied multilevel analysis*. Mahwah, NJ: Erlbaum.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Taylor & Francis.

İsmail, N. A., & Awang, H. (2008). Assessing the effects of students' characteristics and attitudes on mathematics performance. *Problems of Education in 21st Century, 9*, 34-41. Retrieved from https://www.researchgate.net/profile/Noor-Ismail-18/publication/268226049_Effects_of_Teachers_and_Schools_on_Mathematics_AchievementsProblems_of_Education_in_the_21st_Century_Recent_Issues_in_Education/links/5466d1b10cf2397f7829e637/Effects-of-Teachers-and-Schools-on-Mathematics-AchievementsProblems-of-Education-in-the-21st-Century-Recent-Issues-in-Education.pdf

İpekçioğlu-Önal, S. (2015). *TIMSS 2011 cross country comparisons: relationship between student- and teacher -level factors and 8th grade students' achievement and attitude toward science* (Doctoral thesis). Middle East Technical University, Secondary Science and Mathematics Education Department, Ankara.

Karip, E. (2017). T*ürkiye'nin TIMSS 2015 performansı üzerine değerlendirme ve öneriler*. *https://tedmem.org/download/turkiyenin-timss-2015-performansi-uzerine-degerlendirme oneriler?wpdmdl=2515* adresinden erişilmiştir.

Kaya, S. (2008). *The effects of student-level and classroom-level factors on elementary students' science achievement in five countries* (Doctoral dissertation). Florida State University, Florida.

Keeley, T. J. H., & Fox, K. R. (2009). The impact of physical activity and fitness on academic achievement and cognitive performance in children. *International Review of Sport and Exercise Psychology, 2*(2), 198-214. doi: 10.1080/17509840903233822

Khalid, N. (2017). Effects of absenteeism on students' performance. *International Journal of Scientific and Research Publications,7*(9), 151-168. Retrieved from http://www.ijsrp.org/research-paper-0917.php?rp=P696781

Kılıç-Özün, S. (2010). *Hayat bilgisi öğretiminde kavram karikatürü yaklaşımının öğrenci başarısı ve tutumuna etkisi* (Yüksek lisans tezi). Zonguldak Karaelmas Üniversitesi, Sosyal Bilimler Enstitüsü, Zonguldak.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262-273. Retrieved from https://scholar.google.com.tr/scholar?q=Klem,+A.+M.,+%26+Connell,+J.+P.+(2004).+Relationships+matter:&hl=tr&as_sdt=0&as_vis=1&oi=scholart

Kolasa, K. M., Díaz, S. R., & Duffrin, M. W. (2018). Exploring the associations among nutrition, science, and mathematics knowledge for an integrative, food-based curriculum. *Journal of School Health*, *88*(1), 15-22. doi: 10.1111/josh.12576

Korkmaz, F. (2012). *Contribution of some factors to eight grades students' science achievement in Turkey: TIMSS 2007* (Doctoral dissertation). Middle East Technical University, Ankara.

Kunuk, M. (2015). *Okul öncesi eğitimin ilkokul öğrencilerinin akademik başarılarına etkisi (Üsküdar örneği)* (Yüksek lisans Tezi). İstanbul Aydın Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Kyriakides, L. (2006). Using international comparative studies to develop the theoretical framework of educational effectiveness research: A secondary analysis of TIMSS 1999 data. *Educational Research and Evaluation*, *12*(6), 513-534.

Lamb, S., & Fullarton, S. (2002). Classroom and school factors affecting mathematics achievement: A comparative study of Australia and the United States using TIMSS. *Australian Journal of Education, 46*(2), 154-171. doi: 10.1177/000494410204600205

LaRoche, S., & Foy, P. (2016). Sample implementation in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 5.1- 5.170). Retrieved from https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-5.html

Lee, J. W., & Barro, R. J. (2001). Schooling quality in a cross-section of countries. *Economica, New Series, 68*(272), 465-488.

Lee, V. E., & Zuze, T. L. (2011). School resources and academic performance in Sub-Sharan Africa. *Comparative Education Review*, *55*(3), 369-397. Retrieved from https://www.jstor.org/stable/10.1086/660157?seq=1

Lieberman, D. A., Bates, C. H., & So, J. (2009) Young children's learning with digital media. *Computers in the Schools*, *26*(4), 271-283. doi: 10.1080/07380560903360194

Liouaeddine, M., Bijou, M., & Naji, F. (2017). The main determinants of Moroccan students' outcomes. *American Journal of Educational Research,5*(4), 367-383. doi: 10.12691/education-5-4-5

Martin, M. O., Mullis, I. V. S., & Foy, P. (2013). TIMSS 2015 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85-96). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in science*. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A. & O'Connor, K. M. (2000). *TIMSS 1999 international science report.* Retrieved from *https://timssandpirls.bc.edu/timss1999i/pdf/T99i_Sci_All.pdf*

Martin, M. O., Mullis, İ. V. S., Hooper, M, Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1-15.312). Retrieved from https://timss.bc.edu/publications/timss/2015-methods/chapter-15.html

Marzano, R. J. (2003). *What works in school: Translating research into action* (1st ed.). Alexandria, VA: Association for Supervision & Curriculum Development.

Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Glendale, CA: Pyrcz Publishing.

Mohammadpour, E., & Abdul Ghafar, M. N. (2014). Mathematics achievement as a function of within-and-between school differences. *Scandinavian Journal of Educational Research, 58(*2), 189-221. doi: 10.1080/00313831.2012.725097

Mohammadpour, E., Shekarchizadeh, A., & Kalantarrashidi, S. A. (2015). Multilevel modeling of science achievement in the TIMSS participating countries. *The Journal of Educational Research*, *108*(6)1-16. doi: 10.1080/00220671.2014.917254

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 ınternational report: IEA's study of reading literacy achievement in primary school in 35 countries.* Chestnut Hill, M.A.: International Study Center, Boston College.

Nilsen, T., Gustafsson, J. E. & Blömöke, S. (2016). Conceptual framework and methodology of the report. In T. Nilsen & J. E. Gustafsson (Eds.), *Teacher quality, insturactional quality and student outcomes, relationshps accross countries, cohorts and time* (pp. 1-21). Switzerland: IEA Publishing. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-41252-8_1

Organisation for Economic Co-operation and Development. (2013). Education Policy Outlook: Turkey. Retrieved from http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20TURKEY_EN.pdf.

Organisation for Economic Co-operation and Development. (2015). Education at a Glance. Retrieved from https://www.oecd-ilibrary.org/education/education-at-a-glance-2015_eag-2015-en

Olson, J. F., Martin, M. O. & Mullis, I. V. S. (2008). *TIMSS 2007 technical report.* United States: International Study Center, Boston College.

Ozborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment Research anad Evaluation, 8.*

Özdemir, S. (2013). Türk eğitim sistemi ve okul yönetimi. S. Özdemir (Ed.), *Türk eğitim sisteminin yapısı, eğilimleri ve sorunları* içinde (ss. 7- 52). Ankara: Pegem Akademi.

Özden, M. (2007). Problems with science and technology education in Turkey. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(2), 157-161. Retrieved from https://www.ejmste.com/download/problems-with-science-andtechnology-education-in-turkey-4061.pdf

Özgen, C. (2009). *The connection between school and student characteristics with mathematics achievement in Turkey (*Doctoral dissertation). Middle East Technical University, Varsa Enstitü Bilgisi, Ankara.

Raudenbush, S. V., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* London: Sage.

Reddy, V. (2005). Cross-national achievement studies: learning from South Africa's participation in the trends in international mathematics and science study (TIMSS). *A Journal of Comparative and International Education, 35*(1), 63-77. doi: 10.1080/03057920500033571

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78. Retrieved from https://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf

Ryoo, H. (2001). *Multilevel influences on student achievement: An international comparative study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database.

Sarı, M. H., Arıkan, S. ve Yıldızlı, H. (2017). 8. Sınıf matematik akademik başarısını yordayan faktörler-TIMSS 2015. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(3), 246-265. doi: 10.21031/epod.303689

Savaş, E., Taş, S. ve Duru, A. (2016). Factors affecting students' achievement in mathematics. *Inonu University Journal of the Faculty of Education, 11*(1), 113-132. Retrieved from https://dergipark.org.tr/tr/download/article-file/92276

Selçuk, E. (2015). *Müzik dersinde zihin haritalama tekniği kullanımının öğrenci başarısı ve tutumlarına etkisi* (Yüksek lisans tezi). Marmara Üniversitesi Eğitim Bilimleri Enstitüsü, İstanbul. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=oBF44VZN7pDvnhafMH48dw&no=l51CbOtziZRlPV_Nj4M4RQ adresinden erişilmiştir.

Sevgi, S. (2009). *The connection between school and student characteristics with mathematics achievement in Turkey* (A thesis submitted to the graduate). Middle East Technical Universty, Secondary Science and Mathematics Education Department, Ankara.

Sezer, E. (2016). *Öğretmenlerin kişisel ve mesleki niteliklerinin 4 ve 8. Sınıf öğrencilerinin TIMSS 2011 matematik başarısına etkisinin incelenmesi* (Yüksek lisans tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Shaw, S. R., Gomes, P., Polotskaia, A., & Jankowska, A. M. (2015). The relationship between student health and academic performance: Implications for school psychologists. *School Psychology International*, *36*(2), 115-134. doi: 10.1177/0143034314565425

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta- analytic review of research. *Review of Educatiınal Research, 75*(3), 417-453. doi: 10.3102/00346543075003417

Stemler, S. E. (2001). *Examining school effectiveness at the fourth grade: A hierarchical analysis of the third ınternational mathematics and science study (TIMSS).* (Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy). Wesleyan University, Department of Psychology, United States.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics.* Boston: Ally and Bacon.

Taştekinoğlu, E. (2014). *4. sınıf matematik sorularının bilişsel alan kapsamında incelenmesi; TIMSS sınav sorularıyla karşılaştırmalı bir analiz* (Yüksek lisans tezi). İstanbul Aydın Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Toraman, Ç., Akay, E., Özdemir, H. F. ve Karadağ, E. (2011). *Çok düzeyli regresyon modelleri, HLM uygulamaları*. Ankara: Nobel.

Trautwein, U. (2007). The homework achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction, 17*(3), 372-388. doi: 10.1016/j.learninstruc.2007.02.009

Uzun, S., Bütüner, S. Ö. ve Yiğit, N. (2010). A comparison of the results of TIMSS 1999-2007: The most successful five countries-Turkey sample. *Elementary Education Online, 9*(3), 1174-1188. Retrieved from https://dergipark.org.tr/tr/download/article-file/90742

Van de Valle, J. A., Karp, K. S., & Bay-Williams, J. M. (2010). *Elementary and middle school mathematics - teaching developmentally*. USA, Pearson.

van den Broeck, A., Opdenakker, M. C., & Van Damme, J. (2005). The effects of student characteristics on mathematics achievement in Flemish TIMSS 1999 data. *Educational Research and Evaluation, 11*(2), 107-121. doi: 10.1080/13803610500110745

Webster, B. J., & Fisher, D. L. (2000). Accounting for variation in science and mathematics achievement: A multilevel alaysis of Australian data third ınternational mathematics and science study. *School Effectivenes and School Improvement, 11*(3), 339-360.

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality.* Princeton, NJ: Educational Testing Service.

Won, S. J., & Han, S. (2010). Out-of-school activities and achievement among middle school students in the U.S. and South Korea. *Journal of Advanced Academics, 21*(4), 628-661. doi: 10.1177/1932202X1002100404

World Bank (2011). *Improving the quality and equity of basic education ın turkey challenges and options.* The World Bank: Washington D.C.

Wößmann, L. (2003). Schooling resources, educational institutions and pupil performance: international evidence. *Oxford Bull Economics Statistic*s, *65*(2), 117-170. doi: 10.1111/1468-0084.00045

Yaman, İ. (2004). *Modeling the realationship between the science teacher characteristics and eight grade Turkish student science achievement in TIMSS- R* (Yüksek lisans tezi). Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Yatağan, M. (2014). *Fen ve teknoloji dersi öğretim programının öğrenci ve öğretmen özelliklerine göre değerlendirilmesi: TIMSS 2007 ve 2011 verileri ile bir durum analizi* (Doktora tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Yılmaz, H. B., ve Aztekin, S. (2012, Haziran). *Türkiye`deki düzey istatistiki bölge birimlerine göre 15 yaş grubu öğrencilerinin akademik başarılarını etkileyen bazı faktörlerin incelenmesi*. X. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresinde sunulan sözlü bildiri, Niğde, Türkiye.

Yücel, C., ve Karadağ, E. (2016). *TIMSS 2015 Türkiye; Patinajdaki eğitim.* doi: 10.13140/RG.2.2.20445.20964/1

Zhu, Y., & Leung, F. K. S. (2012). Homework and mathematics achievement in Hong Kong: Evidence from the TIMSS 2003. *International Journal of Science and Mathematics Education, 10*(4), 907-92. doi: 10.1007/s10763-011-9302-3

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 97-111

Retrofitting of Polytomous Cognitive Diagnosis and Multidimensional Item Response Theory Models *

Levent YAKAR** Nuri DOĞAN *** Jimmy DE LA TORRE ****

Abstract

In this study, person parameter recoveries are investigated by retrofitting polytomous attribute cognitive diagnosis and multidimensional item response theory (MIRT) models. The data are generated using two cognitive diagnosis models (i.e., pG-DINA: the polytomous generalized deterministic inputs, noisy “and” gate and fA-M: the fully-additive model) and one MIRT model (i.e., the compensatory two-parameter logistic model). Twenty-five replications are used for each of the 54 conditions resulting from varying the item discrimination index, ratio of simple to complex items, test length, and correlations between skills. The findings are obtained by comparing the person parameter estimates of all three models to the actual parameters used in the data generation. According to the findings, the most accurate estimates are obtained when the fitted models correspond to the generating models. Comparable results are obtained when the fA-M is retrofitted to other data or when the MIRT model is retrofitted to fA-M data. However, the results are poor when the pG-DINA is retrofitted to other data or the MIRT is retrofitted to pG-DINA data. Among the conditions used in the study, test length and item discrimination have the greatest influence on the person parameter estimation accuracy. Variation in the simple to complex item ratio has a notable influence when the MIRT model is used. Although the impact on the person parameter estimation accuracy of the correlation between skills is limited, its effect on MIRT data is more significant.

Key Words: Polytomous attribute cognitive diagnosis models, pG-DINA, fA-M, multidimensional item response theory, retrofitting.

INTRODUCTION

Some of the specific measurement procedures used in education and psychology can be applied to one or more attributes. Scales constructed to measure a single skill may also be applied to another, but high correlations between the skills measured may render the scale insensitive to measuring other skills (Reckase, 2007). Consequently, tests may appear to measure only one main skill. However, if the correlations between measured skills are not too high, the main factor may not suppress other factors, particularly in psychological-based measurements. Thus, multiple skills may be measured intentionally or unintentionally.

Various psychometric approaches can be taken when measuring multiple skills. For example, in item response theory (IRT), unidimensional IRT (UIRT) models can be applied multiple times to measure one skill at a time, whereas multidimensional IRT (MIRT) models can be used to measure more than one skill simultaneously.

* This study is based on the first author’s PhD thesis entitled “Retrofitting of cognitive diagnosis and multidimensional item response theory models” [Bilişsel tanı ve çok boyutlu madde tepki kuramı modellerinin karşılıklı uyumlarının incelenmesi] and was written under the supervision of the other authors.

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

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

**** Professor, The University of Hong Kong, Faculty of Education, Hong Kong, China, j.delatorre@hku.hk, ORCID ID: 0000-0002-0893-3863

To cite this article:
Yakar, L., Doğan, N., & de la Torre, J. (2021). Retrofitting of polytomous cognitive diagnosis and multidimensional item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 97-111. doi: 10.21031/epod.778861

Received: 10.08.2020
Accepted: 26.05.2021

Multidimensional Item Response Theory (MIRT)

MIRT models were developed to address the main limitation of UIRT models – they assume a single underlying skill. In contrast, MIRT models can be used when multiple skills interact to determine the probability that an individual will respond correctly to the test items (Ackerman, Geirl & Walker, 2003). These models can produce ability parameter estimates that correspond to the measured skills (Reckase, 2009). MIRT applications have become increasingly common, as test items typically measure more than one skill.

Various MIRT models have been developed and are generally classified as either compensatory or noncompensatory models. In compensatory models, high levels of individual ability in one dimension can make up (i.e., compensate) for lower ability in another dimension. Noncompensatory models are harder to estimate, particularly if exploratory analysis is required (Chalmers & Flora, 2014), and so compensatory models are more commonly used in the field.

MIRT models can also be differentiated based on the number of item parameters involved. If only the item difficulty parameter *d* is involved, the MIRT model will be deemed to belong to the one-parameter model family; for those that belong to two-parameter model family, the item discrimination parameter vector ***a*** will be included in addition to *d*; and for those that belong to the three-parameter model family, the pseudo-guessing parameter *c* will be included in addition to ***a*** and *d*.

The compensatory two-parameter logistic (2PL) MIRT model introduced by McKinley and Reckase (1982) is widely used. Here, the probability of an individual *i* answering item *j* correctly is given by the formula:

$$p(\boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{1}{1+\exp(-D\sum_{k=1}^{m}(a_{jk}\theta_{ik})+d_j)},$$

where D = 1.7 is the measurement constant; θ_{ik} is individual *i*'s *k*th ability parameter; a_{jk} and d_j are the *k*th discrimination parameter and difficulty parameter of item *j*, respectively; and *m* is the number of dimensions.

Cognitive Diagnosis Model (CDM)

Other families of psychometric models called cognitive diagnosis models (CDMs) also available in the pertinent literature. These models were developed to be used in conjunction with cognitively diagnostic assessments (de la Torre & Minchen, 2014). The main purpose of CDM is to determine whether individuals have mastered the attributes or skills measured by the test. As such, CDMs classify individuals based on their mastery profiles, which can be used to identify learning deficiencies. CDM research has recently increased, as CDMs are more effective for measuring finer-grained skills than IRT models (Rupp, Templin & Henson, 2010; von Davier & Lee, 2019).

CDMs classify individuals into latent categories, which are determined by the presence or absence of the measured skills. This classification is based on the individuals' skills, or estimated mastery status, in terms of the measured attributes. The mastery of an attribute is represented by 1, while nonmastery is represented by 0 represents. A correct response to an item signals mastery of the attributes required to correctly respond to the item. A high proportion of correct responses to items requiring a specific attribute may indicate that an individual has already mastered this attribute (Rupp, Templin & Henson, 2010).

The Q-matrix, a common feature of CDMs, is used to define associations between measured attributes and test items. In a Q-matrix, items are placed in rows and attributes in columns. The Q-matrix is essential in a CDM and plays an important role in defining individuals' attribute profiles, as the Q-matrix clarifies the attribute requirements of each item (de la Torre & Minchen, 2014).

CDMs are commonly classified based on whether the measured attributes are dichotomous or polytomous in nature. Dichotomous attributes are those specified as either required (i.e., 1) or not required (i.e., 0) for correct responses to items in the Q-matrix. Similarly, the attribute profile estimates of individuals are represented by either 0 (i.e., nonmastery) or 1 (i.e., mastery) when the measured attributes are dichotomously scored. If the attributes are polytomously scored, different levels of measured attributes may be required for a successful response to an item, and individuals may have mastered the attributes at different levels. For example, for an attribute with three categories, there may be nonmastery (i.e., 0) along with two mastery levels (i.e., 1, 2). Polytomous attributes may thus reflect different levels of item difficulty associated with the different levels of the measured skills.

Models that consider polytomous attributes can be viewed as extended versions of those that consider dichotomous attributes. These extended models are more flexible and can address problems that generally dichotomous models cannot. Thus, dichotomous models have been generalized to polytomous models. The polytomous G-DINA (pG-DINA: Chen & de la Torre, 2013) model is an example of polytomous CDMs, and is the polytomous version of the *generalized deterministic inputs, noisy "and" gate* model (G-DINA; de la Torre, 2011).

Polytomous Generalized Deterministic-Inputs, Noisy "And" Gate (pG-DINA)

General CDMs can be reduced to specific CDMs by applying restrictions. General, unrestricted models are referred to as saturated, and specific restricted models as reduced (de la Torre & Lee, 2013). For example, the G-DINA is a saturated model, from which several reduced models, such as *deterministic inputs, noisy "and" gate* (DINA; Junker & Sijstma, 2001) and *deterministic inputs, noisy "or" gate* (DINO; Templin & Henson, 2006) can be derived. Similarly, the pG-DINA has been proposed as a saturated model and can be reduced to restricted polytomous models through various constraints.

The pG-DINA first reduces the number of possible attribute vectors into *reduced attribute vectors* by considering only the attributes required by an item. It then further reduces the number of attribute vectors into *collapsed attribute vectors* by only considering the levels of the required attributes. The number of reduced attribute vectors is computed by $M^{K_j^*}$, where *M* represents the attribute level and K_j^* the number of attributes required by item *j*. The number of collapsed attribute vectors is equal to the dichotomous G-DINA case and defined by $2^{K_j^*}$. For example, consider an item that measures two of the three *K* = 3 attributes, each with three *M* = 3 levels, as in, 0, 1, and 2. Assume further that this item requires levels 2 and 1 of the first and second attributes, respectively. Thus, the q-vector for this item is (2 1 0). The original, reduced, and collapsed attribute vectors, as defined by Chen and de la Torre (2013), are given in Table 1.

Table 1 shows that $3^{K_j^*} = 3^2 = 9$ reduced attribute vectors are obtained when only considering the attributes that are required by item *j*. Similarly, the collapsed attribute vectors are obtained by comparing the attribute levels in the reduced attribute vectors to those specified in the q-vector of item *j*. When an attribute level of the reduced attribute vector is equal to or higher than the level specified in the q-vector, it is represented by 1 in the collapsed attribute vector. Otherwise, it is represented by 0. The number of collapsed attribute vectors in this example then reduces to $2^{K_j^*} = 2^2 = 4$.

Table 1. Reduced and Collapsed Attribute Vectors for Original Attribute Vectors

Original $\boldsymbol{\alpha}_{lj}$	Reduced Attribute Vector ($\boldsymbol{\alpha}_{lj}$*)	Collapsed Attribute Vector ($\boldsymbol{\alpha}_{lj}^{**}$)
(0,0,0), (0,0,1), (0,0,2)	(0,0)	(0,0)
(1,0,0), (1,0,1), (1,0,2)	(1,0)	
(0,1,0), (0,1,1), (0,1,2)	(0,1)	(0,1)
(1,1,0), (1,1,1), (1,1,2)	(1,1)	
(0,2,0), (0,2,1), (0,2,2)	(0,2)	
(1,2,0), (1,2,1), (1,2,2)	(1,2)	
(2,0,0), (2,0,1), (2,0,2)	(2,0)	(1,0)
(2,1,0), (2,1,1), (2,1,2)	(2,1)	(1,1)
(2,2,0), (2,2,1), (2,2,2)	(2,2)	

The probability of success associated with the collapsed attribute vector or latent group $\boldsymbol{\alpha}_{lj}^{**}$ computed using the pG-DINA function is as follows:

$$P(\boldsymbol{\alpha}_{lj}^{**}) = \delta_{jo} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'>k}^{K_j^*} \sum_{k=1}^{K_j^*} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \cdots + \delta_{j1,\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lK_j^*}^{**}.$$

The interpretations of the model parameters are the same as those for the dichotomous attribute cases in the G-DINA model. Whereas the pG-DINA model uses the collapsed attribute vectors given in Table 1, the fully additive model (fA-M; Yakar, de la Torre, & Ma, 2017), another polytomous CDM, considers the reduced attribute vectors.

Fully Additive Model (fA-M)

If restrictions are applied to the saturated pG-DINA model, it can be reduced to a polytomous additive CDM (pA-CDM; Chen & de la Torre, 2013). This pA-CDM is derived from the pG-DINA by setting all interaction effects to zero. The intercept and the main effects of the mastered attributes required by the item are summed in the pA-CDM to obtain the probability of a correct response to the item. The fA-M can also be considered as an additive and restricted model. The main difference between these two models is the latent classes for which they compute the item response functions. The pA-CDM only considers the collapsed attribute vectors like the pG-DINA model, whereas the fA-M considers reduced attribute vectors.

Although fA-M is a restricted model, incorporating the reduced rather than the collapsed latent class in the item response function distinguishes it from many other CDMs. Rather than all-or-none, the fA-M considers the contributions of all levels (i.e., 0, 1, 2,…), as in, it considers the levels of the polytomous attribute in computing the probability of a correct response. This characteristic indicates that the model mimics the compensatory MIRT model, as the higher level of skills (i.e., attributes) leads to a higher probability of a correct response. The item response function of fA-M is given as

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \sum_{m=1}^{M_k} \delta_{jkm}\alpha_{lk}^{*},$$

where δ_{j0} is the intercept, δ_{jkm} is the (main) effect of the m^{th} level of attribute k, K_j^* is the number of required attributes, and M_k is the highest level of attribute k.

A characteristic common to CDMs and MIRT models is that both can be used with multidimensional scales. In addition, both theories contain compensatory and noncompensatory models (Reckase, 2009; Rupp, Templin, & Henson, 2010). These similarities indicate that these model families can be used to estimate multiple attributes or abilities. The type of the item structure in these models are also common, as they can be simple or complex in both CDMs and MIRT models, which is of particular importance in the analyses. However, these families of models differ in terms of other features, such as item parameters, the nature of the person parameters, which can be continuous or discrete, and the measurement units used.

The similarities between these psychometric models imply that deciding which model to use can be an issue of high consideration. Under some analysis conditions, fitting various models to the data may provide different points of view and lead to a deeper understanding – comparing the results obtained from different models that have similar infrastructures can extend our understanding of the focal phenomenon. Thus, evaluating the outputs of CDMs and MIRT models together can be of value.

To obtain additional information, a model that does not share psychometric properties with the tests used to gather the data may be fitted to the data. This process, referred to as retrofitting, and can be used to obtain potentially different information that supports or refutes existing knowledge about the data. The results of a retrofitting analysis may be much more valuable when the true and retrofitted models have similar structures, as with the CDM and MIRT models.

A literature review reveals that many studies have focused on retrofitting CDM to IRT data, and vice versa. Various CDMs are retrofitted to data obtained via tests that have been developed for IRT purposes (Ardıç, 2020; Chen & Chen, 2016; Chen & de la Torre, 2014; Lee, Park & Taylan 2011; Liu, Huggins-Manley, & Bulut, 2018; Şen & Arıcan, 2015). Other studies (de la Torre & Karelitz, 2009; Wang, 2009) involve reciprocal retrofitting CDMs and MIRT models. However, no retrofitting study that focuses on polytomous CDMs has been identified. Therefore, a significant contribution of the current study is the reciprocal retrofitting of three models: two CDMs and one MIRT model.

Purpose of the Study

The aim of this research was to examine the level of information obtained through retrofitting two specific CDMs and a MIRT model. We addressed this through the following sub-problems:

1- What levels of accuracy can be obtained for the person parameter classification and ability level estimation from the two CDMs and one MIRT model when they are fitted to the MIRT data generated under various item discrimination, item structure, correlation between skills, and test length conditions? Is there a difference between the person parameter estimation accuracy levels of the models?

2- What levels of accuracy can be obtained for the person parameter classification and ability level estimation from the two CDMs and one MIRT model when they are fitted to the CDMs data generated under various item discrimination, item structure, correlation between skills, and test length conditions? Is there a difference between the person parameter estimation accuracy levels of the models?

METHOD

Research Type

Experimental or theoretical studies that do not have any apparent specific application or use, and are primarily carried out to obtain novel information on the basis of phenomena and observable facts are defined as basic research (OECD, 2002). This study can be considered basic research as the aim is to assess the comparability of the results from fitting a MIRT model and two CDMs to various data. The data were generated using the models considered, and the analytic performance of the retrofitted models and the generating models were then examined.

Data Generation

Item discrimination, item structure, test length, and correlation between skills were manipulated to obtain various conditions simulation conditions. Three levels of item discrimination were specified and the generated discrimination parameters were drawn from uniform distributions, as in, $a \sim U(0.6, 0.8)$, $U(0.9,1.1)$, and $U(1.5,1.7)$ for the low, moderate, and high item discrimination conditions, respectively. The item structure was defined in terms of item complexity (i.e., whether the item measures one or more dimensions/attributes). In this research, an item is said to have a simple structure if it measures only one dimension/attribute, and a complex structure, otherwise. Tests with Q-matrices consisted of 20%, 50%, and 80% simple structure items were considered to have mostly complex, equal, and mostly simple item structures, respectively. In terms of the test length condition, the three levels of test length (i.e., short, medium, and long) consisted of 15, 30, and 60 items, respectively. The two levels of correlation (i.e., no relationship and moderate relationship) were created by setting the correlation between skills to .00 and .60. Although the correlation cannot be zero in real data cases and under compensatory models, this value was nonetheless considered because it reflects a situation in which a relationship is not present, which may provide a better understanding of the parameters in its related state. In terms of factor selection and their levels, the conditions used in other similar studies (Chen & de la Torre, 2013; Wang, 2009) and factors affecting model performance were considered. The study was conducted with 25 replications, as analyzing polytomous attribute data takes longer than when using dichotomous attribute data (de la Torre & Douglas, 2004; de la Torre & Douglas, 2008; Huebner & Wang, 2011). Thus, in the three models, three-item discrimination levels, three-item structure levels, three test length levels, and two correlation levels were crossed to yield $3 \times 3 \times 3 \times 3 \times 2 = 162$ conditions. With 25 replications for each condition, a total of $162 \times 25 = 4050$ data were generated and analyzed using the two CDMs and the MIRT model.

Generation of MIRT data

For each level of crossed factors, two-dimensional 2PL MIRT data were generated using the R program. For this data generation, the ability parameters followed a multivariate normal distribution, and the attribute levels in the polytomous CDMs indicated the item difficulty levels. Specifically, the Q-matrix entries of the polytomous CDMs were transformed into the item difficulty parameters. When generating the data, the item difficulty parameters were obtained by multiplying each element of the Q-matrices by 0.67 and subtracting 1.34 from each. Accordingly, the levels of 0, 1, 2, and 3 in the Q-matrix correspond to the difficulty levels of -1.34, -0.67, 0, and 0.67, respectively. As 0 in a Q matrix stands for an unmeasured attribute, the discrimination parameters of the items with the difficulty parameter of -1.34 were set to zero to ensure that the item parameters of the CDMs and the MIRT model were as matched as possible.

The continuous person parameters in MIRT were converted to discrete attribute levels in order to obtain classification accuracy rates that can be compared. By applying the cut-off points (i.e., -0.67, 0, and 0.67) to the MIRT person parameters, discrete values of 0, 1, 2, and 3 were obtained for each dimension resulting in individuals being classified into approximately four equal groups for each dimension. The sample size was set to 5000 to obtain more stable item and person parameter estimates.

Generation of CDM data

Two CDMs were used in this study: fA-M and pG-DINA models. As the item parameters of these CDMs differed in terms of number and structure, a two-dimensional 2PL MIRT data of 100000 examinees was initially generated to obtain related conditions for the CDMs. Item parameters compatible with the fA-M and pG-DINA model were then obtained in the R environment for two attributes. A self-written R code and the GDINA package (Ma & de la Torre, 2016) were used to generate the data for the fA-M and pG-DINA model, respectively.

Data Analysis

The generated data in MIRT were analyzed using the MIRT package (Chalmers, 2012). Person parameter estimates were obtained based on the expected a posteriori (EAP) method. The estimated person parameters were converted into discrete variables, similar to the generated person parameters in order to obtain the classification accuracy rates of the person parameter estimates under MIRT conditions. Analyses of MIRT data in CDM were performed using the GDINA package (Ma & de la Torre, 2016) for pG-DINA cases and through a self-written R code for fA-M cases.

Although the data in the MIRT estimation of the CDM data were originally based on 2PL, the relative fit of 2PL and 3PL MIRT models were both checked. After the data were analyzed in both models, ANOVA tests on deviance indices were conducted through R. If the difference was statistically significant ($p < .05$), the parameters of the 3PL model were considered. In general, the 3PL model was observed to fit better with the data.

After the analyses, the correct vector classification rates (CVCR) of the person parameters were obtained. If the estimated and generating ability/attribute vectors of an examinee matched, the examinee was considered to be accurately classified by the estimating model. The ratio of the number of accurately classified examinees to the total number in a dataset (i.e., 5000) provides the CVCR of the model. The average CVCR of the study was obtained across 25 replications. The significance of the differences between the CVCRs across models was tested through ANOVA. Since violation of the equality of variance assumptions, pairwise comparisons of groups were performed using the Tamhane procedure.

The ability/attribute-level accurate classification rates reflect the degree to which each dimension/attribute level is accurately estimated by the model. It, therefore, reflects the performance of the model at the individual ability/attribute levels. Accordingly, the averages of the correct attribute level classification rates (CALCRs) of the ability/attribute levels of all examinees on two abilities/attributes were obtained.

Data were generated for each model under 54 different conditions by crossing the main factors. As these factors are independent of each other, no interaction between different conditions was identified; thus, the CVCR averages at different levels of the basic conditions were reported rather than at the level of the crossed conditions. The findings can thus be effectively presented and interpreted. The CVCR averages across the conditions and repetitions are presented in Appendix.

RESULTS

This section presents the results of retrofitting the CDMs to the MIRT data, as stated in the first sub-problem, and of retrofitting the MIRT model to the CDM data, as stated in the second sub-problem.

Results of the MIRT Data Analysis

The CVCRs obtained by analyzing the MIRT data are presented in Table 2. The table shows that the highest CVCRs are obtained for the MIRT data when the fitted model was the MIRT model, followed by the fA-M. The CVCRs ranged from .41 to .60 when the MIRT model (i.e., the generating model) was fitted, and from .36 to .52 when the fA-M model was retrofitted to the data. The lowest levels of correct classification rates were observed when the pG-DINA model was retrofitted to the data, where the CVCRs ranged between .26 and .33. These findings suggest that the CVCRs of the MIRT analyses and the fA-M retrofitting results were comparable, which were different from the CVCRs of the pG-DINA analyses.

Table 2. CVCRs Obtained from the MIRT Data

Condition	Level	MIRT	pG-DINA	fA-M
Item discrimination	Low	0.43	0.27	0.40
	Moderate	0.50	0.29	0.45
	High	0.58	0.33	0.48
Item structure	Mostly complex	0.45	0.27	0.38
	Equal	0.52	0.29	0.43
	Mostly simple	0.54	0.33	0.51
Test length	15	0.41	0.28	0.36
	30	0.50	0.30	0.44
	60	0.60	0.32	0.52
Correlation between abilities	0.00	0.48	0.26	0.38
	0.60	0.53	0.33	0.50

As the item discrimination increased, the correct classification rates of all three models also increased. Moving from lower to higher discrimination levels, the increment for the correct classification performance of the MIRT analyses (.15) was larger than those of the CDM cases (.06 for pG-DINA and .08 for fA-M model). Similarly, regardless of the models, higher CVCRs were observed under mostly simple item conditions. The average increments in the CVCRs of the MIRT, pG-DINA, and fA-M model conditions were .09, .06, and .13, respectively.

In terms of the test length condition, an increase in the test length improved the CVCRs – the mean CVCRs increased from .41 to .60 and from .36 to .52, respectively, when the MIRT model and fA-M were fitted to the data. A relatively smaller increase (i.e., from .28 to .32) was observed when the fitted model was the pG-DINA.

The CVCRs also tended to increase when the abilities were correlated. This increase was larger for the retrofitted CDMs, particularly for the fA-M.

The ANOVA test results presented in Table 3 demonstrate the differences among the CVCRs of all three models when they were fitted to the MIRT data. The results indicate a significant difference between the CVCRs obtained through the analysis of the MIRT data [$F(2,4047) = 1592.984$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method reveals that the CVCR of the MIRT is significantly higher than those of the CDMs ($p < .001$). Similarly, the CVCR of the fA-M is significantly larger than that of the pG-DINA model ($p < .001$).

Table 3. Test of the Difference Between CVCRs of the MIRT Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	30.807	2	1592.984*	MIRT>fA-M
Within group	39.132	4047		MIRT>pG-DINA
Total	69.939	4049		fA-M>pG-DINA

**p*<.001

The attribute-level correct classification rates of all three models fitted to the MIRT data are presented in Table 4. The most significant results observed for the pG-DINA models were that the attribute-level CALCRs for levels 0 and 3 were very high (i.e., .96), but the CALCRs of levels 1 and 2 were very low (i.e., .10). Although the CALCRs in attribute levels 0 and 3 were also higher than those in attribute levels 1 and 2 when the MIRT model and fA-M were fitted to the data, the difference was not as dramatic. For attribute levels 0 and 3, the MIRT and fA-M CALCRs are .79 and .74, respectively; the corresponding CALCRs for attribute levels 1 and 2 are .61 and .56, respectively.

Table 4. CALCRs in the Analyses of MIRT Data

Model	Level 0	Level 1	Level 2	Level 3
MIRT	0.79	0.61	0.61	0.79
pG-DINA	0.96	0.10	0.10	0.96
fA-M	0.74	0.56	0.56	0.74

Results of the pG-DINA Data Analysis

The CVCRs obtained from the analysis of the pG-DINA data are presented in Table 5. The table shows that the largest CVCRs are obtained for the pG-DINA data when the fitted model was the generating model (i.e., pG-DINA model), followed by the fA-M. The CVCRs vary from .53 to .90 when the pG-DINA model (i.e., the generating model) was fitted, and from .51 to .89 when the fA-M model was fitted to the data. The lowest correct classification rates are observed when the MIRT model was retrofitted to the data – the CVCRs vary between .31 and .56. These findings suggest that the CVCRs of the pG-DINA model and the fA-M are comparable (i.e., the maximum difference is .02), whereas those of MIRT were quite different.

The correct classification rates for all three models increased with the item discrimination. Moving from lower to higher discrimination levels, the increment for the correct classification performance of the MIRT analyses (.19) was slightly lower than that of the pG-DINA model and fA-M (i.e., .24 and .23, respectively). When the items became simpler, an apparent increase in the CVCRs of MIRT was observed (i.e., .21), whereas at most, a slight increase (i.e., .02 in pG-DINA cases) was observed when CDMs were used in the data analysis. In terms of the test length, the CVCRs increased with the test length – the mean CVCRs improved from .53 to .90 and from .51 to .89 when the pG-DINA model and fA-M were fitted to the data, respectively. A relatively smaller increase (i.e., from .32 to .56) was observed when the fitted model was the MIRT. In addition, the CVCRs also increased when the abilities were correlated, although only to a very limited extent.

Table 5. CVCRs Obtained in the Analyses of pG-DINA Data

Condition	Level	pG-DINA	MIRT	fA-M
Item Discrimination	Low	0.60	0.35	0.59
	Moderate	0.72	0.44	0.72
	High	0.84	0.54	0.82
Item Structure	Mostly Complex	0.73	0.31	0.70
	Equal	0.72	0.50	0.71
	Mostly Simple	0.71	0.52	0.70
Test Length	15	0.53	0.32	0.51
	30	0.73	0.45	0.72
	60	0.90	0.56	0.89
Correlation between Abilities	0	0.71	0.44	0.70
	0.6	0.73	0.45	0.72

Table 6 displays the ANOVA test results of the observed differences between the CVCRs of all three models when they were fitted to the pG-DINA data. The results indicate a significant difference between the CVCRs obtained from the pG-DINA data analysis [$F(2,4009) = 1010.622$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method revealed that the CVCR of the MIRT was significantly lower than those of the CDMs ($p < .001$). In addition, the CVCRs of the fA-M were not significantly different from those of the pG-DINA model ($p > .05$).

Table 6. Test of the Difference between CVCRs of the pG-DINA Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	65.273	2	1010.622*	pG-DINA> MIRT
Within group	129.464	4009		fA-M > MIRT
Total	194.737	4011		

**p*<.001

Table 7 presents the attribute-level correct classification rates of all three models when they are fitted to the pG-DINA data. CALCRs of pG-DINA model and fA-M were significantly larger than the CALCRs of the MIRT model. Although the CALCRs in attribute levels 1 and 2 were lower than those in attribute levels 0 and 3 (i.e., the smallest is .81 and the largest .87), the largest CALCRs were obtained when the fitted model was the generating model (i.e., pG-DINA). These were followed by the CALCRs obtained when the fA-M was fitted, which is more uniform across attribute levels (i.e., the smallest is .80 and the largest .81). The lowest CALCRs were observed when the fitting model was MIRT, and the lowest and highest are .56 and .71, respectively.

Table 7. CALCRs in the Analyses of pG-DINA Data

Model	Level 0	Level 1	Level 2	Level 3
pG-DINA	0.87	0.81	0.80	0.87
MIRT	0.65	0.56	0.60	0.71
fA-M	0.81	0.80	0.80	0.81

Results of the fA-M Data Analysis

The CVCRs obtained from the analysis of the fA-M data are presented in Table 8. The table shows that the largest CVCRs were obtained for the fA-M data when the fA-M was fitted, and these varied from .44 to .80. The minimum and maximum CVCRs of the MIRT model, when it was retrofitted to the fA-M data under various conditions, were .39 and .71, respectively. The lowest correct classification rates were observed when the pG-DINA model was fitted to the data – the CVCRs varied between .24 and .34. The CVCRs of the MIRT and fA-M models were comparable; however, the performance of pG-DINA model was relatively poor.

Table 8. CVCRs Obtained in the Analysis of fA-M Data

Condition	Level	fA-M	MIRT	pG-DINA
Item Discrimination	Low	0.50	0.46	0.26
	Moderate	0.61	0.55	0.29
	High	0.74	0.65	0.34
Item Structure	Mostly Complex	0.59	0.47	0.28
	Equal	0.63	0.57	0.29
	Mostly Simple	0.63	0.61	0.31
Test Length	15	0.44	0.39	0.27
	30	0.61	0.55	0.29
	60	0.80	0.71	0.32
Correlation between Abilities	0	0.61	0.55	0.30
	0.6	0.62	0.55	0.29

In terms of the effects of the examined factors on CVCRs, the correct classification rates of all three models increased with the item discrimination. This increment was largest for the fA-M (i.e., .24), followed by the MIRT model (i.e., .19), and the smallest increment was for the pG-DINA model (i.e., .08). Similarly, regardless of the models, higher CVCRs were observed as the number of simple items in a test increased. The average increment for the CVCRs of the MIRT model (i.e., .14) model was relatively higher than the increments observed under the pG-DINA model (i.e., .03) and fA-M (i.e., .04).

In terms of the test length, an increase in the test length resulted in a rise in the CVCRs. The observed increments in CVCRs were very large in the fA-M and MIRT cases, which increased from .44 to .80 and from .39 to .71 when the fA-M and MIRT models were fitted, respectively. However, the increase for the pG-DINA model was limited (i.e., from .27 to .32). In addition, no remarkable changes in CVCRs were observed when the skills/attributes were correlated.

Table 9 presents the ANOVA test results of the observed differences between the CVCRs of all three models when they were fitted to the fA-M data. The table shows a significant difference between the CVCRs obtained from the analysis of the fA-M data [$F(2.4047) = 1934.53$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method revealed that the CVCR of the fA-M was significantly higher than those of the pG-DINA and MIRT models ($p < .001$). Similarly, the CVCR of the MIRT was significantly higher than that of the pG-DINA model ($p < .001$).

Table 9. Test of the Difference between CVCRs of the fA-M Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	77.937	2	1934.53*	fA-M>pG-DINA
Within groups	81.521	4047		fA-M> MIRT
Total	159.458	4049		MIRT>pG-DINA

**p*<.001

Table 10 presents the attribute-level correct classification rates of all three models when they were fitted to the fA-M data. The results in this table are comparable to those for the MIRT data given in Table 4. In the pG-DINA cases, CALCRs for levels 0 and 3 were very high (i.e., .98), whereas the CALCRs for levels 1 and 2 were very low (i.e., .11). Although the CALCRs for attribute levels 0 and 3 were also higher than those for levels 1 and 2 when the fA-M and MIRT model were fitted to the data, the differences were not as dramatic for attribute levels 0 and 3, the MIRT and fA-M CALCRs were .82 and .85, respectively, whereas the corresponding CALCRs for levels 1 and 2 were .63 and .68, respectively.

Table 10. CALCRs in the Analyses of fA-M Data

Model	Level 0	Level 1	Level 2	Level 3
fA-M	0.85	0.67	0.68	0.84
MIRT	0.82	0.62	0.63	0.81
pG-DINA	0.98	0.11	0.11	0.98

DISCUSSON and CONCLUSION

This study aimed to examine how the MIRT model and polytomous CDMs, the pG-DINA model (Chen & de la Torre, 2013) and the fA-M (Yakar, de la Torre, & Ma, 2017) when retrofitted to data generated with different underlying processes. Data for each model were generated with varying item discrimination, item structure, test length, and correlation between ability conditions. The data were then fitted with all three models, and the CVCRs of the generated person parameters were examined.

For the first sub-problem, the data generated using the MIRT model were, as expected, most accurately estimated by the MIRT; the fA-M estimation was the next best, and the lowest performance was observed for the pG-DINA model. The results from MIRT and fA-M can be explained due to the use of reduced latent groups in fA-M, which follows a similar logic as the MIRT – a higher level of proficiency corresponds to a higher probability of answering the item correctly. The pG-DINA model is processed through the collapsed latent groups, and an increase in attribute level does not always produce an increase in the success probability, unlike in the fA-M, where every increase in attribute level results in an increase in the success probability.

The highest level of efficiency was obtained from the test length condition in the MIRT data analysis, followed by item discrimination and item structure; the effect of the correlation between abilities was limited compared to other factors. However, the item structure was only effective in the MIRT data or estimation. In addition, the findings revealed that the effect sizes of the conditions may differ in the CDMs. The pG-DINA was less sensitive to changes in the conditions, and the fA-M was more affected by item structure and correlations between skills than the MIRT model. Wang (2009) conducted a reciprocal retrofitting of the reduced reparametrized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002) and the MIRT model, and found the estimation accuracy varied according to the item structure and item discrimination. This is consistent with the fA-M results found in the present study. In a different study, where reciprocal retrofitting of one-dimensional IRT and DINA models were examined, de la Torre and Karelitz (2009) found that item discrimination greatly affected the estimation accuracy. Again this is similar to the fA-M results. These results suggest that common factors may affect the performance of different but compatible models in situations involving reciprocal retrofitting.

For the second sub-problem, the analysis of the pG-DINA data indicated that the pG-DINA accurately estimated its own data. The accuracy rates were the highest obtained in the study. The rates obtained from the fA-M were very close to the pG-DINA, and no statistical difference between the results was found. The similar CVCRs of the fA-M and the pG-DINA model when fitted to pG-DINA data is remarkable and provided the best retrofitting results; however, the CVCRs for the MIRT were substantially lower than those of the two CDMs. This finding is consistent with outcomes for the first research problem and suggests that the MIRT model cannot be retrofitted to pG-DINA data, and vice versa.

Although the outcomes were not identical, the successful estimation of the pG-DINA data when fitted with the fA-M may be due to the interaction effects in pG-DINA being substituted with the main effects for each level. The models do not need to have exactly the same item parametrization to produce similar results as different parameters can adjust and fill the gaps when changes in model parametrization occur. To this end, models that contain more item parameters can be more flexible and advantageous. Although the CVCRs of the MIRT were relatively poor for the pG-DINA data, these results were close to the values obtained when fitted to its own data. Thus, at least for the current setup, the MIRT model may not be expected to provide good classification results.

The results of the pG-DINA data analysis revealed that longer test length and higher item discrimination improved the CVCRs of all of the models. In addition, simplifying the item structure resulted in an increase in CVCR of the MIRT model only. Another remarkable result is that the item structure may have limited impact when CDMs are fitted to pG-DINA data.

In the analysis of the fA-M data, it was found that, as in other models, the classification rate was best in the correct model fitted to the data. However, the results of the MIRT model were almost at the same level as those of the fA-M. Similar results were found in the MIRT data analysis. This suggests that the MIRT model and fA-M may be used interchangeably in situations similar to those examined in the current study. As in the MIRT data analysis, the pG-DINA results were again found to have the lowest rates. In terms of the factors considered, all except the correlation between the dimensions had a substantial impact on the CVCRs.

The relatively low CVCRs of the pG-DINA when retrofitted to MIRT and fA-M data were due mainly to the very low CALCRs observed in the two middle attribute levels. A closer inspection (not presented) showed the pG-DINA model had a tendency to misclassify middle attribute levels as extreme attribute levels. In the study, we discretized the continuous abilities to create a uniform distribution. Poor retrofit performance may be worsened if the abilities have a normal distribution. The poor performance of the retrofitted pG-DINA model may be due to the assumption invoked by the model to create the collapsed latent classes.

When fitting the correct model to the data, the MIRT was found to have lower CVCRs than the pG-DINA model or the fA-M. The poor results suggest that estimating MIRT data may be more challenging. Moreover, the original person parameters of the MIRT are continuous but were made discretized for comparison purposes. The loss of information due to this transformation may have negatively affected the results.

It is worth noting that the CVCRs obtained in retrofitting the fA-M to the pG-DINA data were unexpectedly higher than those obtained in fitting the model to its own data. A similar situation was found for MIRT model – the MIRT estimations of the MIRT data were less accurate than those of the fA-M data. In their retrofitting studies, de la Torre and Karelitz (2009) and Wang (2009) found similar results. These results suggest that the underlying processes in generating the different data vary in complexity. To the extent that the findings can be generalized, the underlying process of the pG-DINA model is the simplest, followed by the fA-M, and the MIRT model has the most complex underlying process.

Overall, fitting a model that corresponds to the true underlying process produced the best results, whereas fitting a wrong model can lead to slightly or substantially poorer results depending on the extent of the mismatch. Of the three models, the fA-M was relatively robust to the possible mismatch between the true and fitted models; the same cannot be said of the pG-DINA and MIRT models. Although model-data fit still needs to be evaluated, fitting the fA-M to real data appears to be a safer option.

A limitation of the current study pertains to how the abilities were converted to attributes. Although the fA-M can be used to extract diagnostic information for polytomous attributes from MIRT data, and vice versa, these results may only be true when abilities are discretized in a particular manner. Future studies should consider other ways of establishing the comparability of the MIRT model and fA-M in order to arrive at more general conclusions. It should be noted that this study does not suggest that the MIRT model and fA-M can be used interchangeably – as pointed out repeatedly, fitting the true model will always produce the best results. To this end, further studies are needed to establish which procedures can be used to identify the best model when inferences about polytomous attributes are of interest.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Ardıç, E. Ö. (2020). *Bilişsel tanı ve çok boyutlu madde tepki modellerinin sınıflama doğruluğu ve parametrelerinin karşılaştırılması [Comparison of classification accuracy and parameters of cognitive diagnostic and multidimensional item response models]*. Unpublished PhD Dissertation, Hacettepe University, Ankara.

Chalmers, R. P., (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218-230.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419-437.

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology, 5*(18), 1967-1978.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450-469.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355-373.

de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89-97.

DiBello, L.V. Roussos L. A., & Stout, W. (2007). *Review of cognitively diagnostic assessment and a summary of psychometric models*. Rao, C. Sinharay, S. (Eds.) Handbook of Statistics, Psychometrics. Vol. 26. North-Holland: Amsterdam.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*, Unpublished PhD dissertation, University of Illinois at Urbana-Champaign.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144-177.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357-383.

Ma, W. & de la Torre, J. (2016). GDINA: The generalized DINA model framework. R package version 0.9.2.

McKinley R. L. & Reckase M. D. (1982) *The use of the general Rasch model with multidimensional item response data* (Research Report: ONR 82-1). American College Testing, Iowa City, IA.

Organisation for Economic Co-operation and Development. (2002). Frascati Kılavuzu. Paris: OECD.

Reckase, M. D. (2007). Multidimensional item response theory. Rao, C. Sinharay, S. (Ed.) *Handbook of Statistics,* Psychometrics. Vol. 26. North-Holland: Amsterdam.

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

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Şen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 238-253.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

von Davier, M., & Lee, Y. S. (2019). Introduction: From latent classes to cognitive diagnostic models. In *Handbook of Diagnostic Classification Models* (pp. 1-17). Springer, Cham.

Wang, Y. C. (2009). *Factor analytic models and cognitive diagnostic models: How comparable are they? – A Comparison of R-RUM and compensatory MIRT model with respect to cognitive feedback.* Unpublished PhD dissertation, The Faculty of The Graduate School at The University of North Carolina at Greensboro).

Yakar, L., de la Torre, J., & Ma, W. (2017). *An empirical comparison of two cognitive diagnosis models for polytomous attributes*. In the Annual Meeting of National Council on Measurement in Education. National Council on Measurement in Education (NCME), San Antonio, TX.

Appendix. CVCR Averages for Crossed Conditions

Conditions				True Model	Retrofitted Model		True Model	Retrofitted Models		True Model	Retrofitted Models	
Correlation	Test Length	Item Structure	Item Disc.	MIRT	pG-DINA	fA-M	pG-DINA	MIRT	fA-M	fA-M	MIRT	pG-DINA
0	15	M. Complex	Low	.26	.19	.24	.39	.21	.37	.32	.28	.23
0	15	M. Complex	Moderate	.31	.20	.26	.51	.26	.48	.40	.33	.26
0	15	M. Complex	High	.39	.21	.26	.63	.29	.60	.49	.39	.30
0	15	Equal	Low	.31	.21	.27	.38	.24	.37	.34	.31	.23
0	15	Equal	Moderate	.38	.23	.29	.51	.31	.49	.43	.39	.26
0	15	Equal	High	.47	.25	.27	.66	.43	.63	.55	.49	.30
0	15	M. Basic	Low	.32	.23	.31	.39	.26	.38	.35	.33	.23
0	15	M. Basic	Moderate	.41	.27	.37	.51	.37	.50	.44	.43	.28
0	15	M. Basic	High	.52	.33	.42	.66	.53	.65	.57	.55	.35
0	30	M. Complex	Low	.34	.22	.29	.58	.27	.57	.45	.40	.25
0	30	M. Complex	Moderate	.40	.22	.31	.72	.31	.71	.57	.46	.28
0	30	M. Complex	High	.47	.23	.28	.85	.36	.83	.69	.54	.32
0	30	Equal	Low	.40	.24	.35	.58	.36	.56	.50	.47	.26
0	30	Equal	Moderate	.49	.25	.38	.72	.47	.71	.62	.56	.28
0	30	Equal	High	.59	.28	.33	.86	.63	.85	.76	.68	.33
0	30	M. Basic	Low	.43	.25	.40	.56	.38	.54	.49	.49	.26
0	30	M. Basic	Moderate	.52	.28	.47	.71	.51	.71	.62	.60	.28
0	30	M. Basic	High	.63	.33	.59	.86	.73	.86	.77	.75	.37
0	60	M. Complex	Low	.43	.23	.36	.81	.32	.80	.63	.53	.27
0	60	M. Complex	Moderate	.50	.23	.37	.92	.35	.91	.76	.61	.31
0	60	M. Complex	High	.59	.24	.35	.98	.45	.97	.88	.71	.34
0	60	Equal	Low	.52	.25	.43	.80	.52	.80	.68	.64	.28
0	60	Equal	Moderate	.60	.27	.48	.91	.67	.91	.81	.74	.31
0	60	Equal	High	.69	.30	.40	.97	.82	.97	.92	.84	.37
0	60	M. Basic	Low	.55	.28	.48	.77	.56	.76	.68	.66	.29
0	60	M. Basic	Moderate	.64	.31	.58	.90	.73	.89	.82	.79	.33
0	60	M. Basic	High	.72	.43	.69	.97	.60	.97	.93	.89	.42
0.6	15	M. Complex	Low	.39	.29	.38	.43	.23	.42	.33	.29	.23
0.6	15	M. Complex	Moderate	.43	.31	.41	.55	.25	.54	.41	.33	.26
0.6	15	M. Complex	High	.48	.33	.44	.66	.29	.66	.51	.39	.29
0.6	15	Equal	Low	.39	.30	.38	.41	.25	.41	.35	.32	.24
0.6	15	Equal	Moderate	.45	.31	.43	.53	.31	.52	.44	.40	.26
0.6	15	Equal	High	.52	.34	.48	.68	.43	.66	.57	.49	.30
0.6	15	M. Basic	Low	.39	.30	.38	.40	.26	.39	.35	.33	.24
0.6	15	M. Basic	Moderate	.45	.33	.44	.52	.37	.52	.45	.43	.28
0.6	15	M. Basic	High	.55	.37	.53	.68	.54	.67	.58	.56	.34
0.6	30	M. Complex	Low	.45	.30	.44	.65	.27	.64	.47	.40	.26
0.6	30	M. Complex	Moderate	.49	.31	.48	.77	.30	.77	.59	.46	.28
0.6	30	M. Complex	High	.56	.33	.49	.88	.34	.88	.72	.54	.31
0.6	30	Equal	Low	.47	.31	.46	.63	.37	.62	.51	.47	.26
0.6	30	Equal	Moderate	.54	.33	.52	.75	.48	.75	.63	.56	.29
0.6	30	Equal	High	.62	.36	.57	.88	.64	.88	.78	.68	.33
0.6	30	M. Basic	Low	.47	.32	.45	.58	.39	.56	.50	.49	.26
0.6	30	M. Basic	Moderate	.55	.34	.52	.73	.52	.73	.63	.60	.28
0.6	30	M. Basic	High	.64	.40	.62	.87	.73	.87	.78	.75	.37
0.6	60	M. Complex	Low	.51	.31	.50	.86	.31	.86	.65	.53	.28
0.6	60	M. Complex	Moderate	.56	.32	.54	.95	.34	.95	.78	.61	.30
0.6	60	M. Complex	High	.63	.33	.53	.99	.44	.99	.89	.70	.31
0.6	60	Equal	Low	.56	.32	.52	.84	.53	.84	.69	.63	.27
0.6	60	Equal	Moderate	.62	.34	.59	.93	.67	.93	.83	.73	.31
0.6	60	Equal	High	.70	.38	.64	.98	.83	.98	.93	.83	.34
0.6	60	M. Basic	Low	.57	.33	.52	.79	.57	.78	.69	.67	.28
0.6	60	M. Basic	Moderate	.65	.36	.61	.91	.73	.91	.83	.79	.33
0.6	60	M. Basic	High	.73	.46	.71	.98	.64	.98	.94	.88	.41

"True model" indicates the estimation of data belonging to the models. The subsequent two columns indicate the retrofitting estimations of the true model data.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 112-128

PISA 2015 Reading Test Item Parameters Across Language Groups: A measurement Invariance Study with Binary Variables*

Pelin BAĞDU SÖYLER ** Burak AYDIN *** Hakan ATILGAN ****

Abstract

Large-scale international assessments, including PISA, might be useful for countries to receive feedback on their education systems. Measurement invariance studies are one of the active research areas for these assessments, especially cross-cultural and linguistic comparability have attracted attention. PISA questions are prepared in the English language, and students from many countries answer the translated form. In this respect, the purpose of our study is to investigate whether there is a measurement invariance problem across native English and non-native English speaker groups in the PISA-2015 reading skills subtest. The study sample included students from Canada, the USA, and the UK as the native speaker group and students from Japan, Thailand, and Turkey as the non-native speaker group. Measurement invariance studies taking into account the binary structure of the data set for these two groups revealed that eight of the twenty-eight items in the PISA-2015 reading skills test had possible limitations in equivalence.

Key Words: PISA 2015, measurement invariance (MI), binary variables, reading skills.

INTRODUCTION

Internationally conducted student assessments play an essential role in the educational policies of countries. One of these assessments is administered by the OECD (Organization for Economic Cooperation and Development) (Milli Eğitim Bakanlığı-MEB, 2016). The OECD is an institution that plays a vital role in the regulation of the welfare of the world, economic development, and educational policies; it carries out many studies in line with its goals. One of these studies is the International Student Assessment Program (PISA), which is one of the most extensive educational researches in the world implemented internationally. PISA assessments are carried out regularly in fields of mathematics, science, and reading skills. In PISA, the concept of literacy is handled as special equipment used to fulfill a function in life practices. In this extensive study at the international level, equivalence studies are extremely important for ensuring the validity of the measurement instrument. PISA develops different cognitive measurement instruments to measure student performance at all levels in the fields of science and mathematics and contextual measurement instruments (OECD, 2018). One of the main assumptions in this practice, which closely concerns educational policies by comparing student achievements between countries, is that the measured structures are the same for all participants. Construct validity should be ensured by minimizing bias to make valid comparisons between different language groups and countries. Martin, Mullis, Gonzales, Gregory, Garden, O'Connor, Chrostowski and Smith (2000) emphasize the necessity of neutrality while comparing student achievement among countries. Accordingly, construct validity has distinctive importance.

*This work is based on the first author’s master thesis and preliminary results were presented at AERA 2021 Conference.

**Ministry of Education, İzmir-Türkiye, pelinbagdu@gmail.com, ORCID ID: 0000-0001-8169-2165
***Assoc. Prof., Ege University, Faculty of Education, İzmir-Türkiye, burak.aydin@ege.edu.tr, ORCID ID: 0000-0003-4462-1784
**** Prof., Ege University, Faculty of Education, İzmir- Türkiye, hakan.atilgan@ege.edu.tr ve 0000-0002-5562-3446, ORCID ID: 0000-0002-5562-3446

To cite this article:
Bağdu Söyler, P., Aydın, B., & Atılgan, H. (2021). PISA 2015 reading test item parameters across language groups: A measurement invariance study with binary variables. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 112-128. doi: 10.21031/epod.800697

Received: 29.09.2020
Accepted: 09.06.2021

Baykal and Circi (2010) conducted a material revision study to improve the structure validity of PISA 2006 in science testing, and the authors concluded that the different characteristics of the countries should be taken into consideration in stages of item development and translation into different languages by examining the construct validity. Accordingly, it was seen that in international applications such as PISA, the tests are not understood by all participating countries in the same way. Generally, the active role of PISA in national education policies is based on the general assumption that PISA tests are reliable and valid instruments; therefore, this acceptance provides an international comparison of student performances. Researches on this have shown that there are many factors such as translation, item content, curriculum differences, exam motivation or exam anxiety, writing system, and culture. Linguistic diversity affects the comparability of scores and consequently may limit the validity of these studies (Arffman, 2002; Bonnet, 2002; Grisay & Monseur, 2007; Hambleton, Merenda & Spielberger, 2005; He & van de Vijver, 2012; Kreiner & Christensen, 2014). PISA questions are prepared in English and are used by translating the languages of the countries whose native language is not English. The native language of most of the participating countries is not English, so non-native English-speaking countries use the tests translated into their language. Since PISA significantly affects the educational policies of countries, it is extremely important that the psychometric structure measured between countries and different groups is comparable (Brown, 2006). Scalar equivalence is required to compare the scores obtained from different language versions of the tests in a significant and valid way (Ercikan & Lyons-Thomas, 2013). In order to compare individuals from different cultures and languages in different subject areas, especially in a direct language-dependent area such as reading skills, it is a critical issue to have no equivalence problems in the structures measured by the tests and to ensure the measurement invariance of the tests.

Arffman (2010) identified six types of problems that limit the equivalence of PISA reading texts. These were language-specific differences in grammar, language-specific differences in writing, language-specific differences in meaning, differences in culture, translators' choices and strategies, and problems with editing. Accordingly, it is important that the questions are accessible in terms of examining the factors that limit the equivalence of the items and understanding these problems. Based on the analysis of PISA 2006 reading items, Kreiner and Christensen (2014) pointed out that the validity of the measurement model was inadequate due to items with differential item functioning (DIF). As a result, it was not appropriate for countries to compare as such. Some critics have suggested that the PISA reading texts, to some extent, support Western countries, consistent with previous cultural and linguistic concerns. (Grisay et al., 2007; Grisay, Gonzalez & Monseur, 2009; Oliveri & von Davier, 2011). Since countries with similar linguistic and cultural histories are likely to hold the equivalence in scores, it is predicted that the MI may be a problem for PISA assessments. (Asil & Gelbal, 2012; Kankaras & Moors, 2013).

In the literature, there are many MI studies in PISA student surveys. Asil and Gelbal (2012) investigated MI in terms of culture and linguistics in PISA 2006 student survey. Results revealed that as the cultural and linguistic differences between countries increase, the number of DIF items increases. Segeritz and Pant (2013) studied the Learning Approaches of Students (SAL) scale in the PISA 2003 in Germany sample among ethnic-cultural groups in a country. The findings obtained with the results have shown that the factor structure of the scale Learning Approaches between Germany and two immigrant student groups is comparable.

The equivalence of PISA tests between countries in terms of cultures and language is questionable. The main criticisms point to linguistic and cultural bias, potentially affecting the nature of reading tests. Therefore, the comparisons between countries raise doubts about accuracy. Literacy performance is influenced by a set of characteristics such as the nature of each language, the writing system used to stimulate literacy, the cultural style, teaching and learning approaches, and level of investment in socio-economic development and education (Asil & Brown, 2015).

MI of the cognitive data has been tested, and the cultural comparability correlations of the cognitive data have been examined by taking the technical reports as reference. It was concluded that comparing the total scores across different cultures may lead to incorrect results.

International large-scale applications such as PISA, TIMMS, and PIRLS aim to measure latent structures among participants and compare between groups. However, when these assessments participating in many countries are taken into consideration, some evidence has been obtained that the method is not practical in such large-scale assessments (Rutkowski & Stevina 2013; Ogretmen, 2006). Rutkowski and Stevina (2013) conducted a simulation study to investigate the change depending on the sample size and the number of groups of multi-group confirmatory factor analysis (MG-CFA) performance. In order to mimic real data, the data were simulated ordinal categorical and analyzed with a linear model. In the findings obtained, it was concluded that there is an inconsistent relationship between a sequential categorical data set and the linear model, so this method selection is not an excellent theoretical practice. In the findings obtained, it was concluded that there is an inconsistent relationship between an ordinal categorical data set and the linear model, so this method is not the right choice in theory. Readers are referred to Jöreskog, Sörbom, Toit and Toit (2001), Sirganci, Uyumaz and Yandi (2020), Gregoric, (2006), Salzberg et al. (1999) , Önen (2009), Wu, Li & Zumbo (2007), and van de Schoot et al. (2013) for further reading on MG-CFA.Therefore, there is an operational need for the suitability of comparisons across countries. In PISA 2015, a recent approach has been applied for MI testing using item response theory (IRT) item consistency (OECD, 2016). Thus, the question raised about the reproducibility of these findings in the context of more common analysis techniques.

In order to compare individuals from different cultures with international measurement instruments, it is essential to hold the equivalence of their forms in different languages when the measurement instrument is translated into other languages. Therefore, measurement invariance is one of the most needed studies in cross-cultural comparisons of multiple groups. It is one of the preconditions to make correct decisions in terms of language skills of cultures and cross-language equivalence in a study playing a significant role in the educational policies of countries such as PISA. Thus, the construction validity studies are very important for the evidence of the validity of the measurement instrument. There are several studies in the literature regarding the MI of PISA; however, it is remarkable that many of the MI analyses ignore the binary nature of the PISA's data sets. PISA questions consist of multiple-choice and partial answer items. In assessments involving such items, it is crucial to perform the MI studies carefully using an appropriate method for the binary nature of the data set in order to achieve valid results.

Measurement Invariance with Different Variable Types

MI studies provide evidence of the structural validity of the measuring instrument. The equivalence of the characteristic of a psychological measurement instrument, such as construct validity and reliability, in different groups is defined as the measurement invariance (Herdman,1998). Whether the psychological structure to be measured is comparable between groups in terms of different cultural factors or variables is essential for the validity. MI means that a measurement model has the same structure in multiple groups, and the factor structures and error variances of the items in the scale are equivalent (Bollen, 1989).

Evaluation of MI within common factor linear models is known as factorial invariance. When the linear factorial model is used in data sets involving binary, ordered, and Likert-type variables, the structure of the observed variables are ignored (Elosua, 2011). In order to test the MI, the chi-square difference test is used. However, the models are different for continuous and ordinal categorical datasets, so testing the MI between groups requires testing the parameters for each model (Meredith, 1993). While the related parameters are factor loadings and residual variance in a dataset containing continuous variables, the thresholds are required to compare between groups in an ordinal categorical dataset. Using the maximum likelihood estimation (ML) and continuous linear models to analyze ordered categorical datasets involves some disadvantages and uncertainties about the resource of invariance (Lubke & Muthén, 2004). French and Finch (2006) concluded that the chi-square difference test in evaluating measurement invariance was inadequate in a data set containing multidimensional binary categorical items. Instead of the linear factor analysis commonly used for continuous variables, the variables in the ordered categorical structure can be modeled with MG-CFA in accordance with the threshold structure (Kim & Yoon, 2011). Since linear CFA is not a suitable analysis for ordered categorical data, the MI test cannot be sufficiently compared with linear CFA (McDonald, 1999; Oishi, 2006; Reise et al., 1993). Meade and Lautenschlager (2004) stated that in some cases, the IRT approach could give different and potentially more useful information for modeling MI.

Without modeling the threshold structure, CFA assumes that the underlying distributions of dichotomous or polytomous variables are normal. Threshold values are mathematically related to item difficulty parameters in IRT (Lord & Novick, 1968; Takane & de Leeuw, 1987). Accordingly, ordered categorical CFA with the appropriate analysis method based on IRT to test the MI with ordered categorical variables gives more accurate results than linear CFA without considering the threshold structure (Kim & Yoon, 2011). It should be noted that, especially in PISA assessments, cognitive tests have a binary categorical structure, and attitude scales include Likert-type variables. In other words, analyzing categorical data using methods developed for continuous variables has serious limitations in general (Raykov, Marcoulides & Milsap, 2013).

Measurement Invariance with Binary Variables

It has been demonstrated in recent studies that the methods commonly used in MI studies have limitations. As mentioned previously, the MG-CFA method is frequently used for continuous, and Likert-type scored variables. Raykov, Dimitrov, Li, Marcoulides & Menold (2018) suggested an alternative method for testing the MI with binary scored items. This method aims to determine cases that do not hold the MI with item factor loadings and threshold values. The recent approach does not require defining a reference variable and allows us to study the MI directly with one or two-parameter IRT modeling (Raykov et al., 2018).

IRT suggests that the performance of a person in a test can be predicted according to the item characteristic curve that shows the relationship between the latent traits or abilities (Hambelton and Swaminathan, 1985). IRT is concerned with the participants` responses to each item rather than the total score received from the test. Two item parameters can be used to define the item characteristic curve, which is the basis of IRT. One of these is item difficulty (*b*), and the other is item discrimination (*a*) index. Item difficulty states where the item is functional. For example, while an easy item is more functional for individuals with lower ability, a difficult item is more functional for individuals with higher ability levels. The item discrimination index states how well it characterized individuals who are below the ability level of the item and individuals with an ability level above this point (Baker, 2016).

Assume y = (y_1, y_2,... y_k) represents the components of a psychological scale. In addition, it is assumed that the component *y* discharges the conditions of structural invariance in groups with large samples (Millsap, 2011). In this setting, a factor analysis model has been developed in each group in which *a* parameter with loadings and *b* parameter with thresholds are related. Hence, the necessary conditions for *y* component and MI of the g^{th} group are represented as follows;

$$y_g^* = \Lambda_g \eta_g + \delta_g \tag{1}$$

$$\Lambda_1 = \Lambda_2 = \ldots = \Lambda_g \tag{2}$$

$$\tau_1 = \tau_2 = \ldots = \tau_g \tag{3}$$

The pair of Equations 2 and 3 also represents a necessary condition to study a two-parameter IRT model or the DIF, a special case of it (Muthén, Asparouhov & Morin, 2015). DIF states that the probability of responding to the test item correctly is not an equality case in individuals with the same ability level and from different groups (Adams & Rowe, 1988). DIF analysis aims to investigate whether test scores are affected by variations from different groups and whether these variations give rise to a bias for any subgroup (Algina & Crocker, 1986). If the attribute measured by the test is the same in different subgroups, it can be seen that the items are affected by the same variability and that individuals with the same ability level are similar in the measured structure (Algina & Crocker, 1986). The MI analysis method in the binary scored items used in our study provided to test the MI by determining the items under the two-parameter IRT.

Purpose of the Study

The purpose of this study is to examine whether the PISA 2015 reading skills subtest is equivalent in terms of language skills for countries with native English and non-native English speakers.In order for comparisons and assessments to be valid, equivalence across cultures and languages should hold. Scales developed in a particular culture and language reflect characteristics of that culture and language. Translating a measurement instrument does not warrant that these two scales are equivalent (Sireci & Berberoğlu, 2000). It should be noted that the measurement instrument to be translated or adapted to another language will differ from its original form. These differences should be ensured to be acceptable in terms of psychometric properties (Hambleton & De Jong, 2003). In such a study that plays an essential role in the educational policies of countries, the intercultural equivalence of the tests in terms of language skills is one of the preconditions for making the right decisions (Arffman, 2010; Baykal & Circi, 2010; He, Barrera-Pedemonte & Bucholz, 2018). In this respect, it is very important to investigate construct validity carefully for the proof of the validity of the measuring instrument. Hence in this study, whether the reading skills test of the PISA 2015 assessment has MI problem between the translated language form and the original one has analyzed by statistical analysis methods.

METHOD

Sample sizes of PISA 2015 participant countries included in our study are 14157 from the UK, 5712 from the USA, 20058 from Canada, 6647 from Japan, 8249 from Thailand, and 5895 from Turkey. In PISA, not all students take the same test, and test forms contain common questions as well as different questions (OECD, 2016). A total of 64171 students from selected countries participated in the study. In PISA 2015, 66 different forms were prepared for countries that received computer-based tests. In our study, data from the 41st form were used given that it was the most frequently used form for Canada, UK, the USA, Japan, Thailand, and Turkey. Reading skills achievement was measured in this form with 28 items. The frequencies of the participants who took the 41st form in the sample by country are reported in Table 1.

Table 1 shows that the country with the highest number of participants is Canada with 34.4%, and the USA has the lowest number of participants with 8.9%. The sample of the study consists of 1524 students taken the 41st form from six countries separated out of the countries participating in PISA 2015. The countries included in the research were selected from the countries participating in the PISA 2015 with a computer-based assessment. Therefore, 28 items with the most responded form number 41 selected among 66 different forms were included in the analysis. This form included open-ended and multiple-choice questions. According to the type of question, the items are coded with *0* refers to false responses, *1* refers to partially correct responses, and *2* refers to correct ones. Since the model did not converge with only two partially scored items, the partially correct scores were treated as correct, and items 5 and 6 are re-coded as 0 for incorrect and 1 for correct responses. In our study, the ratio of the missing value to the total sample size was only 6%, considered low (Kline, 2016, p.83) and hence ignorable (Akbaş & Tavşancıl, 2015; Cheema, 2012; Downey and King, 1998; Rubin, 1976; Enders, 2010), and it was decided to exclude the missing data from the analysis to ease the model convergence.

Data Analysis

A single factor model was tested using CFA for each group. The item parameters obtained with separate CFA were examined. The full measurement invariance approach allows the item factor loadings and threshold values between the comparative groups to be the same, and the approximately defined measurement invariance approach allows only small differences in the parameters in question between the compared groups (Kim, Cao, Wang, & Nguyen, 2017). Muthén and Asparouhov (2013) bring in the term of approximate measurement invariance as a stage of measurement invariance, in addition to full invariance and partial invariance, with recent studies (van de Schoot et al., 2013).
Findings obtained in this direction have been reported.

Table 1. Sample Sizes Based On Countries

COUNTRY	N	%
Canada	524	34.4
UK	384	25.2
Thailand	176	11.5
Japan	145	9.5
Turkey	159	10.4
USA	136	8.9
Total	1524	100

The countries included in this study are separated into two groups as native English (UK, Canada, USA) and non-native English speakers (Japan, Thailand, and Turkey). MI for binary scored items was tested using the M*plus* 8.0 (Muthen & Muthen, 2019). In this direction, item loadings and threshold parameters were free for each item in MI analysis. The difference in BIC values (ΔBIC) between the baseline model (M_0) and the free model in each model were studied. The smaller the BIC value, the better the model-data fit (Nylund, Asparouhov & Muthén (2007). The model with ΔBIC> 10 indicates a strong misfit of the model, and such values are considered a threat to MI (Frank J., Fabozzi & Wiley, 2014).

RESULTS

In the first step, CFA was completed in accordance with the nature of binary variables for each group, and the model fit was examined. The model data fit findings obtained with CFA are presented in Table 2.

Table 2. Confirmatory Factor Analysis Results of Reading Skills Test PISA 2015

Group (Countries)	Chi-Square value	n	RMSEA	CFI	TLI
Native English Speakers	409.58*	1044	.03	.96	.97
Non-Native English Speakers	243.86*	480	.03	.97	.98

**p*<.05

When the model fit indices in Table 2 are examined, it is seen that the chi-square value is significant in both groups (*p* <.05). Based on the RMSEA values, it can be understood that the model fits perfectly in both groups since it is .03 for both groups. Concerning CFI and TLI fit indices, it is seen that the CFI value for the native language group indicates a strong fit with .96 and the TLI value with .97. The CFI and TLI values for the non-native English also indicate a strong fit with .97 and .98. CFA results indicated that the one-factor structure of PISA 2015 Reading Skills Test holds for both groups separately. Item factor loadings, threshold values, *a* and *b* parameters obtained as a result of the CFA analysis are showed in Table 3.

Table 3. Item Parameters Regarding CFA Results for the Groups Consisting of PISA 2015 Reading Skills Test Language Variable

	Native English Speakers				Non-Native English Speakers			
Item	ʎ	t	a	b	ʎ	t	a	b
1	1.00	-0.81	0.64	-1.50	1.00	-0.38	0.95	-0.56
2	1.01	-1.13	0.65	-2.07	0.99	-0.35	0.93	-0.51
3	1.06	-1.26	0.70	-2.20	0.88	-0.64	0.76	-1.06
4	1.10	-1.07	0.74	-1.80	0.65	-0.62	0.51	-1.37
5	1.23	-0.90	0.88	-1.36	0.61	-0.41	0.46	-0.97
6	1.25	-0.92	0.91	-1.36	1.03	-0.51	1.02	-0.71
7	1.18	0.67	0.82	1.06	1.03	0.76	1.01	1.06
8	1.00	0.33	0.64	0.61	0.83	0.69	0.71	-1.24
9	1.31	-1.15	0.99	-1.64	0.84	-0.72	0.71	-1.24
10	0.88	0.79	0.54	1.68	0.98	0.98	0.93	1.45
11	1.17	-0.06	0.82	-0.09	0.83	0.22	0.70	0.39
12	0.99	-0.08	0.63	-0.14	0.51	-0.13	0.38	-0.36
13	1.19	-0.79	0.84	-1.23	0.97	-0.54	0.90	-0.81
14	0.90	-0.06	0.56	-0.12	0.70	0.08	0.55	0.17
15	0.50	0.37	0.28	1.36	0.45	0.53	0.32	1.73
16	0.63	-0.24	0.36	-0.70	0.60	0.08	0.46	0.20
17	1.07	-0.86	0.71	-1.48	0.76	-0.76	0.62	-1.26
18	1.31	-0.88	0.99	-1.25	0.76	-0.69	0.61	-1.33
19	0.91	-0.07	0.56	-0.14	0.67	0.10	0.53	0.22
20	1.22	-0.41	0.87	-0.62	1.02	0.17	0.99	0.24
21	0.20	-0.14	0.85	-0.22	1.06	0.30	1.07	0.41
22	0.99	-0.26	0.63	-0.49	0.85	-0.47	0.72	-0.81
23	0.87	-0.83	0.53	-1.77	1.02	-0.34	0.98	-0.48
24	0.81	0.16	0.48	0.36	0.84	0.39	0.71	0.68
25	0.79	-0.10	0.47	-0.21	0.73	0.53	0.58	1.06
26	0.77	-0.94	0.46	-2.26	0.60	-1.15	0.45	-2.78
27	1.02	-0.32	0.66	-0.57	0.53	-0.27	0.40	-0.64
28	1.26	0.58	0.92	0.86	0.96	0.73	0.88	1.10

Note: ʎ= item factor loading, t: threshold , a=item discrimination ,b=item difficulty

The item factor loadings, threshold values, *a* and *b* parameters obtained from the CFA to examine whether the item parameters of each group differ or not are given in Table 3. It is observed that the 21st item has the least factor loading in the group with native language English, whereas the group with non-English has the greatest factor loading. Accordingly, while factor loadings are expected to be approximately equal with each other for both groups, this case indicates that the item does not work in the same way for both groups. It is understood that the 9th and 18th items in the group with native language is English are the ones with the greatest factor loadings. The 15th item is the item with the least factor load (.45) in the group with non-native English and is close to the factor loading (.50) given by the other group in the 15th item. When the factor loadings and the parameters *a* of the 12th item are compared, the item factor loading of the group with native English is .99, and the parameter *a* is .63, whereas the item factor loading of the group with non-native English is .51 and the parameter a is .38. These values are substantially different for the items that are expected to measure the equal characteristic.

When we viewed the item threshold values and *b* parameters, whereas the threshold value is -1.13 for the threshold of the second item in the group with native English, in the group with non-native English, it is-.35, and *b* parameter is -2.07 in the group with native English; the group with non-native English is -0.51. These values are different for an item that should measure the same characteristic in both groups. Similarly, when the parameters of item 23 are compared in both groups, it is understood that the group with native English is -1.77 and -0.48 in the other group. The CFA results performed separately for the two groups are visually examined. It is difficult to say that items 2, 4, 6, 8, 9, 12, 15, 18, 21, 22, 23, 25, 26, 27, and 28 work similarly in psychometric terms. In order to examine whether the 15 differences determined visually are statistically significant, the variation of item parameters and BIC values in 56 different models were examined for the data set consisting of 28 items. The results of MI analysis in binary scored items for the groups with native and non-native English speakers are presented in Tables 4 and 5.

BIC values obtained from 56 different models to be free of item factor loading and thresholds for each item, their differences from the BIC value in the M_0 (ΔBIC) and item factor loadings and thresholds are given in Tables 4 and 5. The BIC values of the M_0 and the BIC values of each model were compared separately. The BIC value was found to be 44745.34 in M_0. The difference of BIC value in each model with BIC value of M_0 was calculated.

Table 4. Measurement Invariance Analysis of PISA 2015 Reading Skills Test Thresholds

				Group 1		Group 2	
Model	Par	BIC	ΔBIC	ʎ	t	ʎ	t
M1	t_1	44748.62	3.28	-	-1.69	-	-1.27
M2	t_2	44708.93	-36.41*	-	-2.35	-	-1.25
M3	t_3	44737.70	7.64	-	-2.69	-	-2.02
M4	t_4	44748.02	2.68	-	-2.21	-	-1.87
M5	t_5	44746.98	1.64	-	-1.84	-	-1.48
M6	t_6	44752.57	7.23	-	-2.14	-	-2.19
M7	t_7	44739.41	-5.93	-	1.66	-	1.00
M8	t_8	44751.38	6.04	-	0.71	-	0.87
M9	t_9	44752.39	7.05	-	-2.76	-	-2.66
M10	t_{10}	44752.62	7.28	-	1.65	-	1.61
M11	t_{11}	44751.68	6.34	-	-0.10	-	-0.24
M12	t_{12}	44731.95	-13.39*	-	-0.14	-	-0.72
M13	t_{13}	44749.29	3.95	-	-1.78	-	-2.10
M14	t_{14}	44748.84	3.50	-	-0.10	-	-0.34
M15	t_{15}	44752.65	7.31	-	0.64	-	0.65
M16	t_{16}	44748.44	3.10	-	-0.42	-	-0.17
M17	t_{17}	44749.87	4.43	-	-1.79	-	-2.05
M18	t_{18}	44745.93	0 .59	-	-2.01	-	-2.44
M19	t_{19}	44751.14	5.80	-	-0.13	-	-0.29
M20	t_{20}	44742.71	-2.63	-	-0.84	-	-0.37
M21	t_{21}	44751.50	6.16	-	-0.21	-	-0.71
M22	t_{22}	44691.77	-53.57*	-	1.13	-	-1.58
M23	t_{23}	44745.36	0.02	-	-1.66	-	-1.30
M24	t_{24}	44752.56	7.22	-	0.31	-	0.27
M25	t_{25}	44722.68	-22.66*	-	-1.15	-	0.58
M26	t_{26}	44725.27	-20.07*	-	-1.75	-	-2.66
M27	t_{27}	44743.97	-1.37	-	-0.59	-	-0.97
M28	t_{28}	44743.49	-1.85	-	1.32	-	0.82

Note: ʎ= item factor loadings; t=threshold; Grup 1: Native English Speakers Grup 2: Non-Native English Speakers

Table 5. Measurement Invariance Analysıs of PISA 2015 Reading Skills Test Item Factor Loadings

				Group 1		Group 2	
Model	Par	BIC	∆BIC	ʎ	t	ʎ	t
M29	ʎ$_{1}$	44747.36	2.02	1.15	-	1.57	-
M30	ʎ$_{2}$	44722.44	-22.90*	1.09	-	2.09	-
M31	ʎ$_{3}$	44741.84	-3.50	1.20	-	1.73	-
M32	ʎ$_{4}$	44752.67	7.43	1.19	-	1.20	-
M33	ʎ$_{5}$	44751.06	5.72	1.36	-	1.17	-
M34	ʎ$_{6}$	44752.55	7.21	1.68	-	1.74	-
M35	ʎ$_{7}$	44749.53	4.19	1.30	-	1.74	-
M36	ʎ$_{8}$	44752.60	7.26	1.09	-	1.14	-
M37	ʎ$_{9}$	44752.26	6.92	1.75	-	1.65	-
M38	ʎ$_{10}$	44746.56	1.12	0.92	-	1.49	-
M39	ʎ$_{11}$	44748.68	4.34	1.40	-	1.05	-
M40	ʎ$_{12}$	44731.21	-24.13*	1.06	-	0.44	-
M41	ʎ$_{13}$	44751.05	5.71	1.52	-	1.33	-
M42	ʎ$_{14}$	44752.08	6.74	0.89	-	0.78	-
M43	ʎ$_{15}$	44752.64	7.30	0.50	-	0.51	-
M44	ʎ$_{16}$	44750.76	5.42	0.59	-	0.77	-
M45	ʎ$_{17}$	44749.77	4.43	1.28	-	1.05	-
M46	ʎ$_{18}$	44736.14	-9.20*	1.75	-	1.15	-
M47	ʎ$_{19}$	44751.27	5.93	0.94	-	0.77	-
M48	ʎ$_{20}$	44749.49	4.15	1.46	-	0.86	-
M49	ʎ$_{21}$	44751.57	6.23	1.46	-	1.69	-
M50	ʎ$_{22}$	44736.79	-8.55*	1.18	-	0.63	-
M51	ʎ$_{23}$	44734.46	-10.88*	0.97	-	1.62	-
M52	ʎ$_{24}$	44748.02	2.68	0.75	-	1.11	-
M53	ʎ$_{25}$	44747.32	1.98	0.78	-	1.17	-
M54	ʎ$_{26}$	44739.60	-5.71	0.95	-	0.44	-
M55	ʎ$_{27}$	44736.40	-8.64*	1.12	-	0.58	-
M56	ʎ$_{28}$	44752.64	7.30	1.43	-	1.40	-

Note: ʎ= item factor loadings; t=threshold; Grup 1: Native English Speakers Grup 2: Non-Native English Speakers

Findings showed that ΔBIC value of the second item is -36.41 (ΔBIC> 10) in Model 2 and ΔBIC value is -22.90 (ΔBIC> 10) in Model 30. It is evaluated that the threshold values of the second item are quite different from each other, as -2.35 for the group (Group 1) with native English speakers and -1.25 for the group with non-native English speakers (Group 2). Accordingly, it can be said that the second item does not show the model fit and is not comparable for both groups. It is evaluated that ΔBIC value of item 18 in Model 46 is a poor fit with -9.20 (6 <ΔBIC <10). Item thresholds and *a*, *b* parameters have different values from each other, as seen in Table 3. Similarly, the ΔBIC value of item 22 in Model 22 is -53.57 (ΔBIC> 10), and in Model 50 this value is -8.55 (6 <ΔBIC <10). Table 3 is indicated that the parameters of these items differ from each other on the basis of both groups. Items 12, 23, 25, 26, and 27 also seem to have poor model fit. Therefore, it is evaluated that ΔBIC values of 8 in 28 items are not in the range of acceptable model fit, and item parameters differ parallel with these results.

DISCUSSION and CONCLUSION

In this study, the MI of the PISA 2015 Reading Skills Test in terms of the language variable between the countries with native English speakers and the countries with non-native English speakers was tested with binary scored items. For two groups with native and non-native English speakers, CFA was performed separately, and model fit was examined, and it was concluded that overall factor structures were confirmed for each group. Item parameters were compared in both groups with the findings obtained with CFA. It was understood that the factor loadings and threshold parameters of some of the items assumed to measure the same ability in both groups of the PISA 2015 Reading Skills test differ considerably from each other. Therefore, it was concluded that there could be a limitation for the comparability of the groups.

When the item thresholds and factor loadings of these items were compared, it was observed that there was a substantial difference. It was evaluated that 8 out of 28 items in the 41st form of PISA 2015 Reading Skills possibly limit the scalar equivalence. Such a limitation in at least one item means that the MI cannot be fully supported for the whole test (Raykov et al., 2018). Therefore, in this test, it can be concluded that the MI cannot be fully defensible without identifying sources that limit the comparison between English and non-native English groups. In the literature, there are similar MI findings. For example, Baykal and Circi (2010) studied the 2006 PISA science test. The authors asked teachers to evaluate the positive and negative properties of the items, an item evaluation form was created, and the items were categorized according to their content. Negative categories were determined according to culture-specific factors reflected in language, grammatical difficulties, unknown words, and expressions of sentences. Item revisions are completed based on the negative categories. A revised test was created by selecting 22 items from the Turkish version of the science test. With the revised science test, the original science test versions were administered to each of two equivalent groups consisting of 30 students. It was concluded that the group that took the language-wise revised test performed better in all the items compared to the group that took the original translation. A similar study by Asil and Brown (2015) compared the English version of the test and its versions translated into other languages of the PISA 2009 reading skills test. The authors reported that socio-economic factors significantly affect the MI, and linguistic factors are relatively less effective.

In international assessments such as PISA, the questions prepared in English are translated into another language by the expert translators and then translated back to English to ensure its equivalence with the original version. In order to study these factors carefully, information about the effects of the differences in culture and their reflections in the language should be obtained in measurement instruments (Goldstein, 2017). Items that are specific to a language and contain expressions causing bias should be excluded from the test. PISA 2015 science test items are not publicly available, the items that limited the MI could not be examined, and the differences between the results could not be studied in detail.

REFERENCES

Adams, R., & Rowe , K. (1988). *Educational research, methodology, and measurement:An international handbook.* Oxford: Pergamon Press.

Akbaş, U. ve Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi., *Journal of Measurement and Evaluation in Education and Psychology*, *6(*1), 38-57

Algina, J. & Crocker, L., (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Arffman, I. (2002). *In search of equivalence: Translation problems in international literacy studies.* Finland.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, *54*(1), 37-59.

Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. *Education ve Science*, 236-249.

Asil M., & Brown, G. (2015). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71-93.

Baker, F. B. (2016). *The basics of item response theory.* Ankara: Pegem Academy.

Baykal, A., & Circi, R. (2010). Item revision to improve construct validity: A study on released science items in Turkish PISA 2006 . *Procedia Social and Behavioral Sciences*, 2(2), 1931-1935.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387–399.

Brown , T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

Cheema, J. (2012). Handling missing data in educational research using SPSS. Unpublished doctoral dissertation. *George Mason University*.

Downey, R., & King, C. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, 175-191.

Elosua, P. (2011). Assessing Measurement Equivalence in Ordered-Categorical Data. *Psicológica*, 403-421.

Enders, C. K. (2010). *Applied missing data analysis.* (1. Ed.). New York: The Guilford Publications, Inc

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting test for use in other languages and cultures. *APA Handbook of Testing and Assesment in Psychology* (s. 545-569). içinde Washington: American Psychological Association.

Frank J. Fabozzi, S. M., & Wiley , J. (2014). Model Selection Criterion: AIC and BIC. *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications.*

French, B. F., & Finch, W. H. (2006). Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance. *Structural Equation Modeling*, *13*(3), 378-402.

Goldstein, H. (2017). Measurement and evaluation issues with PISA. *Routhledge*.

Gregoric, S. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 78-94.

Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249-266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *ERI Monograph Series: Issues and Methodologies In Large-Scale Asssesments*, 2, 63-84.

Hambelton, R. K., & Swaminathan , H. (1985). *Item Response Theory.* Nijhoff Publishing.

Hambleton, R. K., & De Jong, J. A. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing*, 127-134.

Hambleton, R. K., Merenda, P., & Spielberger, C. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence S. Erlbaum Publishers

Herdman M., Rushby J. F., & Badia X. (1998). A Model of Equivalence in The Cultural Adaptation of HRQol Instruments: The Universalist Approach. *Ouality Of Life Research*, 7(4), 323-335.

He, J., Barrera-Pedemonte, F., & Bucholz, J. (2018). Cross-cultural comparability of non-cognitive constructs in TIMSS and PISA. *Assesment in Education:Principles, Policy & Practice*, 26(4), 369-385.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*.

Jöreskog, K.G., Sörbom, D., Du Toit, S.H.C., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd ed.). Lincolnwood, IL: Scientific Software International.

Kankaras, M., & Moors, G. (2013). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, 43(3), 381-399.

Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling*, 212-228.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups:A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford publications.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 210-231.

Lord, F. M., & Novick, M. E. (1968). *Statistical theories of mental test scores* . MA: Addison-Wesley.

Lubke, G. H., & Muthén, B. O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons . *Structural Equation Modeling A Multidisciplinary Journal*, *11*(4), 514-534.

Martin, M., Mullis, I., Gonzalez, E., Gregory, K., Smith, T., Chrostowski, S., O'Connor, K. (2000). TIMSS 2009 International Science Report: Findings from IEA's Repeat of The Third International Mathematics and Science Study at the Eight Grade.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah: NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*,7(4), 361-388.

Milli Eğitim Bakanlığı (2016). *PISA 2015 International report.* Ankara. [Online: https://odsgm.meb.gov.tr/www/2015-pisa-ulusalraporu/icerik/204], 2016.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Pyschometrika*.

Millsap , R. E. (2011). *Statistical approaches to measurement invariance.* New York: US: Routledge/Taylor & Francis Group.

Muthén, B., and Asparouhov, T. (2013). *BSEM measurement invariance analysis. Mplus Web Notes: No. 17*. Available online at: www.statmodel.com

Muthén , B., Asparouhov, T., & Morin, A. J. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal Of Management*.

Muthén, L. K., & Muthén, B. O. (2019). *Mplus user’s guide*. Los Angeles, CA: Muthén & Muthén.

Nylund, K.L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535-569.

Organisation for Economic Co-operation and Development (2016). Online: http://www.oecd.org/education/], 2016

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411-423.

Ogretmen, T. (2006). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği.* Ankara.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and AssessmentModeling*, *53*(3), 315-333.

Onen, E. (2009). *Ölçme değişmezliğinin yapısal eşitlik modellemesi teknikleri ile incelenmesi*. Ankara: Ankara University, Doctoral Thesis.

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. *Educational and Psychological Measurement*, *73*(4), 713-727.

Raykov, T., Dimitrov, D., Marcoulides, G., Li, T., & Menold, N. (2018). Examining Measurement Invariance and Differential Item Functioning With Discrete Latent Construct Indicators: A Note on a Multiple Testing Procedure. *Educational and Psychological Measurement*, *78*(2), 343-352.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552-566.

Rubin, D. B., (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Salzberg, T., Sinkovics, R., & Schlgelmich, B. (1999). Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches. *Australasian Marketing Journal*, 23-38.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, *13*(3), 229-248.

Sirganci, G., Uyumaz, G., & Yandi, A. (2020). Measurement invariance testing with alignment method: Many groups comparison. *International Journal of Assessment Tools in Education*, 657-673.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 1–15.

Wu, D., Li, Z., & Zumbo, B. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assesment, Research & Evaluation*, *12*(3), 1-26.

Appendix A. M*plus* 8.0 Syntax for CFA

```
TITLE: CFA for the first group (native English)
DATA: FILE IS ING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;
```

```
TITLE: CFA for the second group (non-English)
DATA: FILE IS NONING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;
```

Appendix B. M*plus* 8.0 Syntax for the MI with Binary Variables

M_0 base model:

```
TITLE: Raykov (2018) M0
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;
```

Example syntax to relase a threshold (M_1-M_{28}):

```
TITLE: Raykov (2018) M1 (relase first threshold)
!LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u2$1-u28$1](T2-T28);
[u1$1*];
[f1*];
f1*;
```

Example syntax to relase a loading(M_{29}-M_{56}):

```
TITLE: Raykov (2018) M29 (relase first loading)
!LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
f1 BY u1*
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 129-146

Investigation of Psychometric Properties of Likert Items with the Same Response Categories Using Polytomous Item Response Theory Models *

Esra SÖZER **   Nilüfer KAHRAMAN ***

Abstract
The purpose of this study was to investigate within- and between-threshold parameter invariance for items of a fourteen-item Positive Affect Scale developed to assess positive moods (like happy, peaceful, etc.) of university students. To test whether the estimated threshold parameters were as expected (1 to 5, with increments of 1) across all the 14 items, Graded Response, Partial Credit, and Rating Scale Models were fit the response data collected from 326 students. A comparison of the model fit statistics, such as the negative 2log likelihood and chi-square values, revealed that the Graded Response Model had the best fit and that the thresholds estimates for all the items in the Positive Affective Scale were reasonably close to the expected 1 to 5 values with increments of 1. The study illustrates how polytomous response models can be used to test the psychometric quality of items with ordinal rating scales.

Key Words: Item parameters, positive affect, polytomous, threshold, item response theory.

INTRODUCTION

When the response scales of the polytomous scored items are formulated, e.g., Likert scale, it is expected that respondents will choose the category that best describes their state given the measured trait. Even if it can be argued that this is a reasonable expectation, there remain several unanswered questions about how individuals' self-ratings compare amongst themselves, related to potential differences that may exist in the decision-making processes of the individuals when evaluating their state given the scale provided. The study of defining and testing for such individual differences has long been the focus of many scaling studies (e.g., Wang, Wilson, & Shih, 2006), all underlining the importance of a careful analysis of the scale properties of items, especially when subjective assessments are involved (Wang et al., 2006). Even when constructing ordinal scale assessment tools, the main objective of the psychometric work is about deriving the most accurate and meaningful information from the item responses (Wu & Adams, 2006).

Researchers studying traits from the affective domain do often face a greater number of challenges when evaluating the quality of their assessment results when compared to those who study traits from the cognitive domain, yet new methodological advancements rarely target their issues first. In this context, polytomous Item Response Theory (IRT) models, commonly used in calibrating items of most cognitive assessment tools, are yet to gain such common use when it comes to calibrating ordinal rating scale items, which are often used in the evaluation of psychological constructs, such as personality traits (Baker, Rounds, & Zevon, 2000). Given that assessment tools assessing psychological characteristics are, in general, composed of rating scale items, it would be most reasonable that polytomous IRT models are used in estimating non-linear relationships between the

* This study was presented in 6th International Congress on Measurement and Evaluation in Education and Psychology (5-8 September, 2018 in Prizren/Kosovo).

** Res. Assist. PhD., Bartin University, Faculty of Education, Bartin-Turkey, esrsozer@gmail.com, ORCID ID: 0000-0002-4672-5264
*** Prof. PhD., Gazi University, Gazi Education Faculty, Ankara-Turkey, kahramannilufer@gmail.com, ORCID ID: 0000-0003-2523-0155

To cite this article:
Sözer, E., & Kahraman, N. (2021). Investigation of psychometric properties of likert items with the same response categories using polytomous item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 129-146. doi: 10.21031/epod.819927

Received: 02.11.2020
Accepted: 16.04.2021

propensity level of the respondent and the likelihood of responding in a certain category (Embretson & Reise, 2000).

The prototypical Likert-type scale has five categories. These are printed equally spaced and equally sized on the response form (Figure 1). The intention is to convey to the respondent that these categories are of equal importance and require equal attention (Linacre, 2002). Response categories have an explicit and clear continuum and reveal the underlying psychological structures of these categories.

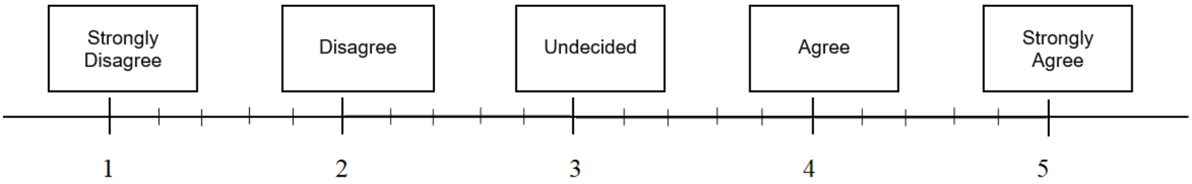


Figure1. Likert-Type Scale Response Categories

According to Linacre (2002), from a measurement perspective, the rating scale may appear in different forms (Figure 2). The rating categories still have a continuum and attempt to measure a psychological construct. Since the psychological construct intended to be measured conceptually is infinitely long, the two extreme categories are also infinitely wide. However, individuals are predominantly in the *agree* category. The size of intermediate categories such as *undecided* is dependent on how they are perceived and used by the respondents. *Agree* categories are usually more attractive than *disagree* categories. Therefore, *agree* categories may be represented by a wider interval for the measured psychological construct.

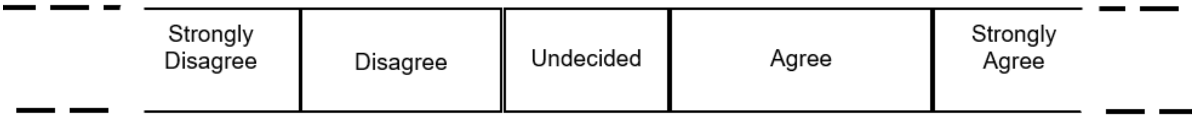


Figure 2. Typical Likert Scale Response Categories from Measurement Perspective

How the variable is divided into categories affects the reliability of a scale (Linacre, 2002). The rating categories with equal intervals as in Figure 1 or ordinal as in Figure 2 can be analyzed with polytomous IRT models. Polytomous IRT models are needed to represent the nonlinear relation between examinee trait level and the probability of responding in a particular category (Embretson & Reise, 2000). Polytomous models allow the use of different item discrimination values in weighting items, the estimation of measurement errors at each ability level, and achieving parameter invariance for the individuals and items (Lord, 1980).

Polytomous models vary based on whether the response categories are ordinal or non-ordered. In this case, in each model, the meaning of the response probability obtained for the response categories will also differ within the context of parameters that the model allows defining. The Graded Response Model (GRM; Samejima, 1969), one of the polytomous models used for modelling ordered response categories, the likelihood of marking each category or an upper category is modelled; while in Partial Credit Model (PCM; Embretson & Reise, 2000), the likelihood of scoring or choosing each category is directly modelled (instead of the category or an upper category).

In this study, category threshold parameters between consecutive categories estimated according to the GRM model used in the estimation of scale item parameters represent the ability level required for responding to the category and above with a probability of .50. According to PCM, the items are assumed to have equal discrimination (slope). In this case, the probability of an individual's responding to a category is computed as a function of the difference between an individual's ability level and the category threshold parameter (step difficulty). Unlike GRM, step difficulty parameters represent the relative difficulty of each step. According to Rating Scale Model (RSM), the last model used in the study, the location parameter estimated separately for each item reflects the relative easiness or difficulty of the particular item. In this model, it is assumed that the same response format is used for all items in the scale; therefore, category threshold values are estimated on an equal basis for all items. In RSM, the response likelihood of an item is determined by location parameter and category threshold parameter (Embretson & Reise, 2000).

Item response categories with different properties are analyzed with different measurement models mentioned above, and model-data fit is assessed. In addition to the assessment of a model-data fit, it is emphasized that the importance of including basic observations to determine to what extent the model fits the psychological reality that underlies the responses (i.e., response format) (Samejima, 1996). For this reason, it is important to determine the characteristics of the analyzed item response categories (whether the categories have a similar order for each item) and to what extent they fit the psychological structure they are trying to measure, in terms of the reliability and validity of the measurement results obtained.

A review of the literature showed that polytomous IRT models are widely used in analyzing psychometric properties of Likert-type rating scales (de Ayala, Dodd, & Koch, 1990; Koch, 1983). These models are also used for analyzing psychometric properties of measurement tools designed for measuring affective skills such as self-esteem (Gray-Little, Williams, & Hancock, 1997), emotional regulation (Rubio, Aguado, Hontangas, & Hernandez, 2007), self-identification (Flannery, Reise, & Widaman, 1995), emotional intelligence (Cho, Drasgow, & Cao, 2015), subjective well-being (Baker et al., 2000), self-reflection (Silvia, 2021), anxiety (Caycho-Rodríguez et al., 2021) as well as of those for measuring cognitive skills (Min & Aryadoust, 2021). Few studies were found in our country which employed polytomous IRT models for analyzing psychometric properties of measurement instruments used for emotional skills. It was found that polytomous IRT models were used for developing and adapting measurement tools like resilience scale (Yaşar & Aybek, 2019), attitude scale (Demirtaşlı, Yalçın, & Ayan, 2016); however, the properties of item response categories were not analyzed in many scale development and adaptation studies. This study focused on the importance of this issue and elucidated how the studies could be conducted in practice by exemplifying through a scale in the context of the use of polytomous IRT models in measuring constructs related to the affective domain such as subjective well-being.

Positive Affect Scale (PAS) used in this study is designed similarly to the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988), but it is a five-point graded (1-5, with increments of 1) Likert scale consisting of 14 positive affect items. These self-report constructs by which individuals assess themselves are considered substantial individual differences' variables for a long time (Hattie, 1992). Determination and improvement of positive affects of individuals such as subjective well-being, happiness, and resilience are among the main objectives of education environments. The responses to polytomous scoring items used for analyzing affective characteristics are based on subjective assessments by which individuals are assumed to select the categories which describe them best. At this point, the satisfaction of the assumption that the order between response categories in the scale used is the same for each item (e.g. evenness of threshold parameters between 1 and 2, 2 and 3, ...) and that the order between items refers to the same meaning is important for a reliable interpretation of measurement results (Koch, 1983).

Purpose of the Study

The purpose of this study was to investigate response categories of rating items (from 1 to 5) in a 14-item PAS scale developed to measure positive affects and to demonstrate the extent of similarities/differences between these categories regarding the items. It was aimed to obtain an estimation of item parameters for polytomous scoring items in PAS scale utilizing different polytomous models, analyze model-data fit and make a comparative evaluation of the measurement precision at different ability levels across the affect scale. Considering the polytomous response format of PAS and theoretical relationship between polytomous models and response processes, whether category threshold parameters used for determining responses to the items were ordered in inter-item was tested through GRM (Samejima, 1969), PCM (Embretson & Reise, 2000) and RSM (Andrich, 1978). Based on the requirements set out by each of these models, the validity of the assumption of invariance of category threshold parameters for all items was analyzed using the data in practice.

METHOD

This study is designed as a descriptive comparative study that analyzed psychometric properties of the PAS according to polytomous Item Response Theory models (Glass & Hopkins, 1984; Kaptan, 1995).

Study Group

The study group comprised 326 volunteer students (pre-service teachers) who studied at the Gazi Faculty of Education in the academic year 2017-2018. The study group included 166 female (51%) and 52 male (17%). The participants were in an age range of 19-35 years. Among these participants, 6 of them were 19 years old (1.8%), 77 were 20 years (23.6%), 92 were 21 years (28.2%), 24 were 22 years (24%), 7 were 23 years (2.1%), 3 were 24 (0.9%), 2 were 25 years (0.6%), 1 participant was 28 years (0.3%), 3 participants were 29 years (0.9%), 2 were 30 years (0.6%) and 1 participant was 35 years (0.3%) old. (Demographic information about the study group was obtained by a separate scale and was not mandatory. Therefore, the values for those whose information could be reached were presented.)

Data Collection Tools

The data used in this study come from a more comprehensive study called Emotion Ruler Field Study (Kahraman, Akbaş, & Sözer, 2019). Positive and Negative Affect Scale consists of 27 positive and negative affects. The individuals were asked to mark the best describe them among the response categories (from 1 for *very slightly or not at all* to 5 for *extremely*). According to the results of Exploratory Factor Analysis (EFA) for factor structure of the scale, a Kaiser-Meyer-Olkin (KMO) value was found to be 0.94. Chi-square ($\chi 2$) statistic and the result of Bartlett's test was statistically significant (χ^2 (351) = 5605.97, p < .05). The data were found to have a two-factor structure with eigenvalues of 11.13 and 3.53. The total variance explained by the factors was 51%. Confirmatory Factor Analysis (CFA) results used for verifying factor structure showed that model-data fit was at an acceptable level, and the scale had a two-factor structure (χ^2 (294) = 838.76, RMSEA= .08, CFI = .87, TLI = .86 and SRMR = .08). The results of Cronbach's Alpha correlation coefficients showed that the reliability for each factor was respectively for positive and negative affects .92 and .91.

In this study, the data came from positive affect items was employed. This sub-factor named PAS consists of 14 items that ask individuals to mark one of the response categories (from 1 for *very slightly or not at all* to 5 for *extremely*) for each item given to them. 14 positive affects included in the scale are as follows (Table 1): Happy, peaceful, contented, open to communication, understanding, motivated, resilience, strong, self-confident, determined, successful, optimistic, brave and energetic. Descriptive statistics for items are given in Table 1. Analyses for the factor structure of PAS are presented in the data analysis section.

Data Collection Procedure

Data for the PAS were collected from the participants through an online application. PAS consists of self-report items whereby individuals are asked to choose one of the response categories appropriate for them.

Table 1. Descriptive Statistics for Positive Affect Scale

Items	Mean	S.D.	Skewness	Kurtosis	r_{ij}*
1. Happy	3.17	0.95	-.34	-.04	.68
2. Peaceful	3.03	1.08	-.25	-.67	.63
3. Contented	2.98	1.06	-.21	-.58	.68
4. Open to communication	3.60	0.99	-.50	-.18	.58
5. Understanding	3.57	0.91	-.41	-.07	.51
6. Motivated	3.10	1.05	-.08	-.49	.74
7. Resilience	3.52	0.99	-.42	-.27	.68
8. Strong	3.48	1.05	-.44	-.41	.66
9. Self-confident	3.31	1.05	-.23	-.42	.66
10. Determined	3.27	1.09	-.28	-.49	.66
11. Successful	3.22	1.00	-.23	-.13	.60
12. Optimistic	3.38	1.03	-.24	.-.54	.60
13. Brave	3.20	1.06	-.12	.-.54	.63
14. Energetic	2.72	1.09	.17	-.62	.62

* r_{ij} = correlation values for item-total test score

Data Analysis

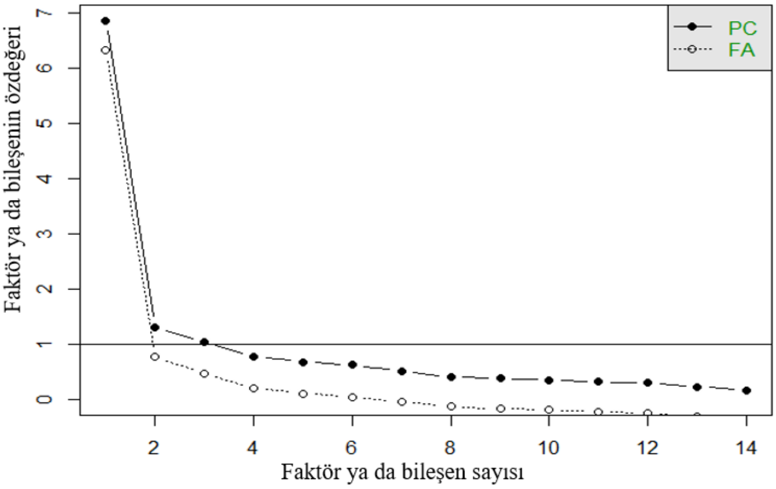
The data were analyzed using the “mirt” package (Chalmers, 2012) in the R (R Core Team, 2016) program. Item parameters for PAS were estimated using GRM (Samejima, 1969), PCM (Embretson & Reise, 2000) and RSM (Andrich, 1978). Descriptive statistics (mean, standard deviation) obtained at the initial data analysis stage, and correlation values for item-total test score (r_{ij}) are given in Table 1. Besides, the factor structure of the scale (unidimensionality assumption) was analyzed using EFA, CFA and parallel analysis. The reliability coefficient for PAS was determined as a Cronbach’s α value of .92. In the evaluation of model-data fit for factor analysis, RMSEA ≤ .08 (Steiger & Lind, 1980); SRMR ≤ .08 (Brown, 2015); CFI ≥ .90 (Hu & Bentler, 1999) and TLI ≥ .90 criteria were considered.

An examination of descriptive statistics given in Table 1 shows that skewness and kurtosis coefficients are in the range of ±1. This points out a normally distribution of the data. In the second stage, IRT models used in parameter estimation and model-data fit process are presented.

Unidimensionality assumption

Unidimensionality which is the fundamental assumption of unidimensional IRT models was analyzed using EFA, CFA and parallel analysis. The KMO value was found to be 0.91, and according to Bartlett's test result, χ^2 value was significant (χ^2 (91) = 2642,29, $p < .05$). The dimensionality of data structure was examined using a scree plot (Figure 3), and a single-factor structure with an eigenvalue of 6.85 was identified. Total variance explained by the factor was 49%, and factor loadings for the items varied between .53 and .77.

Scree plot indicates a rapid decrease in the eigenvalue from the first to the second factor. This shows that PAS had a dominant single-factor structure. At the end of CFA performed to verify factor structure, it was confirmed that model-data fit was at an acceptable level and the scale had a single-factor structure (χ^2 (74) =283.79, RMSEA = .08; CFI = .91, TLI = .88 and SRMR = .06).

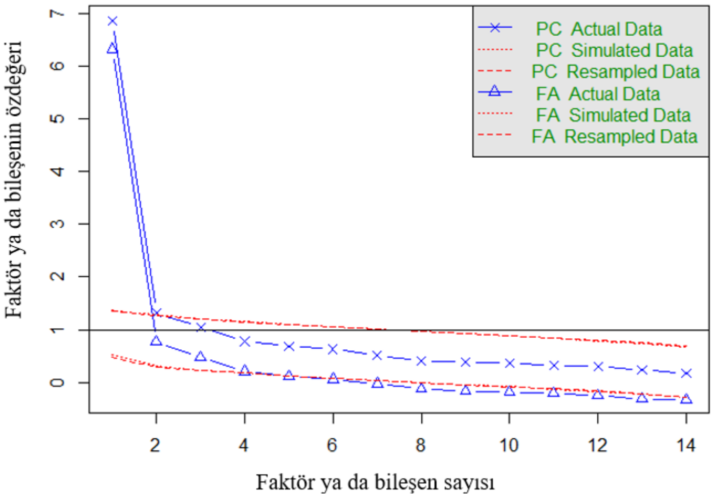


Note. FA: Factor Analysis; PC: Principal Component Analysis

Figure 3. Scree Plot of the PAS Factor Structure

Parallel analysis

Parallel analysis generates random correlation matrices and conducts a factor analysis with these matrices followed by a comparison of eigenvalues obtained through observation of real data with those obtained from simulated data. The fact that eigenvalues obtained from real data are higher than simulated data signals the existence of significant factors.



Note: FA Actual Data: Factor Analysis actual data; FA Simulated data: Factor Analysis simulated data; PC Actual Data: Principal Component Analysis actual data; PC Simulated Data: Principal Component Analysis simulated data

Figure 4. Parallel Analysis Scree Plot

Red-dotted lines in Figure 4 indicate values for simulated data, and blue-dotted lines indicate values for actual data. Blue dots derived from factor analysis up to the red line for simulated data (triangular shape) show factors and components obtained from the data. As a result of the analysis, it was concluded that a single-factor structure was provided.

Local independence assumption

Local independence, given a constant ability level that affects test performance, means that individuals' responses to items are independent of each other. Local independence often occurs when an item is an answer to another item or items depend on a scenario or reading text (DeMars, 2010). Various statistics such as Yen's Q_3 (1984) are suggested for analyzing local independence assumption. The Q_3 statistic proposed by Yen takes into account the relationships between item pairs. First of all, parameters for items and individuals are estimated through an IRT model that is fit for the data. After the estimation of parameters, a residual matrix is formed using the residuals of each item, and correlations between them can be analyzed (DeMars, 2010). If the local independence assumption is confirmed, the items will be independent of each other given an ability level (θ) condition.

It is stated by various studies that if the unidimensionality assumption is met, the local independence assumption is also met (Embretson and Reise, 2000; Hambleton & Swaminathan, 1985). At this point, it was verified by the results of factor analysis that items used in the study displayed a unidimensional structure. Since the unidimensionality assumption was met, it was assumed that the local independence assumption was also met.

Parameter estimation

In the second stage of the analysis, psychometric properties of response categories of 14 items were analyzed using GRM, PCM and RSM. Brief information about the models used in the analysis is given below.

Graded response model (GRM): GRM was used firstly for the estimation of item and test parameters. GRM is appropriate to use when item responses can be characterized as ordered categorical responses. The best advantage of GRM lies in that it provides more information about the ability of individuals compared to dichotomous models. Polytomous items are categorically similar to dichotomous items, but they have more than two response categories. These ordered categories have a k-1 boundary or threshold parameters that separate the categories for an item with k ordered response categories. In comparison with the probability of an individual to respond to any categories lower than a certain category level, they attempt to determine the likelihood to respond to that category or to those above that category (DeMars, 2010).

In the GRM, each scale item (i) is described by two parameters. First, the a_i (discrimination) parameter can be defined as the variation strength of response probability as a function of the latent trait (Rubio et al., 2007). Second, b_i (threshold parameter) refers to the level of latent trait, θ, at which, for each category boundary, the probability of giving a positive response rather than a negative one to that boundary is .5 (Embretson & Reise, 2000).

GRM requires a two-stage procedure to computing the category response probabilities (Embretson & Reise, 2000). In the first step, the estimation of response probabilities involves the computation of k-1 curves for each item of the form given in Equation 1.

$$P_{ik}^{*}(\theta_j) = \frac{e^{Da_i(\theta_j - b_{ik})}}{1+e^{Da_i(\theta_j - b_{ik})}} \tag{1}$$

b_{ik} parameter, for each category boundary, is the level of the latent trait, θ, at which the probability of giving a positive response rather than a negative one to that boundary is .5. $P_{ik}^{*}(\theta_j)$ (operating characteristic curve) refers to the probability of an individual with θ_j to respond above a determined k category boundary. In Equation 2, category characteristic curves are estimated in the second stage, and they represent the probability of an examinee responding in a particular category conditional on trait level. $P_{ik}(\theta_j)$ refers to the probability of an individual under θ_j condition to choose a k category of item i (Embretson & Reise, 2000).

$$P_{ik}(\theta_j) = P_{ik}^{*}(\theta_j) - P_{i(k+1)}^{*}(\theta_j) \tag{2}$$

In this study, the Marginal Maximum Likelihood (MML) method was used for the estimation of GRM item parameters. In GRM, discrimination (slope) for each item and 4 threshold parameters for 5-point response categories were estimated. It is assumed that inter-category threshold (b) parameters for each item are ordered in GRM (Embretson & Reise, 2000).

Partial credit model (PCM): PCM was used secondly in the estimation of item and test parameters. PCM (Muraki, 1992) was developed for items that require responses in multiple steps. It is also used for the analysis of responses to items in scales that measure traits, in which two or more categorical responses are possible such as personality traits (Embretson & Reise, 2000).

It is an extension of the Rasch Model, and raw scores are sufficient for the estimation of ability levels. In this model, the individuals with the same raw scores are at the same ability level. Unlike GRM, the discrimination (a_i) parameter is assumed to be equal for all items. The likelihood of responding to a category can be directly modelled. PCM is a divided-by-total or, as we term it, a direct IRT model (Embretson & Reise, 2000). This means that the probability of responding in a particular category will be written directly as an exponential divided by the sum of exponentials. Assume that item *i* is scored $x = 0 \ldots m_i$ for an item with $K_i = m_i+1$ response categories. For x = *j* the category response curves for the PCM can be written as in Equation 3.

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x}(\theta - \delta_{ij})}{\sum_{r=0}^{mi} \exp \sum_{j=0}^{r}(\theta - \delta_{ij})} \qquad (3)$$

In PCM, different from GRM, step difficulty is defined instead of category threshold parameter. In Equation 3, $\sum_{j=0}^{0}\left(\theta - \delta_{ij}\right) = 0$ terms are called the item step difficulty associated with a category score of j. Step difficulty can be directly interpreted as the point on the latent trait scale at which two consecutive category response curves intersect. Step difficulty can also be defined as the difficulty parameter for passing from one category to the other (Embretson & Reise, 2000).

MML method was also used in PCM for the estimation of item parameters. In PCM, since the discrimination (slope) parameter is considered equal for all items, one discrimination parameter is estimated for all items. *k*-1 step difficulty (b) estimation is obtained for an item with k ordered response categories.

Rating scale model (RSM): It can be used when the items in the scale have the same response format (Embretson & Reise, 2000). In this model, step difficulties of the PCM are defined by location parameter that indicates the place of the item on ability scale and category threshold parameter between consecutive categories. Each item has a single scale location parameter which reflects the difficulty or easiness of the particular item. By the way, the scale location parameter indicates the distance of averages of step difficulties across consecutive categories to zero. It is equivalent to a limited version of PCM where category threshold parameters are equal across items. As is the case in PCM, item discrimination (ai) parameters do not vary across items.

In RSM, the item discrimination parameter is considered equal for all items. k-1 category threshold parameters (b) estimation is obtained for an item with *k* ordered response categories. Since the same scale format is used for all items, category threshold parameters are assumed to be equal for all items. Step difficulty, on the other hand, is defined as the sum of item-specific location parameters and category threshold parameters. MML method was also used in RSM for the estimation of item parameters.

In the RSM model, the step difficulties of the PCM are decomposed into two components, namely, l_i and d_j, where $d_{ij} = (l_i + d_j)$. The l_i is the location of the item on the latent scale and the d_j are the category threshold parameters (Embretson & Reise, 2000). RSM is written as Equation 4.

$$P_x(\theta) = \frac{\exp\left\{\sum_{j=0}^{x}[\theta - (\lambda_i + \delta_i)]\right\}}{\sum_{x=0}^{M} \exp\left\{\sum_{j=0}^{x}[\theta - (\lambda_i + \delta_i)]\right\}} \qquad (4)$$

In PAS with ordered and 5-point Likert type response categories, the same response categories (also in the same number) are used for all scale items. Therefore, item and test parameters were analyzed using GRM, PCM and RSM in an attempt to determine the best fit model to be used for analyzing psychometric properties of the scale.

Model - data fit

For assessment of model-data fit, -2loglikelihood values of polytomous model pairs were compared. Firstly, a comparison was made based on GRM and RSM -2loglikelihood values, χ^2 value and degrees of freedom. AIC and BIC values were also examined. Subsequently, GRM and PCM models were compared. Also, standard error and parameter invariance was investigated. For measurement precision, the amount of information provided by each item across different ability levels was evaluated along with item information functions. The ordinal state of item response categories for each item was examined employing graphical methods.

RESULTS

14 items in PAS were scaled using three different polytomous IRT models. Table 2 displays the model-data fit statistics and Table 3 displays the amount of item information for each model.

Model-data fit was evaluated by comparing in model pairs of lower AIC, BIC and -2loglikelihood values from the models. According to AIC and BIC values in Table 2, the models with the lowest AIC values are GRM, RSM and PCM, respectively, while the models with the lowest BIC values are RSM, GRM and PCM, respectively. These results show that GRM and RSM fitted the data better than PCM.

Table 2. Model-Data Fit Indexes for Polytomous IRT Models

Models	AIC	BIC	χ^2	Degrees of freedom (*df*)
GRM	**11454.84**	11722.66	5763.32	70
RSM	11562.64	**11631.51**	5657.42	19
PCM	11575.21	11793.30	5730.61	57

Table 3 presents item and total test information amount and marginal reliability values derived from different models. The highest amount of total test information was obtained from RSM. Other information amounts were provided by GRM and PCM, respectively. Also, although the reliability coefficient of all three models was close to each other, the highest reliability coefficient was obtained with GRM with a value of .93. Firstly, the values obtained from RSM and GRM which provided the highest amount of total test information were compared to -2loglikelihood, degrees of freedom (*df*) and χ^2 values. The number of parameters varies depending on the different models.

Table 3. Amount of Item and Total Test Information from Polytomous IRT Models

Items	**GRM***	**PCM****	**RSM*****
1. Happy	7.20	3.99	3.99
2. Peaceful	5.44	3.99	4.82
3. Contented	7.08	3.99	5.67
4. Open to communication	4.49	3.99	28.95
5. Understanding	3.70	3.99	22.20
6. Motivated	8.79	3.99	4.15
7. Resilient	6.90	3.99	15.56
8. Strong	6.33	3.99	11.52
9. Self-confident	5.73	3.99	5.08
10. Determined	5.57	3.99	4.53
11. Successful	4.76	3.99	4.14
12. Optimistic	4.64	3.99	6.81
13. Brave	4.89	3.99	4.05
14. Energetic	4.79	3.99	23.54
Total Information	80.35	55.98	145.08
Marginal Reliability	.93	.92	.92

* GRM: Graded Response Model; **PCM: Partial Credit Model; ***RSM: Rating Scale Model

According to RSM, a common a_i parameter, (the number of categories (5) - 1 = 4) category threshold parameters and location parameters for each item were estimated, and the degrees of freedom is (19). In GRM, the a_i parameter for each item and (the number of categories (5) – 1 = 4) category threshold parameters for each item were estimated, and the degrees of freedom was determined as (70). According to this, χ^2 (70, 19) = 5763.32 - 5657.42 = 105.9 and approximate table χ^2 value, χ^2 (51, .05) = 67.50. The difference between the -2loglikelihood χ^2 values from model pairs was found to be significant. Therefore, it can be concluded that GRM is more appropriate for the data.

Secondly, the difference in -2loglikelihood χ^2 values obtained from GRM and PCM was compared with χ^2 statistic using the .05 significance level and degrees of freedom. While the degrees of freedom was determined as (70) for GRM; in PCM, a common a_i parameter for each item and (the number of categories (5) – 1 = 4) category threshold parameters were derived for each item, and the degrees of freedom was determined as (57). In this case, χ^2 (70, 57) = 5657.42 – 5730.61 = -73.19 and, approximate table χ^2 value, χ^2 (13; .05) = 22.36. The difference between the -2loglikelihood χ^2 from model pairs is not significant. This indicates that there is no difference between GRM and PCM. Furthermore, in GRM, the reliability and maximum information values were found to be .93 and 80.35, respectively with a lower AIC value. As a result of model pair comparisons, it was determined that GRM fits the data better, and parameter estimations were performed using GRM. Using GRM, a_i parameter (discrimination) for each item and 4 threshold parameters for 5-point response categories were estimated. Table 4 shows estimated parameters for PAS items.

In GRM calibration, 70 parameters were estimated. Item discrimination parameter refers to the item's power of sorting individuals based on their abilities across latent trait scale. The discrimination level of items is classified as; very low 0.01-0.34, low 0.35-0.64, medium 0.65-1.34, high 1.35-1.69 and very high above 1.70 (Baker, 2001). Item discrimination (a_i) parameters for 14 items vary between 1.25 and 2.66 and with item 6 having the highest and item 5 having the lowest level. Accordingly, it is understood that discrimination values of items are of medium and high levels. In the context of data structure, the a_i parameter can be considered as the numerical value of the psychological uncertainty of an item (Roskam, 1985). Higher a_i parameter values indicate that the item has a well-defined and clear meaning (Ferrando, Lorenzo, & Molina, 2001). As a result, it was concluded that 14 items in the scale were well-defined items with high discrimination.

Table 4. Estimated Item Discrimination and Category Threshold Parameters According to GRM

Items	a_i (se)	b_1(se)	b_2(se)	b_3(se)	b_4(se)
1. Happy	2.21(.20)	-1.94(.15)	-1.06(.18)	0.38(.12)	2.03(.44)
2. Peaceful	1.83(.17)	-1.75(.15)	-0.70(.15)	0.40(.11)	2.14(.34)
3. Contented	2.25(.21)	-1.58(.12)	-0.68(.14)	0.48(.11)	2.00(.22)
4. Open to communication	1.55(.16)	-2.92(.30)	-1.57(.29)	-0.30(.23)	1.36(.40)
5. Understanding	1.25(.14)	-3.71(.46)	-1.97(.40)	-0.26(.32)	1.82(.62)
6. Motivated	2.66(.24)	-1.77(.12)	-0.76(.16)	0.40(.11)	1.56(.03)
7. Resilient	2.18(.20)	-2.42(.19)	-1.30(.23)	0.13(.18)	1.27(.50)
8. Strong	2.08(.19)	-2.27(.18)	-1.14(.21)	-0.12(.16)	1.27(.47)
9. Self-confident	1.92(.18)	-2.18(.17)	-1.07(.20)	0.19(.14)	1.44(.68)
10. Determined	1.92(.18)	-1.98(.16)	-0.99(.18)	0.19(.13)	1.48(.71)
11. Successful	1.64(.16)	-2.23(.20)	-1.21(.21)	0.41(.14)	1.83(.20)
12. Optimistic	1.58(.15)	-2.70(.26)	-1.21(.24)	0.10(.19)	1.56(.59)
13. Brave	1.68(.16)	-2.25(.19)	-0.94(.19)	0.39(.15)	1.70(.89)
14. Energetic	1.67(.16)	-1.56(.13)	-0.30(.11)	0.94(.21)	2.26(.92)

Note: a_i = item discrimination; se = standard error; b_i = category threshold

b_{ik} parameters (b_{i1} and b_{i4}) show the position of items in the latent trait (ability) scale. For example, for item 1, b_{11} = -1.94 refers to the ability level required to respond to category 1 and above with a likelihood of .50. b_{15} = 2.03 refers to the ability level required to respond to category 5 with a likelihood of .50. It is seen that along the latent trait scale, first category threshold parameter values were distributed around -2, second category threshold parameter values around -1, third category threshold parameter values around 0, and fourth category threshold parameter values were distributed around

1.5. This indicates that the scale better differentiates people across with the latent trait scale. Also, category threshold parameter values displayed a hierarchical increase along the ability scale. According to the results, it is understood that it is suitable to use GRM for measuring the psychometric properties of PAS.

Figure 5 presents category threshold parameters estimated for 14 items. a_i (discrimination) parameters obtained in GRM are treated as random effects. Since each item has its discrimination parameter value, graph lines belonging to the category threshold are not parallel to each other. However, it is seen in Figure 5 and Figure 6 that category threshold parameters of 14 items are ordinal for each item.

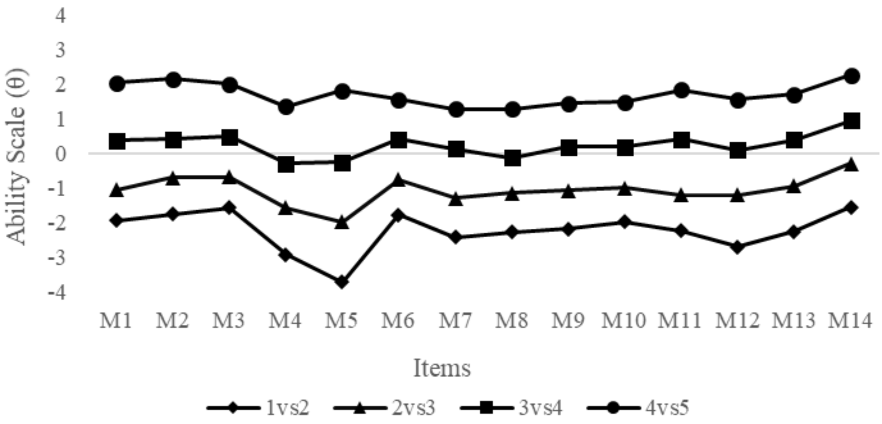


Figure 5. Order of Category Threshold Values for 14 Items Estimated by GRM

Figure 5, horizontal axis denotes 14 items and the vertical axis denotes ability (θ) scale. It is apparent in Figure 5 that category threshold parameters for the items of PAS are in a hierarchical order. In Figure 6, it is exemplified through item 2 and item 6 given that category threshold parameters are ordered based on item.

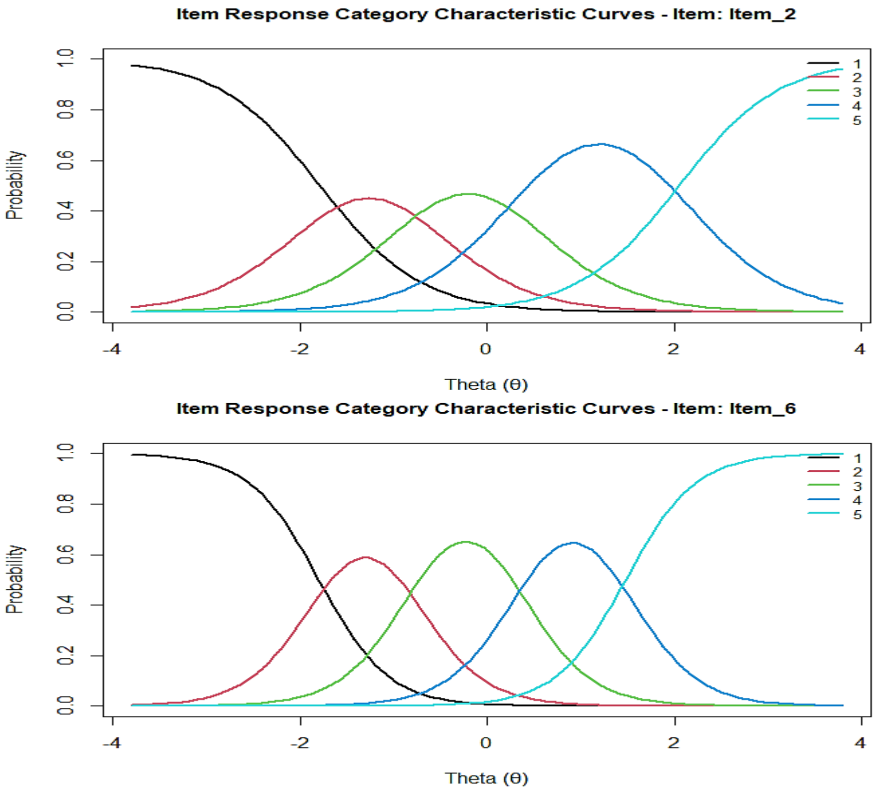


Figure 6. Item Response Category Characteristic Curves for Item 2 and Item 6

In Figure 7, item information functions are given for three items with high (Item 6), medium (Item 2) and low discrimination (Item 5) level. Figure 7 indicates how different discrimination (slope) values affect measurement precision throughout the ability scale. Accordingly, Item 6 with a high discrimination value provided more information than Item 2 and Item 5 all along the scale.

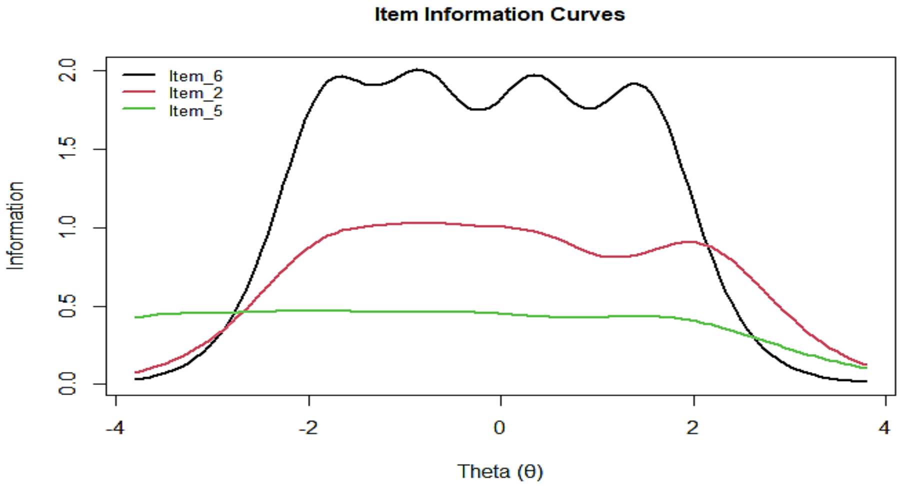
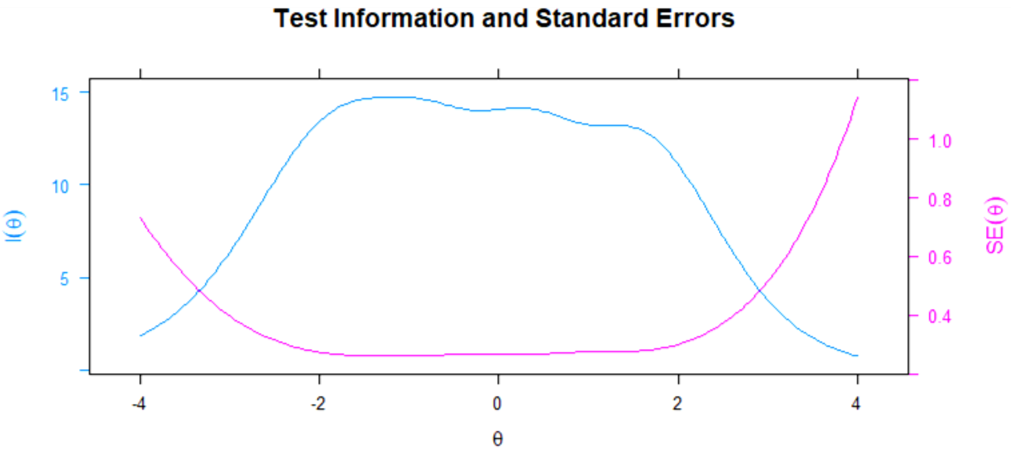


Figure 7. Item Information Curves with Low (Item 5), Medium (Item 2) and High (Item 6) Information Level

Figure 8 shows the relationship between the total test information of PAS based on GRM and the standard error. The amount of information obtained through the ability scale seems to be higher at the ability level within the interval of (-2≤ θ ≤ +2). The figure also shows that standard error estimation is also lower in this ability level interval. It indicates that the amount of maximum information is provided by the scale around the ability level θ = (-1.40).



Note: θ = Latent trait scale (ability scale), blue line indicates total test information function (I(θ)), and the pink line indicates standard error (SE(θ)).

Figure 8. Test Information and Standard Errors for PAS Based on GRM

The sample was randomly divided into two groups to test parameter invariance and, then item discrimination and category threshold parameter values were estimated for each sub-group. Correlation between item discrimination values (a_i) from the two sub-groups is r = .81 (p < .01). Correlations between category threshold (b_{ik}) values were found to be b_{i1} = 0.90, b_{i2}= 0.96, b_{i3}= 0.97,

b_{i4}= 0.83 (p < .01), respectively. The results showed that the correlation values for parameters estimated from different samples were high; in other words, they were analogous, proving that parameter invariance was ensured.

DISCUSSION and CONCLUSION

The review of the literature on scaling reveals that there are many studies of cognitive test structures under IRT models. However, it is a fact that use of IRT-based models in developing scales for measurement of affective traits is relatively limited in our country (Demirtaşlı et al., 2016). The purpose of this study was to investigate whether category threshold parameters, which are used to determine responses to Likert-type polytomous items in measurement tools used particularly for measuring affective traits, were ordered within the items. Responses to polytomous items in Likert-type measurement tools assume that individuals choose the categories which best describe their states. However, differences may occur between assessments as individuals use different decision-making processes when making such decisions. It is important to employ appropriate methods and techniques for developing measurement tools to catch up with this variance between subjective assessments (Wang et al., 2006). The extent to which a psychological construct intended to be measured is represented by response categories of a measurement tool is very important in terms of psychometric properties. This study aimed to test the psychometric properties of the Positive Affect Scale used to determine positive affects across item response categories. The fact that item response categories in the scale are ordered for each item and have similar meanings is of importance for using and interpreting the results of the scale (Messick, 1995).

The ability levels required to respond to each category of each item are estimated separately for measurement tools scaled with IRT models. This allows achieving more reliable and valid results for the measurement of individual differences. The extent of fitness of response format in a measurement tool for the psychological reality which it intends to measure also affects the validity of measurements (Baker et al., 2000). Therefore, selecting the suitable model for the data is important for the interpretability of the inferences from the results. In this study, Samejima's GRM, PCM, and RSM were used for analyzing psychometric properties of item response categories. Results from different IRT models for scaling provide various information about categories. Psychometric properties of item response categories of Likert-type scale items within the scope of this study were evaluated to model-data fit within the context of specific parameters of each model. In particular, the analysis of inter-category psychometric properties of polytomous items used for measuring affective traits will also contribute to the significance of inferences from measurement results. Results based on different models which ensured model-data fit provide different information about the properties of categories.

Application data were used in this study, and the comparability of item parameters of 14-item PAS subject to the application was analyzed using polytomous IRT models. Model comparisons were made to determine the best fit IRT model for PAS items. As a result of analyses, GRM had to the best fit. Since the maximum amount of information provided by GRM and reliability of GRM is higher and its AIC value is lower, parameter estimations were made according to GRM in the analysis of psychometric properties. Similar results were obtained in various studies which examined psychological properties. In the study by Rubio et al. (2007), results that correspond to those of GRM were obtained in the analysis of psychometric properties of emotional adaptation scale, Rosenberg self-esteem scale (Gray-Little et al., 1997). GRM has been frequently used in the analysis of psychometric properties of measurement tools applied for analyzing response categories for positive and negative affects (Baker et al., 2000) and various affective traits (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Demirtaşlı et al., 2016; Köse, 2015).

Item discrimination parameters (a_i) for 14 items estimated based on GRM varied between the values of 1.25 and 2.66. Accordingly, the items had discrimination values of medium and high level. In the analysis, 4 category threshold parameters were estimated for each item. It is seen that along the ability scale, first category threshold parameter values were distributed around -2, second category threshold parameter values around -1, third category threshold parameter values around 0 and fourth category threshold parameter values around 1.5. This shows that the scale well-distinguished people at different ability levels along the latent trait scale.

The information from test and item information functions proved to be higher at the ability levels in ($-2 \leq \theta \leq +2$) interval. The sample was randomly divided into two groups to test parameter invariance, and item parameters were estimated through these groups. Findings support that item parameter invariance was attained.

In scale development or adaptation studies and studies in which measurement tools that intend to measure psychological characteristics are used (in particular for measurement tools used for measuring affective traits), when, in general, evaluating whether measurement tool provides factor structure, analysis of properties of item response categories is often ignored. However, rating level and psychometric properties of item response categories are also important for determining to what extent the measurement tool represents the construct it intends to measure. At this point, the fact that category threshold values are in acceptable intervals for each item and that observed category threshold values are comparable across items indicates that the information obtained from the items can be used in the same way. In computing total scores, it is relatively important that the extent of comparability of a response to an item, for example, a response of 4, with a response of 4 given to another item or the extent of equivalence of the distance between responses of 3 and 4 in an item to the corresponding distance in another item. This study focused on these questions and highlighted the importance of computation of item parameters for measurement tools comprising items that use an ordinal rating scale. It is suggested that model-data fit and item parameters should be studied in detail using models like GRM for ordinal rating scales such as 3-point or 5-point scales.

It is possible to determine at which levels the scale provides more information by obtaining more in-depth information on ability levels upon provision of detailed information on the measurement tool. For future studies, it may be an option to incorporate additional items that will provide more information, particularly on the ability levels for which the scale provided little information. Moreover, ensuring model-data fit for a measurement tool scaling based on IRT allows the estimation of invariant parameters of the scale even if it is applied to different groups. This will provide valid and reliable measurement results in comparisons for the results of the same measurement tool applied to different study groups.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi: 10.1007/BF02293814

Baker, F. B. (2001). *The basis of item response theory*. USA: ERIC Clearing house on Assessment and Evaluation.

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A comparison of graded response and Rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics, 25*(3), 253-270. doi: 10.3102/10769986025003253

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

Caycho-Rodríguez, T., Vilca, L. W., Carbajal-León, C., White, M., Vivanco-Vidal, A., Saroli-Araníbar, D., ..., Moreta-Herrera, R. (2021). Coronavirus anxiety scale: New psychometric evidence for the Spanish version based on CFA and IRT models in a Peruvian sample. *Death Studies.* doi: 10.1080/07481187.2020.1865480

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562. doi: 10.1207/S15327906MBR3604_03

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment*, *27*(4), 1241-1252. doi: 10.1037/pas0000132

de Ayala, R. J., Dodd, B. G., & Koch, W. R. (1990, April). *A comparison of the partial credit and graded response model in computerized adaptive testing*. Paper presented at the AERA Annual Meeting. Boston.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* Oxford: Oxford University Press.

Demirtaşlı, N., Yalçın, S., & Ayan, C. (2016). The development of irt based attitude scale towards educational measurement course. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(7), 133-144. doi: 10.21031/epod.43804

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: LEA publishers.

Ferrando, P., Lorenzo, U., & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25*(1), 3-17. doi: 10.1177/01466216010251001

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Journal of Research in Personality, 29*(2), 168-188. doi: 10.1006/jrpe.1995.1010

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Gray-Little, B., Williams, V., & Hancock, T. (1997). An item response theory analysis of the Rosenberg self - esteem scale. *Personality and Social Psychology Bulletin, 23*(5), 443-451. doi: 10.1177/0146167297235001

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

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

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

Kahraman, N., Akbaş, D., & Sözer, E. (2019). Bilişsel-olmayan öğrenme durum ve süreçlerini ölçme ve değerlendirmede boylamsal yaklaşımlar: Duygu Cetveli Alan Uygulaması örneği. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19*(1), 257-269. doi: 10.17240/aibuefd.2019.19.43815-459831

Kaptan, S. (1995). *Bilimsel araştırma ve istatistik teknikleri* (10. Basım). Ankara: Rehber Yayınevi.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32. doi: 10.1177/014662168300700104

Köse, İ. A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 184-197. Retrieved from https://dergipark.org.tr/tr/download/article-file/17439

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf

Lord, F. M. (1980). *Applications of item response theory practical testing problems.* Hillsdale, NJ: Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1002/j.2333-8504.1994.tb01618.x

Min, S., & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: implications for the dimensionality of language ability. *Studies in Educational Evaluation, 68*. doi: 10.1016/j.stueduc.2020.100963

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. doi: 10.1177/014662169201600206

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roskam, E. E. (1985). Current issues in item response theory. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3-19). Amsterdam: North Holland.

Rubio, V. J., Aguado, D., Hontangas, P. M., & Hernandez, J. M. (2007). Psychometric properties of an emotional adjustment measure: an application of the Graded Response Model. *European Journal of Psychological Assessment, 23*(1), 39-46. doi: 10.1027/1015-5759.23.1.39

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monography No. 17). Retrieved from https://www.psychometricsociety.org/sites/main/files/file-attachments/mn17.pdf?1576606975

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*, 17-35. doi: 10.2333/bhmk.23.17

Silvia, P. J. (2021). The self-reflection and insight scale: applying item response theory to craft an efficient short form. *Current Psychology*. doi: 10.1007/s12144-020-01299-7

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented in Psychometric Society. Iowa City.

Wang, W. C., Wilson, M., & Shih, C. L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*(4), 335-353. doi: 10.1111/j.1745-3984.2006.00020.x

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*(2), 93-113. doi: 10.1007/BF03217438

Yaşar, M., & Aybek, E. C. (2019). Üniversite öğrencileri için bir yılmazlık ölçeğinin geliştirilmesi: Madde tepki kuramı temelinde geçerlilik ve güvenilirlik çalışması. *İlköğretim Online, 18(4),* 1687-1699. Retrieved from https://ilkogretim-online.org/fulltext/218-1597121020.pdf?1618815938

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. doi: 10.1177/014662168400800201

Aynı Tepki Kategorilerine Sahip Likert Maddelerin Psikometrik Özelliklerinin Çok Kategorili Madde Tepki Kuramı Modelleri ile İncelenmesi

Giriş

Ölçme araçlarında yer alan çok kategorili (polytomous) Likert tipi puanlanan maddelere verilen cevaplar, bireylerin durumlarını en iyi tanımlayan kategorileri seçtikleri varsayımıyla, öznel değerlendirmelerine dayanmaktadır. Yapılan bu öznel değerlendirmelere göre bireyler, karar verme süreçlerinde farklı kriterlere göre durumlarını değerlendirerek cevap vermektedir. Bireyler arası bu öznel karar verme farklılıklarını tanımlamak ölçekleme için oldukça önemlidir. Öznel değerlendirmelerdeki bu varyansı yakalayabilmek için ölçme araçlarının uygun yöntem ve teknikler ile incelenmesi önemlidir (Wang, Wilson, & Shih, 2006). Çünkü bireylerin ölçek maddelerine verdiği tepkilerden en doğru ve kullanışlı bilgiler ortaya çıkarmak ölçme ve değerlendirmenin en temel amaçlarındandır (Wu & Adams, 2006). Ölçme modellerindeki yeni gelişmeler ve yaklaşımlar ile ölçme uygulamalarındaki hataların azaltılması, doğruluğun ve etkililiğin arttırılması hedeflenmektedir (Baker, Rounds, & Zevon, 2000). Bu bağlamda kişilik özellikleri gibi psikolojik yapıların değerlendirilmesinde kullanılan farklı cevap formatlarına sahip ölçme araçlarının psikometrik özelliklerinin değerlendirilmesi için çok kategorili Madde Tepki Kuramı (MTK) modelleri geliştirilmiştir. MTK modellerine göre ölçeklendirilen test ve ölçekler ile her bir maddenin her bir kategorisine cevap vermek için gerekli olan yetenek düzeyinin ayrı ayrı kestirimi sağlanmaktadır. Bu da bireysel farklılıkların ölçümü bağlamında daha güvenilir ve geçerli sonuçların elde edilmesine neden olmaktadır. Bir ölçme aracında kullanılan cevap formatının ölçmeye çalıştığı psikolojik gerçekliğe ne derece uygun olduğu, ölçme aracından elde edilen ölçümlerin geçerliğini de etkilemektedir (Baker ve diğerleri, 2000). Dolayısıyla kullanılan veriye uygun bir modelin seçilmesi sonuçlardan elde edilecek çıkarımların anlamlılığı için önem taşımaktadır.

Psikolojik özellikleri ölçen ölçme araçları genelde çok kategorili cevap formatına sahip maddelerden oluşmaktadır. Bu maddelerin incelenmesinde kullanılan çok kategorili puanlanan MTK modelleri, cevaplayıcının yetenek düzeyi ile belli bir kategoride tepki verme olasılığı arasında doğrusal olmayan ilişkiler kuran modellerdir (Embretson & Reise, 2000). Çok kategorili modeller, madde ağırlıklandırmalarında farklı madde ayırt edicilik değerlerinin kullanılması, her bir yetenek düzeyinde ölçme hatası kestiriminin yapılması ve birey ve maddeler için parametre değişmezliğinin elde edilmesini sağlamaktadır (Lord, 1980). Bir ölçeğin ölçmeye çalıştığı yapının kendini temsil eden tepki kategorilerine nasıl ayrıldığı, o ölçeğin güvenirliğini etkilemektedir (Linacre, 2002). Eşit aralıklı veya sıralama düzeyi gibi farklı özelliklere sahip tepki kategorileri MTK içinde yer alan çok kategorili modeller ile incelenebilmektedir. Farklı özelliklere sahip madde cevap (tepki) kategorileri, Aşamalı Tepki Modeli (ATM; Samejima, 1969), Kısmi Puanlama Modeli (KPM; Embretson & Reise, 2000) ve Dereceli Ölçekleme Modeli (DÖM; Andrich, 1978) gibi ölçme modelleri ile incelenmekte ve model-veri uyumları değerlendirilmektedir. Bir modelin veriye uygunluğunun değerlendirilmesinin yanında, modelin yanıtların altında yatan psikolojik gerçekliğe (yani, yanıtların formatı) ne kadar uygun olduğuna dair temel gözlemlerin de dâhil edilmesinin önemi vurgulanmaktadır (Samejima, 1996). Bu nedenle, incelenen ölçme aracının kategorilerine ait özelliklerin neler olduğu (her madde için kategorilerin benzer bir sıraya sahip olup olmadığı) ve ölçmeye çalıştığı psikolojik yapıya ne denli uygun olduğunun belirlenmesi, elde edilen ölçme sonuçlarının güvenirlik ve geçerliği açısından önemlidir. Bu çalışmanın amacı pozitif duygu durumlarının ölçülmesi için geliştirilen 14 maddelik bir Pozitif Duygu Durum (PDD) ölçeğinin içerdiği derecelendirilmiş (1'den 5'e kadar) maddelerin tepki kategorilerini ve bu kategorilerin maddeler arası ne derece benzerlik/farklılık gösterdiğini incelemektir. Bu amaçla, PDD ölçeğinde yer alan çok kategorili puanlanan maddelerin madde parametrelerinin kestiriminin farklı modeller ile elde edilmesi, bu modeller için hesaplanan model-veri uyumunun incelenmesi ve duygu durumu ölçeği boyunca farklı yetenek düzeylerinde elde edilen ölçümlerin ölçme kesinliğinin karşılaştırmalı olarak değerlendirilmesi amaçlanmıştır. PDD ölçeğinin çok kategorili cevap formatına sahip olması ve çok kategorili modellerle cevaplama süreçleri arasındaki teorik ilişki dikkate alındığında, ölçekte yer alan maddelere verilen tepkileri belirlemede kullanılan kategoriler arası eşik (threshold) parametrelerinin maddeler içi sıralı olup olmadığı ATM, KPM ve DÖM ile çalışılmış ve bu modellerin her birinin öngördüğü koşullar üzerinden, maddeler için varsayılan kategori eşik parametrelerinin ölçekteki tüm maddeler için değişmezliği varsayımının geçerliliği, uygulamada bu ölçek için toplanan veriler kullanılarak incelenmiştir.

Yöntem

Bu çalışma, PDD ölçeği'nin psikometrik özelliklerinin MTK modellerine göre incelendiği karşılaştırmalı betimsel bir çalışmadır. Uygulama verisinde 326 gönüllü üniversite öğrencisi yer almaktadır. Bu çalışmada kullanılan veriler Duygu Cetveli Alan Uygulaması (Kahraman, Akbaş, & Sözer, 2019) olarak adlandırılan daha geniş kapsamlı bir çalışmadan gelmektedir. Çalışma verilerinin elde edildiği PDD ölçeği, bireylerden her madde için kendilerine verilen cevap kategorilerinden (*hiç veya çok az* için 1'den *çok* için 5'e kadar) birini işaretlemelerini isteyen 14 maddeden oluşmaktadır. Ölçekte yer alan 14 pozitif duygu durumu şu şekildedir: Mutlu, huzurlu, memnun, iletişime açık, anlayışlı, motive, dayanıklı, güçlü, özgüvenli, azimli, başarılı, iyimser, cesur ve enerjik. Verilerin analizi R (R Core Team, 2016) programında "mirt" paketi (Chalmers, 2012) kullanılarak gerçekleştirilmiştir. PDD ölçeğinden elde edilen verilerin analizinde ATM, KPM ve DÖM kullanılarak madde parametre kestirimleri yapılmıştır. Verilerin analiz aşamasında elde edilen betimleyici istatistikler (ortalama, standart sapma) ve madde-toplam test korelasyon değerleri (r_{ij}) incelenmiştir. Bununla birlikte ölçeğin faktör yapısı (tek boyutluluk varsayımı) Açımlayıcı Faktör Analizi (AFA), Doğrulayıcı Faktör Analizi (DFA) ve paralel analiz ile incelenmiştir. Ölçeğin güvenirlik katsayısı Cronbach's α = .92 olarak belirlenmiştir.

Sonuç ve Tartışma

Madde Tepki Kuramı modellerinin temel varsayımları olan tek boyutluluk ve yerel bağımsızlık incelendiğinde, ölçeğin faktör yapısına ilişkin yapılan analizler sonucunda ölçeğin tek boyutlu bir yapıya sahip olduğu belirlenmiştir. Tek boyutluluk varsayımının sağlanması durumunda yerel bağımsızlık varsayımının da sağlanacağı çeşitli çalışmalar tarafından belirtilmiştir (Embretson ve Reise, 2000; Hambleton & Swaminathan, 1985). Bu noktada, çalışma kapsamında kullanılan maddelerin tek boyutlu bir yapı gösterdiği faktör analizi sonuçlarına göre doğrulanmıştır. Tek boyutluluğun sağlanması nedeniyle yerel bağımsızlık varsayımının da karşılandığı varsayılmıştır.

PDD ölçeğinde yer alan 14 madde, üç farklı çok kategorili MTK modeli kullanılarak analiz edilmiş, model-veri uyum istatistikleri ve her modele göre elde edilen madde bilgi miktarları incelenmiştir. Model-veri uyumu daha düşük AIC, BIC değerleri ve modellerden elde edilen -2loglikelihood değerlerinin çiftler halinde karşılaştırılması ile değerlendirilmiştir. Model-veri uyumu karşılaştırmalarına göre, ATM ve DÖM modellerinin veriye daha iyi uyum sağladığı gözlenmiştir. Madde ve toplam test bilgi miktarları ile farklı modellerden sağlanan marjinal güvenirlik değerleri incelendiğinde en fazla bilgi miktarının DÖM'den elde edildiği gözlenmiştir. Bununla birlikte en yüksek güvenirlik katsayısı .93 olarak ATM modelinden elde edilmiştir. Bu noktada -2loglikelihood, serbestlik dereceleri ve χ^2 değerlerine göre çiftler halinde model karşılaştırmaları yapılmıştır. İkili model karşılaştırmaları sonucunda ATM modelinin veriye daha iyi uyum sağladığı sonucuna ulaşılmıştır. ATM ile her madde için a_i parametresi (ayırt edicilik) ve 5'li tepki kategorileri için 4 eşik parametresi kestirilmiştir.

Aşamalı Tepki Modeli'ne göre elde edilen 14 maddeye ait a_i parametreleri 1.25 ve 2.66 değerleri arasında değişmektedir. Buna göre, maddelerin orta ve yüksek düzeyde ayırt edicilik değerlerine sahip olduğu görülmektedir. Analizde her madde için 4 kategori eşik parametresi kestirimi yapılmıştır. Yetenek ölçeği boyunca birinci kategori kesişim parametre değerleri -2 etrafında, ikinci kategori kesişim parametre değerleri -1, üçüncü kategori kesişim parametre değerleri 0 ve dördüncü kategori kesişim parametre değerleri 1.5 etrafında dağıldığı görülmektedir. Bu da ölçeğin, bireyleri yetenek ölçeği boyunca farklı yetenek düzeylerinde iyi bir şekilde ayırdığını göstermektedir. Test ve madde bilgi fonksiyonları ile ölçekten elde edilen bilginin ($-2 \leq \theta \leq +2$) aralığındaki yetenek düzeylerinde daha fazla olduğu görülmektedir. Parametre değişmezliğinin incelenmesi için örneklem tesadüfi olarak ikiye ayrılmış ve madde parametreleri bu gruplar üzerinden kestirilmiştir. Elde edilen bulgular, madde parametre değişmezliğinin sağlandığını desteklemektedir.

Ölçek geliştirme veya uyarlama çalışmalarında ve psikolojik özellikleri ölçmeye çalışan ölçme araçlarının kullanıldığı çalışmalarda (özellikle duyuşsal becerilerin ölçülmesinde kullanılan ölçme araçları için) genellikle ölçme aracının faktör yapısını sağlayıp sağlamadığı değerlendirilirken madde tepki kategorilerinin özelliklerinin incelenmesinin genelde ihmal edildiği görülmektedir. Oysaki ölçme aracının ölçmeye çalıştığı yapıyı ne derece temsil ettiğinin belirlenmesinde madde tepki kategorilerinin dereceleme düzeyi ve bu kategorilerin psikometrik özellikleri de önem taşımaktadır. Bu noktada, hesaplanan kategori eşik parametrelerinin her madde için kabul edilebilir aralıklarda yer alması ve gözlenen kategori eşik değerlerinin maddeler arası karşılaştırılabilir olması, maddelerden elde edilen bilginin aynı şekilde kullanılabilir olduğunu göstermektedir. Toplam puanların hesaplanmasında, bir maddeye verilen, örneğin, 4 cevabının, diğer bir maddeye verilen 4 cevabı ile ne kadar karşılaştırılabilir veya bir maddedeki 3 ile 4 cevabı arasındaki mesafenin bir diğer maddedeki aynı mesafeye ne kadar denk olduğu oldukça önemlidir. Mevcut çalışma bu sorulara odaklanmakta ve sıralama ölçeği kullanan maddelerden oluşan ölçme araçları için de madde parametrelerinin hesaplanmasının önemli olduğunun altını çizmektedir. Önerilen, 3'lü, 5'li gibi sıralı cevap kategorilerini kullanan maddelerden oluşan ölçekler için ATM gibi modeller ile model uyumu ve madde parametrelerinin detaylı bir biçimde çalışılmasıdır.

Ölçme aracına ilişkin ayrıntılı bilgilerin sağlanması ile yetenek düzeylerine ilişkin daha derinlemesine bilgiler elde edilerek ölçeğin hangi düzeylerde daha fazla bilgi sağladığı belirlenebilmektedir. Gelecek araştırmalarda kullanılacak ölçeğe, özellikle daha az bilgi sağladığı yetenek düzeyleri için daha fazla bilgi sağlayabilecek maddelerin eklenmesi düşünülebilir. Aynı zamanda, MTK'ya dayalı ölçekleme yapılan bir ölçme aracının model-veri uyumunun sağlanması ölçeğin farklı gruplarda uygulansa da değişmez parametre kestirimlerinin elde edilmesini sağlamaktadır. Bu durum, farklı çalışma gruplarına uygulanan aynı ölçme aracının sonuçlarına yönelik yapılacak karşılaştırmalarda geçerli ve güvenilir ölçme sonuçlarının elde edilmesini sağlayacaktır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 147-162

Using Generalizability Theory to Investigate the Reliability of Scores Assigned to Students in English Language Examination in Nigeria *

Olufunke Favour AKINDAHUNSI **   Eyitayo Rufus Ifedayo AFOLABI ***

Abstract
The study investigated the reliability of scores assigned to students in English language in National Examinations Council (NECO). The population consisted of all the students who sat for NECO Senior School Certificate Examination (SSCE) in 2017 in Nigeria. A sample of 311,138 was selected using the proportionate stratified sampling technique. The Optical Marks Record (OMR) sheet containing the responses of the examinees was the instrument for the study. The data was analyzed using lme4 package of R language and environment for statistical computing, factor analysis and Tucker index of factor congruence. The psychometric properties of the data were determined by estimating the generalizability (g) coefficient, phi (Φ) coefficient and construct validity. The results indicated the g-coefficient to be 0.90 and Φ coefficient as 0.87, which is an indication of high reliability of scores. The result also showed that a decrease in the number of the items resulted in a decrease in both g- and phi coefficients in D-study. The construct validity of 0.99 obtained from the result affirms the credibility of the items. Hence, it was concluded that the scores were dependable and generalizable.

Key Words: Reliability, validity, English language, score, Generalizability theory.

INTRODUCTION

Generalizability theory is a statistical method used to analyze the results of psychometric tests, such as performance tests like the objective structured clinical examination, written or computer-based knowledge tests, rating scales, or self-assessment and personality tests (Breithaupt, 2011). It involves separating various sources of error and recognizing that multiple sources of error such as error attributed to items, occasions, and forms may occur simultaneously in a single measurement process, thereby forming the basic approach underlying generalizability theory (g-theory) which is to decompose an observed score into a component for the universe score and one or more error components. Its main purpose is to generalize from an observation at hand to the appropriate universe of observations. It is also advantageous in that it can estimate the reliability of the mean rating for each examinee while simultaneously accounting for both interrater and intra-rater inconsistencies as well as discrepancies due to various possible interactions, which are impossible in Classical Test Theory (CTT) (Brennan, 2001). In generalizability theory, various sources of error contributing to the inaccuracy of measurement are explored. It is a valuable tool in judging the methodological quality of an assessment method and improving its precision. It gives the opportunity of disentangling the error components of measurement and is also interested in the reliability or dependability of behavioral measurement, that is, the certainty that the score is reliable to generalize.

All test scores, just like any other measurement, contain some errors. It is this error that affects the reliability or consistency of test scores. When there are variations in the measurement under the same conditions, then error comes in. Error in measurement can be defined as the difference between a person’s observed score and his/her true score. Error is not a mistake in statistics; it is bound to occur.

* The paper was a part of thesis and the whole study is not published anywhere.

** PhD. Student, Obafemi Awolowo University, Department of Educational Foundations and Counselling, Ile – Ife-Nigeria, olufavour@ymail.com, ORCID ID: 0000-0002-5041-7088
*** Prof., Obafemi Awolowo University, Department of Educational Foundations and Counselling, Ile – Ife-Nigeria, eriafolabi@gmail.com, ORCID ID:0000-0002-0014-0711

To cite this article:
Akindahunsi, O. F., & Afolabi, E. R. I. (2021). Using generalizability theory to investigate the reliability of scores assigned to students in English language examination in Nigeria. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 147-162. doi: 10.21031/epod.820989

Received: 04.11.2020
Accepted: 30.05.2021

Breithaupt (2011) identified two types of measurement errors in the examination of items and test scores: random error and systematic error. The author expressed that random error is a source of bias in scores and an issue of validity while systematic error is a measurement error that can be estimated in reliability studies. Its estimates permit the test developer to determine the possible size and sources of construct irrelevant variation in test scores. Thus, it is assumed that the skill, trait, or ability measured is a relatively stable defined quantity during testing. Therefore, variation in obtained scores is usually attributed to sources of error and thus poses the challenge of determining the psychometric property of a test. The goal of the psychometric analysis is to estimate and minimize, if possible, the error variance so that the observed score (X) is a good measure of the true score (T). Understanding whether the test error is due to high variance is important in measurement. It is generally assumed that the exact or true value exists based on how what is being measured is defined. Though the true value exactly may not be known, attempts can be made to know the ideal value. In CTT any observed score is seen as the combination of a true component and a random error component, even though the error could be from various sources. However, only a single source of measurement error can be examined at any given time. CTT treats error as random and cannot be used to differentiate the systematic error from random error. Generalizability theory also focuses on the universe score, or the average score that would be expected across all possible variations in the measurement procedure (e.g., different raters, forms, or items). This universe score is believed to represent the value of a particular attribute for the object of measurement (Crocker & Algina, 2008). The universe is defined by all possible conditions of the facets of the study. It also gives the opportunity to judge whether the score differences observed between the subject could be generalized to all items and occasions (de Gruijter & van der Kamp, 2008). This means that g-theory helps to know whether the means observed over a sample of items and a sample of occasions could be generalized to the theoretical universe of items and occasions. Since g-theory focuses on the simultaneous influence of multiple sources of measurement error variance, it more closely fits the interest of researchers.

The reliability coefficients under CTT are usually focused on the consistency of the test results. For instance, test-retest reliability considers only the time/occasions of testing, parallel-forms reliability considers only the forms of the test and internal consistency considers the items as the only source of error. Some authors (Mushquash and O’Connor, 2006; Webb, Shavelson, & Haertel, 2006) noted that the effects of various sources of variance can be tested using CTT models within which it is only possible to examine a single source of measurement error at a given time, but that it is impossible to examine the interaction effects that occur among these different sources of error. Generalizability theory is particularly useful in this regard; each feat of the measurement situation is a source of error in test scores and its termed facet. Therefore, the inadequacy of explanation of numerous sources of error as pointed out by several authors (Brennan, 2001; Johnson & Johnson, 2009) and the researchers’ dissatisfaction with CTT’s inability to identify possible sources of error and simultaneously examining them led to the development of g-theory which was an extension of CTT. It offers a broader framework than the CTT for estimating reliability and errors of measurement. Generalizability theory involves two types of study: generalizability study (G-study) and Decision study (D-study). The main purpose of a G-study is to estimate components of score variance that are associated with various sources, while a D-study takes these estimated variance components to evaluate and optimize among alternatives for subsequent measurement. Two types of decision and error variance, relative and absolute, are made in G-study, but only relative decisions are made in CTT (Brennan, 2001; Yin & Shavelson, 2008).

Alkharusi (2012) explained that an observed score for any student obtained through some measurement procedure could be decomposed into the true score and a single error. Since the performances of students in National Examinations Council (NECO) Senior School Certificate Examination (SSCE) are based on the sum of their total scores, that is, CTT, there is a need to consider the psychometric properties (difficulty, discrimination, reliability, validity) of the test in taking decisions on the observable performance of candidates in order to improve upon test construction, administration and analysis. Reliability and validity are two technical properties that indicate the quality and usefulness of tests as well as major factors to be considered in the construction of test items for examinations. Junker (2012) described reliability as the extent to which the test would produce consistent results if it is administered again under the same conditions. It also reflects how dependably a test measures a specific characteristic. This consistency is of three types: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability). Many reasons can be adduced for an individual not getting exactly the same test score every time he or she takes the test. These include the test taker's temporary psychological or physical state, multiple raters and test forms. These factors are sources of chance or random measurement error in the assessment process. If there are no random errors of measurement, the individual will get the same test score, that is, the individual’s true score each time. The degree to which test scores are unaffected by measurement errors is an indication of the reliability of the test.

Reliability is threatened when errors occur in measurement. When a measure is consistent over time and across items, one can conclude that the scores represent what they intend to; meanwhile, there is more to it because a measure can be reliable but not valid. Reliability and validity are therefore needed to assure adequate measurement of the construct of interest. Validity refers to what characteristic the test measures and how well the test measures that characteristic. In other words, it determines the extent to which a measure adequately represents the underlying construct that it is supposed to measure. Valid conclusions cannot be drawn from a test score unless one is sure that the test is reliable. Even when a test is reliable, it may not be valid. Therefore, care should be taken to ensure that any test selected is both reliable and valid for the situation. The accuracy and validity of the interpretation of test results are determined by the inferences made from test scores. Validity of inferences is concerned with the negative consequences of test score interpretation that is traceable to construct under-representation or construct-irrelevance variance. The focus should be on the theoretical dimensions of the construct a test is intending to measure in order to prevent inappropriate consequences from test score interpretation. Generally, in testing, it is necessary to consider how test-takers’ abilities can be inferred based on their test scores. Student marks are affected by various types of errors of measurement which always exist in them, and these reduce the accuracy of measurement. The magnitude of measurement error is incorporated in the concept of reliability of test scores, where reliability itself quantifies the consistency of scores over replications of a measurement procedure. Also, it is often expected that test score variation should only be due to an artifact of test-takers’ differing abilities and task demands. But in reality, it is being proven that test-takers’ scores are most of the time affected by other factors, including test procedures, personal attributes other than abilities, and other random factors. A single score obtained on one occasion on a particular form of a test with a single administration as done by NECO is not fully dependable because it is unlikely to match that person’s average score over all acceptable occasions, test forms, and administrations. A person’s score would usually be different on other occasions, on other test forms, or with different administrators. Which are the most serious sources of inconsistency or error? Where feasible, it is expected that error variances that arise from each identified source be estimated. Regardless of the strengths of g-theory, it has not been widely applied specifically to estimate the dependability of scores of students in secondary school examinations in Nigeria.

In Nigeria, at the end of secondary school education, students are expected to write certification examinations such as the SSCE conducted by the West Africa Examination Council (WAEC) and the NECO, or the National Business and Technical Certificate Education (NBTCE) conducted by the National Business and Technical Examination Board (NABTEB). The NECO conducts the SSCE in June/July and November/December every year. It was established in 1999 to reduce the workload of WAEC, especially to mitigate the burden of testing a large number of candidates. It was also to democratize external examination by providing candidates with a credible alternative. While some Nigerians saw NECO’s arrival as an opportunity for choice of examination body for candidates to patronize, others doubted its capacity to conduct reliable examinations that could command widespread national and international respect and acceptability.

English language education is a colonial legacy that has deeply entrenched in Nigerian heritage and apparently become indispensable. It is widely recognized as an instrument par excellence for socio-cultural and political integration as well as economic development. Its use as a second language as well as the language of education provided a speedy access to modern development in science and technology (Olusoji, 2012). It is for the above reasons that much importance is attached to English Language education nationwide and at all levels of the nation's educational system. To date, the English language remains the major medium of instruction at all levels of education in Nigeria, and no student can proceed to the tertiary level without a minimum of pass in the English language. In addition, considering the importance of the English language as an international language and its influence on Nigerian secondary school students' performance, it is imperative that generalizability theory be used to examine the credibility of secondary school examinations, hence this study.

Purpose of the Study

The objectives of the study are to:

1. Determine the generalizability coefficient of the English Language items;
2. Estimate the phi (dependability) coefficient of the English Language items; and
3. Determine the validity of the English Language items.
4. Conduct a D-study to determine the generalizability and phi coefficients based on the results of G- study.

METHOD

The study adopted the ex post facto research design. This type of design examines the cause and effect through selection and observation of existing variables without any manipulation of existing relations.

Sample

The total population of students who sat for NECO SSCE English Language examination in the year 2017 in Nigeria was 1,037,129, out of which 311,138 candidates constituted the study sample. The sample was selected using a proportionate stratified sampling technique. Thirty percent of the candidates were randomly selected from each state. The detail is presented in Table 1.

Data Collection Techniques

The data used in the study were responses of the candidates (to the 100-item multiple-choice test) who wrote the NECO June/July 2017 English language SSCE in Nigeria as indicated on the Optical Marks Record (OMR) sheets obtained from NECO office.

Instrument

The instrument used for the study was the OMR sheets for the NECO June/July 2017 English language objective items. The OMR sheets contained the responses of examinees to the NECO June/July 2017 English Language objective items paper III. The English Language examination is a dichotomously scored multiple-choice examination consisting of 100 items with five options length. The responses of the examinees were scored 1 and 0 for correct and incorrect responses. The minimum score for an examinee from computation was zero while the maximum score was 100.

Table 1. Population and Sample Size of English Language Candidates Who Sat for NECO Senior School Certificate Examination in 2017

States	Population	Sample size
Abia	10405	3121
Adamawa	37320	11196
Akwa Ibom	23059	6917
Anambra	20509	6152
Bauchi	41413	12424
Bayelsa	4346	1304
Benue	40196	12059
Borno	27439	8232
Cross Rivers	17583	5275
Delta	16647	4994
Ebonyi	10540	3162
Edo	21659	6498
Ekiti	11429	3429
Enugu	26231	7869
FCT	18517	5555
Gombe	25526	7658
Imo	23587	7076
Jigawa	21387	6416
Kaduna	51860	15558
Kano	88227	26468
Katsina	34613	10384
Kebbi	26567	7970
Kogi	28157	8447
Kwara	22079	6624
Lagos	52392	15718
Nasarawa	35950	10785
Niger	33414	10024
Ogun	25212	7564
Ondo	26558	7967
Osun	26126	7838
Oyo	54828	16448
Plateau	34391	10317
Rivers	11484	3445
Sokoto	25379	7614
Taraba	19874	5962
Yobe	17063	5119
Zamfara	25162	7549
Total	**1037129**	**311138**

Data Analysis

The data were analyzed using “lme4” package of R language and environment for statistical computing, factor analysis and Tucker index of factor congruence. The generalizability study was conducted with fitting linear mixed-effect models using lme4 package of R language and environment for statistical computing to find the g-coefficient and phi coefficient. Factor analysis was conducted to identify one dimension underlying the English language test for male and female samples. Thereafter the extracted factor loadings for the test under male and female samples were compared. The comparison of the extracted factor loadings in two samples was made using Tucker index of factor congruence.

RESULTS

One-facet (pxi) design of generalizability theory was adopted to determine the generalizability coefficient. This is because there is a single facet; the items (i) and the persons (p) are the objects of measurement. However, to conduct the analysis under generalizability theory, two levels of analysis were conducted as recommended by Shavelson and Webb (1991). The analysis includes the generalizability (G) study and the decision (D) study. First, the G-study was conducted, and thereafter the D-study was conducted based on the result of the G-study for the extraction of the generalizability coefficient. The analysis was conducted with fitting linear mixed-effect models using lme4 package (Bates, Mächler, Bolker and Walker, 2015) of R language and environment for statistical computing.

Table 2 presents the estimated variances from the G study. The table shows the magnitude of error in generalizing from a candidate's scores on 2017 NECO English language test to the universe score. A useful exploratory approach for interpreting the variances that are estimated in a G study is to calculate the percentage of the total variance that each variance component represents. These percentages are presented in the last column of Table 2.

Table 2. Parameters of G-Study for 2017 NECO English Language Test

Source	Variance Component	Estimated Variance	Percent of Variability
Person	σ_p^2	0.0142	6.0
Item	σ_i^2	0.0747	31.60
Residual	$\sigma_{pi,e}^2$	0.1472	62.30

The table shows that the variance component for candidates (i.e., the universe score variance) accounts for only 0.0142 or 6.0% of all the variance, and this is rather low. Furthermore, the variance component for the items (0.0747, or 31.6% of the total variance) is large relative to the universe score variance but smaller than the residual variance (0.1472 or 62.3% of the total variance).

Figure 1 presents the histogram that calculates the percentage of items that each candidate got correct. The Figure shows that none of the participants got all the items correct or incorrect and that the overwhelming majority of participants got 60% or 70% of the items correct on the test (i.e., 60 to 70 correct answers). This tight clustering accounted for the observed low universe score variance.

Table 3 shows the proportion of correct items obtained by the candidates for the 100 items 2017 NECO English language test. The table shows that the proportion of item correct ranges from .02 to .91, which reflects a lot of variation and corroborates the high percent of variation accounted for by the items. The large residual variance captures both the person by item interaction and the random error (which we are unable to disentangle). Maybe some items were more easily answered by some participants or maybe there was systematic variation such as the physical environment where the test was administered, or possibly other random variation like fatigue during the assessment. Whatever the cases, these sources could not be disentangled from one another in this variance component.

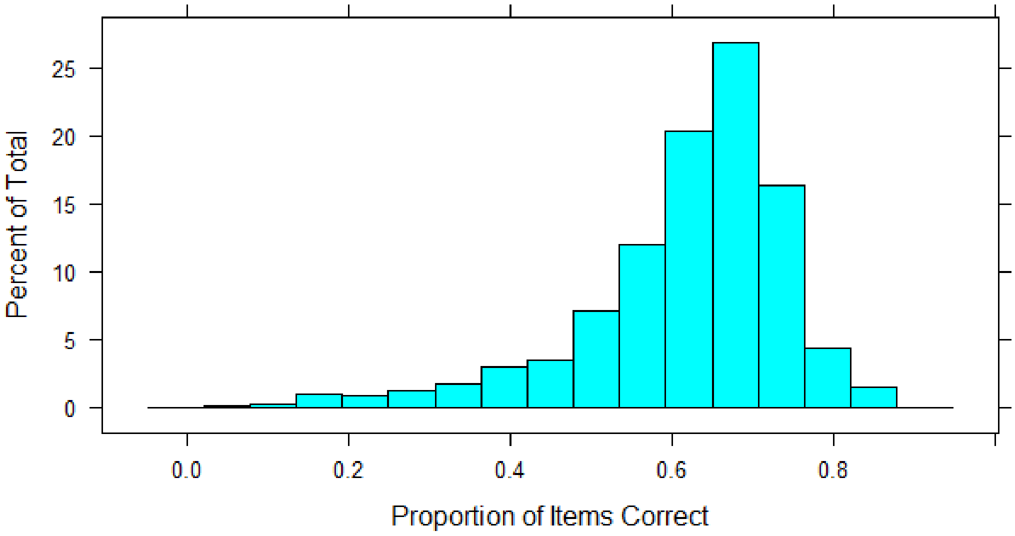


Figure 1. Distribution of Candidates' Proportion of Item Correct

Table 3. Means of 2017 NECO English Language Test Source

Item	Mean	Item	Mean	Item	Mean	Item	Mean
1	.79	26	.21	51	.70	76	.16
2	.32	27	.48	52	.82	77	.81
3	.87	28	.78	53	.89	78	.81
4	.74	29	.84	54	.36	79	.76
5	.79	30	.86	55	.91	80	.34
6	.69	31	.70	56	.81	81	.05
7	.84	32	.83	57	.87	82	.22
8	.81	33	.83	58	.74	83	.28
9	.85	34	.79	59	.29	84	.60
10	.66	35	.61	60	.83	85	.74
11	.33	36	.83	61	.33	86	.79
12	.75	37	.81	62	.83	87	.80
13	.73	38	.84	63	.44	88	.14
14	.86	39	.75	64	.82	89	.13
15	.83	40	.78	65	.84	90	.70
16	.44	41	.27	66	.76	91	.08
17	.88	42	.86	67	.81	92	.72
18	.80	43	.24	68	.40	93	.04
19	.84	44	.83	69	.71	94	.08
20	.71	45	.86	70	.51	95	.09
21	.86	46	.37	71	.36	96	.06
22	.70	47	.84	72	.02	97	.02
23	.74	48	.83	73	.77	98	.11
24	.84	49	.83	74	.85	99	.53

Generalizability Coefficient of 2017 NECO English Language Test

The generalizability coefficient is similar to the reliability coefficient in CTT. It is the ratio of the universe score to the expected observed score variance. For relative decisions and a pxi random-effects design, the generalizability coefficient is calculated as:

$$Ep^2_{X_p} i.u^p = Ep^2 = \frac{E_p(\mu_p-\mu)^2}{E_pE_i(X_{pi}-\mu_i)^2} = \frac{\sigma_p^2}{\sigma_p^2+\sigma_\delta^2} \qquad (1)$$

$$\frac{\sigma_p^2}{\sigma_p^2+\sigma_\delta^2} = \frac{0.0142}{0.0142+0.0015} = 0.9046$$

where σ_p^2 is the variation of students' test scores (the universe-score variance), σ_δ^2 is the relative error variance (Desjardins & Bulut, 2018) Table 4 presents the result.

Table 4. Generalizability Coefficient

Source	Estimate
Variance of person	0.0142
Relative error variance	0.0015
Generalizability coefficient	0.9046

Table 4 shows the parameter used for the estimation of the generalizability coefficient of the 100-item 2017 NECO English language test. The table shows that the generalizability coefficient of the NECO test was .90. The generalizability coefficient of the test was high, suggesting that the test was highly reliable.

To determine the dependability coefficient, D-study was conducted based on the G-study conducted in objective 1. Thereafter, the dependability of the NECO test was extracted from the D-study. As in the case of the generalizability coefficient, lme4 package was used for the analysis. The dependability coefficient is calculated with:

$$Dependability\ coefficient = \Phi = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{abs}^2} \qquad (2)$$

$$\Phi = \frac{0.0142}{0.0142 + 0.0022} = 0.8659$$

where σ_s^2 is the variation of students' test scores (the universe-score variance), and σ_{abs}^2 is the absolute error variance (Desjardins & Bulut, 2018). Table 5 presents the result.

Table 5. Dependability Coefficient

Source	Estimate
Variance for person	0.0142
Absolute error variance	0.0022
Dependability coefficient	0.8659

Table 5 shows the parameter used for the estimation of the dependability coefficient of the 100-item 2017 NECO English Language test. It shows that the dependability coefficient of the NECO test was .87. The result showed that the 2017 NECO English test scores were highly dependable. This implies that candidates' scores obtained on the 2017 NECO English language test were highly dependable in terms of reflecting the ability of the candidates.

Table 6. Decision Study

Number of items	Relative error var.	Absolute error var.	G coefficients	Phi coefficients
90	0.0017	0.0024	.90	.86
80	0.0019	0.0028	.88	.84
70	0.0021	0.0031	.87	.82
60	0.0025	0.0037	.85	.79
50	0.003	0.0044	.81	.76

As can be seen from Tables 4 and 5, the G and phi coefficients for 100-items fully crossed random designs were estimated as .90 and .87 respectively. Table 6 shows the D-study results obtained by reducing the number of items. When the number of items was reduced from 90 to 80, the relative error variance increased from 0.0017 to 0.0019; the absolute error variance also increased from 0.0024 to 0.0028; the g-coefficient decreased from .90 to .88 and phi coefficient also decreased from .86 to .84. The D-study is particularly useful in determining which combination of various measurement methods can be employed to obtain reliable coefficients.

Two levels of analysis were conducted to determine the extent to which the test was able to measure the same trait among male and female students. Factor analysis was conducted to identify one dimension underlying the English language test for male and female samples. Thereafter the extracted factor loadings for the test under male and female samples were compared. The comparison of the extracted factor loadings in two samples was made using Tucker index of factor congruence. The congruence coefficient is the cosine of the angle between two vectors and can be interpreted as a standardized measure of the proportionality of elements in both vectors. It is evaluated as:

$$\phi(x,y) = \frac{\sum_{n=i}^{N} x_i \, y_i}{\sqrt{\sum_{n=i}^{N} x_i^2 \; \sum_{n=i}^{N} y_i^2}} \qquad (3)$$

where x_i and y_i are loadings of variable i on factor x and y, respectively, i = 1, 2, 3, …, n (in this case n = 100). Usually, the two vectors are columns of a pattern matrix. Therefore, how large should the coefficient be before two factors from two samples can be considered highly similar? Lorenzo-Seva and Ten Berge (2006) suggested that a value in the range of .85-.94 corresponds to a fair similarity, while a value higher than .95 implies that the two factors or components compared can be considered equal. The estimated factor loadings and other parameters for the estimation of the congruence index are presented in Appendix.

The table shows the parameters of the Tuckers index for congruence estimation. These parameters were substituted for in Equation 3. The result is presented as follows.

$\sum_{n=i}^{N} x_i\, y_i = 34.06$, $\sum_{n=i}^{N} x_i^2 = 32.31$, $\sum_{n=i}^{N} y_i^2 = 35.98$. Therefore,

$\phi(x,y) = \frac{34.06}{(32.31)(35.98)} = \frac{34.06}{\sqrt{1162.514}} = \frac{34.06}{34.10} = 0.9988$

The result showed that Tucker congruence index of similarity of the factors estimated under male and female candidates' samples was .99. This indicates that the factor underlying the performance of male candidates was almost identical with the factor underlying the female candidates' performance. The implication of the result is that the construct validity of the 2017 NECO English language test was very high and the test measured to a great extent the proficiency of students in the English language, and there was no other nuisance factor(s).

DISCUSSION and CONCLUSION

The findings of this study also showed the magnitude of error in generalizing from a candidate's score on 2017 NECO English language test to a universe score, as shown in Table 2. All 100 dichotomously scored items were analyzed using generalizability theory (G- theory) in a single-facet crossed study of persons (*p*) crossed with items (*i*). The variance component for candidates (i.e., the universe score variance) accounts for a smaller percentage of all the variance, corresponding to the largely similar scores obtained by the examinees. In order to reach more reliable results, it is generally desired that the number of moderate difficult items in the test is higher and the number of easy and difficult items relatively less; most of these items are of moderate difficulty. Therefore, none of the examinees scored all the items correct or incorrect; the majority of them scored between 60% and 70% of the items correct in the test. The tight clustering accounted for the observed low universe score variance. Furthermore, the variance component for the items is large relative to the universe score variance but smaller than the residual variance. The proportion of items that is correct reflects a lot of variations which corroborate the high percentage of variation accounted for by the items. The large residual variance captures both the person by item interaction and the random error, which cannot be disentangled. The high estimated variance component for persons crossed with items and the error is an indicator that almost 2/3 of the variability (random error) lies within this relationship and provides an estimate in the changes in the relative standing of a person from item to item (see Table 2). The result is in agreement with the findings of de Vries (2012) that the majority of error variance for the examination could be due to the interaction of persons with items, and lowering this variance would lead to an increase in dependability.

For relative decisions and a random-effects design, the generalizability coefficient is highly reliable. The dependability coefficient, Φ, an index that reflects the contribution of the measurement procedure to the dependability of the examination was also very dependable. As claimed by Brennan (2003) and Strube (2002), values approaching one (1) indicate that the scores of interest can be differentiated with a high degree of accuracy despite the random fluctuations of the measurement conditions. An important advantage of Φ is that it can be used to determine the sources of error that reduce classification accuracy and the methods to best improve such classifications, although most authors examined variability across facets to determine which one will be of greater benefit to generalizability. These results are consistent with the findings of Gugiu, Gugiu and Baldus (2012), Fosnacht and Gonyea (2018), Tasdelen-Teker, Sahin and Baytemir (2016), Nalbantoglu-Yilmaz (2017), Kamis and Dogan (2018) and Rentz (1987) who reported that the acceptable standards for dependability should be ≥ .70.

The study is also in contrast to the findings of Uzun Aktas, Asiret and Yorulmaz (2018), de Vries (2012) and Solano-Flores and Li (2006), who argued that each test item poses a unique set of linguistic challenges and each student has a unique set of linguistic strengths and weaknesses. Therefore, a certain number of items would be needed to obtain dependable scores. Uzun et al. (2018) and de Vries (2012) also pointed out that increasing the number of raters or occasions would increase the score dependability when rater and occasion are considered as facets. Li, Shavelson, Yin and Wiley (2015) confirmed that increasing the number of items reduces error variance and increases both G and phi coefficients.

Based on the outcome of Tucker congruence index of similarity of the factors estimated under male and female candidates' samples (.99), the factor underlying the performance of male candidates was almost identical with the factor underlying the female candidates' performance. This implies that the examination measures to a great extent proficiency of students in the English Language. The result is in agreement with Zainudin (2012), who reported that the factor loading for an instrument must be higher or equal to .50. Also, Lorenzo-Seva and Ten Berge (2006) suggested that a value in the range of .85-.94 corresponds to a fair similarity, while a value higher than .95 implies that the two factors or components compared can be considered equal.

Conclusion

The study reflected that the reliability was high, which established that the scores assigned to candidates were dependable and generalizable. Also, the item validity was high because it measured the underlying construct, which underscores the good credibility of the items.

Recommendation

Prospective users of a measurement procedure are therefore advised to consider explicitly various sources of variation. They have to state whether they are interested in making absolute or relative decisions and whether they wish to generalize overall or only certain facets of a measurement procedure. However, there is a need to apply this concept to all school subjects to ensure the generalizability of the certification examinations.

REFERENCES

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education, 2*(1), 184-196. doi: 10.5296/jse.v2i1.1227

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01

Breithaupt, K. (2011). *Medical licensure testing: White paper for the assessment review task force of the medical council of Canada*. Retrieved from https://www.mcc.ca/wp-content/uploads/Technical-Reports-Breithaupt-2011.pdf

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

Brennan, R. L. (2003). *Coefficients and indices in generalizability theory* (CASMA Research Report Number 1). Iowa: Centre for Advanced Studies in Measurement and Assessment.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. U.S.A: Cengage Learning.

de Gruijter, D. N., & van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioural sciences*. New York: Chapman & Hall/CRC.

de Vries, I. M. (2012). *An analysis of test construction procedures and score dependability of a paramedic recertification exam* (Master's thesis). Queen's University Kingston, Ontario, Canada.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R* (1st Ed). Parkway, Boca Raton: Chapman and Hall/CRC Press. doi: 10.1201/b20498

Fosnacht, K., & Gonyea, R. M. (2018). The dependability of the updated NSSE: A generalizability study. *Research and Practice in Assessment 13*, 62-73. Retrieved from https://eric.ed.gov/?id=EJ1203503

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of the grades assigned to undergraduate research papers. *Journal of Multidisciplinary Evaluation, 8*(19), 26-40. Retrieved from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/362

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Coventy: Ofqual

Junker, B. W. (2012). *Some aspects of classical reliability theory and classical test theory*. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Kamis, O., & Dogan, C. D. (2018). An investigation of reliability coefficients estimated for decision studies in generalizability theory. *Journal of Education and Learning*, 7(4), 103-113. doi: 10.5539/jel.v7n4p103

Li, M., Shavelson, R. J., Yin, Y., & Wiley, W. (2015). Generalizability theory. In *The* encyclopedia of clinical psychology (pp. 1322-1340). doi: 10.1002/9781118625392.wbecp352

Lorenzo-Seva, U., & Ten Berge, J. U. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. doi: 10.1027/1614-2241.2.2.57

Mushquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavioral Research Methods, 38,* 542-547. doi: 10.3758/BF03192810

Nalbantoglu-Yilmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice*, *17*(2), 395-409. doi: 10.12738/estp.2017.2.0098

Olusoji, O. A. (2012). Effects of English language on national development. *Greener Journal of Social Sciences*, *2*(4), 134-139. doi: 10.15580/GJSS.2012.4.08291255

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research*, *24*(1), 19-28. doi: 10.1177/002224378702400102

Shavelson, R. J., & Webb, N. M.1(991). *Generalizability theory: A Primer.* Newbury Park CA: Sage.

Solano-Flores, G., & Li, M. (2006). The use of generalizability theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, *25*(1), 13-22. doi: 10.1111/j.1745-3992.2006.00048.x

Strube, M. J. (2002). Reliability and generalizability theory. In L.G. Grimm and P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Tasdelen-Teker, G., Sahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences*, *13*(3), 5574-5586. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/4155

Uzun, N. B., Aktas, M. Asiret, S., & Yorumalz, S. (2018). Using generalizability theory to assess the score reliability of communication skills of dentistry students. *Asian Journal of Education and Training, 4*(2), 85-90. doi: 10.20448/journal.522.2018.42.85.90

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*, 81-124. doi: 10.1016/S0169-7161(06)26004-8

Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education*, *21*(3), 273-291. doi: 10.1080/08957340802161840

Zainudin, A. (2012). *Research methodology and data analysis* (5th Ed). Shah Alam: Universiti Teknologi MARA Publication Centre (UiTM Press).

Appendix. Factor Loading of English Test in Male and Female Examinees Groups

Item	Female (X)	Male (Y)	XY	X^2	Y^2
1	0.62	0.65	0.40	0.38	0.42
2	-0.46	-0.47	0.21	0.21	0.22
3	0.63	0.63	0.39	0.39	0.39
4	0.63	0.63	0.40	0.40	0.40
5	0.61	0.63	0.38	0.37	0.39
6	0.61	0.63	0.38	0.37	0.39
7	0.51	0.54	0.28	0.26	0.29
8	0.55	0.57	0.31	0.30	0.33
9	0.58	0.60	0.35	0.33	0.36
10	0.67	0.70	0.47	0.45	0.49
11	-0.48	-0.49	0.23	0.23	0.24
12	0.62	0.67	0.41	0.38	0.44
13	0.70	0.69	0.48	0.49	0.48
14	0.48	0.54	0.26	0.23	0.29
15	0.45	0.48	0.22	0.20	0.23
16	-0.55	-0.57	0.31	0.30	0.32
17	0.50	0.54	0.27	0.25	0.29
18	0.51	0.53	0.27	0.26	0.29
19	0.61	0.60	0.37	0.37	0.36
20	0.72	0.73	0.53	0.52	0.54
21	0.64	0.69	0.44	0.41	0.48
22	0.57	0.59	0.34	0.32	0.35
23	0.61	0.63	0.39	0.38	0.40
24	0.48	0.53	0.25	0.23	0.28
25	0.64	0.69	0.44	0.41	0.47
26	-0.40	-0.43	0.17	0.16	0.18
27	0.78	0.78	0.61	0.60	0.61
28	0.59	0.62	0.37	0.35	0.39
29	0.61	0.67	0.41	0.38	0.45
30	0.65	0.68	0.44	0.42	0.46
31	0.63	0.66	0.42	0.40	0.43
32	0.66	0.70	0.46	0.43	0.49
33	0.68	0.74	0.50	0.47	0.54
34	0.71	0.74	0.53	0.50	0.55
35	0.79	0.81	0.64	0.63	0.66
36	0.60	0.67	0.40	0.36	0.44
37	0.74	0.77	0.57	0.55	0.59
38	0.58	0.67	0.39	0.34	0.44
39	0.63	0.69	0.43	0.40	0.47
40	0.59	0.64	0.38	0.35	0.41
41	-0.39	-0.43	0.17	0.15	0.19
42	0.66	0.68	0.45	0.43	0.46
43	-0.43	-0.45	0.19	0.18	0.20
44	0.60	0.68	0.41	0.36	0.46
45	0.64	0.71	0.45	0.41	0.50
46	-0.44	-0.46	0.20	0.20	0.21
47	0.54	0.62	0.33	0.29	0.38
48	0.53	0.62	0.33	0.28	0.39
49	0.61	0.68	0.41	0.37	0.46
50	-0.51	-0.49	0.25	0.26	0.24
51	0.66	0.72	0.48	0.44	0.52
52	0.54	0.59	0.32	0.29	0.35
53	0.69	0.74	0.51	0.47	0.55
54	-0.55	-0.55	0.30	0.30	0.30
55	0.69	0.75	0.52	0.47	0.56
56	0.51	0.57	0.29	0.26	0.32
57	0.48	0.57	0.27	0.23	0.33
58	0.65	0.69	0.45	0.42	0.48
59	-0.48	-0.48	0.23	0.23	0.23
60	0.41	0.49	0.20	0.17	0.24
61	-0.57	-0.58	0.33	0.32	0.34
62	0.65	0.71	0.46	0.42	0.51

(continued)

Factor Loading of English Test in Male and Female Examinees Groups (continue)

Item	Female (X)	Male (Y)	XY	X^2	Y^2
63	-0.55	-0.56	0.31	0.30	0.32
64	0.65	0.69	0.45	0.42	0.48
65	0.48	0.57	0.27	0.23	0.32
66	0.56	0.61	0.34	0.31	0.37
67	0.45	0.51	0.23	0.20	0.26
68	-0.61	-0.60	0.36	0.37	0.36
69	0.65	0.69	0.45	0.42	0.48
70	0.73	0.76	0.55	0.53	0.57
71	-0.59	-0.58	0.34	0.34	0.34
72	-0.44	-0.46	0.20	0.19	0.21
73	0.67	0.66	0.44	0.44	0.43
74	0.54	0.57	0.31	0.29	0.32
75	0.63	0.67	0.42	0.40	0.45
76	-0.39	-0.37	0.15	0.15	0.14
77	0.60	0.65	0.39	0.36	0.42
78	0.50	0.56	0.28	0.25	0.32
79	0.59	0.60	0.35	0.34	0.36
80	-0.49	-0.50	0.24	0.24	0.25
81	-0.34	-0.37	0.13	0.11	0.14
82	-0.40	-0.40	0.16	0.16	0.16
83	-0.47	-0.46	0.22	0.22	0.21
84	0.70	0.70	0.49	0.48	0.49
85	0.48	0.53	0.25	0.23	0.28
86	0.48	0.53	0.25	0.23	0.28
87	0.50	0.54	0.27	0.25	0.29
88	-0.27	-0.27	0.07	0.07	0.07
89	-0.42	-0.42	0.17	0.17	0.17
90	0.58	0.60	0.35	0.34	0.36
91	-0.44	-0.44	0.19	0.19	0.19
92	0.66	0.68	0.45	0.44	0.46
93	-0.39	-0.40	0.16	0.15	0.16
94	-0.38	-0.40	0.15	0.15	0.16
95	-0.53	-0.53	0.28	0.28	0.28
96	-0.44	-0.49	0.22	0.19	0.24
97	-0.14	-0.18	0.03	0.02	0.03
98	-0.50	-0.50	0.25	0.25	0.25
99	0.69	0.72	0.50	0.48	0.52
100	0.56	0.60	0.33	0.31	0.36
Total			34.06	32.31	35.98

Nijerya’da İngilizce Sınavına Katılan Öğrenci Puanlarının Güvenilirliğinin Genellenebilirlik Kuramı ile İncelenmesi

Giriş

Genellenebilirlik Kuramı, yazılı veya bilgisayar tabanlı gerçekleştirilen bilgi testlerinin, derecelendirme ölçeklerinin veya öz değerlendirme ölçeklerinin ve kişilik testleri gibi performans testlerinin vb. psikometrik testlerin sonuçlarını analiz etmek için kullanılan istatistiksel bir yöntemdir (Breithaupt, 2011). Tek bir ölçüm sürecinde eşzamanlı olarak ortaya çıkan ve sonuçlara karışan birden çok hata kaynağını ayrıştırdığı için genellenebilirlik kuramı, (G-Kuramı) gerçek sonuçlara ulaşmayı hedefler. Gözlemlenen bir puanı, evren puanı için bir bileşene ve bir veya daha fazla hata bileşenine ayrıştırılarak eldeki bir gözlemden uygun gözlem evrenine genelleme yapılması amaçlanır. Klasik Test Teorisinde (KTT) imkânsız olan çeşitli olası etkileşimlerden kaynaklanan tutarsızlıkların yanı sıra hem değerlendiriciler arası hem de görevler arası tutarsızlıkları eş zamanlı olarak hesaba katarken her bir sınava giren kişi için ortalama derecelendirmenin güvenilirliğini tahmin edebilmesi açısından da avantajlıdır (Brennan, 2001). G Kuramında ölçümün gerçek değerinden uzaklaşmasına neden olan çeşitli hata kaynakları araştırılır. Ölçümün hata bileşenlerini çözme fırsatı verir ve ayrıca davranışsal ölçümün güvenilirliği veya güvenilirliği ile ilgilendiği için ölçme ve değerlendirme yönteminin kalitesini değerlendirmede ve kesinliğini geliştirmede değerli bir araçtır.

Tüm test puanları, diğer tüm ölçümler gibi, test puanlarının güvenilirliğini etkileyen bazı hatalar içerir. Aynı koşullar altında ölçümde farklılıklar olduğunda hata devreye girer. Ölçümde hata, kişinin gözlenen puanı ile gerçek puanı arasındaki fark olarak tanımlanabilir. Breithaupt (2011), maddelere ve test puanlarına karışan iki tür ölçüm hatası tanımlamıştır: rastgele ve sistematik hata. Elde edilen puanlardaki çeşitlilik genellikle hata kaynaklarına atfedilir ve bu nedenle bir testin psikometrik özelliğini belirleme zorluğunu ortaya çıkar. Psikometrik analizin amacı, gözlemlenen puanın (X) gerçek puanın (T) iyi bir ölçüsü olması için, mümkünse hata varyansını tahmin etmek ve en aza indirmektir. Gerçek değer tam olarak bilinmese de ideal değer bilinmeye çalışılabilir. KTT, gözlemlenen herhangi bir puan, çeşitli hata kaynaklardan gelse bile gerçek bir bileşen ile rastgele bir hata bileşeninin birleşimi olarak görülür. Bununla birlikte herhangi bir zamanda yalnızca tek bir ölçüm hata kaynağı incelenebilir. Genellenebilirlik Kuramı aynı zamanda evren puanına veya ölçüm sürecindeki tüm olası varyasyonlarda (örneğin farklı puanlayıcılar, formlar veya maddeler) beklenen ortalama puana odaklanır. Bu evren puanının, ölçüm nesnesi için belirli bir özelliğin değerini temsil ettiğine inanılır (Crocker & Algina, 2008). G Kuramı, birden fazla ölçüm hatası varyansının eşzamanlı etkisine odaklandığından araştırmacılara daha fazla geri bildirim sağlamaktadır.

Bazı araştırmalar, (Mushquash & O’Connor, 2006; Webb, Shavelson, & Haertel, 2006) çeşitli varyans kaynaklarının etkilerinin, belirli bir zamanda yalnızca tek bir ölçüm hatası kaynağının incelenmesinin mümkün olduğu KTT modelleri kullanılarak test edilebileceğini belirtmişlerdir. Ancak farklı hata kaynakları arasında meydana gelen etkileşim etkilerini incelemek mümkün değildir. Genellenebilirlik Kuramı, araştırmacılara özellikle bu konuda katkı sağlamaktadır; ölçüm durumunun her bir başarısı, test puanlarında ve onun adlandırılmış boyutunda bir hata kaynağıdır. Bu nedenle, birçok yazarın işaret ettiği gibi (Brennan, 2001; Johnson & Johnson, 2009) çok sayıda hata kaynağının açıklanamaması ve araştırmacıların KTT’nin olası hata kaynaklarını belirleyememesi ve aynı anda inceleyememesi G Kuramı’nın gelişmesini sağlamıştır. G Kuramı iki tür çalışmayı içerir: Genellenebilirlik çalışması (G-çalışması) ve Karar çalışması (D çalışması). Bir G-çalışmasının temel amacı, çeşitli kaynaklarla ilişkili puan varyansının bileşenlerini tahmin etmektir. D-çalışması ise bu tahmin varyans bileşenlerini kullanarak sonraki ölçüm için alternatifleri değerlendirerek optimal sonuca ulaşmaktır.

Alkharusi (2012) herhangi bir öğrenci için bazı ölçüm prosedürleriyle elde edilen gözlenen puanın gerçek puana ve tek bir hataya ayrıştırılabileceğini açıklamıştır. Ulusal Sınav Konseyi (NECO) Kıdemli Okul Sertifika Sınavında (Senior School Certificate Examination-SSCE) öğrencilerin performansları toplam puanlarının yani KTT'nin toplamına dayandığından, karar alırken testin psikometrik özelliklerinin (zorluk, ayırt edicilik, güvenirlik, geçerlik) için testin yapısı, yönetimi ve analizine yönelik iyileştirme çalışmaları için adayların gözlemlenebilir performansının dikkate alınmasına ihtiyaç vardır. Güvenilirlik ve geçerlik, testlerin kalitesini ve kullanışlılığını ve ayrıca sınavlar için test maddelerinin oluşturulmasında dikkate alınması gereken ana faktörleri gösteren iki psikometrik özelliktir. Junker (2012) güvenilirliği, testin aynı koşullar altında tekrar uygulandığında tutarlı sonuçlar üreteceği kapsam olarak tanımlamıştır. Rastgele ölçüm hatası yoksa birey her seferinde aynı test puanını, yani gerçek puanı alacaktır. Güvenilir bir ölçüm geçerli olmayabileceğinden her biri için ayrı ayrı kanıt toplanması gerekmektedir. Ayrıca güvenirlik, geçerlik bir ön koşul olduğundan ölçüm sonuçlarının öncelikle güvenirliğine yönelik kanıtlar toplanabilir. Ölçme sonuçlarına karışan hatalar, öncelikle güvenilirliği etkiler ancak hatalar, geçerliği de tehdit eder. Bu nedenle ilgilenilen yapının yeterli ölçümünü sağlamak için her ikisine yönelik kanıtların toplanmasına ihtiyaç vardır. Öğrenci notları, ölçümün doğruluğunu azaltan çeşitli hata türlerinden etkilenir. NECO tarafından yapılan tek uygulamalı bir testin belirli bir formunda bir seferde elde edilen tek bir puan tamamen güvenilir değildir çünkü o kişinin tüm kabul edilebilir durumlar, test formları ve uygulamalardaki ortalama puanıyla eşleşmesi olası değildir. Bir kişinin puanı genellikle diğer durumlarda, test formlarında veya farklı yöneticilerle farklı olacaktır. En ciddi tutarsızlık veya hata kaynakları hangileridir? Mümkün olduğunda, tanımlanan her bir kaynaktan kaynaklanan hata varyanslarının tahmin edilmesi beklenir. G-Kuramının güçlü yönlerinden bağımsız olarak, Nijerya'da ortaokul sınavlarındaki öğrencilerin puanlarının güvenilirliğini tahmin etmek için özel olarak geniş çapta uygulanmamıştır.

Nijerya'da, ortaokul eğitiminin sonunda, öğrencilerin Batı Afrika Sınav Konseyi (WAEC) ve NECO tarafından yürütülen SSCE veya Ulusal Sınavlar gibi sertifika sınavları yazmaları beklenir. Ulusal İş ve Teknik İnceleme Kurulu (NABTEB) tarafından yürütülen İşletme ve Teknik Sertifika Eğitimi (NBTCE). NECO, SSCE'yi her yıl Haziran/Temmuz ve Kasım/Aralık aylarında yürütür. 1999 yılında WAEC'in iş yükünü azaltmak, özellikle çok sayıda adayı test etme yükünü azaltmak amacıyla kurulmuştur.

İngilizce eğitimi, Nijerya mirasına derinlemesine yerleşmiş ve mevcut durumda vazgeçilmez hâle gelen bir mirastır. Dil eğitimi; ekonomik kalkınmanın yanı sıra sosyo-kültürel ve politik entegrasyon için mükemmel bir araç olarak kabul edilmektedir. Eğitim dilinin yanı sıra İngilizcenin ülkede ikinci bir dil olarak kullanılması, bilim ve teknolojideki modern gelişmelere hızlı bir erişim sağlamıştır (Olusoji 2012). Söz konusu nedenlerden dolayı, ülke çapında ve ülke eğitim sisteminin tüm seviyelerinde İngilizce eğitimine büyük önem verilmektedir.

Bu nedenle, İngilizcenin uluslararası bir dil olarak önemi ve Nijeryalı ortaokul öğrencilerinin performansı üzerindeki etkisi göz önüne alındığında, ortaokul sınavlarının güvenilirliğini incelemek için genellenebilirlik kuramının kullanılması önem taşımaktadır.

Yöntem

Bu araştırma betimsel araştırma yöntemine dayalı olarak yürütülmüştür. Betimsel araştırmalar, mevcut ilişkilerin herhangi bir manipülasyonu olmaksızın, mevcut değişkenlerin seçilmesi ve gözlemlenmesi yoluyla neden ve sonucu ilişkisini incelemektedir. Nijerya'da 2017 yılında NECO SSCE İngilizce Dil Sınavı'na giren toplam 1,037,129 öğrenci bulunmakta olup sınava giren 311,138 aday, çalışmanın örneklemini oluşturmuştur. Örneklem seçkisiz örnekleme yöntemlerinden tabakalı örnekleme tekniği kullanılarak seçilmiştir. Her eyaletten adayların yüzde otuzu rastgele seçilerek çalışma yürütülmüştür. Çalışmada kullanılan veriler, NECO ofisinden alınan OMR sayfalarında belirtildiği gibi Nijerya'da NECO Haziran/Temmuz 2017 İngilizce SSCE yazan adayların (100 maddelik çoktan seçmeli teste) verdiği yanıtlardır. Verilerin analizinde G Kuramına dayalı olarak öncelikle G-çalışması, ardından D-çalışması yürütülmüştür.

Sonuç ve Tartışma

Bu çalışmada öncelikle bir adayın 2017 NECO İngilizce dil sınavındaki puanından bir evren puanına genellemede hatasının büyüklüğü incelenmiştir. Adaylar için varyans bileşeni, tüm varyansın daha küçük bir yüzdesini oluşturmaktadır. Sınava girenlerin aldığı puanların benzer olduğu bulunmuştur. Doğru cevaplandırılan maddelerin oranı, maddeler tarafından açıklanan yüksek çeşitlilik yüzdesini doğrulayan birçok farklılaşmayı yansıtır. Büyük artık varyans, hem kişi bazında madde etkileşimini hem de çözülemeyen rastgele hatayı göstermektedir. Araştırmanın sonuçları, de Vries'in (2012) inceleme için hata varyansının çoğunluğunun kişilerin maddelerle etkileşiminden kaynaklanabileceği ve bu varyansın düşürülmesinin güvenilirlikte bir artışa yol açacağı yönündeki bulgularıyla uyumludur.

Araştırma kapsamında NECO'ya katılan öğrencilerin cevapları doğrultusunda göreceli kararlar ve rastgele etkiler tasarımı için genellenebilirlik katsayısının oldukça yüksek olduğu tespit edilmiştir. Ölçüm prosedürünün muayenenin güvenilirliğine katkısını yansıtan bir indeks olan güvenilirlik katsayısı Φ da güvenilir bulunmuştur. Bu sonuçlar Gugiu, Gugiu ve Baldus (2012), Fosnacht ve Gonyea (2018), Taşdelen-Teker, Şahin ve Baytemir (2016), Nalbantoğlu-Yılmaz (2017), Kamış ve Doğan (2018) ve Rentz'in (1987) güvenilirlik için kabul edilebilir standartların ≥ .70 olması gerektiği bulgusuyla tutarlıdır.

Çalışma aynı zamanda Uzun, Aktaş, Aşiret ve Yorulmaz (2018), de Vries (2012) ve Solano-Flores ve Li (2006), her test maddesinin bir dizi dilsel zorluk oluşturduğunu ve her öğrencinin dilsel olarak güçlü ve zayıf yönlerini ortaya koymaktadır. Bu nedenle, güvenilir puanlar elde etmek için belirli sayıda maddeye ihtiyaç duyulacaktır. Uzun ve diğerleri (2018) ve de Vries (2012) ayrıca, puanlayıcı ve durum birer faktör olarak ele alındığında puanlayıcı veya durum sayısının artırılmasının puan güvenilirliğini artıracağına dikkat çekmiştir. Li, Shavelson, Yin ve Wiley (2015) madde sayısını artırmanın hata varyansını azalttığını ve hem G hem de phi katsayılarını artırdığını doğrulamıştır. Araştırma sonuçları, bu bulgularla tutarlıdır.

Erkek ve kadın adayların örneklemleri altında tahmin edilen faktörlerin benzerliklerine ilişkin Tucker uyum indeksi (0.99) sonucuna göre, erkek adayların performansının altında yatan faktör, kadın adayların performansının altında yatan faktör ile hemen hemen aynı bulunmuştur. Sonuç, bir madde için faktör yükünün .50'ye eşit veya daha yüksek olması gerektiğini bildiren Zainudin (2012) ile uyumludur. Ayrıca Lorenzo-Seva ve Ten Berge (2006), .85-.94 aralığındaki bir değerin makul bir benzerliğe karşılık geldiğini, ancak .95'ten yüksek bir değerin karşılaştırılan iki faktör veya bileşenin eşit kabul edilebileceğini ima ettiğini öne sürmüşlerdir.

Çalışma, güvenilirliğin yüksek olduğunu yansıtmakta ve bu da adaylara verilen puanların güvenilir ve genellenebilir olduğunu ortaya koymaktadır. Ayrıca, maddelerin güvenilirliğinin altını çizen temel yapıyı ölçtüğü için öğe geçerliliği yüksek hesaplanmıştır. Sonuçlar, G-Kuramı ile kestirilerek sonuçlar üzerinde yorumlar yapılmıştır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 163-181

Investigating the Effect of Rater Training on Differential Rater Function in Assessing Academic Writing Skills of Higher Education Students*

Mehmet ŞATA** İsmail KARAKAYA ***

Abstract

This study aimed to examine the effect of rater training on the differential rater function (rater error) in the process of assessing the academic writing skills of higher education students. The study was conducted with a pre-test and post-test control group quasi-experimental design. The study group of the research consisted of 45 raters, of whom 22 came from experimental, and 23 came from control groups. The raters were pre-service teachers who did not participate in any rater training before, and it was investigated that they had similar experiences in assessment. The data were collected using an analytical rubric developed by the researchers and an opinion-based writing task prepared by the International English Language Testing System (IELTS). Within the scope of the research, the compositions of 39 students that were written in a foreign language (English) were assessed. Many Facet Rasch Model was used for the analysis of the data, and this analysis was conducted under the Fully Crossed Design. The findings of the study revealed that the given rater training was effective on differential rater function, and suggestions based on these results were presented.

Key Words: Academic writing, many facet Rasch model, rater training, differential rater function.

INTRODUCTION

Academic writing is defined as a type of text in which thoughts are logically structured and justified (Bayat, 2014). According to another definition, academic writing is defined as explaining the individual's views, ideas, feelings, observations, experiments, and experiences based on his/her world of thought, congruent with the rules of the language by planning them in accordance with the individual's interest towards the chosen subject (Göçer, 2010). It can be seen from these definitions that academic writing requires many skills, and it has a complex process. Academic writing consists of multiple language skills that require the use of mental, motor, and affective skills at the same time (Çekici, 2018). Essays, theses, and research reports written by students in higher education are included in academic writing types (Gillet, Hammond & Martala, 2009). Academic writing aims to convey complex thoughts, abstract concepts, and high-level mental processes (Zwiers, 2008). In this context, when academic writing is considered as the realization of higher-level mental skills, it is important to assess academic writing validly and reliably (Carter, Bishop & Kravits, 2002).

The tools that are used to assess students' academic writing skills must be authentic, which makes it difficult to choose writing tasks. Selected writing tasks need to have a place in students' lives, and if this situation is neglected, there is a risk of under-representation or a bad definition of the structure in the assessment of academic writing skills (Cumming, 2013, 2014). One of the research areas that are frequently studied in the assessment of academic writing skills is the development and assessment of students' academic writing skills in English as a second language (Aryadoust, 2016; Bitchener, Young,

[*] This article is derived from part of his doctoral dissertation titled "The investigation of the effect of rater training on the rater behaviors in the performance assessment process"

**Dr., Faculty of Education, Agri Ibrahim Cecen University, Turkey, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997
***Prof. Dr., Faculty of Gazi Education, Gazi University, Turkey, ikarakaya2002@gmail.com, ORCID ID: 0000-0003-4308-6919

To cite this article:
Şata, M., & Karakaya, İ. (2021). *Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. doi: 10.21031/epod.842094

Received:16.12.2020
Accepted: 30.0.2021

& Cameron, 2005; Storch & Tapper, 2009). The importance of learning a second/foreign language has been increasing every day, yet many difficulties arise in the teaching and learning process. These difficulties stem from both the complex nature of the second/foreign language learning process and the way the learning process is handled and implemented (Baştürk, 2012).

While it is important to develop students' academic writing skills, it is also important to assess these skills validly and reliably. Considering that academic writing skills are high-level mental skills, it has been stated that traditional assessment methods are not suitable; instead, performance-based assessment methods are more appropriate (Johnson, Penny & Gordon, 2008). Several features distinguish performance-based assessment from traditional assessment. While performance-based assessment has features such as being based on real-life, focusing on the process rather than the product, identifying the strong and weak skills of the individual, and prompting the individual to think more and solve problems, the traditional evaluation does not have these features (Brown & Hudson, 1998; Moore, 2009).

It can be stated that one of the important concerns about performance-based assessment is the issue of objectivity in the process of assessing individual performance and determining the situation because it is very difficult to assess objectively with performance-based assessment methods compared to traditional ones (Romagnano, 2001). Many methods have been proposed in the literature to ensure objectivity in performance-based assessment. These methods can be listed as automated scoring (Attali, Bridgeman & Trapani, 2010; Burstein et al., 1998), using more than one rater (Gronlund, 1977, p.85; Kubiszyn & Borich, 2013, p.170), using rubrics (Dunbar, Brooks & Miller, 2006; Ebel & Frisbie, 1991, p. 194; Kutlu, Doğan & Karakaya, 2014, p.51; Oosterhof, 2003, p.81), and rater training (Bernardin & Buckley, 1981; Haladyna, 1997, p.143; İlhan & Çetin, 2014; Lumley & McNamara, 1995). Each of these methods has advantages & disadvantages and strengths & weaknesses compared to each other. Haladyna (1997) emphasized that it was difficult to ensure consistency among raters, regardless of the method used. In other words, regardless of the method used, there is always the possibility that some external variables other than individual performance affect the assessments (interfere with the assessments) in performance assessment. These inconsistencies that occur in the process of assessing individual performance were defined as “rater effect/bias” (Farrokhi, Esfandiari & Vaez Dalili, 2011; Haladyna, 1997, p.139; İlhan, 2015, p.3).

In case that one or more rater errors occur during the assessment process of individual performance, the number of errors regarding the estimations of students' ability levels will be high. In other words, the estimations obtained will not be reliable. Rater errors that occur during the assessment process of individual performance also have negative effects on validity. Rater errors pose a direct validity threat since they are attributed to variance unrelated to the structure (Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi et al., 2011). Therefore, it is important to minimize or control the interference of rater errors in assessments (Kim, 2009; Linacre, 1994). Rater training, which is an effective method in reducing rater errors, was used in this study (Bernardin & Buckley, 1981; Feldman, Lazzara, Vanderbilt & DiazGranados, 2012; Haladyna, 1997; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982). Rater training is widely used to reduce rater errors involved in assessments (Brijmohan, 2016). Many methods/designs regarding rater training were suggested in the literature. In this study, rater error training (RET) and frame of reference training (FRT) were used in the training of raters by combining them.

The main purpose of rater training is to enable rater to develop a common understanding of student performance and assessment criteria (Eckes, 2008; Shale, 1996). In other words, rater training ensures a valid and reliable assessment of individual performance (Moser, Kemter, Wachsmann, Köver & Soucek, 2016). Since the scores students get from an open-ended exam consist of both the performance of the student and the rater's interpretation of the student's performance, it creates constant validity anxiety in the test results (Ellis, Johnson & Papajohn, 2002; McNamara, 1996). When decisions taken based on test results are vital, rater errors should be identified, and these behaviours should be reduced to an acceptable level (Ellis et al., 2002).

In statistically identifying rater errors involved in the measurements during the assessment of performance, generalizability theory and item response theory are often used. The development of package programs in recent years has increased the frequency of using methods based on item response theory. The Rasch model, which is one of the models of item response theory, and the Many Facet Rasch Model (MFRM), which is an extension of this model, are frequently used. The main reason why MFRM is frequently used in the performance assessment process is to consider all sources of variability that are thought to affect the test scores of individuals (Kim, Park & Kang, 2012; Linacre, 1996) and to provide statistics at both individual and group level. In addition, common interactions between variability sources can be determined based on this model (Kassim, 2007). Based on these interactions, differential item functioning (DIF), differentiating individual function (DIF), and differentiating rater function (DRF) are determined (Linacre, 2017).

Differentiating rater function is defined as the tendency of the rater to give higher or lower scores to some individuals than others, depending on various characteristics of the rater, such as gender, age, and cultural factors (Wesolowski, Wind, & Engelhard, 2015). For example, a rater can give more points to successful individuals. Because the interference of differentiating rater function in the measurements is considered a systematic error, it has a negative effect on the validity of the measurements. DRF refers to a situation in which students with the same basic ability level are not likely to receive the same level of scores by raters due to their group membership. Thus, an erroneous (bias) rater prefers or dislikes a particular group of students compared to another group, for example, when scoring students' writing skills. DRF often gets involved in measurements when group memberships are known. However, in some studies, it was stated that DRF was also involved in the measurements when group membership was not known (Jin & Wang, 2017).

When the literature was examined, it was found that raters whose assessments involved severity, leniency, or central tendency error in the process of assessing individual performance, generally exhibited DRF error as well (Johnson et al., 2008; Myford & Wolfe, 2003; Wind & Guo, 2019). It was seen that studies investigating the involvement of DRF in assessing performance are quite limited. Wolfe and McVay (2012) found that 10% of the raters displayed more than one rater error in the process of assessing the essays of 120 students by 40 raters. It was investigated that some raters displayed severity, leniency, and DRF together. The study of Engelhard and Myford (2003) revealed that DRF was involved in the measurements of raters in assessing the academic writing skills of students according to their gender, race, and the language they speak. Wesolowski, Wind, and Engelhard (2015) found that DRF was involved in the measurements of 24 expert raters in assessing the jazz band performances of students. In the study conducted by Kim et al. (2012), it was found that very severe and very lenient raters generally displayed DRF. In Liu and Xie's (2014) study, 12 different scenarios were used in the process of assessing students' second language academic writing skills, and it was determined that raters showed DRF according to the scenarios. Schaefer (2008) found that errors of severity, leniency, and DRF were all involved in the process of assessing student essays. In the process of assessing performance, it was seen that rater training was used to reduce this error because DRF was frequently involved in the measurements. Bijani's (2018) study showed that the rater training given in the process of assessing students' oral presentation skills was effective. Fahim and Bijani's (2011) study revealed that rater training given in the process of assessing students' academic writing skills in the second language decreased rater x criterion interactions. On the other hand, in the study conducted by Kondo (2010), it was found that rater training given in the process of assessing second language academic writing skills did not have a significant effect on DRF. In this context, it was noticed that different results were obtained depending on the rater training pattern used and the assessed performance.

Purpose of the Study

It was observed that DRF was frequently involved in measurements in the process of assessing performances such as academic writing skills. It is significant to determine the rater errors involved in the process of assessing academic writing skills of students, such as through student essays, especially when these assessments are used in taking critical decisions such as passing a grade or getting hired in an institution. In addition, determining rater effects such as rater severity and leniency is not sufficient by itself; it is also important to determine DRF, which is a systematic error and has a significant effect on validity. In this context, the main objective of this study is to determine the differentiating rater function and to examine the effect of rater training on DRF to provide evidence for the validity of the measurements in assessing the academic writing skills of students in higher education in second/foreign language.

METHOD

Research Design

The study was conducted with a pre-test and post-test control group quasi-experimental design (Büyüköztürk, 2011). While this pattern is an unrelated design due to the comparison of the measurements belonging to different groups, it was also defined as a relational design due to the comparison of the pre-test and post-test measurements of the same group (Howitt & Cramer, 2008).

Study Group

The research consists of a total of 45 raters, 23 from the control group and 22 from the experimental group. The raters are pre-service English teachers studying at a university's English Language Teaching Department. It was assumed that the participating pre-service teachers could assess academic writing skills since they were in the last year of their education. The average age of the raters was 21.84. A personal information form was prepared to determine whether the participants have been rater and they participated in a rater training program before, and they were asked some demographic questions. It was investigated that the participants did not participate in any rater training program before, their rating experiences were similar, and they were all inexperienced in rating. Since the efficiency of the experimental process is examined rather than the purpose of generalization to the universe in experimental studies, a universe and a sample that represents the universe have not been chosen. The scorers assessed the essays written by 39 students who were continuing their education in the first year of the same department. These students took the advanced writing and reading courses in their first year, and they were all at B1 level. The essays were collected by an academician working in the same department from the students in her course, and the students participated in the study voluntarily. While the students were writing the essays, they were informed that these essays would not be graded, and they were asked not to write their names, student numbers, or ID numbers on the papers.

Data Collection Tools

Writing task

The student essays within the scope of the research were obtained by using the opinion-based writing task published as an example by the International English Language Testing System (IELTS) (Appendix A) (IELTS, t.y.). These writing tasks are prepared in many different areas to improve students' academic writing skills in English. The main purpose here is to help students reach the level in a short time that they can write essays. These writing tasks are prepared in two different categories, academic and general, and the individual chooses one of them according to his / her area of interest. The main reason for choosing this writing task stems from the idea that it will contribute to the validity and reliability of the measurements in the process of assessing the performance of the individual since it represents real-life situations. Students were given 40 minutes for the writing task, and they were asked to write an essay consisting of at least 250 words. The essays written by the students were numbered randomly, reproduced, and distributed to the raters.Rubric (for academic writing)

In the process of assessing student essays, the analytical rubric developed by the researchers was used. A systematic process was followed in the development of the rubric, and in this way, it was aimed that it would contribute to the validity and reliability of the measurements. In this context, suggestions of

Goodrich (2000), Haladyna (1997), Kutlu et al. (2014), and Moskal (2000) were taken into consideration in the rubric development process. The literature was reviewed while determining the rubric's criteria, and sample rubrics in the studies of Weigle (2002), Hughes (2003), Brown (2004), Brown (2007), and Brookhart (2013) were comprehensively examined. After the literature review, a draft form consisting of a total of 20 sub-criteria under seven fundamental criteria was prepared, and the opinions of 11 experts in academic writing skills were consulted. The Lawshe (1975) approach was used to provide evidence for the content validity of the measurements obtained from the rubric, and the content validity rate (CVR) was calculated for each criterion. When the CVR calculated for each criterion is 0.591 and above, it was accepted that the relevant criterion has sufficient content validity (Wilson, Pan & Schumsky, 2012). In line with the opinions of the field experts, the final version of the rubric consisting of six basic criteria and 16 sub-criteria was obtained (Appendix B). Because most students did not give a title to their essays even though they were told to do it, the sub-criterion of 'Title of Essay' was not included in the many facet Rasch analysis.

After collecting the evidence for the content validity of the measures obtained from the rubric, exploratory factor analysis was performed for the construct validity. For the exploratory factor analysis, the assumptions were tested, and it was investigated that the assumptions were met (for the relevant data CVR = 0.70; χ^2 (sd) = 956.427 (105) for the Barlett sphericity test; p = 0.000). In the data set, there were no extreme values and missing data, and the relationship between the criteria was found to be linear, and except for two of them, the criteria showed a normal distribution. When the literature on how big the sample should be in the exploratory factor analysis was reviewed, it was seen that there are many different opinions. Guadagnoli and Velicer (1988) stated that all these different views were not based on a theory and that there were no experimental studies, and they emphasized that the factor loadings of the variables were important rather than the sample size in their Monte Carlo simulation study, which they conducted for the sample size required for exploratory factor analysis. Accordingly, it was stated that variables with a sample size of less than 50 people and with a factor load of 0.80 and higher, regardless of the number of variables, would produce consistent results (Guadagnoli & Velicer, 1988). Although the sample size was less than 50 participants in this study, it was found appropriate to perform an exploratory factor analysis for the data set since the factor load of all variables, except three, was greater than 0.80. Exploratory factor analysis was conducted by taking the average of the scores given by 45 raters to 39 essays. As a result of the analysis, it was found that the criteria were collected under a single factor and explained 70.05% of the variance (the factor loadings of the criteria for the relevant data set are as follows; 0.842; 0.855; 0.936; 0.968; 0.644; 0.860; 0.960; 0.987; 0.945; 0.605; 0.911; 0.891; 0.899; 0.861 and 0.622).

As a result of the exploratory factor analysis, since the factor load obtained for each criterion was different (congeneric measurements), the McDonald ω coefficient (McDonald, 1999) was used for the reliability evidence of the measurements because it gave consistent results (Osburn, 2000) as a reliability determination method. As a result of the analysis, McDonald ω coefficient was found to be 0.971 (95% Confidence Interval: 0.956-0.980). Considering the reliability and validity evidence obtained for the analytical rubric, it can be argued that the measurements obtained using this measurement tool are reliable, and the inferences made based on these measurements are valid.

Experimental Process

Before starting the experimental process, to determine the starting levels of the experimental and control groups, the students' essays were distributed to the raters and the scores they gave were taken as a pre-test, and the cases of statistical differentiation were examined with the independent samples t-test and the Many Facet Rasch Model. As a result of the analysis, it was found that both groups exhibited similar rater errors in the process of assessing student essays, and the rater errors involved in the measurements were close to each other. In addition, before starting the experimental process, the analytical rubric developed for the experimental and control groups was introduced, and how to use it in the scoring process was explained. Later, both groups were explained what academic writing skill is, what its general characteristics are, and its connection with the developed rubric. These procedures were carried out to ensure that the experimental and control groups reach a similar level at the beginning. Thus, in the process of assessing academic writing skills, the mixing of different variance sources (such as measurement tools) in the measurements was tried to be minimized. It was aimed that the raters did not know whether they were in the experimental or control group. Then, the student essays were distributed to the experimental and control groups, and they were given one week to assess the essays. One week later, student essays were collected, and they were analysed on the computer.

Rater training

To create a common understanding between raters while assessing individual performance, rater error training (RET) and frame of reference training (FRT), which are recommended in the literature, were combined. The two selected trainings were combined because of the inability of RET in defining rater behaviors and errors, but not being effective on rater accuracy, and the success of FRT on rater accuracy (Murphy & Balzer, 1989; Sulsky & Day, 1992). In other words, both rater training patterns were chosen because they were complementary to each other. The basic assumption of the RET design is that familiarity with common rater errors and encouraging raters to avoid these errors will result in a direct reduction of rater errors and, therefore, more effective performance assessment. (Woehr & Huffuct, 1994). Although rater errors such as rater severity and leniency decreased in the RET pattern, findings indicate that rating accuracy also decreases (Bernardin & Pence, 1980). In the FRT pattern, it is taken as a basis that the performance assessed is multidimensional (Selden, Sherrier & Wooters, 2012). Therefore, all sub-dimensions of performance should be defined, and behavioural examples representing these dimensions should be given to the raters. The basic principle in the FRT pattern is to train the raters to ensure that the performance dimensions assessed have certain standards. Thus, a match can be made between the scores given by the rater and the actual scores of the student (Woehr & Huffuct, 1994). The rater training was completed in four weeks in total, giving one hour each week in the measurement and evaluation course.

In the first week, the purpose, scope, and importance of rater training were introduced within the framework of RET. Then, the target audiences and the methods used were introduced in the rater training, and the first stage was completed. The second stage included information about the most common rater errors of the performance assessment process and the effects of these errors on validity and reliability. Finally, for rater training, in-group discussions were made based on a few examples. Thus, the first week of rater training was completed.

In the second week, the possible sources of rater errors involved in the measurements in the performance assessment process were explained, and the actions to be taken to reduce these errors were specified. These suggestions were determined by reviewing the literature, and the sample applications were shared with the experimental group. With this process, the RET part of the rater training was completed, and the FRT part was started. First, the academic writing skill, which was assessed by the raters, was defined. The sub-dimensions of this skill and which criteria correspond to the sub-dimensions in the rubric were explained. Then, the raters in the experimental group were asked to give representative behaviours regarding the dimensions of academic writing skill. They were then asked to discuss these representative behaviours in the group.

In the third week, as a continuation of the second week, examples regarding the dimensions of academic writing skill were given, and in-group discussions continued. After completing this stage, based on the pre-test results of the raters, the best, middle, and low-level student compositions were determined. These compositions were multiplied and distributed to the raters in the experimental group, and they were asked to be re-assessed. The raters were not informed about whether the essays were good or bad. After the assessment process, raters were randomly selected and asked about the scores they gave and the reasons for giving these scores. Later, the same question was asked to other raters in the experimental group. This process was carried out considering the criteria with the highest standard error according to the pre-test measurements. The main goal is to create a common understanding among raters. Also, based on the pre-test measurements, written feedback was given to each rater regarding his / her ratings.

In the last week, the activities of the third week were continued with different raters. The compositions of three students, which were determined beforehand according to the pre-test results, were assessed by an academician. Raters were asked to explain how many points the field expert (academician) gave according to the determined criteria; thus, in-group discussions were made conducted. After all stages, rater training was completed, and students' compositions (39) were given to the experimental and control groups again for the post-test measurements (the duration for assessment was one week). Participation in all stages of the experimental process and scoring was voluntary. Also, additional points were added to the final grades to encourage these students.

Data Analysis

During the data analysis process, EFA and Lawshe techniques were applied in order to provide evidence for the validity of the measurements obtained from the first developed measurement tool. Then, many facet Rasch analyses were performed, and Mann Whitney U test was run based on the logit values obtained as a result of this analysis. At first, EFA was performed because the scoring of the raters showed a normal distribution. Then, since the logit values obtained by MFRM were not normally distributed, the Mann Whitney U test was used. The analysis of MFRM was preferred because it gives the common interaction between facets at the individual level. Since all raters assessed the compositions of students over all criteria, MFRM was conducted under a completely crossed-out pattern. Detailed information about MFRM was presented below.

Many facet Rasch model

MFRM has emerged as an extension of the basic Rasch model. Unlike the basic Rasch model, many variability sources (facets) such as rater, item, task, individual, time are placed on a single scale (Kim et al., 2012; Linacre, 1993; Linacre, 1996). Also, interactions between MFRM and sources of variability can be examined (Kassim, 2007). MFRM is a linear model that calibrates all parameters and converts the observations in the ranking scale to an equidistant logit scale (Bond & Fox, 2015). The logistic transformation of the log odds ratios allows independent variables such as peer assessment, status determination criteria, and open-ended items to be seen as dependent variables (Esfandiari, 2015).

Another advantage of MFRM is that it offers information that classical test theory and generalizability theory cannot provide (Lunz, Wright & Linacre, 1990). MFRM can provide the researcher with detailed information about each facet. For example, a lot of information can be obtained such as which of a group of raters assessing the performance of individuals, what the scoring is (observed value), and what the scoring should be (expected value). As MFRM provides detailed feedback, it is possible to determine which rater is good or bad and what kind of intervention is required. Based on these advantages of MFRM, the rater errors can be determined before the rater training; therefore, training can be arranged for these errors. Thus, the validity and reliability of the measurements can be increased.

Considering *rater x student composition (pxb)* interactions, the measurement model is defined as follows;

$$\ln\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x - I_{pb} \qquad (1)$$

where

$\ln (P_{bkpx} / P_{bkpx-1})$ = the probability that Performance b rated by Rater p on Item k in receives a rating in category x rather than category x-1,

θ_b = the logit-scale location (e.g., achievement) of Performance b,

β_k = the logit-scale location (e.g., difficulty) of Item k,

α_p = the logit-scale location (e.g., severity) of Rater p,

τ_x = the point of equal probability on the latent variable between categories

x-1 and x and

I_{pb} = Interaction term between rater facet and student composition facet.

The interaction (bias) index has an important place in determining rater errors in MFRM (Engelhard, 2002; Linacre, 2017).

Since MFRM belongs to the Rasch model family, it must meet the assumptions in the Rasch models (Eckes, 2015; Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi et al., 2011). The assumptions to be met for MFRM are unidimensionality, local independence, and model data fit. As stated in the data collection tools, the rubric had a single factor structure. For the local independence assumption, the G^2 statistics proposed by Chen and Thissen (1997) were used. The standardized LD χ2 values were found to range from -0.4 to 4.5. The marginal fit χ^2 values were close to zero, and local independence was found. Standardized residual values were examined for model-data fit. The total number of observations for the pre-test application was 39x45x15 (composition x rater x criterion) = 26.325. it was observed that model-data fit was achieved for the pre-test application since the number of standardized residual values outside the ± 2 range was 1.067 (4.05%) and the number of standardized residual values outside the ± 3 range was 164 (0.62%). While the total number of observations for the post-test application was 26.322 (3 missing data), the number of standardized residual values outside the ± 2 range was 995 (3.78%), and the number of standardized residual values outside the ± 3 range was 186 (0.71%).

RESULTS

Findings were presented under two headings as before (pre-test) and after (post-test) rater training. MFRM analysis was given by presenting group statistics firstly, then individual statistics.

Investigating DRF Status of Raters in Experimental and Control Groups Before Rater Training

The estimated chi-square value for the statistical indicator of *rater x student compositions (pxb)* interactions at the group level was found to be significant (χ2(sd) = 5 298.40 (1755), $p < 0.05$). According to the significance of the chi-square value, the rater function that differed at the group level was mixed up in the measurements during the assessment of student compositions. After determining that DRF was involved in the measurements at the group level in *pxb* interaction, the statistics at the individual level were examined. T statistics are used for interactions that are significant in interaction between sources of variability in MFRM. Statistical significance is tested by comparing the t-value obtained as a result of MFRM interaction analysis with the critical t-value. Interactions with a t-value outside the ± 2 range indicate differential rater function (Linacre, 2017). The number of possible interactions in the control group was 897 (23x39), and the number of significant interactions was 203 (22.63%). The number of possible interactions in the experimental group was 858 (22x39), and the number of significant interactions was 160 (18.65%). When the t statistic takes a negative value, it is defined as differential rater severity; when it takes a positive value, it refers to differential rater leniency. Table 1 presented the frequency and percentages of the raters in the experimental and control groups regarding the type of significant interactions.

Table 1. Frequencies and Percentages of Significant Interactions Regarding Pre-test Measurements in pxb Interaction

Group	Differential Rater Severity		Differential Rater Leniency		Total	
	f	%	f	%	f	%
Experimental	83	9.67	77	8.98	160	18.65
Control	111	12.37	92	10.26	203	22.63

Table 1 showed that the interference levels of the DFR of the experimental and control groups in the measurements were close to each other. The statistical significance of the differential rater severity and leniency of the raters in the control and experimental groups was tested using the bias size values obtained in MFRM interaction analysis, and analysis results were given in Table 2.

Table 2. The Results of the Mann Whitney U Test Regarding the Differentiation of Significant Interactions Regarding the Pre-test Measurements in the Experimental and Control Groups

Type of DRF	Group	N	Average rank	Z	U
DRS	Control	111	90.88	-1.90	3872.00
	Experimental	83	106.35		
DRL	Control	92	87.55	-0.74	3307.00
	Experimental	77	81.95		

* *p*<0,05; DRS = Differential Rater Severity, DRL = Differential Rater Leniency

As is seen Table 2, the interference levels of the DRF of the raters in the experimental and control groups before the rater training were statistically similar (for DRS, U = 3872.00; Z = -1.90 $p > 0.05$; for DRL, U = 3307.00; Z = -0.74; $p > 0.05$).

Investigating DRF Status of Raters in Experimental and Control Groups after Rater Training

After the experimental procedure, the estimated chi-square values for the statistical indicator of *rater x student compositions* (pxb) interactions at the group level were found to be significant ($\chi 2$(sd) = 4 084.90 (1755), $p < 0.05$). This finding shows that, despite rater training, the differential rater function in the performance assessment process of the raters interfered with the measurements.

Statistics at the individual level were examined since DRF was involved in group-level measurements. Therefore, t statistics regarding pxb interactions were examined. While 163 of 897 possible interactions (18.17%) of the control group were significant, 110 (12.82%) of 858 possible interactions of the experimental group were found to be significant. Table 3 presented the frequency and percentage values of the raters in the experimental and control groups related to the differential rater function involved in the measurements during the performance assessment process after the rater training.

Table 3. Frequency and Percentages of Significant Interactions Regarding Post-test Measurements in pxb Interaction

Group	Differential Rater Severity		Differential Rater Leniency		Toplam	
	f	%	f	%	f	%
Experimental	59	6.88	51	5.94	110	12.82
Control	95	10.59	68	7.58	163	18.17

The interference levels of the DRF of the raters in the experimental and control groups differed after the rater training while assessing student compositions. The statistical significance of the differential rater severity and leniency of the raters in the control and experimental groups was tested using the bias size values obtained in MFRM interaction analysis, and analysis results were given in Table 4.

Table 4. The Results of the Mann Whitney U Test Regarding the Differentiation of Significant Interactions Regarding the Post-test Measurements in the Experimental and Control Groups

Type of DRF	**Group**	**N**	**Average rank**	**Z**	**U**	***p***	**d**
DRS	Control	95	69.82	-2.72	2072.50*	0,007*	0.22
	Experimental	59	89.87				
DRL	Control	68	56.21	-1.38	1476.50	0,167	--
	Experimental	51	65.05				

* *p*<0,05; DRS = Differential Rater Severity, DRL = Differential Rater Leniency

After rater training, the interference level of the differential rater severity in the measurements in the performance assessment process was found to be statistically significant, while the interference level of the differential rater leniency was insignificant (for DRS, U = 2072.50; Z = -2.72 p < 0.05; for DRL, U = 1476.50; Z = -1,38; p > 0.05). According to this result, rater training had a small effect (r = 0.22) on differential rater severity, but no effect on differential rater leniency.

To observe the effect of rater training on pxb interactions, significant interaction numbers of the raters in the experimental group according to the pre and post-tests were given in Table 5.

Table 5**.** Significant pxb Interactions Regarding Raters in the Experimental Group

Test		P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11
Pre-test	f	9	7	7	13	7	11	7	5	13	9	2
	%	23.1	18.0	18.0	33.3	18.0	28.2	18.0	12.8	33.3	23.1	5.1
Post-test	f	2	9	10	5	4	2	2	6	6	1	8
	%	5.1	23.1	25.6	12.8	10.3	5.1	5.1	15.4	15.4	2.6	20.5
		P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
Pre-test	f	3	9	7	13	3	5	7	5	4	6	9
	%	7.7	23.1	18.0	33.3	7.7	12.8	18.0	12.8	10.3	15.4	23.1
Post-test	f	4	5	8	8	2	2	6	8	3	3	9
	%	10.3	12.8	20.5	20.5	5.1	5.1	15.4	20.5	7.7	7.7	23.1

As is seen in Table 5, while assessing student compositions after rater training, the significant interactions of 14 raters (1, 4, 5, 6, 7, 9, 10, 13, 15, 16, 17, 18, 20, and 21) decreased (positively affected by the training); the significant interactions of 7 raters (2, 3, 8, 11, 12, 14, and 19) increased (negatively affected by the training), and the significant interactions of 1 rater (22) remained constant. To make Table 5 more understandable, the graphical representation of pxb interactions was given in Figure 1.

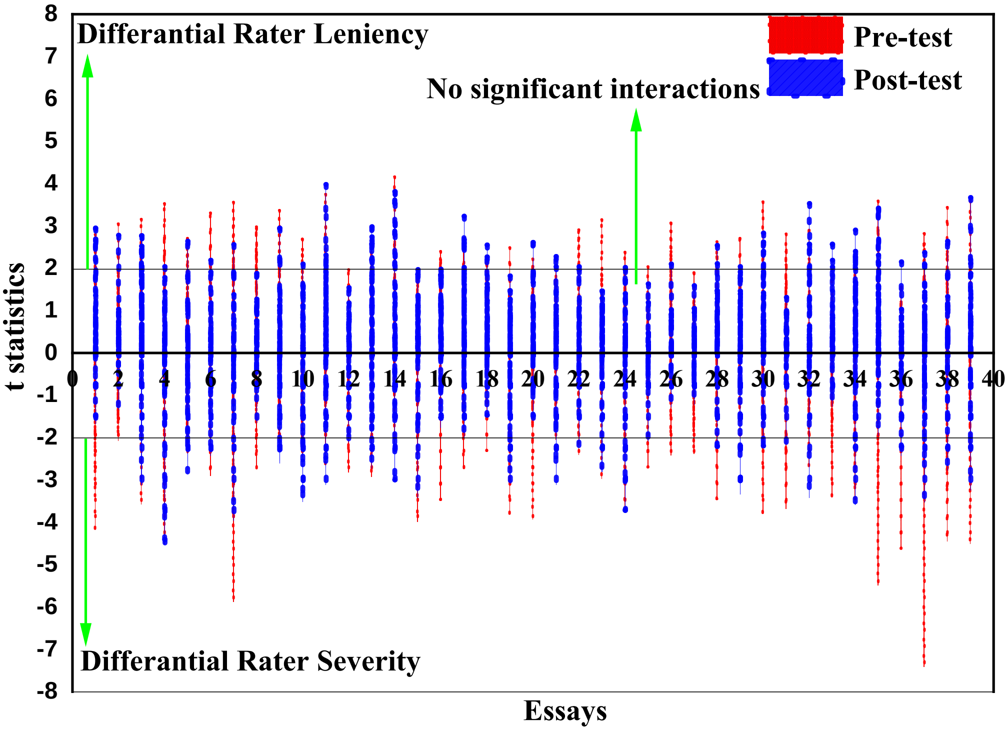


Figure 1. pxb Interactions for All Raters in the Experimental Group

As seen in Figure 1, the red lines representing the raters' pre-test were mostly outside the ± 2 range. After rater training, blue lines representing raters' ratings were observed less outside the ± 2 range. According to Figure 1, some compositions were subject to more rater bias than other compositions. For example, the raters were more severe in assessing composition numbered 37 than the other compositions. Besides, it can be said that the given rater training had a positive effect on rater errors in general, and as a result, contributed to the validity of the measurements.

DISCUSSION and CONCLUSION

This study aimed to investigate the effect of rater training on DRF, which is involved in measurements while assessing second language academic writing skills. In this context, the findings obtained before and after rater training were examined. Before rater training, DRF effect involved in the measurements was similar in both the experimental and control groups while assessing the compositions of students. Similar DRF effects were found in both group level and individual statistics. Approximately one-fifth of pxb interactions in the experimental and control groups were observed to be DRF. Research supports this finding, indicating that DRF is frequently involved in measurements in the performance assessment process (Liu & Xie, 2014; Schaefer, 2008; Wesolowski et al., 2015; Wolfe & McVay, 2012). While assessing the compositions of students, DRF involved in the measurements appeared in two ways: differential rater severity and differential rater leniency. This study found that raters mostly showed differential rater severity. The literature advocates that DRF involved in the measurements during the performance assessment process is a combination of both severity and leniency behavior, and DRF generally occurs due to too severe or too lenient raters (Kim et al., 2012). Considering that there are more severe raters in the current study, the abundance of differential rater severity confirms the literature.

During the process of assessing student compositions after the rater training, the involvement level of DRF in the measurements was examined. While the amount of change in the control group was minimal, a significant change was found in the experimental group. Although the level of interference of the two types of DRF in the experimental and control groups in the measurements was statistically similar before the rater training, it differed statistically after the rater training. It was found that the differential rater leniency was not affected by the experimental process, but the differential rater severity was affected. In other words, rater training was effective on the differential rater severity of DRF. Considering the studies conducted by Bijani (2018), Fahim and Bijani (2011), and May (2008) and Yan (2014), rater training was effective on DRF. Van Dyke (2008) found that the differential rater leniency in the performance assessment process interfered with the measures, but the differential rater severity did not interfere. There are two main reasons for the difference between the current study and the one conducted by Van Dyke (2008): The first reason may be that the raters consisted of different groups, and the second one is that the performance assessed was different.

The results of this study can be summarized as follows;

- During the process of assessing compositions. DRF was involved in the measurements and accounted for approximately one-fifth of pxb interactions.
- Raters in the experimental and control groups exhibited similar DRF before rater training.
- Rater training had an impact on the different types of rater severity of DRF, and rater training had a small effect size on DRF.

Based on these results, some suggestions were made for future studies and researchers;

- In the present study, two different rater training patterns were combined. Considering that there are many different rater training patterns in the literature, different combinations can be made to examine the effects of rater training on DRF.
- A large experimental group was used in this study. The literature emphasizes that the training of smaller (n = 5-6) groups is more effective. Thus, it may be useful to use small groups in future studies.
- The effect of rater training on DRF can be used to train raters and contribute to the validity and reliability of the measurements during the performance assessment process utilized in placement and selection exams.

REFERENCES

Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, *36*(10), 1742-1770. https://doi.org/10.1080/01443410.2014.950946

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment, 10*(3), 1-16. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603

Baştürk, M. (2012). İkinci dil öğrenme algılarının belirlenmesi: Balıkesir örneği. *Balikesir University Journal of Social Sciences Institute*, *15*(28-1), 251-270. Retrieved from http://dspace.balikesir.edu.tr/xmlui/handle/20.500.12462/4594

Bayat, N. (2014). Öğretmen adaylarının eleştirel düşünme düzeyleri ile akademik yazma başarıları arasındaki ilişki. *Eğitim ve Bilim*, *39*(173), 155-168. Retrieved from http://eb.ted.org.tr/index.php/EB/article/view/2333

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: New response sets and decreasing accuracy. *Journal ofApplied Psychology*, *65*, 60-66. https://doi.org/10.1037/0021-9010.65.1.60

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*(2), 205-212. Retrieved from https://journals.aom.org/doi/abs/10.5465/amr.1981.4287782

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, *5(1)*, 1-20. https://doi.org/10.1080/2331186X.2018.1460901

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL students. *Journal of Second Language Writing*, *14*, 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York and London: Routledge. https://doi.org/10.4324/9781315814698

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176. https://doi.org/10.1177/0013164495055002001

Brijmohan, A. (2016). *A many-facet Rasch measurement analysis to explore rater effects and rater training in medical school admissions.* (Doktora Tezi). Retrieved from http://www.proquest.com/

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* Alexandria, Virginia: ASCD.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. New York: Pearson Education.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, *32*(4), 653-675. https://doi.org/10.2307/3587999

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). *Automated scoring using a hybrid feature identification technique.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada. https://doi.org/10.3115/980845.980879

Büyüköztürk, Ş. (2011). *Deneysel desenler- öntest-sontest kontrol grubu desen ve veri analizi*. Ankara: Pegem Akademi Yayıncılık.

Carter, C., Bishop, J. L., & Kravits, S. L. (2002). *Keys to college studying: becoming a lifelong learner*. New Jersey: Printice Hall.

Çekici, Y. E. (2018). Türkçe'nin yabancı dil olarak öğretiminde kullanılan ders kitaplarında yazma görevleri: Yedi iklim ve İstanbul üzerine karşılaştırmalı bir inceleme. *Gaziantep Üniversitesi Eğitim Bilimleri Dergisi*, *2*(1), 1-10. Retrieved from http://dergipark.gov.tr/http-dergipark-gov-tr-journal-1517-dashboard/issue/36422/367409

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. https://doi.org/10.3102/10769986022003265

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163-178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. https://doi.org/10.1080/15434303.2011.622016

Cumming, A. (2014). Assessing integrated skills. In A. Kunnan (Vol. Ed.), *The companion to language assessment: Vol. 1,* (pp. 216–229). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla131

Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 115-128. https://doi.org/10.1007/s10755-006-9012-x

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Jersey: Prentice Hall Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Ellis, R. O. D., Johnson, K. E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233. https://doi.org/10.2307/3588333

Engelhard Jr, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series*, i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates

Esfandiari, R. (2015). Rater errors among peer-assessors: applying the many-facet Rasch measurement model. *Iranian Journal of Applied Linguistics*, 18(2), 77-107. https://doi.org/10.18869/acadpub.ijal.18.2.77

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16. Retrieved from http://www.ijlt.ir/portal/files/401-2011-01-01.pdf

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79-101. Retrieved from https://jalt-publications.org/files/pdf-article/jj2012a-art4.pdf

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, 15(11), 76-83. Retrieved from https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf

Feldman, M., Lazzara, E. H., Vanderbilt, A.A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286. https://doi.org/10.1002/chp.21156

Gillet, A., Hammond, A. & Martala, M. (2009). *Successful academic writing*. New York: Pearson Longman.

Göçer, A. (2010). Türkçe öğretiminde yazma eğitimi. *Uluslararası Sosyal Araştırmalar Dergisi*, *12*(3), 178-195. Retrieved from http://www.sosyalarastirmalar.com/cilt3/sayi12pdf/gocer_ali.pdf

Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational Leadership*, *54*(4), 14-17.

Gronlund, N. E. (1977). *Constructing achievement test.* New Jersey: Prentice-Hall Press

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin*, *103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

Haladyna, T. M. (1997). *Writing test items in order to evaluate higher order thinking*. USA: Allyn & Bacon.

Hauenstein, N. M., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, *25*(3), 253-266. https://doi.org/10.1111/ijsa.12177

Howitt, D., & Cramer, D. (2008). *Introduction to statistics in psychology*. Harlow: Pearson Education.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

IELTS (t.y). *Prepare for IELTS*. Retrieved from https://takeielts.britishcouncil.org/prepare-test/free-sample-tests/writing-sample-test-1-academic/writing-task-2

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi*. (Doktora Tezi). Retrieved from https://tez.yok.gov.tr

İlhan, M., & Çetin, B. (2014). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz. *Journal of European Education*, *4*(2), 29-38. https://doi.org/10.18656/jee.77087

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate behavioral research*, *52*(3), 391-402. https://doi.org/10.1080/00273171.2017.1299615

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

Kassim, N. L. A (2007). *Exploring rater judging behaviour using the many-facet Rasch model.* Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Universiti Utara, Malaysia. Retrieved from http://repo.uum.edu.my/3212/

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, 11(3), 179-197. Retrieved from http://ejournals.ukm.my/gema/article/view/49

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly*, *29*(4), 346-365. https://doi.org/10.1123/apaq.29.4.346

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doktora Tezi). Retrieved from http://www.proquest.com/

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1-23. Retrieved from https://eric.ed.gov/?id=EJ920513

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement.* New Jersey: John Wiley & Sons Incorporated.

Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayıncılık.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, *7*(1), 283-284. Retrieved from https://www.rasch.org/rmt/rmt71h.htm

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: Mesa Press.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, *3*, 85-98. Retrieved from https://files.eric.ed.gov/fulltext/ED364573.pdf

Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: MESA

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing*, *4*(1), 50-65. Retrieved from https://cdn.ov2.com/content/ijlte_1_ov2_com/wp-content_138/uploads/2019/07/422-2014-4-1.pdf

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, *71*(3), 297-313. https://doi.org/10.1177/1080569908321431

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. (Doktora Tezi). Retrieved from http://www.proquest.com/

Moser, K., Kemter, V., Wachsmann, K., Köver, N. Z., & Soucek, R. (2016). Evaluating rater training with double-pretest one-posttest designs: an analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 1-23. https://doi.org/10.1080/09585192.2016.1254102

Moskal, B.M. (2000). *Scoring rubrics: What, when and how?.* Retrieved from http://pareonline.net/htm/v7n3.htm

Murphy, K.R. & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624. https://doi.org/10.1037/0021-9010.74.4.619

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422. Retrieved from http://psycnet.apa.org/record/2003-09517-007

Oosterhof, A. (2003). *Developing and using classroom assessments*. New Jersey: Merrill-Prentice Hall Press.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods, 5*(3), 343. http://dx.doi.org/10.1037/1082-989X.5.3.343

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher, 94*(1), 31-37. Retrieved from http://peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity2.pdf

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Selden, S., Sherrier, T., & Wooters, R. (2012). Experimental study comparing a traditional approach to performance appraisal training to a whole-brain training method at CB Fleet Laboratories. *Human Resource Development Quarterly*, *23*(1), 9-34. https://doi.org/10.1002/hrdq.21123

Shale, D. (1996). Essay reliability: Form and meaning. In: White, E. Lutz, W. & Kamusikiri S. (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76–96). New York: MLAA.

Stamoulis, D.T. & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*(6), 994-1003. https://doi.org/10.1037/0021-9010.78.6.994

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207-223. https://doi.org/10.1016/j.jeap.2009.03.001

Sulsky, L.M., & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*(4), 501-510. https://doi.org/10.1037/0021-9010.77.4.501

Van Dyke, N. (2008). Self-and peer-assessment disparities in university ranking schemes. *Higher Education in Europe*, *33*(2/3), 285-293. https://doi.org/10.1080/03797720802254114

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170. https://doi.org/10.1177/1029864915589014

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and psychological measurement*, *79*(5), 962-987. https://doi.org/10.1177/0013164419834613

Woehr, D.J., & Huffuct, A.I. (1994). Rater training for performance appraisal. A qantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189-205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

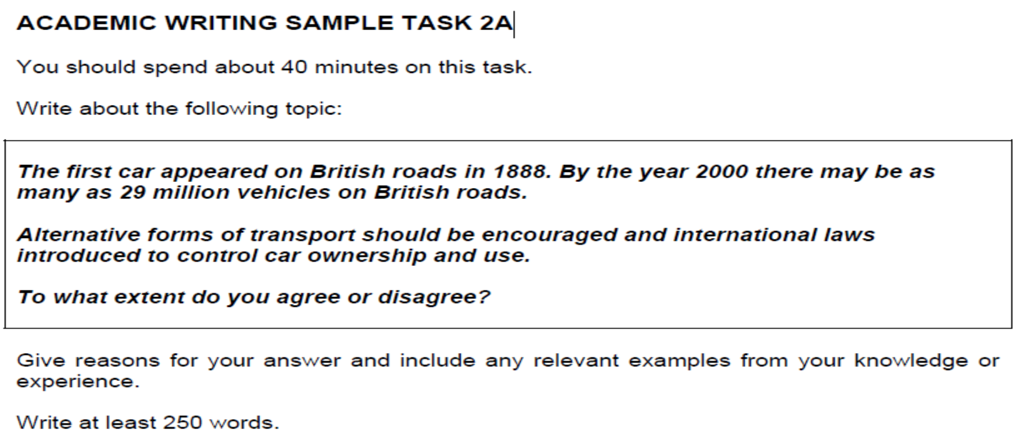
Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501-527. https://doi.org/10.1177/0265532214536171

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, *67*(6), 752-758. https://doi.org/10.1037/0021-9010.67.6.752

Zwiers, J. (2008). *Building academic language: Essential practices for content classrooms*. San Francisco: Jossey-Bass.

Appendix A. Academic Writing Sample



ACADEMIC WRITING SAMPLE TASK 2A

You should spend about 40 minutes on this task.

Write about the following topic:

The first car appeared on British roads in 1888. By the year 2000 there may be as many as 29 million vehicles on British roads.

Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

Appendix B. Rubric (For Academic Writing)

	ORGANIZATION					CONTENT	
Point	Introduction-Body-Conclusion	Thesis Statement	Topic Sentence	Supporting Sentences	Appropriate Length	Topic Relevance	Idea Development
4	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length*.	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.
2	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words.* It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	COHESION	GRAMMAR		VOCABULARY		MECHANICS	
Point	Coherence	Linking	Accuracy of Grammatical Forms	Syntactic Complexity	Word Choice	Lexical Range	Spelling	Punctuation
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.
1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequently* misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 182-191

An Alternative to Likert Scale: Emoji

Abdullah Faruk KILIÇ * İbrahim UYSAL ** Bilal KALKAN ***

Abstract

In the twenty-first century, the wide use of emojis in communication platforms has emerged. As a result, emojis have started to be used in scales. However, there are a limited number of studies in the literature that focuses on the effect of using emojis instead of Likert-type response categories in scales. Therefore, the focus of this study is to examine the differences that may arise from using emoji and Likert-type response categories in scales. For this purpose, the 3, 5, and 7-point Likert-type and 3, 5, and 7 emoji response categories Psychological Well-Being Scale was applied to 341 students studying at two state universities located in different regions of Turkey. Exploratory and confirmatory factor analyses and reliability analyses were carried out on the data of the participants who answered the six forms with different response categories. As a result, it was determined that there were no significant differences in exploratory and confirmatory factor analyses and reliability analyses. However, when correlational analyses were examined, it was observed that as the number of reaction categories increased, the correlation scores of emoji and Likert-type scales decreased.

Key Words: Emoji, likert scale, scale development, response category, validity and reliability.

INTRODUCTION

Researchers frequently adopt scaling techniques such as Thurstone (1927), Guttman (1941), and Likert (1932) when developing self-report scales (Dwyer, 1993). The Thurstone scale has a structure that consists of many items, and the items are rated by experts. In this scale, participants indicate whether they agree or disagree with each item (Payne & Payne, 2004). On the other hand, Guttman scaling technique is a response-based technique, and people can respond to a large number of items. However, they are evaluated according to the answer they give to the strongest item in terms of the feature examined. Items are scaled according to the amount or importance of the feature being measured (Price, 2017). Guttman scales differ from Thurstone scales in their cumulative aspect. In Guttman scales, a positive response to one level of the scale demonstrates a positive response to all items below that level, and with this aspect, it differs from Thurstone scales. Thurstone and Guttman scales are prepared to represent all levels of the feature, but in Likert-type scales, the items are close to the endpoints of the measured feature (Anderson, 1988/1991). In a Likert-type scale, which is a person-oriented method, participants indicate their degree of agreement on many items. The rating can be made as strongly disagree, disagree, neutral, agree, and strongly agree (Price, 2017), and they can be formed as three, four, five, and seven categories. In the scale, there may be an indecision option to choose when there is no positive or negative emotion regarding the item. Likert-type scales do not need an expert view in the scoring process contrary to the Thurstone scale. This situation allows for eliminating errors caused by experts (Bayat, 2014). Likert-type scales are considered to be practical and reliable. However, in recent years, as a reflection of digitalization, it has been observed that emojis are used as reaction categories to the items in the scales. In emoji, *e* represents pictures, and *moji* represents characters. When we look at the history of emojis, we see that they were created in 1998 by a Japanese communicator, and the widespread use of them has been around since 2010. In 2015, an emoji (face with tears of joy [😂]) was chosen as the word of the year by the Oxford Dictionary, which

* Res. Assist. Dr., Adıyaman University, Faculty of Education, Adıyaman-Turkey, abdullahfarukkilic@gmail.com, ORCID ID: 0000-0003-3129-1763
** Res. Assist. Dr., Bolu Abant İzzet Baysal University, Faculty of Education, Bolu-Turkey, ibrahimuysal06@gmail.com, ORCID ID: 0000-0002-6767-0362
*** Assist. Prof. Dr., Adıyaman University, Faculty of Education, Adıyaman-Turkey, kalkanbilal@gmail.com, ORCID ID: 0000-0002-5010-4639

To cite this article:
Kılıç, A. F., Uysal, İ., & Kalkan, B. (2021). An alternative to Likert scale: Emoji. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 182-191. doi: 10.21031/epod.864336

Received: 19.01.2021
Accepted: 11.05.2021

demonstrates that emojis have gained an important place in communication and personal expression. Hence, the increasing importance of emojis in social areas and communication has been acknowledged, and emojis have become a new spelling code (Danesi, 2017). The reflection of this trend in the digital world on scientific researches has been inevitable.

When the literature is examined, a limited number of studies were found on the use of emojis in scales. Alismail and Zhang (2018) examined the use of emoji in electronic user experience in their research. Deubler, Swaney-Stueve, Jepsen, and Su-Fern (2020), in consumers' emotional response to products, and Marengo, Giannotta, and Settanni (2017), on personality assessment, examined the effect of using emojis instead of verbal response categories. Alismail and Zhang (2018) made inferences on the advantages and difficulties of using emojis through semi-structured interviews. Marengo et al. (2017) obtained concurrent validity between emojis and a personality test consisting of verbal response categories. Deubler et al. (2020) made inferences about the validity of the scale data in which emojis are used as response categories. When the studies of Marengo et al. (2017) and Deubler et al. (2020) are considered, it can be understood that emojis can be used instead of verbal response categories. Even though there is evidence relating to the validity of the data obtained with the use of emojis in questionnaires, the studies are not sufficient. Besides, there is no study that compares verbal response categories with emojis. Considering that the use of emojis provides important results about the psychological states of individuals, it seems that more research is needed on the subject. For this reason, this study focuses on the validity and reliability of data obtained with emoji and verbal response categories. In this respect, it will provide inferences about the effects of using emojis. Also, using instruments with 3, 5, and 7 Likert type verbal categories, there is a tendency to choose the highest or lowest category, avoid choosing extreme categories, and respond similarly to items that have close meaning (Albaum, 1997). It is important to determine the occurrence of the same situation when using emojis. Seeing that there is no detailed research in the literature on this subject, this study aims to examine whether the data obtained from scales with emoji and Likert-type response categories differ from each other. Studies also stated that there was a difference between men's and women's emoji use (see Chen et al., 2017; Prada et al., 2018). Therefore, it is important to examine whether the use of emojis as a response category in the instruments makes a difference between men and women in terms of the structure of the scale. Hence, this paper examines the following research questions:

1. Do the factor loadings and proportions of explained variance in the result of the exploratory factor analysis (EFA) of the data obtained with 3-point, 5-point, and 7-point Likert-type verbal response categories and emojis differ?
2. Do the factor loadings and model-data fit indexes in the result of confirmatory factor analysis (CFA) of the data obtained with the 3-point, 5-point, and 7-point Likert-type verbal response categories and emojis differ?
3. How do the relationships between 3-point, 5-point, and 7-point Likert-type verbal response categories and, respectively, 3-point, 5-point, and 7-point emojis differ according to gender?
4. What are the reliabilities of the data sets obtained with 3-point, 5-point, and 7-point Likert-type verbal response categories and emojis?

METHOD

This study utilized a cross-sectional and non-experimental survey research design. In survey research, data are collected from the sample in a single session. The main way to collect data is to ask questions, and it is a method used to examine certain characteristics (belief, attitude, ability, etc.) (Fraenkel, Wallen, & Hyun, 2012). In this study, different response categories of the questions asked students about their psychological well-being were compared. Hence, the survey research method was adopted.

Population and Sample

The accessible population of the research consisted of undergraduate students studying at two state universities, one in the Southeastern Anatolia and the other in the Black Sea region. In the study, no inference was made about the feature examined; only the use of emoji and verbal expressions as a response category were compared. For this reason, the convenience sampling method was adopted. In convenient sampling, a non-random sampling method, researchers reach out to the most accessible participants in order to prevent excessive time and energy loss and to reduce study costs (Fraenkel et al., 2012). The sample group consisted of 341 students, and the demographic characteristics of the students were shown in Table 1.

Table 1. Demographic Characteristics of the Students in The Sample

Variable	f	%	Variable	f	%
Woman	252	73.9	Adıyaman University	165	48.4
Man	89	26.1	Bolu Abant İzzet Baysal University	176	51.6
Faculty of Education	283	83.0	First Grade	66	19.4
Faculty of Science and Literature	15	4.4	Second Grade	148	43.4
Faculty of fine arts	11	3.2	Third Grade	63	18.5
Vocational School of Social Sciences	19	5.6	Fourth Grade	57	16.7
Other	13	3.8	Other	7	2.0
Sum				341	100

When Table 1 is examined, it is seen that 79.9% (*n* = 252) of the university students in the sample were female and 26.1% (*n* = 89) were male. The ages of the participants range between 18 and 41, with an average of 21.6 and a median of 21. Of all the participants, 83% (*n* = 283) studied at the faculty of education, 5.6% (*n* = 19) at social sciences vocational school, 4.4% (*n* = 15) at the faculty of science and literature, 3.2% (*n* = 11) at the faculty of fine arts, and 3.9% (*n* = 13) at other faculties (dentistry, pharmacy, economics and administrative sciences, health sciences, tourism) and institutes (natural sciences). The sample consisted of 19.4% (*n* = 66) first year, 43.4% (*n* = 148) second year, 18.5% (*n* = 63) third year, 16.7% (*n* = 57) fourth year, and 2% (*n* = 6) other year (preparatory year and fifth year) students.

Data Collection Tools

The data collection tools consisted of a questionnaire inquiring the participants about their genders, universities, faculties, and years, as well as the Psychological Well-being Scale. The scale was developed by Diener et al. (2010) and adapted to Turkish culture by Telef (2013). When the psychometric properties of the Turkish form of the Psychological Well-Being Scale were examined, it was seen that the scale was unidimensional, and the explained variance was 42%. The factor loadings of the items varied between .54 and .76. The Cronbach Alpha reliability coefficient of the scale scores was .80, and the test-retest reliability coefficient was .86. In order to obtain evidence of criterion validity, the correlation of a different psychological well-being and a needs satisfaction scale was examined. As a result, correlation values of .56 and .73 were found with the psychological well-being and needs satisfaction scales, respectively. The Psychological Well-Being Scale consists of eight items, and the items are rated as 1 *strongly disagree*, 2 *disagree*, 3 *slightly disagree*, 4 *neutral*, 5 *slightly agree*, 6 *agree*, and 7 *strongly agree*.

Data Collection Procedure

The demographic information form and Psychological Well-Being Scale which was formed as 3-point (disagree, neutral, agree), 5-point (strongly disagree, disagree, neutral, agree, strongly agree), and 7-point (strongly disagree, disagree, slightly disagree, neutral, slightly agree, agree, absolutely agree) Likert-type response categories and 3-point (🙁, 😐, 🙂), 5-point (😭, 🙁, 😐, 🙂, 😂), and 7-

point (😭, 😢, ☹️, 😐, 🙂, 😂, 🤣) emoji reaction categories were turned into online forms and applied to university students in a single session.

Data Analysis

Before the analysis, the data set was examined, and it was observed that there was no missing data. This study was carried out to compare the results of the exploratory (EFA) and confirmatory factor analysis (CFA) of the scales with Likert-type and emoji response categories. First, it was analyzed whether the data sets met the assumptions of the factor analysis. For that purpose, it was investigated whether there were multivariate extreme values in the data set obtained with both Likert-type and emoji response categories from 341 participants, and Mahalanobis distances were calculated. Among the obtained Mahalanobis distances, those giving significant results at α = .001 were excluded from the data sets. Also, whether there is multicollinearity in the data sets was examined through tolerance value (TV), variance inflation factor (VIF), and condition index (CI) values. Whether the data sets provided multivariate normality was analyzed through Mardia's coefficient of multivariate kurtosis. The suitability of the data sets for EFA was investigated through the use of KMO and Bartlett test of sphericity. All values obtained according to the data sets regarding the assumptions were presented in Table 2.

Table 2. Examination of Data Sets in Terms of Factor Analysis Assumptions

Response Type	Number of Categories	Number of Multivarite Outlier	TV (min-max)	VIF (min-max)	CI (min-max)	Mardia's Kurtosis Coefficient	KMO	Bartlett Test
Likert	3	0	.43 - .83	1.21 - 2.33	1 - 27.68	14.51*	.85	1229.6*
	5	10	.34 - .60	1.67 - 2.92	1 - 23.78	15.77*	.92	1940.3
	7	14	.27 - .44	2.25 - 3.71	1 - 25.47	20.56*	.93	2318.0*
Emoji	3	5	.43 - .83	1.21 - 2.33	1 - 27.68	14.51*	.85	1403.0*
	5	4	.37 - .58	1.74 - 2.72	1 - 24.36	17.48*	.91	1830.1*
	7	15	.26 - .55	1.81 - 3.88	1 - 26.41	21.21*	.91	2643.1*

*p < .05

In Table 2, it is seen that the number of multivariate extreme values in data sets varies between 0 and 15. These extreme values were extracted from the data sets of 341 people. It was observed that the tolerance values of all data sets were greater than .01, the variance inflation factor was less than 10, and the condition indexes were less than 30. Accordingly, it can be argued that there is no multicollinearity in data sets (Kline, 2011; Tabachnick & Fidell, 2013). When KMO values and Bartlett's sphericity test results were examined, KMO values were between .85 and .93. The acceptable minimum KMO value for factor analysis is specified as .60 (Kaiser, 1974). Accordingly, the data sets have a sufficient sample size for EFA (Kaiser & Rice, 1974). Bartlett's sphericity test results were significant in all data sets. So, it can be said that the correlation matrices obtained from the data sets were different from the identity matrix. Since the multivariate normal distribution assumption was not provided to perform EFA, the stronger unweighted least squares (ULS) factor extraction method was used against the violation of this assumption (Brown & Moore, 2012). In CFA, the mean and variance adjusted unweighted least squares (ULSMV) estimation method was used. EFA and CFA were carried out by using a polychoric correlation matrix. Factor 10.10.03 (Lorenzo-Seva & Ferrando, 2020) was used for the EFA, and Mplus (Muthén & Muthén, 2012) software was used for CFA.

Ethics Committee Approval

In this study, all rules stated to be followed within the scope of Higher Education Institutions Scientific Research and Publication Ethics Directive were followed. None of the actions stated under the title of Actions Against Scientific Research and Publication Ethics, were taken.

Name of the board conducting the ethical review: Bolu Abant İzzet Baysal University Social Sciences Human Research Ethics Committee

Date of the ethical assessment decision: 15.04.2020 (Session 2020/03)

Ethics assessment document issue number: 2020/81

RESULTS

In this section, findings were given according to the order in the research questions.

Comparison of EFA Results of Data Obtained from Emoji and Likert Type Response Categories

EFA results of the data obtained from the scales with Likert-type and emoji response categories were compared in terms of the variance ratio explained and the factor loadings of the items. The results obtained were presented in Table 3.

Table 3. EFA Results of The Data Obtained from Emoji and Likert Type Rating Scales

	Number of Categories											
	3				5				7			
	Response Type											
Item	Likert	Emoji	Likert	Emoji	Likert	Emoji	Likert	Emoji	Likert	Emoji	Likert	Emoji
No	Factor Loadings		Explained Variance		Factor Loadings		Explained Variance		Factor Loadings		Explained Variance	
1	.86	.93	54.30%	54.47%	.90	.84	67.52%	65.32%	.89	.90	73.48%	75.23%
2	.66	.73			.81	.85			.87	.88		
3	.64	.66			.81	.79			.87	.88		
4	.58	.49			.74	.71			.78	.79		
5	.58	.61			.68	.75			.78	.84		
6	.80	.75			.86	.85			.88	.94		
7	.65	.70			.73	.67			.79	.69		
8	.73	.63			.80	.75			.83	.85		

In Table 3, factor loadings of the items in scales rated in emoji and Likert type were presented. When EFA results of the data obtained from scales rated in Likert and emoji type were examined, it can be said that the factor loadings were very close to each other, and the explained variance rates were very similar. As the number of response categories increased, the explained variance rate increased. However, the EFA results of the data obtained from the scales rated in Likert and emoji type with the same number of categories were very similar.

The Wilcoxon signed-rank test was applied to examine whether the factor loadings of the data obtained from scales rated in Likert and emoji type differ significantly or not. As a result, no significant difference was found between the factor loadings of the data sets obtained with the Likert-type and emoji response categories of both 3-point (Z = -.70, p = .94) and 5- point (Z = -.84, p = .40) as well as 7-point scales (Z = -1.40, p = .16).

Comparison of CFA Results of Data Obtained from Emoji and Likert Type Response Categories

CFA results obtained from data sets whose response categories are Likert-type and emoji were compared with regard to factor loadings of the items. Accordingly, the results obtained were presented in Table 4.

When Table 4 is reviewed, the factor loadings of the scales with both Likert-type and emoji response categories obtained from CFA results can be seen. Findings showed that the factor loadings of the data obtained from the scales with the Likert-type and emoji response category with the same number of categories were very similar.

Table 4. CFA Factor Loading Results of Data Obtained From Emoji and Likert Type Rated Scales

Item No	Number of Categories					
	3		5		7	
	Response Type					
	Likert	Emoji	Likert	Emoji	Likert	Emoji
1	.86	.93	.90	.84	.89	.90
2	.66	.73	.81	.85	.87	.88
3	.64	.67	.81	.79	.87	.88
4	.58	.48	.74	.71	.78	.79
5	.58	.60	.68	.75	.78	.84
6	.80	.76	.86	.85	.88	.94
7	.65	.70	.73	.67	.79	.69
8	.73	.64	.80	.75	.83	.85

The Wilcoxon signed-rank test was applied to examine whether the factor loadings differed in the data obtained from scales rated in emoji and Likert type. As a result, it was found that Likert-type rating with emoji does not reveal a significant difference between factor loadings for both 3-category (Z = .00, p = 1.00) and 5-category (Z = -.84, p = .40) as well as 7-category scored scales (Z = -1.40, p = .16). Table 5 included the fit indices obtained from CFA.

Table 5. Fit Indices in CFA of Data Sets Obtained from Emoji and Likert Type Rating Scales

Number of Categories	Response Type	CFI	ΔCFI	TLI	ΔTLI	RMSEA	ΔRMSEA	90 % CI	RMSEA p-value	Chi-Square	Chi-Square/df	Chi-Square p-value
3	Likert	.98	.04	.97	.06	.05	-.03	.03 - .08	.38	39.29	1.96	.01
3	Emoji	.94		.92		.09		.06 - .11	.00	69.39	3.47	.00
5	Likert	.97	.00	.96	.00	.13	-.00	.11 - .16	.00	137.24	6.86	.00
5	Emoji	.97		.96		.13		.11 - .16	.00	140.27	7.01	.00
7	Likert	.98	.01	.98	.01	.12	-.00	.10 - .14	.00	6116.71	305.84	.00
7	Emoji	.98		.97		.17		.15 - .19	.00	8834.76	441.74	.00

When the scales rated with Likert and emoji had 3 categories, CFI values were obtained as .98 for Likert-type and .94 for emoji. It is stated that the CFI change is important when the difference between these two CFI values is greater than .01 (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Hereunder, when examined in terms of the CFI index, a 3-point Likert-type rating fits the data better than a 3-point emoji rating. However, when the ΔCFI values are examined for the 5 and 7-point, it is observed that these values are less than .01.

When examined in terms of RMSEA, it is stated that the difference is important when the value of ΔRMSEA is greater than .01 (Chen, 2007). Accordingly, in terms of RMSEA, it can be concluded that the Likert-type 3-point rating fits the data better than the 3-point emoji rating. There are no similar comparisons for TLI and Chi-Square (Vandenberg & Lance, 2000). On the other hand, statistics obtained from Likert and emoji type scales are not at a level that will affect the model-data fit decision. In other words, if the model-data fit is provided in the data set obtained from Likert-type scales, it is also provided in the data set obtained from emoji type scales. Similarly, if the model-data fit is not provided in the Likert-type scale, it is not provided in the emoji-type scale, as well. For instance, when the results obtained from 3-point data sets are compared, while the CFI value for the emoji type scale is .94, for the Likert-type scale, it is .98. Since it is stated that CFI and TLI are greater than .90 indicates that model-data fit is achieved (Hair, Black, Babin, & Anderson, 2009; Vandenberg & Lance, 2000), it does not affect the decision about whether model-data fit is achieved in emoji or Likert type scales.

Investigation of The Relationships Between the Scores Obtained from Emoji and Likert Type Response Categories

The relationships between the scores obtained from the data sets, the reaction categories of which are Likert-type and emojis, were examined by gender. Results were presented in Table 6.

Table 6. Correlation Between Scores Obtained from Emoji and Likert Type Rated Scales According to Gender

	Women (*n* = 252)			Man (*n* = 89)		
Response Type	3 Categories Emoji	5 Categories Emoji	7 Categories Emoji	3 Categories Emoji	5 Categories Emoji	7 Categories Emoji
3 Categories Likert	.75**	-	-	.80**	-	-
5 Categories Likert	-	.69**	-	-	.81**	-
7 Categories Likert	-	-	.54**	-	-	.72

** *p* < .01

In Table 6, the correlations between the scores obtained from the emoji and Likert type rated scales varied between .54 and .75 for females and .72 and .80 for males. It can be stated that as the number of categories increases for both males and females, the correlations between the scores obtained from emoji and Likert type rating scales decrease.

Comparison of Reliability of Scores Obtained from Emoji and Likert Type Response Categories

The Cronbach Alpha coefficients obtained from the data sets whose response categories are Likert-type and emojis were presented in Table 7.

Table 7. Cronbach Alpha Coefficients of Data Obtained from Emoji and Likert Type Rated Scales

Response Type	Cronbach Alfa Coefficient
3 Categories Likert	.81
3 Categories Emoji	.81
5 Categories Likert	.90
5 Categories Emoji	.89
7 Categories Likert	.93
7 Categories Emoji	.93

Table 7 shows the Cronbach Alpha coefficients of the data obtained from emoji and Likert-type rated scales. It can be stated that as the number of categories increases, the reliability coefficient increases, and this is already an expected result. It can also be indicated that the reliability of the scores obtained from the Emoji and Likert type rating scales is very close to each other.

DISCUSSION and CONCLUSION

The current study was conducted to examine the structures of scales consisting of Likert and emoji response categories. It was observed that the structures were similar as a result of EFA and CFA obtained from the data of scales with the same number of categories. As the number of categories increased as a result of EFA, the variance rate also increased. However, similar results were obtained from emoji and Likert type data. When EFA was conducted to see factor loads, there was not enough evidence that the factor loads were statistically significantly different from each other. Therefore, the construct validity of the scales consisting of Likert and emoji response categories in terms of EFA was found to be sufficient. Based on this result, it can be argued that emoji response categories can be used instead of Likert response categories.

When CFA was conducted, results showed that fit indices were sufficient for both emoji and Likert type scale data. However, the fit indices decreased as the number of categories increased. Moreover, the number of categories of fit indices has changed, but the differences between Likert and emoji type response categories were not significant. When CFA factor loadings were examined, results showed that the factor loads obtained from the emoji and Likert type data did not differ significantly. Therefore, the current study results showed that the construct validity of the data obtained from both scale types was sufficient.

When the correlations of emoji and Likert type scales were examined, it was seen that the correlation scores decreased with the increase in the number of categories. Results also showed that the highest correlation indicated a moderate relationship. Therefore, the same scale in Likert and emoji categories may not measure the same structure, or it may cause different reactions in participants. In particular, when female participants' seven-category Likert and emoji scales data were examined, the correlation decreased to .54, suggesting that different characteristics are measured with the same items. Similar results were found by Setty, Srinivasan, Radhakrishna, Melwani, and Dr (2019) when they used 3 different scales (emoji scale, Venham picture test, and facial image scale) to measure dental anxiety in children aged 4-14. The correlation between the emoji scale and the Venham Picture test was .73, and the correlation with the facial image scale was .87. Unlike these findings, Swaney-Stueve, Jepsen and Deubler. (2018) compared liking and emotions and stated that the correlation between 9-point Likert and 7-point emoji scale was .99. This difference may have occurred since the comparison was made on different scales. Also, since the comparison was carried out with individuals in the 8-14 age group, the difference with the current study results may be occurred due to population and age differences. On the other hand, in a study conducted by Alismail, and Zhang (2018), it was stated that individuals interpreted the same emojis differently. For instance, some individuals rated the neutral facial expression (😐) as sad. The number of emoji used increases with the increase in the number of categories. Therefore, it can be stated that individuals do not perceive emojis in the same way as Likert-type verbal expressions. As a result, low correlation results were found.

According to the research findings, there is no obstacle to the use of emoji type response categories in scales. It was observed that scales with the same number of categories were very similar in terms of reliability coefficients of construct validity and internal consistency. Therefore, emoji type response categories can also be used in scale development studies. However, the relationships between the total scores were at a medium level. These differences may be because of differences in measured structures or because the reaction categories of emoji and Likert-type caused different reactions in individuals. In the current study, it is seen that 3-emoji reaction categories can be used instead of 3-Likert response categories. However, the correlation results of the 5 and 7 emoji and Likert response categories were different. Since the use of emoji response categories is still new, in order to contribute to the literature and practitioners, the similarities or differences of the results obtained from the present study should be compared with samples from different age groups and different scales. Based on the findings of the current study, it can be stated that the data obtained from university students with Likert type or emoji response categories have similar construct validity. However, it should be acknowledged that this study is limited to the instrument and the sample used.

According to the present study findings, when the results obtained from the scales consisting of 3, 5 and, 7 emoji and Likert response categories are examined, it was seen that women and men attribute different meanings to the same emoji. In future studies, research should be conducted to examine the reasons for those attributions. In addition, this differentiation can be examined in depth with different age groups and equal/close numbers of gender groups. However, it should be kept in mind that this study is limited to the data obtained from the Psychological Well-Being Scale.

Considering that the use of emoji response categories in scales is new, future studies need to be conducted to examine whether the situations of indecision, which can be experienced in scales with 7 or more Likert response categories (verbal and numerical), can be prevented. Moreover, preschool and primary school students' literacy level and limitations need to be considered, and it should be investigated whether a more valid result can be obtained by using emoji reaction categories among these populations. Additionally, questions may also be read to illiterate individuals, and researchers may ask them to indicate the answers by showing emojis to obtain first-hand data. This is because it is easier to collect data from these individuals and the validity and reliability of the collected data can be increased.

REFERENCES

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society, 39*(2), 331-342. doi:10.1177/147078539703900202

Alismail, S., & Zhang, H. (2018, January). *The use of emoji in electronic user experience questionnaire: An exploratory case study*. Paper presented at 51st Hawaii International Conference on System Sciences, Hawaii. doi: 10.24251/hicss.2018.427

Anderson, L. W. (1991). Attitudes and their measurement (N. Çıkrıkçı, Trans.). *Ankara University Journal of Educational Sciences*, *24*(1), 241-250. doi: 10.1501/Egifak_0000000734 (Original work published 1988).

Bayat, B. (2014). Scaling, scales and "Likert" scaling technique in applied social science researches. *Ankara Hacı Bayram Veli University, Journal of the Faculty of Economics and Administrative Sciences, 16*(3), 1-24. Retrieved from https://dergipark.org.tr/tr/pub/gaziuiibfd/issue/28309/300829

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). New York: Guilford.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. doi: 10.1080/10705510701301834

Chen, Z., Lu, X., Shen, S., Ai, W., Liu, X., & Mei, Q. (2017). *Through a gender lens: An empirical study of emoji usage over large-scale android users*. Retrieved from https://arxiv.org/abs/1705.05546

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Danesi, M. (2017). *The semiotics of emoji*. London: Bloomsbury Publishing.

Deubler, G., Swaney-Stueve, M., Jepsen, T., & Su-Fern, B. P. (2020). The k-state emoji scale. *Journal of Sensory Studies, 35*(1), 1-9. doi: 10.1111/joss.12545

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research, 97*(2), 143-156. doi: 10.1007/s11205-009-9493-y

Dwyer, E. E. (1993). *Attitude scale construction: A review of the literature* (Report No. ED359201). Retrieved from https://eric.ed.gov/?id=ED359201

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

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 321-348). New York: Social Science Research Council.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Upper Saddle River: Prentice Hall.

Kaiser, H. (1974). An index of factor simplicity. *Psychometrika, 39*(1), 31-36. doi: 10.1007/BF02291575

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*(1), 111-117. doi: 10.1177/001316447403400115

Kline, R. B. (2011). *Principles and practise of structural equating modeling* (3rd ed.). New York: The Guilford Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5-55. Retrieved from https://legacy.voteview.com/pdf/Likert_1932.pdf

Lorenzo-Seva, U., & Ferrando, P. J. (2020). *Factor* (Version 10.10.03) [Computer software]. Tarragona: Universitat Rovira i Virgili.

Marengo, D., Giannotta, F., & Settanni, M. (2017). Assessing personality using emoji: An exploratory study. *Personality and Individual Differences, 112*(1), 74-78. doi: 10.1016/j.paid.2017.02.037

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0*. Los Angeles, CA: Muthén & Muthén.

Payne, G., & Payne, J. (2004). *Key concepts in social research*. London: Sage Publications.

Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B., & Gaspar, R. (2018). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics, 35*(7), 1925-1934. doi: 10.1016/j.tele.2018.06.005

Price, L. R. (2017). *Psychometric methods*. New York: The Guilford Press.

Setty, J. V., Srinivasan, I., Radhakrishna, S., Melwani, A. M., & Dr, M. K. (2019). Use of an animated emoji scale as a novel tool for anxiety assessment in children. *Journal of Dental Anesthesia and Pain Medicine, 19*(4), 227-233. doi: 10.17245/jdapm.2019.19.4.227

Swaney-Stueve, M., Jepsen, T., & Deubler, G. (2018). The emoji scale: A facial scale for the 21st century. *Food Quality and Preference, 68*, 183-190. doi: 10.1016/j.foodqual.2018.03.002

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). New Jersey: Pearson.

Telef, B. B. (2013). The adaptation of psychological well-being into Turkish: A validity and reliability study. *Hacettepe University Journal of Education, 28*(3), 374-384. Retrieved from https://dergipark.org.tr/tr/download/article-file/87222

Thurstone, L. L. (1927). Three psychophysical laws. *Psychological Review*, *34*(6), 424-432. doi: 10.1037/h0073028

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. doi: 10.1177/109442810031002

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 192-211

The Comparison of the Equated Tests Scores by Various Covariates using Bayesian Nonparametric Model *

Meltem YURTÇU ** Hülya KELECİOĞLU *** Edward L. BOONE ****

Abstract
This research is based on obtaining equated scores by using covariates in the Bayesian nonparametric model. As covariates in the study, gender, mathematics self-efficacy scores, and common item scores were used. The distributions were obtained for all score groups. Hellinger Distance was calculated to obtain the distances between the distributions of equated scores by using covariates and the distribution of the target test scores. These distances were compared with the distributions of equated scores obtained from methods based on Item Response Theory. The study was conducted on Canadian and Italian samples of Programme for International Student Assessment (PISA) 2012. PARSCALE and IRTEQ were used for classical methods, and R was used for Bayesian nonparametric model. When gender, mathematics self-efficacy scores, and common item scores were used as covariates in the model, distance values of obtained equated scores to target test scores were close to each other, but their distributions were different. The closest distribution to target test scores was achieved when gender and mathematics self-efficacy scores were used together as covariates in the model, and the farthest distributions were obtained from item response theory methods. As a result of the research, it was determined that the model is more informative than the classical methods.

Key Words: Test equating, Bayesian nonparametric model, covariates, equated scores, score distribution.

INTRODUCTION

It is very important to compare the scores of the individuals evaluated by the tests. Equating is used to compare the scores obtained from different test forms that serve the same purpose. One of the most important steps of equating is the selection of the equating method, which differs regarding the use of common items or common individuals. The methods involving common individuals can be classified as single group design, counterbalanced design, and equivalent group design, whereas the method involving common items in non-equivalent groups is named as Non-Equivalent groups with Anchor Test (NEAT) (Branberg & Wiberg, 2011). NEAT is used when there is no chance of applying another questionnaire and the data required to reveal the difference between the groups were obtained from common items/tests (Liou, Cheng, & Li, 2001; Moses, Deng, & Zhang, 2010). The selection of the common tests is crucial in the design, and the selected test should have a similar mean and item difficulty with the tests in question and should represent this test in terms of content (Dorans, Moses, & Eignor, 2010; Kolen, 1988; Kolen & Brennan, 2014; Mittelhaeuser, Beguin, & Sijtsma, 2011; Sinharay & Holland, 2006; Wei, 2010; Wiberg & von Davier, 2017). The common test should be one-dimensional, should have a high correlation with the scores of the other tests to be equated, and should reflect the exact structure of the test forms (Wallin & Wiberg, 2017). In addition, the use of common tests that address the trends over time in NEAT design may be appropriate only for certain individuals, which may create a bias for equating. If the common tests/items fail to satisfy these conditions, the

* This study is based on Meltem Yurtçu's doctoral thesis titled "The Comparison of the Equated Tests Scores by Using Various Covariates using Bayesian Nonparametric Model".

** Assist. Prof, Inonu University, Faculty of Education, Malatya-Turkey, meltem.yurtcu@gmail.com, ORCID ID: 0000-0003-3303-5093
*** Prof. Ph.D., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934
**** Prof. Ph.D., Virginia Commonwealth University, Department of Statistical Science and Operations Research, Richmond-USA, ELBoone@vcu.edu, ORCID ID: 0000-0003-0755-6899

To cite this article:
Yurtçu, M., Kelecioğlu, H., & Boone, E. L. (2021). The comparison of the equated tests scores by various covariates using bayesian nonparametric model. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 192-211. doi: 10.21031/epod.864744

Received: 19.01.2021
Accepted: 17.06.2021

reliability of equating and other processes associated with common tests/items will be negatively affected (Wei, 2010; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). Moreover, the tests to be equated may not have any common items or tests. In this case, the bias and mean standard error can be reduced by adding variables associated with the test scores to the test equating process, which allows to explain the difference between the groups (Branberg & Wiberg, 2011; Liou et al., 2001; Oh, Guo, & Walker, 2009; Wiberg, 2015; Wiberg & Branberg, 2015), and to increase the accuracy of the estimation (Branberg & Wiberg, 2011; Kim, Livingston, & Lewis, 2009, 2011; Livingston & Lewis, 2009; Oh et al., 2009; Wiberg & Branberg, 2015). Wiberg and Branberg (2015) stated that using a single common variable that has a high correlation with the test scores could give results similar to a common test. Liou et al. (2001) also suggested that the variables selected from historical data of the individuals may give better results than common tests.

In recent years, Non-Equivalent Groups with Covariates (NEC) design, which uses common variables/covariates in the absence of common items, has been added to the literature (Branberg & Wiberg, 2011; Wiberg & Branberg, 2015). The design involving the use of both common item/s and covariate/s is called NEATNEC (Wiberg & Branberg, 2015).

The most important assumption of NEC design is that covariates are able to explain the difference between groups. The most important step of this design is that the situational distributions of the test scores should be the same in both groups in terms of covariates categories. This is an indication that the achievement of individuals is evaluated according to their categorical characteristics. However, if the test scores to be equated were obtained at different time periods (i.e., equating a new test with an old test), this hypothesis may not be valid because the characteristics of the test scores and the covariates may have changed over time (Wiberg & Branberg, 2015).

Although many researchers have described covariates in different terms, they emphasized that these variables are related to test scores, and they can explain the difference between groups (Branberg & Wiberg, 2011; Kim et al., 2009; Liou, 1998; Liou et al., 2001; Wiberg & Branberg, 2015; Wright & Dorans,1993). In the literature, the variables such as age, gender, and educational status were observed to be included as covariates (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana,2015a; Karabatsos & Walker, 2009; Liou et al., 2001; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). The accuracy of the prediction may increase with the increase of the number of covariates added to the study, which makes the number of covariates added to the study important. Another important issue is the number of covariate categories. As the number of covariate categories increases, the number of individuals falling into each relevant category may decrease. Therefore, limiting the number of variable categories will give more appropriate results and will strengthen the prediction (Wallin & Wiberg, 2017; Wiberg & Branberg, 2015).

Equating methods are based on various theories and assumptions, which are classified in the literature as Classical Test Theory and Item Response Theory (IRT). However, in recent years, Bayesian approach has come to the fore in test-equating studies.

Bayesian Approach

In the classical approach, the *p*-value is used to test the significance of null hypotheses, which varies according to the sample and purpose of the researcher (Berger, Boukai, & Wang, 1997; Kruschke, 2010; Kruschke, Aguinis, & Joo, 2012; Lee & Boone, 2011; Rounder, Morey, Speckman, & Province, 2012). This can be considered as a disadvantage because point estimation affects the outputs in terms of reaching an accurate result. The confidence interval used in Bayesian approach carries more information than point estimation. The confidence intervals for posterior inferences generated by Bayesian approach can be expressed with the mean and 95% confidence interval (highest density interval/HDI). The points falling in this range are more accurate than the points that are outside (Kruschke, 2010).

Bayesian approach provides well-defined probabilistic models for observed data and unknown values. There are two types of Bayesian approaches. Parametric Bayesian approach uses a limited number of parameters, but it has some limitations, whereas the flexible use of the number of parameters in the models constitutes the basis of Bayesian nonparametric approach (De Iorio, Müller, Rosner, & MacEachern, 2004, Müller & Quintana, 2004; Orbanz & Teh, 2010; Shah & Ghahramani, 2013). Dirichlet Process (DP) Model is one of the models that have a central role in Bayesian nonparametric approaches (De Iorio et al., 2004; Gonzalez et al., 2015a; Petrone, 1999a). This model allows the inclusion of the covariates in equating process. The randomness effect of the variables on the distribution of the test scores will appear as dependency, which is explained by the Dependent Dirichlet Process (DDP), an extension of the DP model (Barrientos, Jara, & Quintana, 2016; MacEachern, 1999, 2000). However, the selection of prior distributions in Bayes nonparametric approaches is usually very difficult. Petrone (1999a, 1999b) suggested using Bernstein-Dirichlet Prior (BDP) model to eliminate this limitation. In their studies, Barrientos et al. (2016) expanded the model further and developed Dependent Bernstein Polynomial Process (DBPP) model. Barrientos et al. (2012, 2016) discussed two specific types of DBPP. In this study, DBPP involving a dependent stick-breaking process with common weights and predictor-dependent support points was employed. This type is called single-weight DBPP (wDBPP). Z represents covariate space, and F_z represents covariate-dependent random probability distributions.

For $\forall z \in Z, \{F_z: z\epsilon Z\}$, wDBPP can be formulated as;

$$f_{(z)}(\cdot) = \sum_{j=1}^{\infty} w_j \beta\big(z\big|\lceil k\theta_j(\mathbf{z})\rceil, k - \lceil k\theta_j(\mathbf{z})\rceil + 1\big)$$

This model, which represents an infinite set of beta distributions, suggests that the test scores have covariate-dependent sample densities. This model can be shown as:

$\{Fz; z \in Z\} \sim$ wDBPP$(\alpha, \lambda, \psi, H)$.

Where $= \{\boldsymbol{h_z}; \boldsymbol{z} \in \boldsymbol{Z}\}$; $\quad \boldsymbol{v_1}, \boldsymbol{v_2}, \dots \dots .., \alpha > 0$ are independent, random variables whose distribution is defined by $\beta(1, \alpha)$; k is a discrete random variable with a distribution indexed to a finite-dimensional parameter λ, $\boldsymbol{\theta}_{\mathbf{j(z)}} = \boldsymbol{h_z}\left(\boldsymbol{r}_{\mathbf{j}}(\boldsymbol{z})\right), \boldsymbol{r_1}, \boldsymbol{r_2}$ **....,** are independent and identically distributed real-valued stochastic processes indexed by the parameter ψ. This model provides a covariate-dependent equating transformation (Gonzalez, Barrientos, & Quintana, 2015b).

In this study, the accuracy of the predictions and their contribution to the test equating process were analyzed by comparing the equated scores obtained from Bayesian Nonparametric Model (BNP) by using various covariates at NEC design.

METHOD

The research was conducted with real data. The distribution of equated scores obtained from the scaling methods based on IRT was compared with the distributions of equated scores obtained from the BNP model.

Sample

The data used in the research was obtained from PISA 2012. In order to carry out the equating process in non-equivalent groups, two countries with different success levels were selected. According to PISA 2012 math results, the data of Canada, which was ranked as 13th with an average score of 518, and Italy, which was ranked as 32nd with an average score of 485, were taken from the database published by OECD (http://www.oecd.org/pisa/data). The records with missing data were removed, and Italian data with a sample size of 908 and Canadian data with a sample size of 931 were used in the analysis.

Data Collection Tools

In PISA 2012, a cognitive test measuring students' mathematics literacy and a student questionnaire were used. The data of the research is comprised of the Italian students' responses to booklet 5 and Canadian students' responses to booklet 6 of the mathematics sub-test. Booklets 5 and 6 were selected to be used in the research because of the equal number of math questions and the high number of common items. There were 12 common items in the booklets.

Gender and mathematics self-efficacy score (MATHEFF) were used as covariates in the analysis, where gender is a two-category variable and MATHEFF is a continuous variable. In addition, the anchor item scores were taken as the covariate in the BNP model. The reason for using MATHEFF is that it is defined as the variable that explains the mathematics achievement (Ayotola & Adedeji, 2009; Hackett & Betz, 1989; Koğar, 2015; Schulz, 2005; Siegle & McCoach, 2007; Thien & Darmawan, 2016). This variable was derived from the sum of the item scores, where a higher score indicates lower self-efficacy. MATHEFF scores varied between 8-32. But, since the scores range between 0-1 in the model, MATHEFF scores were also converted into the 0-1 range, showing the change within one unit.

Another covariate used in the NEC design of the BNP model was gender. There are many studies in the literature using gender as covariate (Branberg & Wiberg, 2011; Gonzalez et al., 2015a, 2015b; González & Wiberg, 2017; Liou et al., 2001). In addition, in many studies, gender is considered as a variable that creates differentiation among groups (Martin, Mullis, Foy, & Stanco, 2012; Yıldırım, Yıldırım, Yetişir, & Ceylan, 2013).

Regarding the equating studies performed in non-equivalent groups, the number of common items in the tests should be equal to at least 20% of the number of questions to minimize the equating error (Angoff, 1971). The study was carried out with 24 items in NEC design, and the total score of the common items was used as the covariate. In NEAT design, 12 items were taken as external commonitems, and the study was carried out with 36 items. To avoid them from affecting the model as a different criterion, partially scored items in the booklets were converted into two category-scores.

Data Analysis

In the research, IRT-based scale conversion methods and the analyses using the BNP model were carried out separately. First of all, unidimensionality and local independence were tested for IRT. Factor 10.3 analysis software was used to test unidimensionality, which was analyzed over 36 items. The unidimensionality of 36-item in booklets was taken as the proof of the unidimensionality of the 24-item version. As a result of the factor analysis, Kaiser Mayer Olkin (KMO) value of booklet 5 was found to be .95, whereas Bartlett's value was 7086.60 (df = 630; p < .001). Regarding booklet 6, KMO value was .94 and Bartlett's value was 6427.00 (df = 630, p < .001). KMO values indicated the sufficiency of the sample sizes for the analysis, and Bartlett's value indicated the factorizability of the data set. Regarding these values, it can be said that the tests were unidimensional.

The unidimensionality of the booklets provided insight about local independence assumption. Moreover, in order to test the local independence assumption, the correlation between the items was calculated for the top and bottom 27% of the data (Kelley, 1939). The correlation between the top and bottom groups was found to be lower than the overall correlation; therefore it was concluded that the local independence assumption was met.

Parameter estimation

The two test forms to be scaled in the study are parallel. The parameters obtained from these forms were estimated from different individuals, and the mean and standard deviations of the groups were different; therefore the estimations were made using separate calibration methods.

Equating by NEAT design was performed using ability parameters. The -2loglikelihood values obtained for 2 parameter logistic model (PLM) and 3 PLM were tested by chi-square test and 3 PLM model was found to be significant. Therefore, the parameters were estimated according to 3 PLM method. Parscale 4.1 program was used in the estimation of item parameters.

Scale conversion

Common items were taken as external common items in NEAT design to allow a comparison with NEC design. IRTEQ software was used to convert the parameters taken from the PARSCALE software to the same scale. Since IRT true-score equating is more accurate and precise (Li, Jiang, & von Davier, 2012), this process was carried out on true-score. In the study, booklet 6 was taken as the target test, whereas booklet 5 was taken as the basic test.

Test equating by Bayes nonparametric approach

In order to make accurate statistical predictions, Markov chain Monte Carlo (MCMC) sampling method was used to obtain a sample representing the universe (Kruschke, 2015; StataCorp, 2015). In this study, MCMC method was used to estimate population parameters (k, γ, w) of the BNP model. General information about the population can be obtained using covariates. MCMC processes were performed separately for Canada and Italy data sets. The covariates and parameters compatible with the data are combined in the files prepared in MCMC sampling by using DBPP.

The covariates used in the research were added to the model as anonymous priors. This fact prevented the bias that may arise from the effects of these variables on the posterior distributions of the scores and ensured a more objective evaluation.

Prior distribution specification: The distributions of wDBPP based on MCMC method were given as:

$\boldsymbol{h_z}(\cdot) = \frac{\boldsymbol{exp}\{\cdot\}}{\boldsymbol{1+exp}\{\cdot\}}$, $\boldsymbol{r_j}(\boldsymbol{z}) = \boldsymbol{z^T\gamma_j}$ and $\boldsymbol{\gamma_j}|\ \boldsymbol{\mu}, \boldsymbol{S} \sim^{iid} \boldsymbol{N_p}(\boldsymbol{\mu}, \boldsymbol{S})$, $\boldsymbol{j = 1, 2}, \ldots.$

Here; $\boldsymbol{v_j}|\ \boldsymbol{\alpha} \sim \boldsymbol{\beta}(\boldsymbol{1}, \boldsymbol{\alpha}), \boldsymbol{k}|\ \boldsymbol{\lambda} \sim \boldsymbol{Poisson}(\boldsymbol{\lambda})\ \|_{\{\boldsymbol{k}>1\}}, \boldsymbol{\mu}|\ \boldsymbol{m_0}, \boldsymbol{S_0} \sim \boldsymbol{N_p}(\boldsymbol{m_0}, \boldsymbol{S_0}), \boldsymbol{S}|\boldsymbol{v},\ \boldsymbol{\psi} \sim \boldsymbol{IW_p}\ (\boldsymbol{v}, \boldsymbol{\psi})$. In equation I, $\boldsymbol{W_p}\ (\boldsymbol{v}, \boldsymbol{A})$; scale matrix A represents *p*-dimensional inverted-Wishart distribution with degrees of freedom v. The values that Gonzalez et al. (2015a) found to be significant in their study, were also included in their study of 2015b, therefore the following values were used while generating the prior distribution $\boldsymbol{\lambda = 25}, \boldsymbol{m_0 = 0_p}, \boldsymbol{S_0 = 2.25 * I_p}, \boldsymbol{v = p + 2}$, and $\boldsymbol{\alpha = 1}$. MCMC algorithm was run to explain the posterior distribution of wDBPP model and to obtain the posterior distribution samples of all model parameters.

Posterior inference: All computations were coded and performed in R 3.2.1 statistics software. The posterior probability distribution was given by:

$$\begin{aligned}
&p(v, k, w, \gamma | y, z) \\
&\propto \prod_{i=1}^{n} \left[\sum_{j=1}^{10} w_j \beta \left(y_i \middle| \left| k \frac{e^{Z_i^T \gamma_j}}{1 + e^{Z_i^T \gamma_j}} \right|, k - \left| k \frac{e^{Z_i^T \gamma_j}}{1 + e^{Z_i^T \gamma_j}} \right| + 1 \right) \right] \left[\prod_{j=1}^{10} \beta(v_j | 1{,}1) \right] \\
&\times \left[\frac{25^k e^{-25}}{k!\,(1 - e^{-25})} \right] \left[\prod_{j=1}^{10} (2\pi) |S|^{-\frac{1}{2}} e^{-0.5(\gamma_j - \mu)^T S^{-1} (\gamma_j - \mu)} \right] (2\pi) |S_0|^{-\frac{1}{2}} e^{-0.5(m_0)^T S_0^{-1} (m_0)} \\
&\times \frac{|\psi|^2}{2^2 \Gamma_2(2)} |S|^{\frac{7}{2}} e^{-\frac{1}{2} tr(\psi S^{-1})}
\end{aligned}$$

The posterior predictive distribution was given as below:

$$p(T | y_i, z_i) = \int p(v, k, w, \gamma | y, z)\, L(T | v, k, w, \gamma) dv dk dw d\gamma$$

Where

$$L(T|v,k,w,\gamma) = \sum_{j=1}^{10} w_j \beta \left(T \middle| \left\lceil k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}} \right\rceil, k - \left\lceil k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}} \right\rceil + 1 \right)$$

shows the sum obtained for the identified distributions.

The number of iterations was first set as 5000 to test the parameters in the generated files. Then, MCMC number was set as 150 000, and the analyses were performed by repeating 10 times for each file in order to obtain a proper distribution. The analyses of the test forms were carried out simultaneously, which took around 10 hours and 23 minutes for each file.

The algorithm of Gibbs and Metropolis-Hastings sampling method was as follows. It was used to explain the posterior distribution obtained by gathering the covariables with the model in MCMC files:

An initial $v^{*}\sim p(v|v^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there were 10 v values in the research).

An initial $\gamma^{*}\sim p(\gamma|\gamma^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there were 20 γ values in the research).

An initial $k^{*}\sim p(k|k^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there was 1 k value in the research).

After completing this stage, the equated scores were obtained using cumulative distributions of the test scores.

The transformation functions are as follows, where T is score distribution; $\boldsymbol{t_x}$ represents the scores obtained from test X, $\boldsymbol{t_y}$ represents the scores obtained from test Y, and z represents the covariates;

$$\begin{array}{lcl} \boldsymbol{t_x} = \boldsymbol{F^{x^{-1}}}(\cdot) & & \\ & \Longrightarrow & \boldsymbol{t_y} = \boldsymbol{\varphi}(\boldsymbol{t_x}) = \boldsymbol{F^{y^{-1}}}(\boldsymbol{F^{x}}(\cdot)) \\ \boldsymbol{t_y} = \boldsymbol{F^{y^{-1}}}(\cdot) & & \end{array}$$

$$\begin{array}{lcl} \boldsymbol{t_{z_x}} = \boldsymbol{F^{z_x^{-1}}}(\cdot) & & \\ & \Longrightarrow & \boldsymbol{t_{z_y}} = \boldsymbol{\varphi}\left(\boldsymbol{t_{z_x}}\right) = \boldsymbol{F^{z_y^{-1}}}(\boldsymbol{F^{z_x}}(\cdot)) \\ \boldsymbol{t_{z_y}} = \boldsymbol{F^{z_y^{-1}}}(\cdot) & & \end{array}$$

The analyses conducted to obtain equated scores were completed in 7 days and 6 hours. The equating process was completed by putting the generated profile distributions into the percentiles determined for covariate categories.

DBPP model defines continuous distribution functions in (0-1) range. Therefore, the score estimations were made in this range as Gonzalez et al. (2015b) have done in their study. After equating, the scores were converted to the scale-of-100 so that the highest score will be 100. This is considered as the best scaling method in equating studies involving the tests with different ranges (Livingston, 2004). Therefore, the continuous variables used in the distributions were converted and analyzed in (0-1) range, then the graphics and distributions obtained for equated scores were converted to the scale-of-100 and interpreted.

Comparison criteria

In traditional equating methods, standard criteria such as Root Mean Square Error (RMSE), Mean Square Error (MSE), bias, and standard errors (SE) are used to assess parameter estimation error.

However, it is difficult to compare the results obtained by the methods based on different models such as IRT and BNP (Wiberg & Gonzalez, 2016). Therefore, in this study, the comparison of the results using the criteria such as RMSE and MSE was not possible. Hellinger Distance, which provides statistical information, was used in this study to compare the equated scores obtained by BNP and IRT methods to target test's scores. This distance is the sum of the distances between the points of each distribution. There are many forms of Hellinger distance. Hellinger Distance used to compute the distance between two distributions f and g (Boone, Merrick, & Krachey, 2012) is formulated as;

$$\widehat{H}(f,g) = \left[\frac{1}{2}\int\left(\sqrt{\hat{f}(x)} - \sqrt{\hat{g}(x)}\right)^2 dx\right]^{1/2} \approx \left[\frac{1}{2}\sum_{l=1}^{k}\left(\sqrt{\hat{f}(x)} - \sqrt{\hat{g}(x)}\right)^2 (x_l - x_{l-1})\right]^{\frac{1}{2}}$$

The distances between the distributions of the scores were computed according to the method above, and the distributions are shown through graphics in the results part. One of the titles (participants, sample, or working group) should be used with respect to the group formation procedure used in the study. The information about the sampling procedure and the group should be given in this part.

RESULTS

In PISA 2012, the mean score and standard deviation of 908 Italian students, who answered booklet 5, was 51.51 and 20.72, respectively. Whereas the mean score and standard deviation of 931 Canadian students who answered booklet 6 was 52.27 and 22.06 respectively.

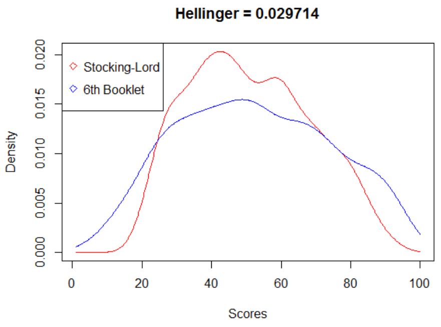
Equating errors occurred as a result of scaling according to IRT methods in the NEAT design were computed, and the score distributions obtained from various methods were analyzed.

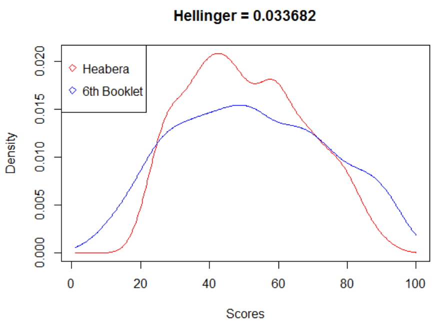
In the two booklets, answers taken by two non-equivalent groups were used for scaling. RMSE values were calculated.

Table 1. RMSE Values Obtained According to IRT Methods

Mean – Mean	Mean-Sigma	Stocking-Lord	Heabera
0.149	0.13	0.20	0.18

The lowest error was obtained from Mean-Sigma method and the highest error from Stocking-Lord method. New ability parameters were computed, and item parameters of the target test were used for finding true scores. Probability density distributions of each method and their distance from the target test were calculated using Hellinger distances.





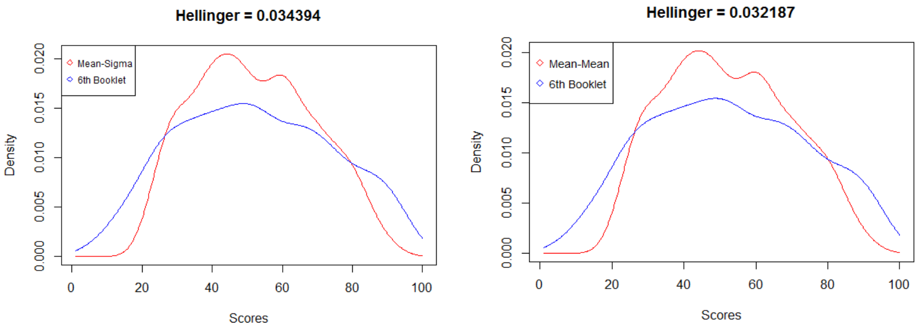
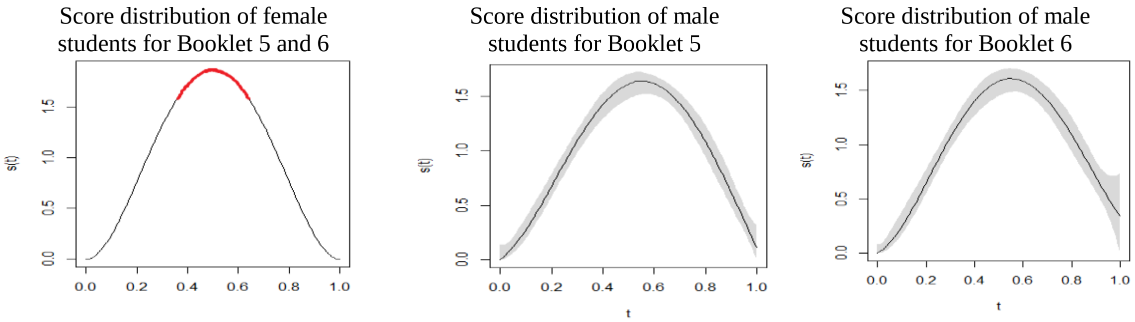


Figure 1. The Distribution of Obtained Scores and Their Distance from the Target Test's Scores

Regarding the probability density distributions of the predicted scores in Figure 1, the distributions of the scores were observed to be similar and to be at approximately similar distances to the target test's distribution according to the Hellinger distance. Although Mean-Sigma method gave the lowest RMSE, the distributions obtained from the characteristic curve methods were closer to the distribution of the target test. According to Hellinger distance, Stocking-Lord method was the closest distribution with 0.029714.

The distance between the distribution of equated scores obtained by using gender as covariate in the BNP model and the distribution of target test's scores

Gender was taken as covariate, and students' scores were gathered with this variable. Distributions were first examined according to the booklets. Figure 2 shows the distribution of the scores and confidence intervals that best reflect the population for each gender.


Score distribution of female students for Booklet 5 and 6

Score distribution of male students for Booklet 5

Score distribution of male students for Booklet 6

Note. The confidence interval is shown in red to female because it was very narrow.

Figure 2. Score Distributions and Confidence Intervals according to Gender for Booklets.

The distributions were observed to be similar. Especially, the distribution of female students was the same for both booklets. The accuracy of the score estimation was checked through confidence intervals. Confidence intervals of female students' score distributions were found to be quite narrow, whereas male students' confidence intervals were wide, which may indicate uncertainty in the estimation of these scores. The decrease in the accuracy may be due to the low number of students in the sample used for the estimation of scores, or due to the fact that the scores of the students having the same profile were distributed in a wide range.

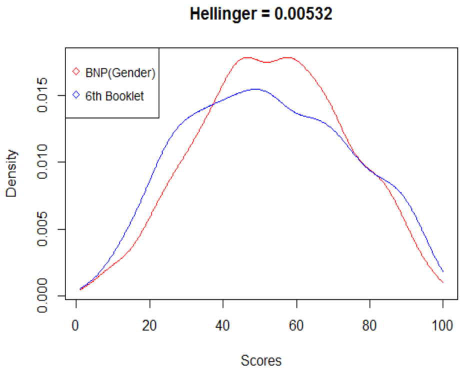


Figure 3. Distribution of the Target Test’s Scores and the Scores Equated with Gender

The score equated with gender covariate was calculated for each student. The distributions of equated scores and target test’s scores were compared. The distance between these distributions was calculated by Hellinger distance. As can be seen from Figure 3, the distribution of equated scores was observed to be sharper than the distribution of the target test’s scores. The distance between these two curves was 0.00532, which was approximately one-fifth of the distance obtained by IRT methods.

The distance between the distribution of equated scores obtained by using MATHEFF as covariate in the BNP model and the distribution of target test’s scores

MATHEFF was taken as the covariate, and students’ scores were associated with this variable. The score distributions that best reflect the population according to MATHEFF levels were computed. The distributions of scores at different MATHEFF levels were analyzed according to booklets.

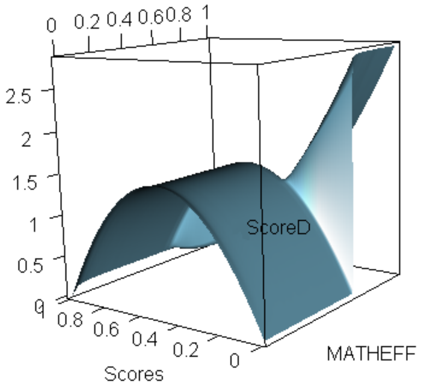


Figure 4. The Distributions of the Scores Equated with MATHEFF

Students at different MATHEFF levels had different profiles. The distribution of each profile was computed. Test score distributions of booklets 5 and 6 according to MATHEFF levels of the students were similar, therefore they are shown in a single graph in figure 4. As students' self-efficacy levels decrease (or for higher values of MATHEFF), the intensity of their scores decreases. Based on these distributions in each profile, it was also possible to see at which scores the students' distribution changed and how this change was affected for both booklets.

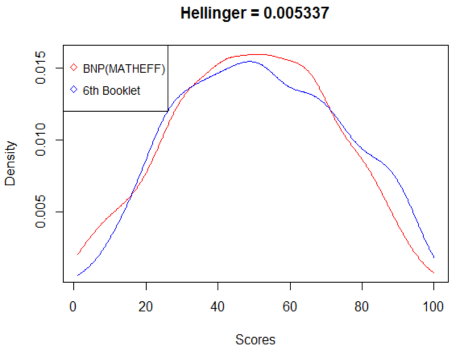


Figure 5. Distribution of the Target Test's Scores and the Scores Equated with MATHEFF Levels

In the BNP model, the distribution of equated scores was very close to the distribution of the target test' scores. Hellinger distance was calculated as 0.005337. This distance is significantly lower than the distance obtained from IRT methods and the distance of the model obtained using gender. Compared to the BNP model using gender, the distributions were observed to approach and differentiate from the target test at different points. In the model using MATHEFF, the distribution of equated scores moved away from the target test at the ends, whereas in the model using gender, the distribution of equated scores differed from the target test in average values.

The distance between the distribution of equated scores obtained by using both gender and MATHEFF as covariates in the BNP model and the distribution of target test's scores

Students' MATHEFF scores were examined according to gender. The distributions obtained for female students were similar to males for booklets 5 and 6, therefore, graphs are shown for both genders in figures. Figure 6 and 7 shows the distributions of the students for booklets 5 and 6.

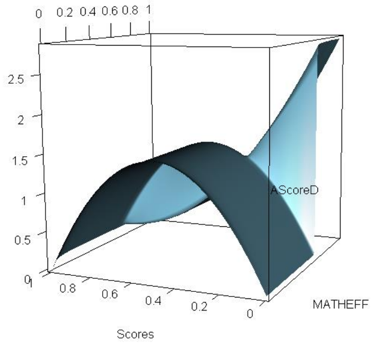


Figure 6. Distributions for Booklet 5

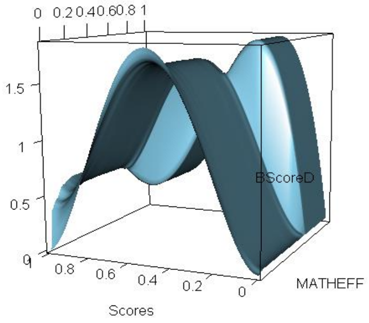


Figure 7. Distributions for Booklet 6

Regarding booklet 5, it was observed that the intensity of high scores of both genders' students with low mathematics self-efficacy decreased. In booklet 6, the students of both genders with low mathematics self-efficacy were observed to be clustered around 20. As can be seen from these distributions, students' intensity around high scores decreased as MATHEFF scores get higher, which indicates lower mathematics self-efficacy levels.

So, it can be concluded that booklet 6 was easier than booklet 5 for both female and male students. In addition, the differentiation of the distributions in booklets may indicate that using these two covariates was effective in revealing the differences between the booklets. Equated scores were obtained using the cumulative distributions of these distributions generated by combining covariates and individuals' scores. The probability distributions of equated scores and target tests were examined together in Figure 8.

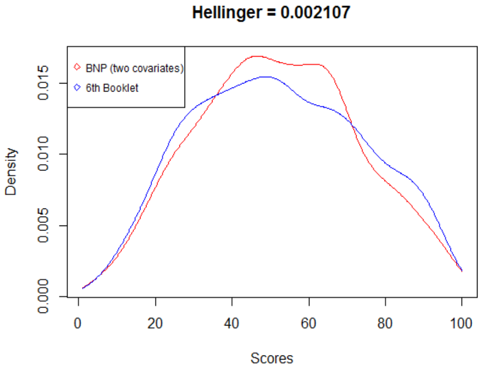


Figure 8. Distribution of the Target Test’s Score and the Scores Equated with both Covariates

The distribution of equated scores is very close to the target test when both covariables were included in the model; Hellinger distance is also relatively small (0.002107) compared to other models. From Figure 8, it can be seen that equated scores obtained by using two covariates got closer to the target test. In particular, the approximation of distributions to the extreme values might indicate that the model could be used to tolerate the error in extreme values.

The distance between the distribution of equated scores obtained by using common items as covariate in the BNP model and the distribution of target test’s scores

In the first part of the study, equated scores were obtained from common items according to IRT scaling methods. In this section, the scores obtained from the sum of common items were used as a covariate. The distributions obtained from the combination of student scores and covariates are shown in Figures 9 and 10.

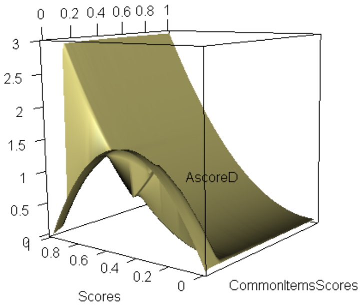


Figure 9. Distributions for Booklet 5

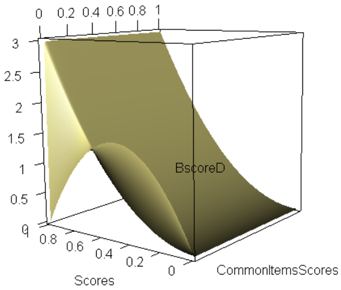


Figure 10. Distributions for Booklet 6

In order to check whether common items reflect the tests or not, the correlation between common test scores and test scores was examined. These correlations were found to be .79 for booklet 5 and .75 for booklet 6. Accordingly, it can be said that common items represent the tests statistically.

According to Figure 9, if common item scores were not included in the model as covariate or they contributed to the model with very low scores in booklet 5, the density of students was observed to increase on average scores and densities towards the end scores decreased. With the increase of common item scores, the shapes of distributions differed from first distributions, and it was observed that low score densities decreased and high score densities increased.

Regarding Figure 10, which shows the analysis results for booklet 6, if common item scores were not included in the model as a covariate or contributed to the model with very low scores, students are concentrated around the mean. The distributions of students were quite similar for other score levels. Therefore, regarding the individuals with other scores than low common item scores, the distributions are similar for both booklets. The differences in common item scores failed to explain the difference in the math achievement of the students. Booklet 6 was observed to be easier than booklet 5.

Equated scores were obtained according to common item scores of students. The probability distributions of these scores and target test were examined together, and their distributions are given in Figure 11.

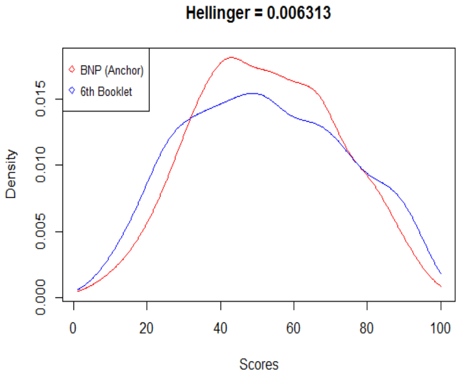


Figure 11. Distribution of the Target Test’s Scores and The Scores Obtained from BNP Model with Common Items

Hellinger distance between the distribution of equated scores obtained by using common items as covariate and the distribution of target test scores was calculated as 0.006313. This distance was smaller than the one of the IRT methods, but it was greater than the values obtained from BNP models with other covariates. The distribution of equated scores obtained using common items is similar to the distribution of the equated scores obtained using gender. Both distributions diverged from target test’s distribution at the ends. Although the numerical values of Hellinger distances were insufficient, their shapes supported the information given about these distributions.

DISCUSSION and CONCLUSION

In this study, equated scores were computed using the BNP model, bringing a different perspective than classical methods. Gender, mathematics self-efficacy scores, and the sum of common items scores were used as covariates. Equated scores were computed for different covariates, and the distances between these scores’ distributions and the distribution of the target test’s scores were examined. The explanation of mathematics achievement by the variables and the differences between booklets were interpreted using the BNP model. The results obtained from IRT and BNP models and their interpretation are given below.

The scores taken from common items were considered as the external common test in IRT equating methods; the minimum error was obtained from Mean-Sigma method, whereas the maximum error from the Stocking-Lord method. Therefore, it was concluded that external common items caused more error than moment methods in reducing the difference between items’ characteristic curves; and the difference between the discriminant parameters obtained from common tests applied to the groups was less than the difference in characteristic curves. Regarding the distances between the distribution of true scores obtained by IRT scaling methods and distribution of target test’s scores, the closest distribution was obtained from Stocking-Lord method. This fact can be expressed as that Stocking-Lord method produced closer values, even though it generated more erroneous predictions than other IRT methods.

In the BNP model, similar score distributions were obtained from female and male students for each booklet when gender was considered as the only covariate. Although gender was seen to be insufficient in showing the difference between the booklets, it was found that booklet 6 was comprised of easier questions than booklet 5. In spite of similar distributions, the confidence intervals of male and female students’ distributions were different. Since the same distributions were obtained for the students of both genders, it was concluded that gender has no significant effect on mathematics performance/achievement. There are various studies supporting this fact in the literature (Hall & Hoff, 1988; Lindberg, Hyde, Petersen, & Linn, 2010; Thien & Darmawan, 2016).

In the BNP model, when MATHEFF was taken as the covariate, the distributions of the students with medium and high scores were similar. The distributions of both booklets varied according to the MATHEFF level; therefore, it was found that MATHEFF was effective on mathematics achievement. Thus, it can be concluded that MATHEFF explains mathematics achievement. The literature contains studies showing that MATHEFF explains mathematics achievement (Ayotola & Adedeji, 2009; Ding, 2016; Hackett & Betz, 1989; Koğar, 2015; Schulz, 2005; Siegle & McCoach, 2007; Thien & Darmawan, 2016). In traditional equating, if the knowledge of individuals is not included, score distributions of each student group would be considered to be the same. In this study, the differentiation in the score distribution of the students in various sub-groups was kept under control, and equated scores of each sub-group were computed. Regarding the model in which MATHEFF was used, it was concluded that the distribution of equated scores approaches the distribution of target test's scores. The most important assumption of NEC design is that the distribution categories obtained from covariates should be the same for the sub-groups (Wiberg & Branberg, 2015). The differences between booklets can be observed using this assumption. Since MATHEFF distributions were similar in both booklets, it was concluded that either this variable could not fully explain the difference between booklets, or the booklets were very similar. However, even in this case, it could be said that booklet 5 contained more difficult questions than booklet 6.

When both MATHEFF and gender were used as covariates in the BNP model, the information obtained from the model was more detailed than the models with a single covariate. If two covariates are used in the model, it is possible to distinguish the variables affecting the distributions of students' mathematics achievement and the magnitude of this effect. The distributions in booklets were the same for both genders. In our case, different distributions were obtained for different booklets and MATHEFF levels. The use of these variables together revealed that they could explain both the difference between booklets and mathematics achievement levels. The distribution of equated scores obtained using two covariates was observed to approach the distribution of target test's scores more than other models.

When the sum of common item scores in the BNP model was used as a covariate, only the distributions of low-score students varied, and the range was quite small. Therefore, the distribution of medium- and high-score students was observed to remain the same. In other words, it was concluded that common items were at the same level and uniform; otherwise they would change the distribution of test scores directly. The same result was obtained for both booklets. The correlation of common item scores was higher for booklet 5 and caused more distributional variations for this booklet. This fact showed that common items were more similar to the questions in booklet 5 and made more distinctions between the sub-groups with different scores in this booklet. Since the distributions obtained from common item scores did not differ significantly according to the booklets, it was concluded that common items don't adequately explain mathematics achievement. The distance between the distribution of the scores equated with common item scores and the distribution of the target test's scores showed the effectiveness of the method but using two covariates in the model was more effective. There are studies supporting the use of covariates for achieving more positive results in equating process, in cases where common items do not possess the properties required for equating or the assumptions of test equating are not satisfied (Dorans & Holland, 2000; Liou et al., 2001; Wright & Dorans, 1993).

When only MATHEFF and only gender were used as a covariate, the distributions did not differ significantly according to booklets. In the model where two covariates were used, distribution differences were observed according to booklets. In the model where the common item scores were used, distribution differences were observed in the low-score student group. This result suggested that in BNP models, common item scores explained the difference between the booklets more than MATHEFF scores. Despite different covariate types used in BNP models, booklet 6 was observed to be easier than booklet 5. Likewise, it is possible to say that the questions in booklet 5 were more distinctive.

Regarding the distributions of equated scores and the distances of these distributions to target test, the comparison between IRT methods and BNP models was straightforward. The distributions of equated scores obtained from the BNP model were closer to the distributions of the target test. The distances between the distributions of equated scores using the BNP model and the distribution of target test's scores were smaller. The closest distance was obtained from the distribution of the BNP model using two covariates together. Therefore, it can be said that more precise estimations are obtained by using BNP model. There are many studies supporting that the Bayesian method makes better predictions than classical methods, and it can be used to obtain much useful information (Karabatsos & Walker, 2009; Kruschke et al., 2012; van de Schoot, et al., 2013).

It was very difficult to compare BNP models that use different covariates according to Hellinger distances. Even though the numerical values obtained from Hellinger distance between BNP models is not sufficient for decision making, the shape of the distributions supported the information about the distance to the target test. Since BNP model uses score distributions for equating, it doesn't require any limitation such as having a same number of individuals in the basic test and target test. Moreover, there is no need to limit the number of individuals in the sub-groups involved in the tests. In the study, the low number of individuals in some sub-groups and the inclusion of covariates to the model as missinformation caused large confidence intervals. However, in spite of large confidence intervals, BNP models would yield more useful and informative results.

As BNP model keeps group invariance under control, the irregularities and discontinuities of the distributions have been eliminated. For this reason, there is no need for pre-smoothing, the selection of the bandwidth parameter, and the derivation of the standard error of equating used in other equating methods (Gonzalez et al., 2015b). This is an indication of the importance of the model (Karabatsos & Walker, 2009).

In future research, researchers may use the model for test equating without using any covariate. When covariate is used in the model, the study can be carried out to determine the items with DIF (Differential Item Functioning) according to variable/s' categories. In the model, equated scores can be obtained using different continuous and discrete covariates such as socioeconomic status, age, etc.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Ayotola, A., & Adedeji, T. (2009). The relationship between mathematics self-efficacy and achievement in mathematics. *Procedia Social and Behavioral Science, 1*, 953-957. Retrieved from https://cyberleninka.org/article/n/1232855.pdf

Barrientos, A. F., Jara, A., & Quintana, F. (2012). On the support of MacEachern's dependent dirichlet processes and extensions. *Bayesian Analaysis*, 7(2), 277-310. Retrieved from https://projecteuclid.org/download/pdfview_1/euclid.ba/1339878889

Barrientos, A. F., Jara, A., & Quintana, F. (2016). *Fully nonparametric regression for bounded data using Bernstein polynomials.* Retrieved from http://www.mat.uc.cl/~ajara/Publications_files/DependentBernstein.pdf

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unied frequentist and bayesian testing of a precise hypothesis. *Statistical Science, 12*(3), 133-160. Retrieved from https://www2.stat.duke.edu/~berger/papers/statsci.pdf

Boone, E. L. Merrick, J. R. W., & Krachey, M. J. (2012). A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, *84*(4), 833-849. doi: 10.1080/00949655.2012.729588

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440. doi: 10.1111/j.1745-3984.2011.00153.x

De Iorio, M., Müller, P., Rosner, G., L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. J*ournal of the American Statistical Association, 99*(465), 205-215. doi: 10.1198/016214504000000205

Ding, Y. (2016). *How do students' mathematics self-efficacy, mathematics self-concept and mathematics anxiety influence mathematical literacy?-A comparison between Shanghai-China and Sweden in PISA 2012* (Master thesis). University of Gothenburg, Faculty of Education, Gothenburg, Sweden.

Dorans, J. N., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281-306. doi: 10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS RR-10-29). New Jersey: ETS, Princeton.

González J., & Wiberg M. (2017) Recent developments in equating. In J. González & M. Wiberg (Eds.), *Applying test equating methods: Methodology of educational measurement and assessment* (pp. 157-178). Switzerland: Springer, Cham

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015a). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis 89*, 222-244. doi: 10.1016/j.csda.2015.03.012

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015b). A dependent Bayesian nonparametric model for test equating. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W-C. Wang, (Eds.) *Quantitative psychology research* (pp. 213-226). New York: Springer Cham Heidelberg New York Dordrecht London.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20*(3), 261-273. doi: 10.2307/749515

Hall, C. W., & Hoff, C. (1988). Gender differences in mathematical performance. *Educational Studies in Mathematics 19*(1988) 395-401. Retrieved from https://link.springer.com/content/pdf/10.1007%2FBF00312455.pdf

Karabatsos, G., & Walker, S. G. (2009). A bayesian nonparametric approach to test equating. *Psychometrika*, *74*(2), 211-232. doi: 10.1007/S11336-008-9096-6

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

Kim, S., Livingston, S. A., & Lewis, C. (2009). *Effectiveness of collateral information for improving equating in small samples*. New Jersey: ETS, Princeton.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education*, *24*(4), 302-323. doi: 10.1080/08957347.2011.607057

Koğar, H. (2015). PISA 2012 matematik okuryazarlığını etkileyen faktörlerin aracılık modeli ile incelenmesi. *Eğitim ve Bilim, 40*(179), 45-55. doi: 10.15390/EB.2015.4445

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice*, *7*(4), 29-36. doi: 10.1111/j.1745-3992.1988.tb00843.x

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3nd. ed.). New York: Springer.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews; Cognitive Science*, *1*(5), 658-676, doi: 10.1002/wcs.72

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (Second Ed.)*: A tutorial with R, JAGS, and Stan.* Waltham, MA: Academic Press / Elsevier.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4) 722-752. doi: 10.1177/1094428112457829

Lee, A. H., & Boone, E. L. (2011). A frequentist assessment of Bayesian inclusion probabilities for screening predictors. *Journal of Statistical Computation and Simulation, 81*(9), 1111-1119. doi: 10.1080/00949651003702135

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Mesurment*, *49*(2), 167-189. doi: 10.1111/j.1745-3984.2012.00167.x

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. doi: 10.1037/a0021276

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica, 8,* 669-690. Retrieved from http://www3.stat.sinica.edu.tw/statistica/oldpdf/A8n33.pdf

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, *25*(2), 197-207. doi: 10.1177/01466210122032000

Livingston, S. A. (2004). *Equating test scores (Without IRT).* Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Livingston, S. A., & Lewis, C. (2009). Small-sample equating with prior information. (ETS Research Rep. No. RR-09-25). New Jersey: ETS, Princeton.

MacEachern, S. N. (1999). *Dependent nonparametric processes*. Retrieved from https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf

MacEachern, S.N., (2000). *Dependent Dirichlet processes* (Tech. rep). Department of Statistics, The Ohio State University. Retrieved from https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Boston College, MA, USA: International Study Center.

Mittelhaeuser, M.-A., Beguin, A. A., & Sijtsma, K. (2011). *Comparing the effectiveness of different linking design: The internal anchor versus the external anchor and pre-test data* (Measurement and Research Department Reports, 1). Arnhem: Cito.

Moses, T., Deng, W., & Zhang, Y.-L. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (RR-10-23). New Jersey: ETS, Princeton.

Müller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*(1), 95-110. doi: 10.1214/088342304000000017

Oh, H. J., Guo, H., & Walker, M. E. (2009). *Impraved reability estimates for small samples using empirical Bayes teshniques* (RR-09-46). New Jersey: ETS, Princeton.

Orbanz, P., & Teh, Y. W.(2010). Bayesian nonparametric models. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning.* Boston, MA: Springer. doi: 10.1007/978-0-387-30164-8_66

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics 27*(Varsa sayı no) 105-126. Retrieved from https://www.jstor.org/stable/pdf/3315494.pdf?refreqid=excelsior%3A7e6e0614f5a5f181dfd25d2ad6947bc6

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics 26*, 373-393. Retrieved from https://www.jstor.org/stable/pdf/4616563.pdf?refreqid=excelsior%3A801798d1ac07988dafb6e83769c949b2

Rounder, J. N., Morey, R. D., Speckman, P. L., & Province, M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(2012), 356-374, doi: 10.1016/j.jmp.2012.08.001

Schulz, W. (2005, April). *Mathematics self-efficacy and student expectations: Result from PISA 2003*. Annual Meetings of the American Educational Research Association in Montreal. Retrieved from https://files.eric.ed.gov/fulltext/ED490044.pdf

Shah, A., & Ghahramani, Z. (2013, September). *Determinantal clustering process- A nonparametric bayesian approach to kernel based semi-supervised clustering.* Proceedings of the TwentyNinth Conference on Uncertainty in Artificial Intelligence. Retrieved from http://auai.org/uai2013/prints/papers/200.pdf

Siegle, D., & McCoach, D. B. (2007). Increasing student mathematics self-efficacy through teacher training. *Journal of Advanced Academics, 18*(2), 278-312. Retrieved from https://files.eric.ed.gov/fulltext/EJ767452.pdf

Sinharay, S., & Holland, P. W. (2006). *Choice of anchor test in equating* (RR-06-35). New Jersey: ETS, Princeton.

StataCorp. (2015). *Stata Bayesian analysis reference manual release 14*. College Station, TX: StataCorp LLC. https://www.stata.com/manuals14/bayes.pdf

Thien, L. R., *&* Darmawan, I. G. N. (2016). Factors associated with Malaysian mathematics Performance in PISA 2012. In L. M. Thien, N. A. Razak, J. Keeves, & I. G. N. Darmawan (Eds.), *What can PISA 2012 data tell us?: Performance and challenges in five participating Southeast Asian countries* (pp. 81-105). Rotterdam: Sense Publisher.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2013). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 1-19. doi: 10.1111/cdev.12169

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology – 81st annual meeting of the psychometric society, Asheville, North Carolina*. New York: Springer.

Wei, H. (2010, May). *Impact of non-representative anchor items on scale stability*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Denver, CO.

Wiberg, M. (2015). Anote on equating test scores with covariates. In E. Frackle-Fornius (Ed.), *Festschrift in honor of Hans Nyquist on the occasion of his 65th birthday* (pp. 96-99). Stockholm: Department of Statistics Stockholm University, Sweden.

Wiberg, M., & Gonzalez, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement. 53*(1), 106-125. Retrieved from: http://www.mat.uc.cl/~jorge.gonzalez/papers/TR/Assess_TR.pdf

Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing, 17*(2), 105-126. doi: 10.1080/15305058.2016.1277357

Wiberg, M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361. doi: 10.1177/0146621614567939

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (RR-93–04). New Jersey: ETS, Princeton.

Yıldırım, H. H., Yıldırım, S., Yetişir , M. İ., & Ceylan, E. (2013). *PISA 2012 ulusal ön raporu*. Ankara: MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü (YeğiTek).

Parametrik Olmayan Bayes Yöntemiyle Ortak Değişkenlere Göre Yapılan Test Eşitlemelerinin Karşılaştırılması

Giriş

Denk olmayan gruplarda ortak test deseninde ortak testin seçimi oldukça önemli olup bu test, eşitlenecek olan testler ile benzer ortalama, madde zorluğuna sahip olmalı ve bu testleri içerik olarak temsil etmelidir (Dorans, Moses, & Eignor, 2010; Kolen, 1988; Mittelhaeuser, Beguin, & Sijtsma, 2011; Sinharay & Holland, 2006; Wei, 2010; Wiberg & von Davier, 2017). Ancak ortak testler bu tür özelikleri her zaman sağlayamayabilir. Ortak testlerin tek boyutlu olmaması, diğer testlerdeki puanlarla yüksek oranda ilişki vermemesi, test formlarındaki yapıyı tam olarak ölçmede yetersiz kalması (Wallin & Wiberg, 2017) veya uygulamasından kaynaklı hataların olması (Liou, Cheng, & Li, 2001) eşitlenmedeki güvenirliği ve ortak testlere bağlı diğer süreçleri etkilemektedir (Wiberg & von Davier, 2017; Wei, 2010). Bu durumlara ek olarak, sadece zaman içerisindeki eğilimleri ele alan ortak testlerin denk olmayan gruplarda ankor madde (NEAT) deseninde kullanılması, sadece belirli bireyler için uygun olabilir ki bu durumda eşitleme için bir yanlılık oluşturabilir. Bu da testlerin güvenirliklerini olumsuz yönde etkileyecektir (Wiberg & Branberg, 2015; Wiberg & von Davier, 2017; Wei, 2010). Ayrıca birçok büyük uygulamaları gerektiren sınavlarda ortak madde veya ortak test bulunmamaktadır. Bu durumda test puanları ile ilişkili ve gruplar arasındaki farkı açıklayabilen değişkenlerin kestirim sürecine ek bilgi olarak veya ortak testlerin yerine eklenmesi ile yanlılık ve ortalama standart hata azaltılabilir (Branberg & Wiberg, 2011; Liou ve diğerleri, 2001; Oh, Guo, & Walker, 2009; Wiberg, 2015; Wiberg & Branberg, 2015). Böylece kestirimin doğruluğunu arttırabileceği için eşitleme çalışmaları birçok yönden incelenebilecektir (Branberg & Wiberg, 2011; Kim, Livingston, & Lewis, 2009, 2011; Livingston & Lewis, 2009; Oh ve diğerleri, 2009; Wiberg & Branberg, 2015). Son yıllardaki çalışmalarda ortak maddelerin olmadığı durumda ortak değişkenlerin kullanılması ile Denk Olmayan Gruplarda Ortak değişken (Non-equivalent Groups with Covariates /NEC) (Branberg & Wiberg, 2011; Wiberg & Branberg, 2015) ve hem ortak madde hem de ortak değişkenlerin kullanılması ile NEATNEC deseni literatüre eklenmiştir (Wiberg & Branberg, 2015). Bu çalışma NEC deseni üzerinden yürütülmüştür.

NEC deseninin en önemli varsayımı, ortak değişkenlerin gruplar arasındaki farklılığı açıklayabildiğidir. Test puanlarının durumsal dağılımlarının, ortak değişkenlerin kategorilerine göre her iki grupta da aynı olması bu desen için en önemli adımdır (Wiberg & Branberg, 2015). Bu adımın en önemli parçası olan ortak değişkenlerin seçimi ise oldukça önemlidir. Birçok araştırmacı ortak değişkenleri farklı terimlerle ifade etmiş olsa da bu değişkenlerin test puanları ile ilişkili olması ve gruplar arasındaki farkı açıklayabilecek nitelikte olmasına vurgu yapmıştır (Branberg & Wiberg, 2011; Kim ve diğerleri, 2009; Liou, 1998; Liou ve diğerleri, 2001; Wiberg & Branberg, 2015; Wright & Dorans, 1993). Alanyazında ortak değişken olarak genellikle yaş, cinsiyet, eğitim durumu gibi değişkenlerin yer aldığı görülmektedir (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana, 2015a; Karabatsos & Walker, 2009; Liou ve diğerleri, 2001; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017).

Ortak testin kullanımından daha iyi sonuç vermesi için ortak değişkenlerin sayısı arttırılabilir. Ancak ortak değişken sayısı arttıkça, bu değişkenlerin kategorilerine düşen birey sayısı azalacağından dolayı değişkenlere ait kategori sayılarının sınırlandırılması daha uygun sonuçlar verecektir (Wiberg & Branberg, 2015; Wallin & Wiberg, 2017).

Son yıllarda Bayes yaklaşımı da test eşitleme çalışmalarında öne çıkmaktadır. Özellikle Parametrik olmayan Bayes yaklaşımı (BNP) ortak değişkenlerin modele eklenmesini olası hale getirmektedir. Bu araştırmada iki farklı ortak değişken kullanılarak NEC deseninde BNP modeline göre elde edilen eşitlenmiş puanlar Madde Tepki Kuramı (MTK) yöntemleri ile karşılaştırılarak test eşitleme sürecine katkısı incelenmiştir.

Yöntem

Araştırmada ortak maddelerin bulunmadığı NEC deseninde farklı ortak değişkenler ile BNP modeli kullanılmıştır. Modellere göre elde edilmiş olan puan dağılımları ve eşitlenmiş puanların hedef teste olan uzaklığı Hellinger uzaklığı ile incelenmiştir. Araştırma gerçek veri üzerinde yürütülmüş olup BNP modeline göre elde edilen eşitlenmiş puanlara ait dağılımlar ile madde tepki kuramına dayalı olarak ölçekleme yöntemlerinden elde edilen eşitlenmiş puanların dağılımları karşılaştırılmıştır.

Araştırmanın evreni ve örneklemi

Denk olmayan gruplar arasında eşitleme yapmak için PISA 2012 verilerinden yararlanılmıştır. Kayıp ve eksik veriler temizlendikten sonra, 5. kitapçık için 908 kişilik İtalya verisi, 6. kitapçık için 931 kişilik Kanada verisi kullanılmıştır.

Veri toplama araçları

PISA 2012 kapsamında öğrencilere uygulanan matematik okuryazarlığını ölçen bilişsel testten ve öğrenci anketinden yararlanılmıştır. NEC deseni için cinsiyet, matematik öz yeterlik puanı (MATHEFF) ve ortak madde puanları ortak değişken olarak alınmış ve ortak değişkenlerin kullanılması ile elde edilen sonuçlar birbirleri ile karşılaştırılmıştır. Çalışma NEC deseninde 24 madde üzerinden yürütülmüş olup, ortak maddelerin toplam puanı ortak değişken olarak kullanılmıştır. NEAT deseninde ise 12 madde dış ortak madde olarak alınmış olup 36 madde üzerinden çalışma yürütülmüştür.

Verilerin analizi

Araştırmada MTK kuramına dayalı ölçek dönüştürme yöntemleri ve BNP modeli için analizler ayrı ayrı sürdürülmüştür. İlk olarak MTK varsayımlarından tek boyutluluk ve yerel bağımsızlık test edilmiş ve testlerin tek boyutlu olduğu sonucuna varılmıştır. Alt ve üst gruplardaki korelasyon ile toplam gruptaki korelasyon birlikte incelenerek yerel bağımsızlık varsayımı desteklenmiştir.

Parametre kestiriminde veri seti ile uyumlu model olarak 3 PLM anlamlı bulunmuş ve analizler bu yönteme göre kestirilmiştir. Madde parametrelerinin kestirimi için Parscale 4.1 programından yararlanılmıştır. Kalibre aşamasında Bayes modellerini temel alan modellerden Expected A Posteriori (EAP) yöntemi kullanılmıştır.

Ölçek Dönüşümü için NEC deseninde ortak değişkenler ortak madde yerine kullanılarak 24 madde üzerinden analizleri gerçekleştirilecektir. NEAT deseninde de ortak maddeler, NEC deseni ile karşılaştırmayı sağlayabilmek için, dış ortak madde olarak alınmıştır. IRTEQ programı ile ölçekleme yapılmıştır. Araştırmada 6. kitapçık hedef test olarak belirlenmiştir. 5. kitapçık temel test olarak alınmış ve gerçek puan hesaplanmıştır.

Parametrik olmayan bayes (bnp) yaklaşımına göre test eşitleme: BNP yöntemi kullanılarak yapılan eşitleme çalışmaları ile eski ve yeni test puanları arasında kurulabilecek ilişki ortak değişkenlerin sürece katılması ile şekillendirilmiştir. Modelde yer alan parametrelerin kestirimlerinde uygun sonuçlar elde edebilmek için MCMC yöntemi kullanılmıştır. MCMC örnekleme süreci ile hazırlanan dosyalarda DBPP modeli kullanılarak veriye uygun parametreler ve ortak değişkenler birleştirilmektedir. Kanada ve İtalya veri setleri için ayrı ayrı MCMC süreçleri yürütülmüştür. Daha sonra ise eşitleme fonksiyonundan yararlanarak eşitlenmiş puanlar elde edilmiştir. Çalışmada elde edilen puan dağılımları ile birlikte güven aralıklarına da yer verilmiştir.

BNP modeli için, Gonzalez ve diğerlerinin (2015a, 2015b) çalışmalarında kullanmış olduğu formüllerden yararlanılarak R 3.2.1 programında kodlar oluşturularak analizler gerçekleştirilmiştir.

Karşılaştırma kriteri: Çalışmada, MTK yöntemleri ile BNP Yöntemi ile elde edilen eşitlenmiş puanları karşılaştırmak için istatistiksel bilgi veren ve eşitlenmiş puanlara ait dağılımların hedef teste olan uzaklıklarını inceleyen Hellinger Uzaklığı kullanılmıştır.

Sonuç ve Tartışma

Araştırmada ortak maddelerden elde edilen puanlar dış ortak test olarak alınmıştır. Ortak maddelerin parametreleri üzerinden yapılan ölçekleme sonucunda Stocking-Lord yönteminin diğer MTK yöntemlerine göre daha hatalı kestirim yapmış olsa dahi gerçek puan olarak hedef teste daha yakın değerler ürettiği şeklinde ifade edilebilir. Li, Jiang ve von Davier (2012) de araştırmasında MTK gerçek puan eşitleme ile elde edilen puanların daha doğru ve kesin olduğunu vurgulamaktadır.

BNP modelinde ortak değişken olarak sadece cinsiyet ele alındığında, kız ve erkek öğrenciler için kitapçıklarda benzer dağılımlar elde edilmiştir. Cinsiyet değişkenin kitapçıklar arasındaki farkı göstermede yetersiz olduğu sonucu görülse de 6.kitapçığın 5.kitapçıktan daha kolay sorular içerdiği sonucu elde edilmiştir. Ortak değişken olarak cinsiyetin kullanıldığı araştırmaları literatürde görmek mümkündür (Branberg & Wiberg, 2011; Gonzalez & Wiberg, 2017; Gonzalez ve diğerleri, 2015a, 2015b; Liou ve diğerleri, 2001). Aynı kitapçığı almış olan kız ve erkek öğrenciler için güven aralıkları farklılık gösterse de dağılımları oldukça benzer olup cinsiyetin matematik performansı üzerinde önemli bir etkisinin olmadığını göstermektedir. Literatürde bu durumu destekleyen benzer çalışmaların yer aldığını görmek mümkündür (Hall & Hoff, 1988; Lindberg, Hyde, Petersen, & Linn, 2010; Thien & Darmawan, 2016).

BNP modelinde ortak değişken olarak MATHEFF alındığında tüm düzeylerdeki bireylere yönelik üç boyutlu bir dağılım grafiğine yer verilmiştir. Orta ve yüksek puana sahip bireylere ait dağılımlar benzerlik göstermiş, düşük düzeydeki puana sahip bireylere ait dağılımlar ise farklılaşmıştır. Kitapçıkların her ikisi için de dağılımlar MATHEFF puan düzeyinde göre değişim gösterdiğinden, MATHEFF değişkeninin matematik performansında bireyler arasındaki farkı ortaya koyduğu sonucuna ulaşılmaktadır. Dolayısı ile MATHEFF ortak değişkeninin matematik başarısını açıkladığı sonucuna ulaşılabilir. Literatürde MATHEFF değişkeninin matematik başarısını açıkladığını gösteren çalışmalar yer almaktadır (Ayotola & Adedeji, 2009; Ding, 2016; Hackett & Betz, 1989; Koğar, 2015; Thien & Darmawan, 2016; Schulz, 2005; Siegle & McCoach, 2007). Geleneksel yöntemle yapılan eşitleme çalışmalarında bireylere ait önsel bilgilere yer verilmemesi durumunda her birey için eşitleme dağılımları aynı olarak alınacaktır. Bu çalışma ile bireylere ait puan dağılımlarının alt gruplarda farklılaşması kontrol altında tutularak, alt gruplara göre eşitlenmiş puanlar elde edilmiştir. MATHEFF değişkenin modelde kullanılması ile eşitlenmiş puanlardan elde edilen dağılımın, hedef testteki puanlara yaklaştığı sonucunu ortaya çıkarmaktadır. NEC deseninde ortak değişkenlerden elde edilen dağılımlara ait kategorilerin alt gruplar için aynı olması (Wiberg & Branberg, 2015) varsayımdan yararlanılarak kitapçıklar arasındaki farklar gözlenebilmektedir. MATHEFF değişkeninin her iki kitapçıkta da benzer dağılımlar vermiş olması ile kitapçıklar arasındaki farkı tam olarak açıklayamadığı veya kitapçıkların birbirlerine oldukça benzer oldukları söylenebilir. Fakat bu durumda dahi, bu alt problem için elde edilen sonuçlarda 5.kitapçığın, 6.kitapçığa kıyasla zor sorular içerdiği ifade edilebilir.

MATHEFF ve cinsiyet birlikte ortak değişken olarak BNP modelinde kullanıldığında daha önceki alt problemlere kıyasla modelde daha detaylı bilgiler elde edilmiştir. Bu alt problem ile hangi değişkenin bireylerin matematik başarısına ait dağılımlarını ne kadar değiştirdiğini görmek mümkündür. Bu iki değişken birlikte ele alındığında, her kitapçık ve MATHEFF değişkenindeki her puan düzeyi için farklı dağılımlar oluşturduğundan, bu değişkenlerin birlikte hem kitapçıklar arasındaki farkı hem de matematik başarısını açıklayabildiği sonucunu ortaya koymuştur. İki ortak değişken kullanımı ile elde edilen eşitlenmiş puanların dağılımının hedef test puanlarına ait dağılıma yaklaştığı sonucu gözlemlenmiştir.

BNP modelinde ortak madde puanları ortak değişken olarak alındığında bireylere ait elde edilen puan dağılımları sadece düşük puanlarda ve çok az bir ranjda değişmektedir. Dolayısı ile ortak maddelerden yüksek puan alan bireyler ile düşük puan alan bireylerin puan dağılımları benzerlik göstermektedir. Bu da farklı düzey ortak madde puanına sahip öğrencilerin matematik başarıları arasında net bir ayrım yapılmadığını göstermektedir. Yani ortak maddelerin aynı düzey ve tek tip olduğu veya direkt test puanlarına etki ederek dağılımlarını değiştirdiği sonucunu ortaya çıkarmaktadır. İki kitapçık için de bu durum benzer şekildedir. Ancak ortak madde puanlarının 5.kitapçıkla daha yüksek korelasyon vermesi ve bu kitapçıktaki dağılımlarda daha çok değişim yapmış olması, ortak maddelerin 5.kitapçıktaki sorulara daha çok benzediği ve bu kitapçıktaki farklı puan almış alt gruplar arasında daha fazla ayrım yaptığını göstermektedir. Ortak madde puanlarından elde edilen dağılımların kitapçıklara göre büyük bir farklılık göstermemesi, ortak maddelerin matematik başarısını yeterli düzeyde açıklamadığı sonucunu ortaya çıkarmıştır. Ortak madde puanlarının kullanılması ile elde edilen eşitlenmiş puanlar ile hedef teste ait dağılım arasındaki uzaklık yöntemin etkili olduğunu ancak iki ortak değişken kullanılmasının ortak maddelerden daha etkili olduğu sonucunu ortaya çıkarmıştır. Ortak maddelerin eşitleme için gereken özellikleri taşımadığı veya test eşitleme için varsayımların ihlal edildiği durumlar için, ortak değişkenlerin kullanılmasının eşitleme sürecinde daha uygun sonuçlar vereceğini destekleyen çalışmalar literatürde yer almaktadır (Dorans & Holland, 2000; Liou ve diğerleri, 2001; Wright & Dorans,1993).

Sadece MATHEFF ve sadece cinsiyet değişkeni kullanıldığında dağılımlar kitapçıklara göre aşırı bir farklılık göstermemektedir. İki ortak değişkenin kullanıldığı modelde dağılımların kitapçıklara göre farklılıkları açık bir şekilde görülmekte; ortak madde puanlarının kullanıldığı modelde ise düşük ortak madde puanlarında kitapçıklara göre dağılımların farklılaştığı görülmektedir. Bu durum BNP modellerinde; ortak madde puanlarının kitapçıklar arasındaki farkı, sadece MATHEFF değişkeni kullanıldığı modelden daha çok açıkladığı sonucunu ortaya çıkarmaktadır.

Bütün BNP modellerinde farklı ortak değişkenler kullanılsa dahi 6.kitapçığın 5.kitapçıktan daha kolay olduğu ve bu kitapçıkta bireylerin yüksek puan olma yoğunluğunun fazla olduğu sonucu ortaya çıkmaktadır. Aynı şekilde yine her model için 5.kitapçıktaki soruların daha ayırıcı olduğunu söylemek mümkündür.

Eşitlenmiş puanlara ait dağılımlar ve bu dağılımların hedef teste uzaklıkları incelendiğinde MTK yöntemleri ve BNP modelleri arasında karşılaştırma yapmak kolaydır. BNP modeli ile elde edilmiş olan eşitlenmiş puanlar için hesaplanan Hellinger Uzaklığı, MTK ölçek dönüştürme yöntemlerine göre oldukça düşük olup, bu dağılımlar hedef teste daha yakındır. Bu dağılımlardan en yakın uzaklığı iki ortak değişkenin kullanıldığı BNP modeli vermiştir. Dolayısı ile eşitlenmiş puan-hedef teste ait dağılımların birbirlerine MTK yöntemlerine kıyasla yakınlaştığı ve bu model kullanılarak daha kesin kestirimler elde edildiği sonucuna ulaşılmıştır. Bayes yönteminin klasik yöntemlerden daha iyi kestirim yaptığını ve daha yararlı bilgiler için de kullanılabileceğini ifade eden çalışmalar bu sonucu desteklemektedir (Karabatsos & Walker, 2009; Kruschke, Aguinis, & Joo, 2012; van de Schoot ve diğerleri, 2013).

BNP modeli ile grup değişmezliği kontrol altında tutulduğu gibi dağılımların düzensizliği ve süreksizliği de giderilmiş olduğundan; diğer eşitleme yöntemlerinde kullanılan ön-düzgünleştirme, bant genişliği parametresinin seçimi ve eşitlemenin standart hatasının türetilmesine ihtiyaç duyulmamaktadır (Gonzalez ve diğerleri, 2015b). Bu durum ise modelin önemliliğinin bir göstergesidir (Karabatsos & Walker, 2009).

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 212-225

Impact of Retrofitting and Item Ordering on DIF

Lokman AKBAY *

Abstract
Richer diagnostic information about examinees' cognitive strength and weaknesses are obtained from cognitively diagnostic assessments (CDA) when a proper cognitive diagnosis model (CDM) is used for response data analysis. To do so, researchers state that a preset cognitive model specifying the underlying hypotheses about response data structure is needed. However, many real data CDM applications are adds-on to simulation studies and retrofitted to data obtained from non-CDAs. Such a procedure is referred to as retrofitting, and fitting CDMs to traditional test data is not uncommon. To deal with a major validity concern of item/test bias in CDAs, some recent DIF detection techniques compatible with various CDMs have been proposed. This study employs several DIF detection techniques developed based on CTT, IRT, and CDM frameworks and compares the results to understand the extent to which DIF flagging behavior of items is affected by retrofitting. A secondary purpose of this study is to gather evidence about test booklet effects (i.e., item ordering) on items' psychometric properties through DIF analyses. Results indicated severe DIF flagging prevalence differences for items across DIF detection techniques employing Wald test, Raju's area measures, and Mantel-Haenzsel statistics. The largest numbers of DIF cases were observed when the data were retrofitted to a CDM. The results further revealed that an item might be flagged as DIF in one booklet, whereas it might not be flagged in another.

Key Words: Differential item functioning, DINA model, retrofitting, booklet affect, cognitive diagnosis models.

INTRODUCTION

In educational practice, many large-scale tests focus on summative assessment, and their formative features are limited. Tests developed to diagnose examinees' strengths and weaknesses may provide rich information toward formative assessment and are referred to as cognitively diagnostic assessments (de la Torre & Minchen, 2014). To obtain diagnostic information, examinee responses obtained from such assessment procedures may be analyzed via statistical models known as cognitive diagnosis models (CDMs). Such diagnostic information may be considered as valuable feedback for students, teachers, and educational programs. Generally, CDMs are used to estimate examinees attribute-profiles that are defined by the mastery or nonmastery status of measured attributes. Rather than being just a coarse indicator of how examinees think about and complete educational tasks, CDM enables practitioners to identify and report finer grained attributes examinees use to complete such tasks.

As the test development procedure and response data hold the characteristics of cognitively diagnostic assessment (CDA), then, a successful CDM application providing detailed information to facilitate the explanation of examinee performance might be possible. In other words, a cognitive model specifying a structure of the data by means of theories or hypotheses is needed and must be set a priori (Gierl & Cui, 2008; Rupp & Templin, 2008). However, as reported by Gierl, Alves, and Majeau (2010), many CDM applications are adds-on to simulation studies and retrofitted to previous test data. Cognitive diagnosis retrofitting refers to the application of CDM as a psychometric model to response data from traditional testing procedures (Gierl & Cui, 2008).

More often than not, we come across the studies retrofitting traditional test responses to CDMs to determine examinee attribute-profiles. Examples of real data retrofitting studies include Choi, Lee, & Park (2015) and Terzi & Sen (2019). For a recent comprehensive review of the CDM applications, including retrofitting studies, readers may refer to Sessoms and Henson (2018). In conducting large-

* Asst. Prof., Istanbul University-Cerrahpasa, Istanbul-Turkey, lokmanakbay@istanbul.edu.tr, ORCID ID : 0000-0003-4026-5241

To cite this article:
Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 212-225. doi: 10.21031/epod.886920

Received: 25.02.2021
Accepted: 28.03.2021

scale tests, it is aimed to reveal the cognitive ability levels of individuals in their study areas. One of the primary concerns in large-scale exams is the validity of assessment (Kane, 2013). The validity of a measurement tool is the degree to which it serves specified purposes and that it does not involve other features (Messick, 1995). Test bias is one of the severe factors threatening the validity of a test. Bias is observed when examinees' test scores in different subgroups contain group-dependent systematic errors (Camilli & Shepard, 1994). Differential item functioning (DIF) detection is a useful tool for identifying item bias. DIF is defined as the differentiation of the probability of answering an item correctly among individuals who are in different subgroups but have the same ability level (Zumbo, 2007). In other words, DIF arises when an item's response function differs from one group to another.

When an item is diagnosed by a specific DIF technique, content domain and measurement experts examine the items to understand whether the item offers a systematic advantage in favor of any subgroup. This systematic advantage is referred to as item bias, and DIF analysis is a crucial step in item bias examination. Various statistical DIF detection techniques based on classical test theory (CTT) and item response theory (IRT) are used to identify DIF items. These techniques include Mantel-Haenszel (Holland & Thayer, 1988), Logistic Regression (Swaminathan & Rogers, 1990), IRTLR tests (Thissen & Steinberg, 1988), Lord's χ2 test (1980), and the MIMIC model (Jöreskog & Goldberger, 1975; Woods, 2009). Recently, DIF detection techniques for cognitive diagnosis modeling framework have also been proposed (Hou, Terzi & de la Torre, 2020; Ma, Terzi & de la Torre, 2021). For example, Hou, de la Torre, and Nandakumar (2014) proposed a DIF detection method based on the Wald test that is compatible with the deterministic inputs, noisy "and" gate (DINA: Junker & Sijtsma, 2001) model. In this study, DIF detection techniques developed based on CTT, IRT, and CDM frameworks are employed. Namely, Mantel-Haenszel (Holland & Thayer, 1988), Raju's (signed) area measures (1988, 1990) and Wald test for DIF (Hou, de la Torre & Nandakumar, 2014) are employed.

In light of the above discussion, the primary purpose of this study is to examine the psychometric properties of a test through DIF analyses. Specifically, DIF flagging patterns of three DIF detection techniques, namely Mantel-Haenszel, Raju's area measures, and Wald test for DIF, are examined in terms of pattern consistency/similarity when the cognitive model specifying the data structure and psychometric model directing the psychometric analysis are different. In other words, DIF flagging patterns of the three DIF detection techniques were examined when response data are retrofitted. For this purpose, real data from a large-scale assessment are used. The data were collected using two booklets (i.e., Booklets A and B), and the subgroups of DIF analyses were based on variables gender and booklet type.

Another important issue on large-scale testing is the use of different booklets in test administration. Regarding the effect of using different types of booklets on the examinee achievement, testing agencies such as Measurement, Selection, and Placement Center (ÖSYM) argue that random assignment of test items to the booklets does not have any impact on examinees' achievement (2011). On the contrary, some experts claim that the positions of the items in the booklet could affect examinee performance by affecting anxiety and motivation levels, from which the estimates of test's psychometric properties may be affected (Middle East Technical University-METU, 2011; Ankara University, 2011). Although revealing the effect of the booklet on a single examinee is not feasible, the booklet effect on estimates of tests' psychometric properties can be statistically examined. Then, the secondary purpose of this study is to examine impact of the booklet on DIF analyses. Specifically, gender DIF flagging pattern of items across Booklets A and B is documented. Therefore, both the booklet effects and impact of retrofitting on real testing situations are examined, and the compatibility of Wald test based DIF detection under DINA model with more traditional DIF detection techniques is emphasized.

Purpose of the Study

Below research problems are addressed in this study:

- Do the DIF detection techniques developed based on CTT, IRT, and CDM frameworks yield compatible results (focusing on the cases where data are retrofitted)?

- Do the DIF flagging items differ across test booklets with different item ordering? In other words, do DIF analysis results get affected by the order of test items?

Dif Detection Techniques

Mantel-Haenszel technique for DIF detection

This CTT based DIF detection technique was proposed by Holland and Thayer (1988) using the statistic developed by Mantel and Haenzsel (1959). This technique is referred to as Mantel-Haenzsel DIF technique and examines whether item responses are independent of group membership after conditioning on the observed total score. The test statistic in this technique asymptotically follows a chi-square (χ^2) distribution with 1 degrees of freedom so that the statistic is compared against the chi-square distribution. To obtain the test statistic (χ^2_{MH}), for all total scores from 1 to $J-1$, N_m examinees are classified into 2×2 contingency tables, where J is the total number of items in the test and N_m is the number of examinees obtained a total score of m.

Table 1. A 2×2 Contingency Table Conditioned on the Total Score of *m*

Correct response to item *j*	Incorrect response to item *j*	Total response to item *j*
C_{Fm}	I_{Fm}	N_{Fm}
C_{Rm}	I_{Rm}	N_{Rm}
$N_{Cm} = C_{Fm} + C_{Rm}$	$N_{Im} = I_{Fm} + I_{Rm}$	$N_m = N_{Fm} + N_{Rm} = N_{Cm} + N_{Im}$

Note. C_{Fm} is the number of examinees who correctly responded to item *j* in the focal group; I_{Fm} is the number of examinees incorrectly responded to item *j* in the focal group; N_{Fm} is the total number of examinees with a total score of *m* in the focal group; C_{Rm} is the number of examinees correctly responded to item *j* in the reference group; I_{Rm} is the number of examinees incorrectly responded to item *j* in the reference group; N_{Rm} is the total number of examinees with a total score of *m* in the reference group; N_{Cm} is the total number of examinees with a total score of *m* who correctly responded to item *j*; N_{Im} is the total number of examinees with a total score of *m* who incorrectly responded to item *j*; and N_m is the total number of examinees with a total score of *m*.

Based on the information obtained from 2×2 contingency tables, the below formula is used to obtain test statistic:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_{m=1}^{J-1}[C_{Rm} - E(C_{Rm})]\right| - 0.5\right\}^2}{\sum_{m=1}^{J-1} Var(C_{Rm})}, \tag{1}$$

where

$$E(C_{Rm}) = \frac{N_{Rm}N_{Cm}}{N_m} \tag{2}$$

and

$$Var(C_{Rm}) = \frac{N_{Rm}N_{Fm}N_{Cm}N_{Im}}{N_m^2(N_m - 1)}. \tag{3}$$

Raju's (Signed) area measures for DIF detection

This DIF detection technique is based on item response curves (IRCs) defined by the item parameters obtained under one- two-, or three parameter logistic models. For a dichotomously scored item, unidimensional three-parameter logistic model is defined as

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\left[1 + \exp\{-1.7\alpha_j(\theta - \beta_j)\}\right]^{-1}, \tag{4}$$

where $P_j(\theta)$ is the probability of correctly answering item *j* when examinee's continuous ability level is θ; γ_j is the pseudo-guessing parameter of item *j*; α_j is the discrimination parameter of item *j*; θ is the continuous ability level; and β_j is the difficulty parameter of item j. Two- parameter logistic model can be derived from the above function by setting γ_j to zero. Similarly, one-parameter logistic model is derived by setting γ_j to zero and α_j to an estimated constant. This estimated discrimination parameter is fixed for all items in the test.

For one- two-, or three-parameter logistic models, Raju's (signed) area measure is the area between the IRCs defined by the estimated item parameters of focal and reference groups (Raju, 1988, 1990). As stated by Raju (1988, 1990) when the pseudo-guessing parameters of the IRF of subgroups for three-parameter logistic models are not equal, the area between the two item characteristic curves becomes infinite. Therefore, to avoid this problem, he suggests constraining the lower asymptote (i.e., pseudo-guessing parameter) to a fixed value. Based on this technique, DIF is examined by comparing the computed area between the item response curves to the determined critical values.

Given the item response functions of focal and reference groups,

$$F_F(\theta) = \gamma_{Fj} + (1 - \gamma_{Fj})[1 + \exp\{-1.7\alpha_{Fj}(\theta - \beta_{Fj})\}]^{-1} \tag{5}$$

and

$$F_R(\theta) = \gamma_{Rj} + (1 - \gamma_{Rj})[1 + \exp\{-1.7\alpha_{Rj}(\theta - \beta_{Rj})\}]^{-1}, \tag{6}$$

the area between the curves determined by the functions is calculated by taking the integral of the absolute differences

$$Area = \int_{-\infty}^{\infty} |(F_R - F_F)| d\theta. \tag{7}$$

Then, based on the null hypothesis that the true area is zero, a test statistic Z corresponding to the measured area is computed and compared against standard normal distribution. Readers may refer to Raju (1990) for details on the computation of the Z statistics.

Wald test for DIF detection under DINA model

One of the most parsimonious CDMs is the DINA model (Junker & Sijtsma, 2001), which is used to predict the probability of correctly answering an item as a function of individuals' discrete attributes' mastery status and item parameters (Li, 2008). Based on the DINA model, examinees' attribute profiles indicating mastered and nonmastered attributes are estimated. Regardless of the number of attributes measured by the test and the number of attributes required by an individual item, for DINA model, two item parameters are estimated. These parameters are referred to as guessing and slip parameters (de la Torre, 2009). Guessing parameter of item j (g_j) is the probability of successful response of an examinee who has not mastered at least one of the attributes that are required to correctly answer item j. Likewise, the slip parameter of item j (s_j) is the probability of incorrectly responding to item j when an examinee has already mastered all required attributes required by the item (de Carlo, 2012; de la Torre, 2009). These two parameters are mathematically defined as

$$g_j = p\left[X_{ij} = 1 | \eta_{ij} = 0\right] \tag{8}$$

and

$$s_j = p\left[X_{ij} = 0 | \eta_{ij} = 1\right], \tag{9}$$

where g_j is guessing parameter of item j; s_j: slip parameter of item j; η_{ij} is ideal response (i.e., when $s_j = g_j = 0$) of examinee i to item j.

Given the item parameters, the DINA model item response function (i.e., probability of correctly responding to given item) is defined as

$$P(X_{ij} = 1 | \boldsymbol{\alpha}_l) = g_j^{(1-\eta_{jl})} (1 - s_j)^{\eta_{jl}}, \tag{10}$$

where X_{ij} is the observed response of examinee i to item j; $\boldsymbol{\alpha}_l$ is attribute vector l among 2^K attribute vectors formed by K measured attributes; η_{il} is the ideal response of an examinee when his/her attribute vector is $\boldsymbol{\alpha}_l$.

First of all, in CDM context, DIF refers to the difference in the success probability of reference and focal groups with the same attribute mastery patterns (Hou et al., 2014). Under the DINA model, DIF is observed for item j when $\Delta g_j = g_{Fj} - g_{Rj} \neq 0$ and/or $\Delta s_j = s_{Fj} - s_{Rj} \neq 0$, where F and R stand for focal and reference groups, respectively. When Δg_j and Δs_j have the same sign, the DIF referred to as uniform; otherwise, it is called non-uniform DIF. Wald test DIF for the DINA model tests the significance of the joint differences between the item parameters of the subgroups:

$$W_d = (C\hat{v}_j)'\left(C\hat{\Sigma}_j C'\right)^{-1}(C\hat{v}_j), \tag{11}$$

where $\hat{v}_j$ is an item parameter column vector of $(g_{Fj}, s_{Fj}, g_{Rj}, s_{Rj})^T$; $\hat{\Sigma}_j$ is asymptotic variance-covariance matrix associated with the subgroups' item parameter estimates; and C is the contrast matrix of $\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$. In this test, W_d asymptotically follow a chi-square (χ^2) distribution with 2 degrees of freedom, and the tested null hypothesis is $C\hat{v}_j = 0$.

METHOD

Sample

The data used in this study were obtained from a 19-item mathematic section of the high school admission exam (TEOG). More specifically, the data are the responses of high school applicants who took the test in 2013 in Ankara, Turkey. It should be noted here that rather than answering any specific research questions raised about this specific exam, this study employed this data set to mimic real life conditions where the data analysis may or may not flag DIF items. In other words, this dataset is used in this simulation-like study rather than using simulated data that may not truly reflect real life conditions. For the current study, 100 datasets were randomly drawn from the entire data, which consist of 39,146 male and 37,318 female examinees' responses to 19 multiple-choice mathematics items. The sample size for each data was fixed to 1,000 in order to obtain stable item parameter estimates under the DINA and IRT models for both focal and reference groups. This sample size is sufficient for unbiased and accurate estimation of the DINA model parameters (see De la Torre, Hong, & Deng, 2010) as well as unidimensional three-parameter logistic (3PL) model parameters (de Ayala, 2009, p. 130). In the study, Ox-Edit program (Doornik, 2003) was used for random sample drawings, and DIF analyses were conducted via R-programming (R Core Team, 2016).

Table 2: Descriptive Statistics by the Booklet Type

	Booklet A			Booklet B		
	Male	*Female*	*Total*	*Male*	*Female*	*Total*
Number of examinees	20,076	18,869	38,945	19,070	18,549	37,619
Number of items	19	19	19	19	19	19
Mean	8.49	9.499	8.979	8.801	9.776	9.28
Variance	26.099	25.471	26.048	24.854	23.742	24.558
Standard deviation	5.108	5.047	5.104	4.988	4.873	4.955
Skewness	-0.694	-0.908	-0.894	-0.755	-0.97	-0.893
Kurtosis	0.552	0.288	0.417	0.599	0.35	0.447

As stated above, because this study has no specific interest in examining either test items in detail or examinee achievement, descriptive statistics are not thoroughly discussed. Rather, descriptive statistics for each gender group for the A and B test booklets are summarized in Table 2.

Dimensionality

To be able to apply Raju's area statistic based on the unidimensional IRT model, the data need to be unidimensional. So, dimensionality was checked through exploratory factor analysis conducted via SPSS, and the results confirmed the unidimensionality. The results of this analysis are presented in Table 3.

Table 3. Findings of Exploratory Factor Analysis

	1st Dimension	2nd Dimension	3rd Dimension	4th Dimension	5th Dimension
Explained variance	.33	.06	.05	.04	.04
Cumulative explained variance	.33	.39	.44	.48	.52

Model Selection

To be able to retrofit the data to a CDM, an item-attribute specification matrix, namely, Q-matrix was developed after establishing the attributes measured by the test. The attributes were set, and the Q-matrix was constructed by mathematics education experts. The model fits statistics indicated an acceptable fit of the data to the DINA model so that Wald test based DIF detection under the DINA model was conducted. In terms of unidimensional models, data were fitted to the Rasch, 1PL, 2PL, and 3PL IRT models for model selection. It should be recalled that the only difference between the Rasch model and 1PL model is the common item discrimination index. In particular, item discrimination is fixed to 1.00 for all items under the Rasch model. On the contrary, under the 1PL model, a common discrimination parameter is estimated from the data and fixed across all items in the test. Model selection yielded that the 3PL model best fitted to the data, and the model selection results were presented in the results section.

Analysis

In order to facilitate the analyses and interpretation of the analyses, the order of the items in different booklets was rearranged before conducting the analyses for which booklet A was taken as reference. Each of the 100 datasets was obtained from the entire examinee response data, and these data sets were analyzed through the Wald test, Raju's area measures, and Mantel-Haenzsel DIF detection techniques for gender groups. To understand the impact of booklet type on estimated item parameters (i.e., the impact of item ordering on psychometric properties of a test), DIF analyses were conducted on booklet A and B separately, and the results were compared. To perform the analyses, Ox-Edit program for the Wald test cases and the difR package (version 4.6) developed by Magis, Beland, and Raiche (2015) were used for Raju's area measures and Mantel-Haenzsel DIF detection cases. Comparing the obtained test statistics to corresponding relevant statistical distributions, *p*-values were computed and reported to compare and contrast DIF detection results of different techniques and their variation by test booklets. Therefore, by comparing and contrasting the obtained p-values to the significance levels of α = .01 and α = .05, DIF flagging rates across two booklets and different DIF detection techniques were examined.

RESULTS

To determine which IRT model to employ for the Raju's area measure DIF technique, a model selection analysis was conducted to select one from one-, two-, and three-parameter logistic models. Because all four models are nested, a deviance test (i.e., likelihood ratio test) test is also conducted along with consideration of Akaike's information criterion (AIC) and Bayesian information criterion (BIC) for model selection. The test statistics and the test results are given in Table 4, which indicate that 3PL model is the best fitting model among all four. As discussed by Raju (1988, 1990), area measures for DIF detection are computed after fixing the lower asymptote. For this study, because all items in the test were multiple-choice with four options, theoretically constraining the pseudo-guessing parameter to 0.25 was meaningful. Accordingly, for the purpose of employing Raju's area measures DIF detection technique, 3PL model pseudo-guessing parameters were set to 0.25 across all items.

Table 4. Data-Model Fit Statistics

Model	AIC	BIC	Loglikelihood	-2*x*Loglikelihood	*df*
Rasch	820747.5	820910.3	-410354.7	-----------	---
1PL	811908.6	812080.0	-405934.3	8840.85*	1
2PL	796224.6	796550.3	-398074.3	15719.99*	18
3PL	788745.1	789233.6	-394315.5	7517.56*	19

Note: * p<.001, AIC is Akaike information criterion; BIC is information criterion; and *df* stands for degrees of freedom.

One of the main aims of this study was to examine the variation in DIF-flagging prevalence of the test items when analyzed under different psychometric models. This study especially focused on the variation in DIF analysis results when the data were retrofitted to a CDM such as DINA model. Thus, DIF flagging rates of three DIF techniques employed for CTT, IRT, and CDM-based psychometric models were examined, and the results at α = .05 and α = .01 levels were summarized in Table 5 and 6, respectively. For example, at α-level of .05, item-1 was flagged as DIF-item by Raju's area measures 22 out of 100 times in booklet A and 32 out of 100 times in booklet B conditions. Likewise, the number of times this item was flagged as DIF-item at α-level of .01 were 5 and 14 under booklet A and B, respectively.

Table 5. Null Hypotheses Rejection Rates of the DIF Detection Techniques at α = .05

	Psychometric models used as a basis for DIF analyses					
	Wald test for DINA		*Raju's area for 3PL*		*Mantel-Haenzsel for CTT*	
Items	Booklet A	Booklet B	Booklet A	Booklet B	Booklet A	Booklet B
Item 1	.51	.79	.22	.32	.12	.40
Item 2	.16	.25	.23	.32	.05	.05
Item 3	.20	.23	.55	.64	.63	.69
Item 4	.39	.39	.41	.27	.48	.47
Item 5	.10	.15	.02	.06	.16	.30
Item 6	.91	.79	.26	.29	.42	.35
Item 7	.65	.62	.17	.20	.06	.02
Item 8	.49	.27	.00	.00	.05	.01
Item 9	.31	.56	.07	.04	.33	.44
Item 10	.62	.38	.11	.14	.17	.13
Item 11	.38	.12	.24	.21	.05	.05
Item 12	.53	.66	.45	.25	.86	.82
Item 13	.34	.55	.17	.15	.07	.25
Item 14	.92	.92	.09	.11	.66	.61
Item 15	.19	.17	.12	.37	.06	.07
Item 16	.48	.35	.46	.39	.06	.04
Item 17	.96	.81	.49	.64	.70	.44
Item 18	.66	.76	.37	.43	.53	.65
Item 19	.07	.09	.68	.77	.26	.24

The rejection rates of the null hypotheses given in Tables 5 and 6 were obtained by comparing the observed p-values of the analyses to the critical values of .05 and .01, respectively. Thus, it is not clear whether the null hypotheses were rejected with a p-value of .051 or .999. Therefore, in addition to the null hypotheses rejection rates presented in the abovementioned tables, boxplots were also created based on the p-values obtained from analyses of 100 data sets for each of the booklets. These boxplots are presented in Figure 1, in which horizontal lines indicate the null hypothesis rejection levels of .01 and .05.

By looking at the tables, severe differences in the prevalence of DIF flagging for an item can be observed across different DIF techniques. First of all, numbers of DIF cases are the largest for Wald test DIF detection under the DINA model with grand mean ratios of .47 and .31 when $\alpha = .05$ and $\alpha = .01$, respectively. Although they are not quite different from the Mantel-Haenzsel results, the smallest grand means for DIF flagging rates (mean rates of .28 and .11 when $\alpha = .05$ and $\alpha = .01$, respectively) are observed for Raju's area measures under 3PL model. Lastly, the Mantel-Haenzsel DIF technique yielded a grand mean null hypotheses rejection rates of .31 and .16 under $\alpha = .05$ and $\alpha = .01$, respectively.

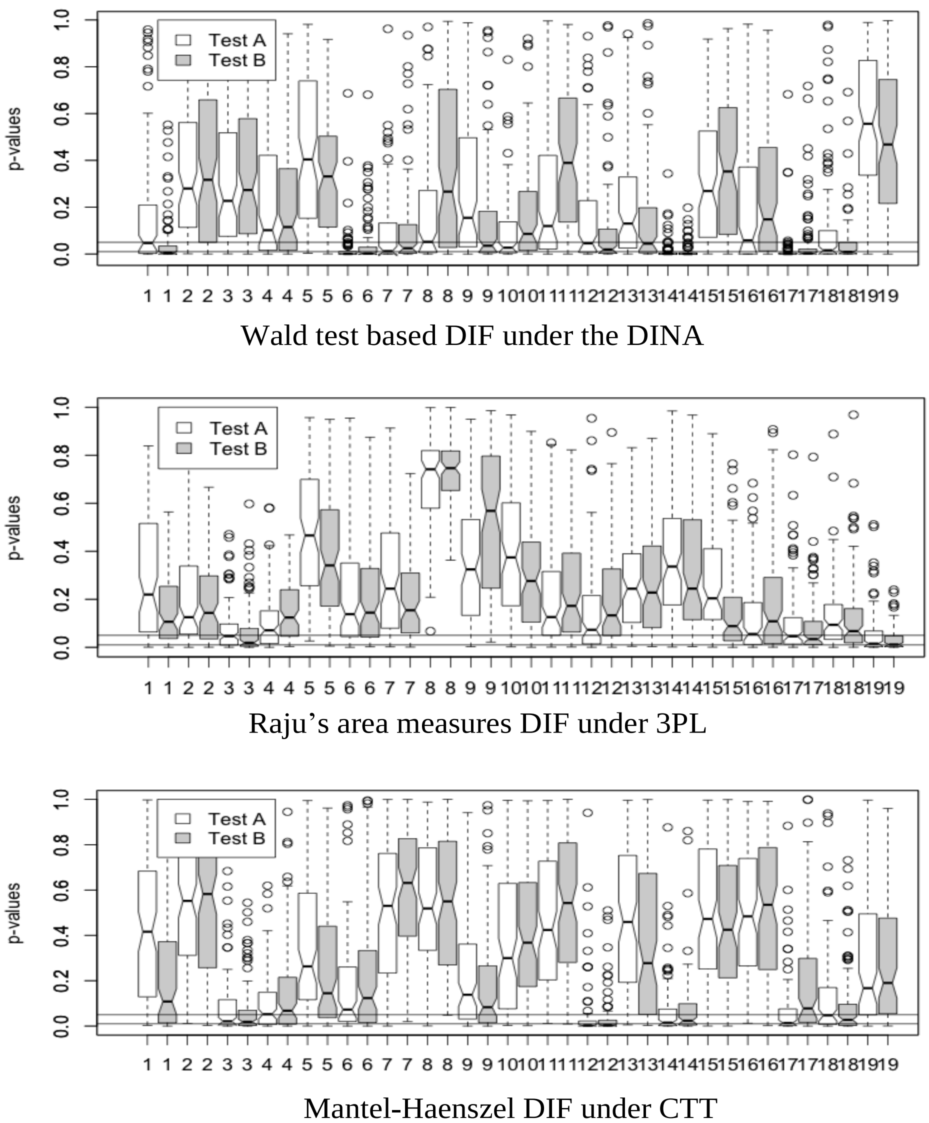
In terms of pairwise comparisons of DIF techniques, the largest differences in the DIF flagging ratios were observed between the Wald test and Raju's area measures. Relatively large differences in the prevalence of DIF flagging are observed for 13 out of 19 items (items 1, 3, 6, 7, 8, 9, 10, 12, 13, 14, 17, 18, and 19). For this comparison, the largest difference was observed for items 14A and 14B with differences of $.92 - .09 = .83$ and $.81 - .02 = .79$ for $\alpha = .05$ and $\alpha = .01$ cases, respectively. Further, in comparison of the DIF flagging ratios for the Wald test and Mantel-Haenzsel techniques, large differences were observed for 11 items (items 1, 3, 6, 7, 8, 10, 12, 13, 14, 16, and 17). In this comparison, the largest differences in ratios were observed for items 7B and 6A with differences of $.62 - .02 = .60$ and $.74 - .17 = .57$ when $\alpha = .05$ and $\alpha = .01$, respectively. When comparing the rejection rates of Raju's area measures and Mantel-Haenzsel techniques, the gaps between the ratios were relatively smaller. Nevertheless, five items (items 9, 12, 14, 16, and 19) were reported to have large differences in terms of the ratio of being flagged as DIF items. In this case, the largest ratio differences were reported for item 12B with a difference of $.82 - .25 = .57$ and $.63 - .09 = .54$ for $\alpha = .05$ and $\alpha = .01$ conditions, respectively.

Table 6. Null Hypotheses Rejection Rates of the DIF Detection Techniques at α = .01

	Psychometric models used as a basis for DIF analyses					
	Wald test for DINA		*Raju's area for 3PL*		*Mantel-Haenzsel for CTT*	
Items	Booklet A	Booklet B	Booklet A	Booklet B	Booklet A	Booklet B
Item 1	.32	.62	.05	.14	.03	.19
Item 2	.01	.09	.08	.07	.00	.01
Item 3	.05	.09	.26	.36	.37	.42
Item 4	.21	.19	.18	.05	.32	.28
Item 5	.01	.05	.00	.01	.06	.14
Item 6	.74	.68	.09	.12	.17	.19
Item 7	.48	.33	.06	.05	.01	.00
Item 8	.31	.17	.00	.00	.01	.00
Item 9	.16	.34	.01	.00	.10	.21
Item 10	.33	.20	.02	.03	.06	.04
Item 11	.19	.07	.07	.06	.00	.01
Item 12	.31	.40	.19	.09	.58	.63
Item 13	.16	.37	.03	.03	.01	.07
Item 14	.78	.81	.04	.02	.41	.34
Item 15	.11	.08	.07	.16	.02	.03
Item 16	.44	.23	.24	.21	.00	.01
Item 17	.79	.66	.20	.21	.47	.15
Item 18	.46	.54	.14	.15	.31	.37
Item 19	.04	.04	.39	.48	.08	.09

The secondary purpose of this study was to investigate the booklet effect, if any, on estimated item parameters via DIF detection techniques. Because the DIF is examined through variations of items' psychometric properties, variation in observed DIF results across test booklets may be considered as empirical evidence to argue that item order in a test affects items' estimated parameters. When the Wald test DIF results for the DINA cases were examined, clear variations in DIF flagging rates of this technique for two test booklet conditions were observed. Specifically, when $\alpha = .05$ was considered, DIF flagging rates of seven items (items 1, 8, 9, 10, 11, 13, and 16) were substantially different. Even though the significance level was reduced to α = .01, five out of these seven items (items 1, 8, 9, 13, and 16) were flagged as DIF-items with notably different flagging rates. Similarly, Raju's area measures DIF flagging rates of four items (4, 12, 15, and 17) were relatively different across two test booklet conditions. Even under a more conservative α-level (i.e., α = .01), items four and 12 were still slightly diversified. Lastly, when detecting DIF items via the Mantel-Haenzsel technique, the difference in DIF flagging rates of four items (items 1, 5, 13, and 17) came to the forefront. Among these four, items 1 and 17 remained diversified in terms of being flagged as DIF items under the α-level of .01.

Furthermore, Figure 1 was also used to explore the relationships between the booklets with respect to DIF flagging behavior. Boxplots in this Figure were plotted with notches, where lack of overlap between the notches of the boxplots for booklets A and B indicates that the median scores specified in these box plots are different (Chambers, Cleveland Kleiner, & Tukey, 1983). These plots in Figure 1 yielded compatible results from those presented in Tables 5 and 6. Specifically, the notches of the boxplots for booklets A and B did not have any overlap for items 1, 8, 9, 10, 11, and 13 when the DIF detection technique was the Wald test DIF for DINA. Similarly, when Raju's area measure and Mantel-Haenzsel DIF detection techniques were employed, boxplot notches did not overlap for items 1, 4, 9, and 15; and items 1, 5, 13, and 17, respectively. Based on the above results, it is evident that booklet type yielded different outcomes from DIF analyses.



Wald test based DIF under the DINA

Raju's area measures DIF under 3PL

Mantel-Haenszel DIF under CTT

Figure 1. Boxplots of the p-values computed for DIF hypothesis testing.

DISCUSSION and CONCLUSION

In practice, many large-scale tests focus on summative assessments providing coarse test scores that provide limited formative information. Analyzing the data collected from cognitively diagnostic assessments (CDA) by CDMs may offer richer diagnostic information about examinees' cognitive strengths and weaknesses. Specifically, CDM enables practitioners to identify and report finer grained attributes examinees use to complete cognitive tasks. However, Gierl and Cui (2008) and Rupp and Templin (2008) state that a cognitive model specifying theories or hypotheses related to the structure of the data must be set. Yet, many real data CDM applications are adds-on to simulation studies and *retrofitted* to data already collected (Gierl, Alves, & Majeau, 2010; Terzi & Sen, 2019). Therefore, more often than not, practitioners fit CDMs to traditional test responses.

A major validity concern arises in large-scale assessments when item/test bias occurs, and DIF detection is a useful method for dealing with this validity thread. Various statistical techniques based on CTT and IRT are used to identify DIF-items. Up to date, DIF detection techniques that are compatible with CDMs, such as Wald test DINA DIF detection technique (Hou, de la Torre, & Nandakumar, 2014; Hou, Terzi, & de la Torre, 2020), have been proposed. In this study, DIF detection techniques developed based on CTT, IRT, and CDM frameworks are employed, and the results are compared to derive conclusions about the compatibility of the results. It is particularly important to understand how tests' psychometric properties are affected in retrofitting. Therefore, this study aimed to examine the psychometric properties of a test through DIF analyses. For this purpose, real data from a large-scale assessment were used. Because the dataset was collected via two test booklets with different item ordering, this study also examined the booklet impact on estimated item parameters through DIF analyses across gender groups were conducted on booklet A and B.

Results indicated severe DIF flagging prevalence differences for items across different DIF techniques. The largest numbers of DIF cases were observed under the DINA retrofitting, whereas comparably less frequent DIF cases observed when Raju's area measures under 3PL model and Mantel-Haenzsel DIF detection technique based on CTT were employed. One of the presumptive reasons for this result is that the original exam was not developed for CDA purposes. Specification of attributes to be measured by the test, development of items assessing the attribute set, and construction of the Q-matrix to establish a precise relationship between items and attributes are the key points for obtaining accurate information from a test in the CDA framework. Thus, the alignment of items and attributes in a test is a crucial step for enhancing the benefit of diagnostic assessment. In many cases, not specific for the test and data used in this study, psychometric properties of a test may not be accurately determined when data are collected via an achievement test that was not developed based on CDA.

Further results were obtained with respect to the booklet effect on items' psychometric properties through DIF detection techniques. When the Wald test DIF results for the DINA were examined, clear variations in DIF flagging rates of this technique for the two test booklet conditions were observed. Although the alterations of DIF analysis results across two booklets were not as high, DIF flagging rates of Raju's area measures and Mantel-Haenzsel techniques resulted in a similar pattern. Thus, it may be concluded that different booklets have an impact on the estimated psychometric properties of items such that these differences produce variant DIF patterns on a test. In the literature, there are studies suggesting that changes in item positions change the difficulty level of the items (Kingston & Dorans, 1984). In addition, it is also known that the speed responding to an item, fatigue, and exam experience can also lead to DIF. Thus, variations in items response speed, strategies used for response generation, cognitive effort exertion rate, and fatigue across subgroups may yield variation in estimated item parameters as item order changes in a test. Therefore, as the differences in the estimated item parameters for the subgroups increases due to the sequence of items in a test, items may be flagged by DIF detection techniques. Therefore, even if item ordering changes across booklets, these changes in item locations should not be dramatic to minimize item order effect on DIF and eventually on test scores.

REFERENCES

Ankara University (2011). *Ankara Üniversitesi Eğitim Bilimleri Fakültesi'nin YGS Hakkında Görüşü.* Retrieved November 30, 2015, form https://dahilibellek.wordpress.com/2011/04/12/ankara-ebf-ygs/

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Chambers, J.M., Cleveland, W.S., Tukey, P.A., Kleiner, B. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group, the University of Michigan.

Choi, K. M., Lee, Y.-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science, & Technology Education, 11,* 1563–1577.

de Ayala, *R. J. (2009). The theory and practice of item response theory.* The Guilford Press, New York, NY.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 47,* 115-127

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement 47,* 227–249.

de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis framework. *Psicologia Educative 20,* 89-97

De Carlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36,* 447-468.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 10,* 318-341. doi:10.1080/15305058.2010.509554

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement, 6,* 263-275.

Holland, P. W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hou, L., Terzi R., & de la Torre, J. (2020). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education, 7(2),* 145-158.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631-639.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8(2),* 147-154.

Li, F. (2008) *Modified higher-order DINA model for de11tecting differential item functioning and differential attribute functioning.* Unpublished doctoral dissertation University of Georgia, USA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting Differential Item Functioning Using Multiple-Group Cognitive Diagnosis Models. *Applied Psychological Measurement, 45(1),* 37-53.

Magis, D., Beland, S., & Raiche, G., (2015). *difR: Collection of methods to detect dichotomous differential item functioning (DIF).* R package version 4.6.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Middle East Technical University. *(2011). 2011 Yılı Yükseköğretime Geçiş Sınavı Hakkında ODTÜ Eğitim Fakültesi Görüşü.* Retrieved November 30, 2015, form *http://fedu.metu.edu.tr/sites/fedu.metu.edu.tr/files/ygs2011hkegitimfakultesigorusu_28_4_2011_v2.pdf*

Measurement, Selection, and Placement Center. *(2011). Adaya özgü soru kitapçığı.* Retrieved December 29, 2015, file://localhost/from http/::www.osym.gov.tr:belge:1-12431:adaya-ozgu-soru-kitapcigi-21032011.html

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. *URL https://www.R-project.org/.*

Raju, N. S. (1988). The area between two item characterıstıc curves. *Psychometrika, 3,* 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-207.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6(4),* 219-262.

Sessoms, J. & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A literatüre review and critical commentary. *Measurement: Interdisciplinary research and persperctives, 1,* 1-17.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 361-370.

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *SAGE Open, 9(1),* 1-11.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99(1), 118.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32(7),* 511-526.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223-233.

Veriye Sonradan Model Eklemenin ve Madde Sıralamasının DMF Üzerindeki Etkileri

Giriş

Çoğu geniş ölçekli testler özetleyici değerlendirmeye yönelik olup genel ve özet puanlarla ölçülen özelliğin testi alanlardaki seviyesini ortaya koymakta ve biçimlendirici değerlendirme çerçevesinde oldukça sınırlı bilgi sağlayabilmektedir. Bilişsel tanılama yapabilmek adına geliştirilen testlerin sonuçları bilişsel tanı modelleri (BTM) aracılığıyla analiz edildiğinde ise testi alanların bilişsel niteliklere sahip olma ya da olmama durumları ile ilgili zengin tanısal geri dönütler elde edilebilir. BTM ile yapılan analizler, testi alanların test içerisinde sunulan bilişsel görevleri tamamlamak için kullandıkları küçük boyutlu ve ayrıntılı bilişsel niteliklerin tanımlamasını ve testi alanlarda bulunup bulunmama durumlarının belirlenmesini sağlar. Gierl ve Cui (2008) ile Rupp ve Templin (2008) tarafından belirtildiği üzere, BTM odaklı bir test oluşturmak için, test maddelerine verilen cevapların nasıl oluştuğunu ve elde edilen verinin yapısını açıklayan kuram veya hipotezleri barındıran bilişsel bir model temel alınmalıdır. Ancak, literatüre bakıldığında, birçok gerçek veri kullanımına bağlı BTM uygulamasının simülasyon çalışmalarına ek olarak ortaya koyulduğu ve halihazırda toplanan verilere sonradan model ekleme (retrofitting) faaliyetlerinin ağırlıkta olduğu görülmektedir (Gierl, Alves ve Majeau, 2010).

Ölçme-değerlendirme süreçlerinde madde/test yanlılığı önemli bir geçerlilik sorunu olarak karşımıza çıkmaktadır (Kane, 2013). Bu sorunla başa çıkmak adına değişen madde fonksiyonu (DMF) tespiti yararlı bir yöntem olarak değerlendirilmektedir. DMF-maddelerini belirlemek için klasik test kuramını (KTK) ve madde tepki kuramını (MTK) temele alan DMF belirleme teknikleri ortaya koyulmuştur. Son zamanlarda, BTM çerçevesinde DMF belirleme teknikleri de literatüre kazandırılmaktadır. Yaygın kullanımı olan BTM'lerden DINA (the deterministic input, noisy "and" gate: Junker & Sijtsma, 2001) modelin veri analizinde kullanıldığı durumlar için Wald testine bağlı olarak DMF belirleme tekniği geliştirilmiştir (Hou, de la Torre ve Nandakumar, 2014). Bu çalışmada, KTK, MTK ve BTM tabanında geliştirilmiş DMF belirleme teknikleri kullanılmış ve sonuçların uyumluluğu değerlendirilmiştir. Özellikle, BTM çerçevesinde geliştirilmemiş olan testlerden elde edilen verilerin sonradan eklenen bir BTM ile analizi sonucunda maddelerin DMF gösterme durumları incelenmiştir. Bu analizlerle testin geliştirilmesinde dikkate alınan ve test sonuçlarının analizinde kullanılan psikometrik modellerin aynı olmadığı durumlarda cinsiyet gibi bağımsız değişkenlerce oluşturulacak alt gruplar için maddelerde DMF görülme durumunun farklılaşıp farklılaşmadığının incelenmesi hedeflenmektedir. Bu çalışmanın ikincil amacı kitapçık türünün psikometrik özellikleri (örneğin madde parametreleri) üzerindeki etkisinin DMF belirleme teknikleri aracılığıyla incelenmesidir. DMF maddelerin psikometrik özelliklerinin alt gruplara göre farklılık göstermesi neticesinde oluştuğundan, test kitapçıklarında (maddelerin sıralaması değiştiğinde) gözlemlenen DMF analiz sonuçlarındaki varyasyon testteki maddelerin sıralamasının kestirilen parametreleri etkilediğine yönelik ampirik kanıt olarak sunulacaktır.

Yöntem

Yukarıda belirtilen hedefler çerçevesinde, bu çalışmada 2013 yılında Ankara ilinde TEOG sınavına girmiş olan 39146 erkek ve 37318 kadın adayın 19 çoktan seçmeli matematik maddesine verdiği cevaplardan seçkisiz örnekleme yöntemi ile oluşturulan örneklemler kullanılmıştır. Verilerin elde edilmesinde kullanılan sınav A ve B kitapçığı olmak üzere sınava giren adaylara sunulmuştur. Bu kitapçıklarda maddelerin sıralaması (konumları) farklılık göstermektedir. Bu testten elde edilen toplam veri setinden, 1000 öğrencinin verisini içeren seçkisiz örnekleme ile 100 tane örneklem oluşturulmuştur. Bu örneklemler, cinsiyete göre yukarıda bahsi geçen üç farklı DMF belirleme tekniği ile analiz edilmiş ve elde edilen istatistikler ilgili istatistiksel dağılımlarla karşılaştırılarak 'kadın ve erkek öğrenciler için maddenin fonksiyonu değişmemektedir' şeklinde ifade edilebilecek yokluk hipotezleri test edilmiştir. Test sonuçları, her bir teknik ve test kitapçığı türü için hipotezin reddedilme oranı olarak rapor edilerek ve ayrıca elde edilen p-değerleri kutu-grafiği olarak karşılaştırılmıştır.

Sonuç ve Tartışma

Yokluk hipotezleri reddedilme oranlarına bakıldığında, farklı DMF tekniklerinde maddelere DMF tanısı konulma oranlarında ciddi farklılıklar gözlemlenmektedir. Öncelikle belirtilmelidir ki Wald teste bağlı olarak DINA model ile veriler analiz edildiğinde ortalama DMF gözlemlenme oranları, sırasıyla α = .05 ve α = .01 anlamlılık düzeylerinde, .47 ve .31 olarak ortaya hesaplanmıştır. Bu haliyle DINA modeli üzerinden Wald test DMF belirleme tekniği en yüksek DMF sonuçlarını doğurmuştur. Mantel-Haenzsel sonuçlarından çok da farklı olmada dahi, Raju'nun alan ölçüleri tekniğiyle DMF analizi yapıldığında elde edilen maddelerde DMF görülme oranlarının ortalaması en düşük seviyede seyretmiştir (α = .05 ve α = .01 olduğunda sırasıyla .28 ve .11). Son olarak, Mantel-Haenzsel DMF belirleme tekniği, α = .05 ve α = .01 altında, sırasıyla, maddelerde .31 ve .16 oranlarında DMF rapor etmiştir. Böylesine bir sonucun olası nedenlerinden biri, orijinal sınavın BTM'ye bağlı olarak geliştirilmemiş olması olarak düşünülebilir. Test tarafından ölçülecek niteliklerin belirlenmesi, nitelik setini ölçen maddelerin geliştirilmesi ve maddeler ile nitelikler arasında doğru bir ilişkinin kurulması için Q-matrisinin oluşturulması, BTM çerçevesinde hazırlanan testten maksimum düzeyde bilgi elde etmek için kilit adımlardır. Bu nedenle, bir testte yer alan maddelerin ve niteliklerin doğru şekilde ilişkilendirilmesi, bilişsel tanıya yönelik değerlendirmenin etkililiğini artırmak için çok önemli bir adım olacaktır. Bu çalışmada kullanılan test ve verilere özgü olmaksızın, genel olarak, bilişsel tanı modellemesi çerçevesinde hazırlanmamış testlerden elde edilen veriler üzerinde sonradan eklenen bir BTM ile analizine yönelik atılacak adımlarda, testin ve test maddelerinin psikometrik özellikleri (örneğin madde parametreleri) hatalı kestirilebilecektir.

DINA model ile yapılan analizler için Wald testine bağlı olarak DMF sonuçları incelendiğinde, kitapçıklar arasında bu tekniğin DMF belirleme oranlarında açık farklılıklar gözlenmiştir. Detaylandırılacak olursa, α = .05 düzeyinde, yedi maddenin DMF gösterme eğilimleri büyük ölçüde farklılaşmıştır. Anlamlılık seviyesi α = .01'e düşürülmüş olsa bile bu yedi maddeden beşi hala belirgin şekilde DMF gösterme eğilimlerinde farklılıklar sergilemişlerdir. Benzer şekilde Raju'nun alan ölçüleri ve Mantel-Haenzsel DMF teknikleri ele alındığında ise dörder maddede DMF gösterme eğiliminde kitapçıklar arasında yüksek farklılıklar ortaya çıkmıştır. Yokluk hipotezlerinin reddedilme oranlarından yola çıkarak yaptığımız değerlendirmede sunulan oranlar analizlerde raporlanan gözlenen p-değerleri sırasıyla .05 ve .01 kritik değerleriyle karşılaştırılarak elde edilmiştir. O halde, yokluk hipotezlerin .051 mi yoksa .999 gibi bir p-değeriyle mi reddedildiği bilinememektedir. Bu nedenle, yokluk hipotezi reddetme oranlarına ek olarak, her bir kitapçık için ele alınan 100 veri setinin analizlerinden elde edilen p-değerleri kutu-grafikleri olarak sunulmuştur ve bu grafikler DMF teknikleri ve kitapçık türleri arasında maddelerde DMF gözlemlenme eğilimlerinin kıyaslanmasında kullanılmıştır.

Kitapçık türlerinden alınan örneklemler üzerinde her üç DMF tekniğiyle cinsiyet grupları açısından maddelerin DMF gösterime eğilimlerinin kutu grafikleriyle incelenmesi sonucunda yukarıda açıklanan bulgularla benzer sonuçlar elde edilmiştir. Dolayısıyla, farklı kitapçıkların maddelerin psikometrik özelliklerinin kestirimi üzerinde bir etkiye sahip olduğu, bir diğer ifadeyle, maddelerin test içerisindeki sıralamalarının maddelerin kestirilen parametrelerine etki ettiğine yönelik ampirik bulgulara ulaşılmıştır. Maddelerin sıralamalarındaki değişikliler farklı alt gruplar için farklı sonuçlar doğurmuş ve dolayısıyla alt gruplar arasında (bu çalışmada cinsiyet grupları arasında) maddenin kestirilen parametrelerinde farklılıklar ortaya çıkmıştır. Alanyazın incelendiğinde, madde konumlarındaki değişikliklerin maddelerin zorluk seviyelerini değiştirdiğini öne süren çalışmalar bulunmaktadır (Kingston ve Dorans, 1984). Bu nedenle, bir testte madde sırası değiştikçe, madde yanıtlama hızında, yanıt oluşturma stratejilerinde, bilişsel çaba harcama oranında ve alt gruplardaki yorgunluk seviyesinde meydana gelebilecek farklılıklar, madde parametrelerinin kestirilen değerlerinde değişikliğe ve dolayısıyla alt gruplar açısından DMF'ye sebebiyet verebilmektedir. Bu bulgular çerçevesinde, maddelerin konumları kitapçıklar arasında değişiklik gösterse dahi, bu konum değişikliklerin DMF'ye ve sonunda test puanları üzerinde ciddi farklılıklara sebep olacak kadar büyük olmaması önem taşımaktadır.