



**Volume 6**

**Issue 2**

**2019**

International Journal of  
Assessment Tools in Education

**International Journal of  
Assessment Tools in Education**

International Journal of  
Assessment Tools in Education

<http://ijate.net/>

e-ISSN: 2148-7456



e-ISSN 2148-7456

<http://www.ijate.net/index.php/ijate/index>

**Volume 6**

**Issue 2**

**2019**

**Dr. İzzet KARA**

Editor in Chief

International Journal of Assessment Tools in Education

Pamukkale University,

Education Faculty,

Department of Mathematic and Science Education,

20070, Denizli, Turkey

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : [ijate.editor@gmail.com](mailto:ijate.editor@gmail.com)

Publisher : İzzet KARA

Frequency : 4 issues per year starting from June 2018 (March, June, September, December)

Online ISSN: 2148-7456

Website : <http://www.ijate.net/index.php/ijate>

<http://dergipark.org.tr/ijate>

Design & Graphic: IJATE

### **Support Contact**

Dr. İzzet KARA

Journal Manager & Founding Editor

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : [ikara@pau.edu.tr](mailto:ikara@pau.edu.tr)

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).



## International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed online journal. IJATE accepts original theoretical and empirical English-language manuscripts in psycho-educational assessment. Theoretical articles addressing new developments in measurement and innovative applications are welcome. IJATE publishes articles appropriate for audience of educational measurement specialists and practitioners.

There is no submission or publication process charges for articles in IJATE.

### **IJATE is indexed in:**

- Emerging Sources Citation Index (ESCI) (Web of Science Core Collection)
- TR Index (ULAKBIM),
- ERIH PLUS,
- DOAJ,
- Index Copernicus International
- SIS (Scientific Index Service) Database,
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,
- International Scientific Indexing

**Editor in Chief**

Dr. Izzet Kara, *Pamukkale University, Turkey*

**Editors**

Dr. Eren Can Aybek, *Pamukkale University, Turkey*

Dr. Özen Yıldırım, *Pamukkale University, Turkey*

**Section Editor**

Dr. H.İbrahim Sari, *Kilis 7 Aralık University, Turkey*

**Editorial Board**

Dr. Hafsa Ahmed, *National University of Modern Languages, Pakistan*

Dr. Beyza Aksu Dünya, *Bartın University, Turkey*

Dr. Murat Balkıs, *Pamukkale University, Turkey*

Dr. Gül ah Ba ol, *Gaziosmanpa a University, Turkey*

Dr. Bengü Börkan, *Bo aziçi University, Turkey*

Dr. Kelly D. Bradley, *University of Kentucky, United States*

Dr. Okan Bulut, *University of Alberta, Canada*

Dr. Javier Fombona Cadavieco, *University of Oviedo, Spain*

Dr. William W. Cobern, *Western Michigan University, United States*

Dr. R. Nükhet Çıkrıkçı, *İstanbul Aydın University, Turkey*

Dr. Safiye Bilican Demir, *Kocaeli University, Turkey*

Dr. Nuri Do an, *Hacettepe University, Turkey*

Dr. Erdiñç Duru, *Pamukkale University, Turkey*

Dr. Selahattin Gelbal, *Hacettepe University, Turkey*

Dr. Anne Corinne Huggins-Manley, *University of Florida, United States*

Dr. Violeta Janusheva, *"St. Kliment Ohridski" University, Republic of Macedonia*

Dr. Francisco Andres Jimenez, *Shadow Health, Inc., United States*

Dr. Nicole Kaminski-Öztürk, *University of Illinois at Chicago, United States*

Dr. Orhan Karamustafaoglu, *Amasya University, Turkey*

Dr. Yasemin Kaya, *Atatürk University, Turkey*

Dr. Hulya Kelecioğlu, *Hacettepe University, Turkey*

Dr. Hakan Ko ar, *Akdeniz University, Turkey*

Dr. Sunbok Lee, *University of Houston, United States*

Dr. Froilan D. Mobo, *Ama University, Philippines*

Dr. Ibrahim A. Njodi, *University of Maiduguri, Nigeria*

Dr. Jacinta A. Opara, *Kampala International University, Uganda*

Dr. Nesrin Ozturk, *Ege University, Turkey*

Dr. Turan Paker, *Pamukkale University, Turkey*

Dr. Abdurrahman Sahin, *Pamukkale University, Turkey*

Dr. Ragip Terzi, *Harran University*, Turkey

Dr. Hakan Türkmen, *Ege University*, Turkey

Dr. Hossein Salarian, *University of Tehran*, Iran

Dr. Kelly Feifei Ye, *University of Pittsburgh*, United States

**English Language Editors**

Dr. Hatice Altun, *Pamukkale University*, Turkey

Dr. Ça la Atmaca, *Pamukkale University*, Turkey

Dr. Sibel Kahraman, *Pamukkale University*, Turkey

Arzu Kanat Mutluo lu - *Pamukkale University*, Turkey

**Copy & Language Editor**

Anıl Kandemir, *Middle East Technical University*, Turkey

## Table of Contents

### *Research Article*

---

1. [The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model](#)  
Pages 170 - 192  
Gökhan Aksu, Cem Oktay Güzeller, Mehmet Taha Eser
2. [An Empirical Study for the Statistical Adjustment of Rater Bias](#)  
Pages 193 - 201  
Mustafa LHAN
3. [Development of Perceived School Counselor Support Scale: Based on the ASCA Mindsets and Behaviors](#)  
Pages 202 - 217  
Mehmet Akif Karaman, Cemal Karada , Javier Cavazos Vela
4. [Investigating a new method for standardising essay marking using levels-based mark schemes](#)  
Pages 218 - 234  
Jackie Greatorex, Tom Sutch, Magda Werno, Jess Bowyer, Karen Dunn
5. [Teaching Game and Simulation Based Probability](#)  
Pages 235 - 258  
Timur Koparan
6. [Explanatory Item Response Models for Polytomous Item Responses](#)  
Pages 259 - 278  
Luke Stanke, Okan Bulut
7. [Thematic Content Analysis of Studies Using Generalizability Theory](#)  
Pages 279 - 299  
Gül en Ta delen Teker, Ne e Güler
8. [Examination of the Extreme Response Style of Students using IRTree: The Case of TIMSS 2015](#)  
Pages 300 - 313  
Münevver İgün dibek
9. [Analyzing the Views of Teachers and Prospective Teachers on Information and Communication Technology via Descriptive Data Mining](#)  
Pages 314 - 329  
Ozge Can Aran, Ahmet Selman Bozkir, Bilge Gok, Esed Yagci

## The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model

Gökhan Aksu <sup>1\*</sup>, Cem Oktay Güzeller <sup>2</sup>, Mehmet Taha Eser <sup>3</sup>

<sup>1</sup> Adnan Menderes University, Vocational High School, Aydın, Turkey

<sup>2</sup> Akdeniz University, Faculty of Tourism, Antalya, Turkey

<sup>3</sup> Akdeniz University, Statistical Consultation Center, Antalya, Turkey

### ARTICLE HISTORY

Received: 07 November 2018

Revised: 19 February 2019

Accepted: 20 March 2019

### KEYWORDS

Artificial Neural Networks,  
Prediction,  
MATLAB,  
Normalization

**Abstract:** In this study, it was aimed to compare different normalization methods employed in model developing process via artificial neural networks with different sample sizes. As part of comparison of normalization methods, input variables were set as: work discipline, environmental awareness, instrumental motivation, science self-efficacy, and weekly science learning time that have been covered in PISA 2015, whereas students' Science Literacy level was defined as the output variable. The amount of explained variance and the statistics about the correct classification ratios were used in the comparison of the normalization methods discussed in the study. The dataset was analyzed in Matlab2017b software and both prediction and classification algorithms were used in the study. According to the findings of the study, adjusted min-max normalization method yielded better results in terms of the amount of explained variance in different sample sizes compared to other normalization methods; no significant difference was found in correct classification rates according to the normalization method of the data, which lacked normal distribution and the possibility of overfitting should be taken into consideration when working with small samples in the modelling process of artificial neural network. In addition, it was also found that sample size had a significant effect on both classification and prediction analyzes performed with artificial neural network methods. As a result of the study, it was concluded that with a sample size over 1000, more consistent results can be obtained in the studies performed with artificial neural networks in the field of education.

## 1. INTRODUCTION

The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data (Mannila, 1996). Decision tree, nearest neighborhood, support vector machine, Naive Bayes classifier and artificial neural networks are among the main classification methods and they are supervised learning approaches (Neelamegam & Ramaraj, 2013). Educational data

CONTACT: Gökhan Aksu ✉ [gokhanaksu1983@hotmail.com](mailto:gokhanaksu1983@hotmail.com) 📍 Adnan Menderes University, Vocational High School, Aydın, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

mining is concerned developing methods for predict student's academic performance and their behaviour towards education by the data that come from educational database (Upadhyay, 2016). It aims at devising and using algorithms to improve educational results and explain educational strategies for further decision making (Silva & Fonseca, 2017). Artificial Neural Networks (ANN) is one of the essential mechanisms used in machine learning. Due to their excellent capability of self-learning and self-adapting, they have been extensively studied and have been successfully utilized to tackle difficult real-world problems (Bishop 1995; Haykin 1999). Compared to the other approaches, Artificial Neural Networks (ANN), which is one of the most effective computation methods applied in data mining and machine learning, seems to be one of the best and most popular approaches (Gschwind, 2007; Hayashi, Hsieh, & Setiono, 2009). The word "Neural" (called as neuron or node, as part of this study the term "node" was used) included in the name Artificial Neural Network, indicates that the learning structure of human brain was taken as the basis of learning within the system. For a programmer, ANN is the perfect tool to discover the patterns that are very complex and numerous. The main strength of ANN lies on predicting multi-directional and non-linear relationships between input and output data (Azadeh, Sheikhalishahi, Tabesh, & Negahban, 2011). ANN, which can be used as part of many disciplines, is frequently used in classification, prediction and finding solutions to learning problems that involve the minimization of the disadvantages of traditional methods. Non-linear problems can also be solved through ANN, besides linear problems (Uslu, 2013).

Fundamentally, there are three different layers in an artificial neural network; namely input layer, hidden layers and output layer. Input layer communicate with the outer environment that contributes neural network to have a pattern. Input layer deals only with the inputs. Input layer should represent the condition where the neural network would be trained. Each input node should represent some independent variables that have an effect on the output of the neural network. Hidden layer is the layers on which the nodes executing activation function are gathered, they are located between input layer and output layer. Hidden layer is formed by many layers. The task of the hidden layer is processing the input obtained from the previous layer. Therefore, hidden layer is the layer that is responsible for deriving requested outcomes using input data (Kriesel, 2007). Numerous studies have been conducted to determine the number of the nodes included in the hidden layer but none of these researches were successful in finding the correct result. Moreover, an ANN may contain more than one hidden layer. There are no single formulas for computing the number of the hidden layers and the number of nodes in each hidden layer, various methods are used for this purpose. The output layer of an ANN collects and transmits the data considering the design to which the data will be transferred. The design represented by the output layer can be directly tracked up to the input layer. The number of nodes in an output layer should be directly associates to the performance of the neural network. The objective of the relevant neural network should be considered while determining the number of nodes in the output layer.

Artificial Neural Networks, is made of artificial neural network cells. An artificial neural network cell is built on two essential structures, namely neurons and synapses. A node (neuron) is a mathematical function that models the operation of a biologic node. In theory, an artificial node is formed by a transformation function and an activation function along with a group of weighted input. A typical node computes the weighted average of its input and this sum is usually processed by a non-linear function (i.e. sigmoid) called as activation function. The output of a node may be sent as input to the nodes of another layer that repeats the same computation. The nodes constitute the layers. Each node is connected to another node through a connection. Each connection is associated with a weight, including information about the input signal. Being associated with a weight is one of the most useful information for the nodes while solving a problem because the weight usually triggers or blocks the transmitted signal. Each node has an implicit status called as activation signal. The produced output signals are



allowed to be sent to the other units after combining input signal with the activation rule (Hagan, Demuth, Beale, & Jesus, 2014).

Main operating principle of an artificial neural network is as below:

- 1) Input nodes should represent an input based on the information that we attempt to classify.
- 2) A weight is given to each number in the input nodes for each connection.
- 3) In each node located at the next layer, the outputs of the connections coming to this layer are triggered and added and an activation function is applied to the weighted sum.
- 4) The output of the function is taken as the input of the next connection layer and this process continues until the output layer is reached (O'Shea & Nash, 2015).

Artificial Neural Networks was built inspiring from biological neural system, in other words human brain's working pattern. Since the most important characteristic of human brain is learning, the same characteristic was adopted in ANN as well. Artificial Neural Networks is a complex and adaptive system that can change its inner structure based on the information that it possesses. Being a complex, adaptive system, the learning of ANN is based on the fact that input/output behavior may vary according to the change occurring in the surrounding of a node. Another important feature of neural networks is they have an iterative learning process in which data status (lines) are represented to the network one by one and the weights associated with input values are modified at every turn. Usually the process restarts when all cases are represented. A network of learning stage learns by modifying the weights so that the correct class definitions of input samples are predicted. Neural network learning is also called as "Learning to make a connection" because of the connections among the nodes (Davydov, Osipov, Kilin, & Kulchitsky, 2018).

The most important point in the application of artificial neural networks to real-world problems is to be able to understand the solution that will be determined without being complicated, easy to interpret and in a practical way to the real world. The common point of these three features is very closely related to how the data is managed and processed. Normalization plays a very critical role, especially in the context of intelligibility and easy interpretation in the most critical point of data management (Weigend & Gershenfeld, 1994; Yu, Wang, & Lai, 2006). The normalization process, in which the data is sensible and reassembled in a much smaller interval, arises as a need in the case of a method usually used on very large data sets, such as artificial neural networks. In the case of artificial neural networks, the number of nodes in the input, the number of nodes in the hidden layer, and the number of nodes in the output are very important elements, and the connection for any two layers is called positive or negative weight (Hagan, Demuth, Beale, & Jesus, 2014). The algorithm used in the artificial neural network-based model established when different ranges are used for the variables in the data set will most likely not be able to discover the possible correlation between the variables. At the same time, the fact that there are different intervals for the variables in the data set causes these weights to be affected in different meanings. And at the same time, the use of variables with very different intervals is eliminated in the geometric sense, and the results obtained from the experiments or analyzes and the results obtained from the experiments in the artificial neural network are eliminated in a smaller and specific range. normalization is needed to make interpretations much easier for the total of variables (Lou, 1993; Weigend & Gershenfeld, 1994; Yu, Wang, & Lai, 2006). And in normal neural network based studies, which are used on normalization process, especially on the methodological data, the number of variables can be high and the practical benefits of real life are desired, it is more needed in artificial neural network based studies.

A network gets ready to learn after being configured for a certain application. The configuration process of a network for a certain application is called as "Preliminary preparation process".

Following the completion of the preparation belonging to the preliminary process, either training or learning starts. The network processes the records of the training data at a time using the weights and functions in the hidden layers, then compare the outputs with desired outputs. Afterwards, the errors are distributed backwards in the system, which allows the system to modify application weights for the subsequent records to be processed. This process takes place continuously as the weights are modified. The same data sample may be processed many times since the connection weights are continuously refined during the training of a network (Wang, Devabhaktuni, Xi, & Zhang, 1998).

The preliminary data processing of an artificial neural network modelling is a process having broad applications, rather than a limited definition. Almost all theoretical and practical research involving neural networks focus on the data preparation for neural networks, normalizing data for conversion and dividing the data for training (Gardner & Dorling, 1998; Rafiq, Bugmann, & Easterbrook, 2001; Krycha & Wagner, 1999; Hunt, Sbarbaro, Bikowski, & Gawthrop, 1992; Rumelhart, 1994; Azimi Sadjadi & Stricker, 1994). In some studies, neural networks were used for modelling purposes without any data preparation procedure. For these studies, there is an implicit assumption indicating that all data were prepared in advance so that they can be directly used in modelling. Regarding the practice, it cannot be said that the data is always ready for analysis. Usually there are limitations about the integrity and quality of the data. As a result, complex data analysis process cannot be successful without performing a preliminary preparation process to the data. Researches revealed that data quality has significant impact on artificial neural network models (Famili, Shen, Weber, & Simoudis, 1997; Zhang, Zhang, & Yang, 2003). Smaller and better-quality data sets, which may significantly improve the efficiency of the data analysis, can be produced through preliminary data processing process. Regarding ANN learning, data preparation process allows the users to take decisions about how to represent the data, which concepts to be learned and how to present the outcomes of the data analysis, which makes explaining the data in the real world much easier (Redman, 1992; Klein & Rossin, 1999; Zang et al., 2003).

Applying a preliminary preparation process to the data is an important and critical step in neural network modelling for complex data analysis and it has considerable impact on the success of the data analysis performed as part of data mining. Input data affects the quality of neural network models and the results of the data analysis. Lou (2003) emphasized that the deficiencies in the input data may cause huge differences on the performance of the neural networks. Data that was subject to preliminary processing play a major role in obtaining reliable analysis outcomes. In theory, data lacking preliminary process makes data analysis difficult. In addition, data obtained from different data sources and produced by modern data collection techniques made data consumption a time-consuming task. 50-70% the time and effort spend on data analysis projects is claimed to be for data preparation. Therefore, preliminary data preparation process includes getting the data ready to analysis for improving complex data analysis (Sattler, 2001; Hu, 2003; Lou, 2003).

There are few parameters affecting the learning process of an artificial neural network. Regarding the learning of the nodes as part of learning process, if a node fails, the remaining nodes may continue to operate without any problem. The weights of the connections located in an artificial neural cell vary, which plays a role in the success of the neural network and in the formation of the differences on the values involving the learning of the neural network. In addition to the weights, the settings about the number of nodes in the hidden layers and learning rate parameters affect neural network learning process as well. There is not a constant value for the mentioned parameters. Usually expert knowledge plays a major role in determining these parameters (Anderson, 1990; Lawrance, 1991; Öztemel, 2003). Sample size is also one of the parameters that affect learning process. According to “Central Limit Theorem”, each unbiased

samples coming from a universe with normal distribution, formed by independent observations, shows normal distribution provided that sample size is over 30. In addition, regardless of the universe, the shape of the distribution approaches to normal distribution as the sample size increases and therefore the validity and reliability of the inferences to be made for the parameters increase (Dekking, Kraaikamp, Lopuhaä & Meester, 2005; Roussas, 2007; Ravid, 2011). There is no rule indicating that at the end of the learning process the nodes will definitely learn; some networks never learn.

Number of nodes and learning rate are not the only factors playing a role in making the execution of certain preliminary data processing more effective as part of the neural network learning. The normalization process of the raw input is as important as the other preliminary data processes (reducing the size of the input field, noise reduction and feature extraction). In many artificial neural network applications, raw data (not processed or normalized prior to use) is used. As a result of using raw data, multi-dimensional data sets are employed and many problems are experienced, including longer analysis duration. The normalization of the data, which scales the data to the same range, minimizes the bias in the artificial neural network. At the same time the normalization of the data speeds up the process involving the learning of the features covered in the same scale. In theory, the purpose of the normalization is rescaling the input vector and modify the weight and bias corresponding to the relevant vector for obtaining the same output features that have been obtained before (Bishop, 1995; Elmas, 2003; Ayalakshmi & Santhakumaran, 2011). In general, machine learning classifiers cannot compute Euclidian distance between features. Euclidian distance is the linear distance between two points (vectors of the nodes) located in Euclidian space, which is simply two or three dimensional. Therefore, the features should be normalized in order to prevent the bias that may occur in the model built with artificial neural network (Lou, 1993; Weigend & Gershenfeld, 1994; Yu, Wang, & Lai, 2006).

In many cases normalization improves the performance but considering the normalization as mandatory for the operation of the algorithm is wrong. In case of a trained data set, whose model is unseen, using raw data may be more useful. There are many data normalization methods. Among them the most important ones are Z-score, min-max (feature scaling), median, adjusted min-max and sigmoid normalization methods. As part of the research, different normalization methods used in the process of modelling with Artificial Neural Networks (Z-score, min-max, median, adjusted min-max) were applied the learning, test, validation and overall data sets and the results were compared. Below, the normalization methods used in the research are summarized:

- 1) Z-score Method: Mean and standard deviation of each feature are used across a series of learning data to normalize the vector of each feature included in the input data. Mean and standard deviation are calculated for each feature. The equality used in the method is as below where  $x'$  indicates normalized data,  $x_i$  input variable,  $\mu_i$  arithmetic mean of the input variable and  $\sigma_i$  standard deviation of the input variable.

$$x' = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

This procedure sets the mean of each feature in the data set equal to zero and standard deviation to one. As a part of the procedure, first the normalization is applied to the feature vectors in the data set. The mean and standard deviation are calculated for each feature over the training data and it is kept for using as weight in the final system design. In short, this procedure is a preliminary processing within the artificial neural network structure.

- 2) Min-Max Method: The method is used as an alternative to Z-score Method. This method rescales the features or the outputs in any range into a new range. Usually the features are scaled between 0-1 or (-1)-1. The equality used in the method is as below where  $x_{m\ n}$  indicates minimum value,  $x_m$  maximum value,  $x_i$  input value and  $x'$  normalized data:

$$x' = \frac{x_i - x_m}{x_m - x_m} \quad (2)$$

When min-max method is applied, each feature remains the same while taking place in the new range. This method keeps all relational properties in the data.

- 3) Median Method: As part of median method, the median of each input is calculated and it is used for each sample. The method is not affected by extreme variations and it is quite useful in case of computing the ratio of two samples in hybrid form or to get information about the distribution. The equality used in the method is as below where  $x'$  indicated normalized data,  $x_i$  input variable:

$$x' = \frac{x_i}{M} \quad (a_i) \quad (3)$$

- 4) Adjusted Min-Max Method: The fourth normalization method is adjusted min-max method. For the implementation of the method, all the data are normalized between 0.1 and 0.9, with the equality used as part of the method. With the normalization, the data set gets a dimensionless form. The equality used in the method is as below where  $x'$  indicated normalized data,  $x_i$  input variable,  $x_m$  maximum value of the input variable and  $x_m$  minimum value of the input variable:

$$x' = 0.8 * \frac{x_i - x_m}{x_m - x_m} + 0.1 \quad (4)$$

In adjusted min-max method, the results obtained in the previously given formula are multiplied by a constant value of 0.8 and a constant value of 0.1 is added.

The variables used by the researchers working in the field of educational sciences can be summarized as situations related to the student in terms of the starting point, the situations related to the personnel, the situations related to the administration and the situations related to the school. All these cases reveal large data sets that need to be analyzed. These large data sets are data sets that consist of too many variables and too many students (participants). In recent years, the concepts of machine learning, which are related to algorithms working in the background of data mining and data mining methods, are frequently mentioned in Educational Sciences. The analysis of the data sets formed by many variables and too many participants from the databases related to Educational Sciences brought with it the concept of Educational Data Mining (Gonzalez & DesJardins, 2002; Scumacher, Olinsky, Quinn, & Smith, 2010; Romero & Ventura, 2011). Nowadays, in the context of educational data mining, studies on modeling of education and training programs, predictive and classification based models on student and teacher are carried out. By using these purposes, artificial neural networks, decision trees, clustering and Bayesian based algorithms are used in the background (Gerasimovic, Stajenovic, Bugaric, Miljkovic, & Veljovic, 2011; Wook, Yahaya, Wahab, Isa, Awang, & Seong, 2009).

Artificial neural network is a non-linear model that is easy to use and understand compared to other methods. Most other statistical methods are evaluated within the scope of parametric

methods which require a statistical history. Artificial neural networks are often used to solve problems related to estimation and classification. Artificial neural networks alone are insufficient to interpret the relationship between input and output and to cope with uncertain situations. However, these disadvantages can easily be overcome by the structure of artificial neural networks designed to be integrated with many different features (Schmidhuber, 2015; Goodfellow, Bengio, & Courville, 2016). Regarding all of these, the purpose of the research will be to determine the differentiation that different normalization methods employed in model developing process exhibit at different sample sizes. In the study, the changes on the prediction results obtained from data sets of 250, 500, 1000, 1500 and 2000 cases, through different normalization methods were analyzed and the classification level of the normalization method that had best prediction results was evaluated. Determining the number of sample sizes the study conducted by Finch, West and Mackinnon (1997) in determining the number of samples, it was determined that there were differences in the estimations in different sample sizes. In addition, Fan, Wang and Thompson (1996) in their study showed that the calculation methods in different sample sizes differed and this difference was significant especially in small samples. For this reason, within the framework of the specified objectives, the problem statement of the research was set as “Does the sample size affects the normalization method used in predicting science literacy level of the students using work discipline, environmental awareness, instrumental motivation, science self-efficacy, and weekly science learning time variables in PISA 2015 Turkey sample”. The following research questions were addressed within the framework of the general purpose specified according to the main problem of the study:

1. Does sample size affect Z-score normalization method in the process of modelling with ANN?
2. Does sample size affect min-max normalization method in the process of modelling with ANN?
3. Does sample size affect median normalization method in the process of modelling with ANN?
4. Does sample size affect adjusted min-max normalization method in the process of modelling with ANN?
5. Does sample size affect the best normalization method in the process of modelling with ANN, in case of a two-category output variable?

Allowing input values and output values to be at the same range through the normalization of the research data has vital importance for the determination of very high or very low values in the data (Güzeller & Aksu, 2018). Moreover, very high or very low values in the data, which may be originated from various reasons such as wrong data entry, may cause the network to produce seriously wrong outputs; thus, the normalization of input and output data has significant importance for the consistency of the results.

## **2. METHOD**

### **2.1. Research Model**

This study is accepted as a basic research because it is aiming to determine the normalization method giving the best result by testing various methods used in modelling process where Artificial Neural Networks were applied in different sample sizes (Frankel & Wallen, 2006; Karasar, 2009). Basic researches aim to add new knowledge to the existing one, in other words improving the theory or testing existing theories (OECD, 2015).

### **2.2. Data Collection**

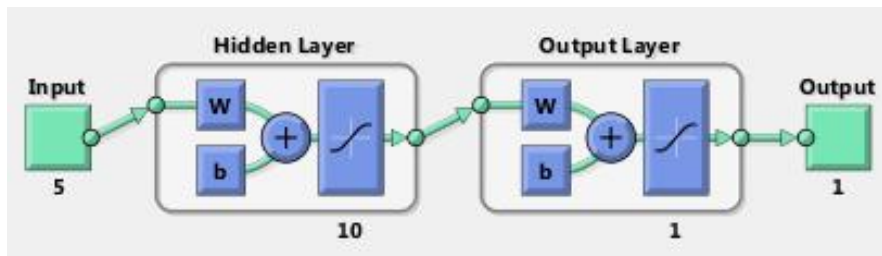
The data used within the scope of the study were obtained from PISA 2015 test (MEB, 2016), which has been organized by OECD. The data obtained from 5895 students who have participated in the test from Turkey universe were divided into groups of 250, 500, 1000, 1500

and 2000 through systematic sampling method. Students’ work discipline, environmental awareness, instrumental motivation, science self-efficacy, and weekly science learning time variables were used as the input variables, whereas students’ science literacy score was used as the output variable. The names and codes of the input and output variables covered in the study are illustrated in Table 1.

**Table 1.** Variables Used in the Analysis

Variable Type	Variables	Data Set
Output Variables	PISA 2015 Science Literacy ( <i>PVISCIE</i> )	Output
Input Variables	Work Discipline ( <i>DISCLISCI</i> )	Input
	Environmental Awareness ( <i>ENVAWARE</i> )	
	Instrumental Motivation ( <i>INSTSCIE</i> )	
	Science Self-Efficacy ( <i>SCIEEFF</i> )	
	Weekly Science Learning Time ( <i>SMINS</i> )	

Hastie, Tibshiranni and Friedman (2017) stated that there is not an ideal ratio for dividing the whole data into training, test and validation data sets; researchers should consider signal noise levels and model-data fit. Therefore, since the best results of the model were obtained when the proportion of training, test and validation data sets were respectively 60%-20%-20% in the model developed with Artificial Neural Networks, 60% of the data set of 1000 students was used for the training of the model, whereas 20% was used for testing and 20% for validation. The theoretical model established by the researchers in the MATLAB program with Artificial Neural Networks to test four different normalization methods covered in the study is illustrated in Figure 1.



**Figure 1.** The theoretical model developed with Artificial Neural Networks

As can be seen from Figure 1, the number of input variables is 5, number of hidden layers is 10, number of output layer is 1 and the number of output variables is 1. Sigmoid function, one of the most common used activation functions, is used to determine between neurons nonlinear activation (Namin, Leboeuf, Wu, & Ahmadi, 2009).

### 2.3. Data Analysis

First of all, regarding the data obtained from PISA survey, both input variables and output variable were normalized in Excel according to Z-score conversion, min-max, median, and adjusted min-max methods, using relevant formulas. In the analysis the following figures were kept constant: number of iterations – 500, layer number – 2 and number of nodes – 10. These parameters are default values determined by the matlab program (Matlab, 2002). Regarding constant parameters, Levenberg-Marquardt (TRAINLM) was set as the training function and adaptive learning (LEARNINGDM) method as the learning function. In data analysis, the changes occurred in the normalization methods for 250, 500, 1000, 1500 and 2000 sample sizes were analyzed. The amount of explained variance and correct classification ratio were used in the

comparison of the normalization methods discussed in the study, for different sample sizes. Data analysis were performed in Matlab2017b software and both prediction and classification algorithms were used in the study. Students who have achieved a score under 425,00, which was Turkey average, were coded as unsuccessful (0), whereas those who have achieved a higher score were coded as successful (1). The success rates of the methods were determined by means of confusion matrix for the two-category output variable.

### 3. RESULTS

In the study, the performance of the outcomes obtained from four different normalization methods on training, test and validation data sets were determined first, then their overall success rates were compared. But, normality tests were performed before the analysis, to check the normality of the data and the results of the analysis are illustrated in [Table 2](#).

**Table 2.** Test for the Suitability of the Data to Normal Distribution

Method	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistics	SD	p	Statistics	SD	p
Work discipline	.096	1000	.000	.970	1000	.000
Environmental awareness	.096	1000	.000	.952	1000	.000
Instrumental motivation	.142	1000	.000	.938	1000	.000
Science self-efficacy	.120	1000	.000	.934	1000	.000
Weekly science learning time	.162	1000	.000	.936	1000	.000
Science literacy	.035	1000	.005	.994	1000	.000

[Table 2](#) revealed that both input variables and science literacy scores, which was taken as the output variable, were not distributed normally ( $p < .01$ ). Based on this result, it was concluded that normalization methods can be applied to the data used as part of the study.

#### 3.1. Findings about Z-Score Normalization

*nntool* command was used for the introduction of the data set obtained by normalizing five input data and one output data, which have been covered in the study, to Matlab software and for the regression analysis that would be carried out by means of Artificial Neural Networks., Analysis results from different sample sizes are illustrated in [Table 3](#); they were obtained after the introduction of the input and output data sets to the program, and the execution of tansig conversion function in the network that was defined as 2-layer and 10-neuron.

**Table 3.** Equations Obtained as a Result of Z-Score Normalization

Sample Size	Training		Test		Validation		Overall	
	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>
<b>N=250</b> Gradient <sup>†</sup> =5.56 iterations=11	y=0.27x-0.17	55.13	y=0.03x-0.17	8.14	y=0.18x-0.20	33.08	y=0.23x-0.18	45.34
<b>N=500</b> Gradient=2.67 iterations=9	y=0.16x-0.19	38.58	y=0.04x-0.28	10.77	y=0.20x-0.16	44.62	y=0.15x-0.20	36.21
<b>N=1000</b> Gradient=6.33 iterations=9	y=0.17x-0.01	44.91	y=0.15x+0.04	40.57	y=0.16x-0.02	44.37	y=0.17x-0.01	44.24
<b>N=1500</b> Gradient=8.67 iterations=13	y=0.24x-0.00	49.29	y=0.22x+0.04	42.87	y=0.26x-0.04	51.79	y=0.24x-0.01	48.84
<b>N=2000</b> Gradient=10.30 iterations=27	y=0.23x-0.01	48.33	y=0.26x-0.03	51.23	y=0.25x-0.07	46.92	y=0.24x-0.02	48.49

<sup>†</sup> It is the square of the slope of the error function whose weight and bias are unknown. It is used as the measure of error in Matlab.



The review of Table 3 revealed that regarding the results of Z-score normalization method, the sample size resulting with: the highest explained variance for the training data set was 250 ( $R^2=55.13$ ); the highest explained variance for the test data set was 2000 ( $R^2=51.23$ ); the highest explained variance for the validation data set was 1500 ( $R^2=51.79$ ); and the highest explained variance for the whole data set was 1500 ( $R^2=48.84$ ). When examined in a holistic manner, it is seen that the sample sizes of 250 and 500 have the lowest explained variance. For the sample size of 2000, the scattering of the output variable predicted from the input variables in two-dimensional space is illustrated in Figure 2 as an example.

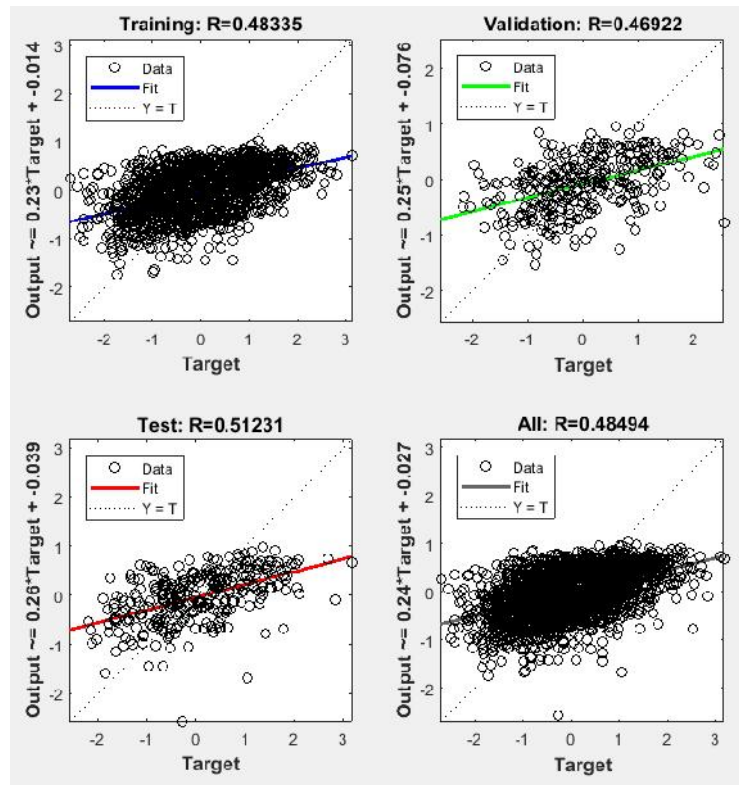


Figure 2. The outcomes of Z-Score Normalization in different data sets.

### 3.2. Findings about Min-max Normalization

The results of regression analysis obtained by Artificial Neural Networks, after the normalization of five input and one output data, which have been covered as part of the study, based on maximum and minimum values are illustrated in Table 4. In addition, it was found that the sample size of 250 and 500 had the lowest explained variance for every data set. The review of Table 4 revealed that regarding the results of Min-max normalization method, the sample size resulting with: the highest explained variance for the training data set was 2000 ( $R^2=54.99$ ); the highest explained variance for the test data set was 1000 ( $R^2=52.41$ ); the highest explained variance for the validation data set was 1000 ( $R^2=50.75$ ); and the highest explained variance for the whole data set was 2000 ( $R^2=51.74$ ). When examined in a holistic manner, it is seen that the sample sizes of 250 and 500 have the lowest explained variance. For the sample size of 2000, the scattering of the output variable predicted from the input variables in two-dimensional space is illustrated in Figure 3 as an example.

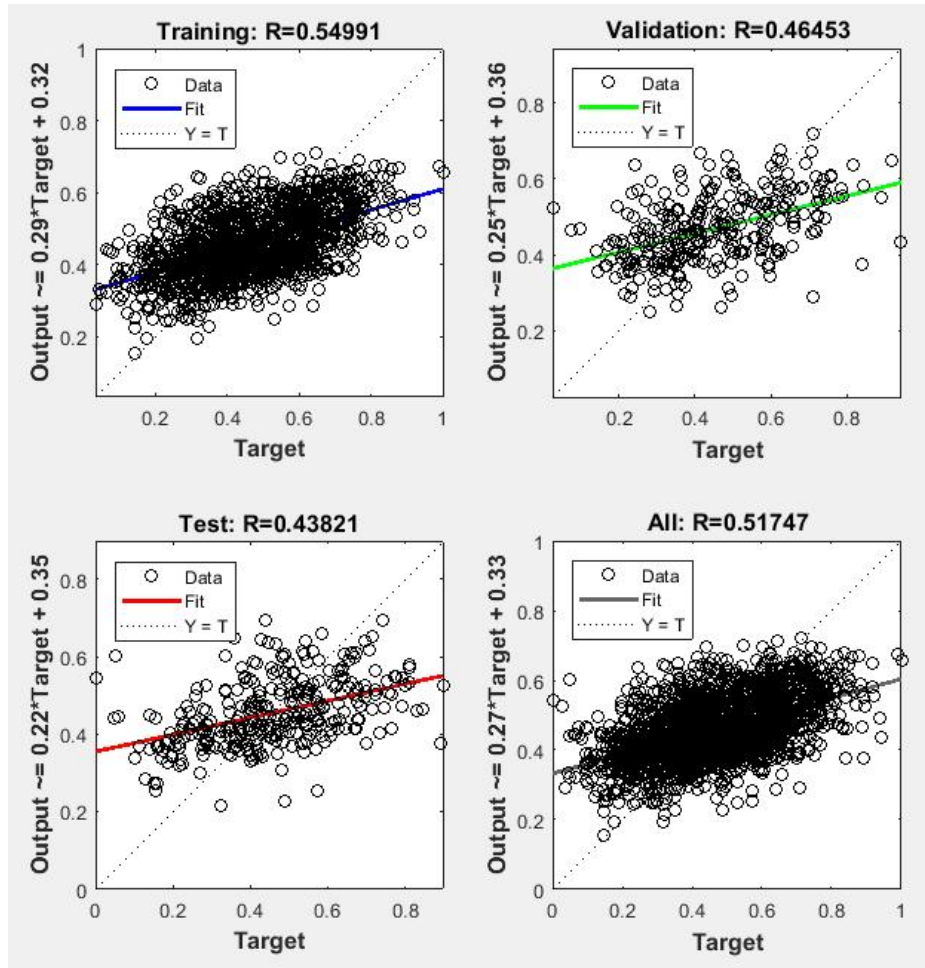


Figure 3. The outcomes of Min-max Normalization in different data sets

### 3.3. Findings about Median Normalization

The results of regression analysis obtained by Artificial Neural Networks, after the normalization of five input and one output data, which have been covered as part of the study, based on median values are illustrated in Table 5.

**Table 4.** Equations Obtained as a Result of Min-max Normalization

Sample Size	Training		Test		Validation		Overall	
	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>
<b>N=250</b>								
Gradient=0.09 iteration=10	y=0.13x+0.38	33.05	y=0.03x+0.41	9.01	y=0.12x+0.41	38.21	y=0.12x+0.39	29.98
<b>N=500</b>								
Gradient=0.08 iteration=10	y=0.18x+0.36	46.98	y=0.01x+0.43	4.05	y=0.06x+0.40	17.21	y=0.15x+0.37	37.19
<b>N=1000</b>								
Gradient=0.18 iteration=9	y=0.23x+0.36	49.48	y=0.25x+0.36	52.41	y=0.26x+0.34	50.75	y=0.24x+0.35	50.15
<b>N=1500</b>								
Gradient=0.14 iteration=10	y=0.23x+0.36	49.39	y=0.24x+0.36	48.48	y=0.21x+0.37	47.09	y=0.23x+0.36	48.93
<b>N=2000</b>								
Gradient=0.24 iteration=16	y=0.29x+0.32	54.99	y=0.22x+0.35	43.82	y=0.25x+0.36	46.45	y=0.27x+0.33	51.74

**Table 5.** Equations Obtained as a Result of Median Normalization

Sample Size	Training		Test		Validation		Overall	
	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>
<b>N=250</b>								
Gradient=0.12 iteration=11	y=0.19x+0.77	42.92	y=0.33x+0.64	46.90	y=0.34x+0.62	50.03	y=0.23x+0.73	43.99
<b>N=500</b>								
Gradient=0.44 iteration=12	y=0.15x+0.81	42.22	y=0.14x+0.81	34.76	y=0.13x+0.83	39.34	y=0.15x+0.81	40.87
<b>N=1000</b>								
Gradient=0.41 iteration=11	y=0.25x+0.75	50.37	y=0.22x+0.79	40.90	y=0.26x+0.73	51.75	y=0.25x+0.76	48.85
<b>N=1500</b>								
Gradient=0.36 iteration=13	y=0.29x+0.71	53.56	y=0.29x+0.71	50.27	y=0.24x+0.76	45.78	y=0.28x+0.72	51.88
<b>N=2000</b>								
Gradient=0.40 iteration=15	y=0.28x+0.73	53.49	y=0.25x+0.77	47.79	y=0.28x+0.73	52.16	y=0.27x+0.73	52.43

The review of Table 5 revealed that regarding the results of Median normalization method, the sample size resulting with: the highest explained variance for the training data set was 1500 ( $R^2=53.56$ ); the highest explained variance for the test data set was 1500 ( $R^2=50.27$ ); the highest explained variance for the validation data set was 2000 ( $R^2=52.16$ ); and the highest explained variance for the whole data set was 2000 ( $R^2=52.43$ ). In addition, it was found that the sample size of 500 had the lowest explained variance for every data set. For the sample size of 2000, the scattering of the output variable predicted from the input variables in two-dimensional space is illustrated in Figure 4 as an example.

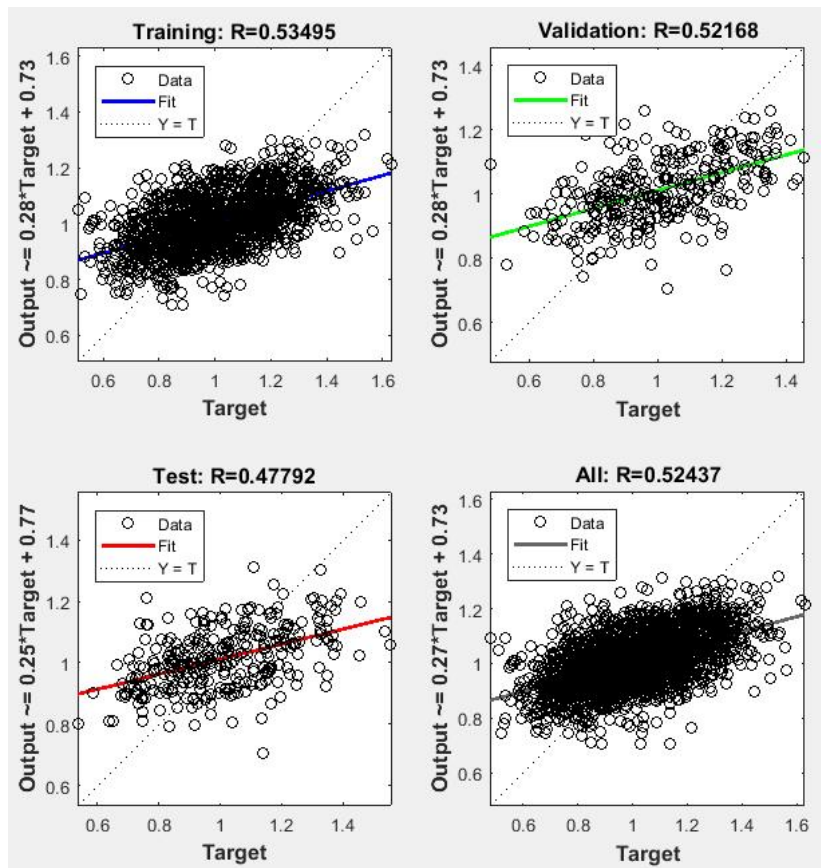


Figure 4. The outcomes of Median Normalization in different data sets

### 3.4. Findings about Adjusted Min-Max Normalization

The results of regression analysis obtained by Artificial Neural Networks, after the normalization of five input and one output data, which have been covered as part of the study, based on maximum and minimum values and processed by an adjustment function are illustrated in Table 6.

**Table 6.** Equations Obtained as a Result of Adjusted Min-Max Normalization

Sample Size	Training		Test		Validation		Overall	
	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>	Regression equation	R <sup>2</sup>
<b>N=250</b> Gradient=0.06 F=12	y=0.28x+0.32	51.08	y=0.59x+0.20	63.86	y=0.50x+0.22	61.26	y=0.34x+0.30	53.55
<b>N=500</b> Gradient=0.21 iteration=14	y=0.19x+0.36	47.58	y=0.07x+0.40	16.69	y=0.16x+0.37	38.87	y=0.17x+0.36	41.92
<b>N=1000</b> Gradient=0.19 iteration=10	y=0.23x+0.36	48.94	y=0.22x+0.37	44.18	y=0.26x+0.34	52.61	y=0.23x+0.36	48.67
<b>N=1500</b> Gradient=0.17 iteration=14	y=0.28x+0.34	53.96	y=0.28x+0.34	50.49	y=0.23x+0.36	47.07	y=0.27x+0.34	52.38
<b>N=2000</b> Gradient=0.19 iteration=23	y=0.30x+0.33	54.84	y=0.24x+0.36	45.01	y=0.29x+0.33	52.96	y=0.29x+0.33	53.09

**Table 7.** Classification Outputs for Raw Data and Normalized Data

Sample Size	Iteration	<i>A hie</i>		<i>A hie</i>		<i>A hie</i>		<i>A hie</i>	
		<i>t<sub>1</sub></i>		<i>t<sub>1</sub></i>		<i>v</i>		<i>o</i>	
N=250	6	%51.10		%63.20		%76.30		%56.80	
N=500	15	%62.60		%62.70		%56.00		%61.60	
N=1000	14	%66.90		%61.30		%60.00		%65.00	
N=1500	21	%67.00		%63.60		%66.20		%66.40	
N=2000	25	%67.90		%67.30		%64.30		%67.30	

The review of Table 6 revealed that regarding the results of Adjusted min-max normalization method, the sample size resulting with: the highest explained variance for the training data set was 2000 ( $R^2=54.84$ ); the highest explained variance for the test data set was 250 ( $R^2=63.86$ ); the highest explained variance for the validation data set was 250 ( $R^2=61.26$ ); and the highest explained variance for the whole data set was 250 ( $R^2=53.55$ ). In addition, it was found that the sample size of 500 had the lowest explained variance for every data set. At the same time, the explained variance for test, validation and overall data sets were found to be the highest for the smallest sample size (250). For the sample size of 2000, the scattering of the output variable predicted from the input variables in two-dimensional space is illustrated in Figure 5 as an example.

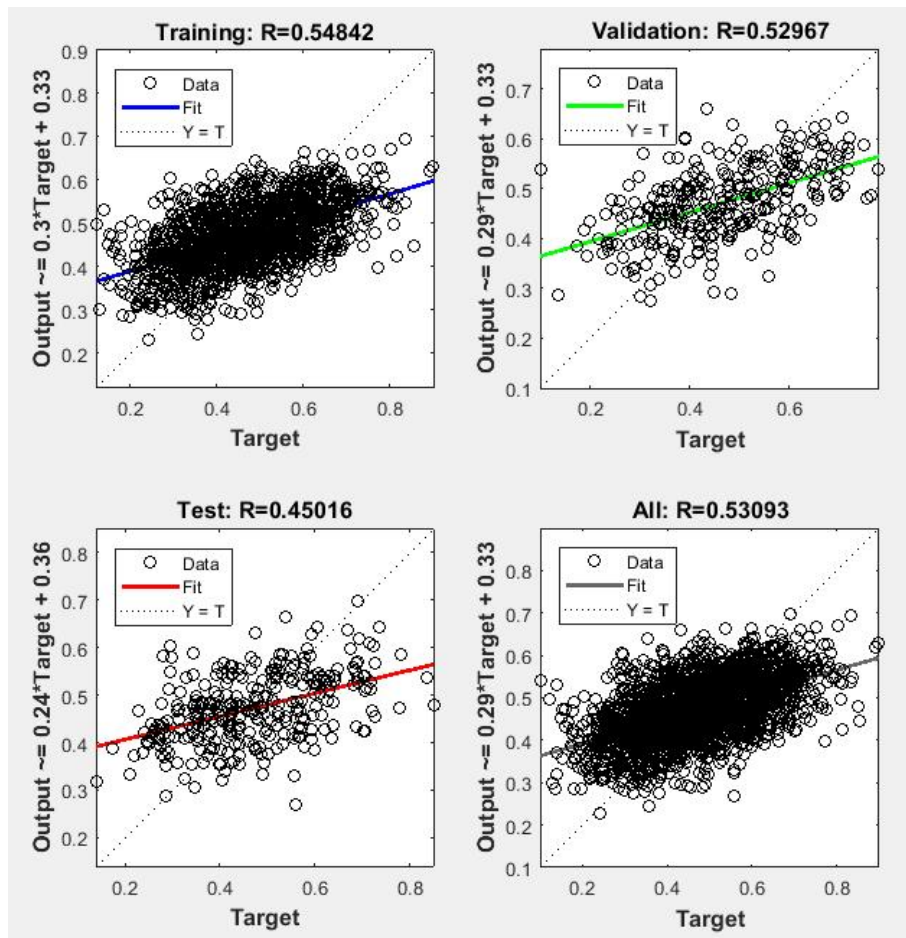


Figure 5. The outcomes of Adjusted min-max Normalization in different data sets

The review of Figure 5 revealed that, for the sample size of 2000, ANN prediction method achieved the highest success in training data set, followed by validation and test data sets. The evaluation of the outputs obtained from training, test and validation data sets as a whole resulted with 53.09% as the rate of correct prediction.

### 3.5. Findings Obtained in case of 2-category Output Variable for the most Successful Normalization Method

After determining that Adjusted Min-Max Normalization method is the best method for the prediction of PISA science literacy score, it was attempted to predict the class of the students in terms of achievement using the input variables covered in the study. The comparison of the classification methods obtained by adjusted min-max method for different sample sizes is illustrated in Table 7.

Table 7 revealed that no significant difference was observed in the test data set with the normalization of the raw data, however differences were observed in the training and validation data sets. Taking the outcomes obtained from training, test and validation data sets into account as a whole indicated that sample size created a significant difference in the correct classification rates of the students from the input variables ( $Z_{\text{computed}}=0.64 < Z_{\text{critical}}=1.96$ ). For the sample size of  $N=2000$ , the confusion matrix of the obtained classification outcomes is illustrated in Figure 6 as an example.



Figure 6. Classification Outcomes Obtained with Raw Data

According to Figure 6, the evaluation of training, test and validation data sets together showed that when students are classified in terms of their PISA achievement as successful or unsuccessful regarding the average score, 67.30% of the students were classified correctly, whereas 32.80% of the students were classified incorrectly.

#### 4. CONCLUSION, DISCUSSION and SUGGESTIONS

With this study Z-score, min-max, median, and adjusted min-max methods, which are employed in the process of modelling via Artificial Neural Networks, were compared in different sample sizes. We tried to find the best normalization method for predicting science literacy level by using statistical normalization methods included in the literature. Based on the evaluation of normalization methods, which have been applied to training, test, validation and overall data sets, as a whole in terms of the amount of explained variance, it was concluded that the highest amount of explained variance was achieved in the data set to which adjusted min-max method was applied. Regarding correct classification percentage, no significant difference was found between research data that was not normally distributed and the data normalized using adjusted min-max method.

In the study, the comparison was performed after setting constant parameter values for each normalization method and it was concluded that adjusted min-max method was the most suitable method for the relevant data set. It was also concluded that for each data set, min-max and median normalization methods have given similar results in terms of average error and explained variance. After determining the normalization method that provided the best performance in the prediction of numeric value, it was found that normalization didn't played

a role in the classification of the students as successful or unsuccessful. For this purpose, artificial neural network's classification results were obtained using raw data, then they were compared with the results obtained with normalized data and it was found that there was no significant difference among them. Accordingly, the normalization method used had an important effect on the prediction of the numeric values, but it had not a significant effect on the classification outcomes. In other words, the normalization method had a significant effect if the output variable obtained through artificial neural networks was numeric, whereas it had not a significant effect if the output variable was categoric (classification).

Regarding the provision of the best results by adjusted min-max normalization method, the results of the research are parallel to the results of the similar researches in the literature. Yavuz and Deveci (2012), have analyzed the impact of five different normalization methods on the accuracy of the predictions. They have tested adjusted min-max, Z-score, min-max, median, and sigmoid normalization methods. According to the results of the research, it was found that considering the average error and average absolute percent error values, the highest prediction accuracy has been obtained from the data set to which adjusted min-max method was applied, whereas the lowest prediction accuracy has been obtained from sigmoid normalization method. Ali and Senan (2017), have analyzed the effect of normalization on achieving best classification accuracy. For this purpose, they have observed the effect of three different normalization methods on the classification rate of multi-layer sensor for three different numbers of hidden layers. In the study, adjusted min-max normalization method, min-max normalization method in [-1, +1] range, and Z-Score normalization method has been tested for three different situations where backpropagation algorithm has been used as the learning algorithm. According to the results of the research, adjusted min-max normalization method has given the best outcomes (97%, 98%, 97%) in terms of correct classification ratio for the three cases where the number of hidden layers has been 5, 10 and 20. It has been observed that min-max normalization method in [-1, +1] range has been the second best normalization method in terms of correct classification ratio (57%, 55%, 59%), whereas Z-score method is the third best normalization method (49%, 53%, 50%). Vijayabhanu and Radha (2013), have analyzed the effect of six different normalization methods on prediction accuracy. For this purpose, they have tested Z-Score normalization method, min-max normalization method, biweight normalization method, tanh normalization method, double sigmoidal normalization method and dynamic score normalization with mahalanobis distance. According to the results of the research, the normalization methods have been ranked as follows with the relevant prediction accuracies: dynamic score normalization with mahalanobis distance (86.2%) has been first followed by Z-score normalization (84.1%), min-max normalization (82.6%), tanh normalization (82.3%), beweight normalization (81.2%), and double sigmoidal normalization (80.5%).

The review of the literature revealed the presence of other researches that are not parallel to this research. Özkan (2017), has analyzed the effects of three different normalization methods on the accuracy of classification. For this purpose, he has tested Z-Score normalization method, min-max normalization method and decimal scaling normalization method. Considering the accuracy of classification, sensitivity and selectivity values, it has been observed that Z-Score normalization method has provided the best outcomes in general, followed by decimal scaling normalization and min-max normalization methods. Panigrahi and Behera (2013), have analyzed the effect of five different normalization methods on forecast accuracy. For this purpose, they have tested min-max normalization method, decimal scaling normalization method, median normalization method, vector normalization method, and Z-Score normalization method. It has been observed that decimal scaling and vector normalization methods have provided better forecast accuracy compared to median, min-max and Z-Score normalization methods. Cihan, Kalıpsız and Gökçe (2017), have analyzed the effect of four different normalization methods on classification accuracy. For this purpose, they have tested



min-max normalization method, decimal scaling method, Z-Score method and sigmoid method. According to the results of the research the best classification has been obtained with 0.24 sensitivity, 0.99 selectivity and 0.36 f-measurement, by applying sigmoid normalization method, whereas the worst classification has been obtained with 0.21 sensitivity, 0.99 selectivity and 0.32 f-measurement, by applying Z-Score Normalization method. Mustaffa and Yusof (2011), have analyzed the effect of three different normalization methods on prediction accuracy. For this purpose, they have tested min-max normalization method, Z-Score normalization method and decimal point normalization method. In the study, least squares support vector machine model and neural network model have been used as the prediction model of the research. According to the results, considering the effect of normalization methods on prediction accuracy and error percentages, it has been found that the outcomes of least squares support vector machine model had better outcomes than neural network model. At the same time, it has been observed that for both least squares support vector machine model and neural network model, the best outcomes have been obtained as a result of the preliminary data processing processes performed with decimal point, min-max and Z-Score normalization methods respectively. Nawi, Atomi and Rehman (2013), have analyzed the effect of three different normalization methods on classification accuracy. For this purpose, they have tested min-max normalization method, Z-Score Normalization method and decimal scaling method. According to the results of the research, it has been found that different normalization methods have provided better outcomes under different conditions and in general the process of normalization has improved the accuracy of artificial neural network classifier at least 95%. Suma, Renjith, Ashok and Judy (2016), have compared the classification accuracy outcomes of discriminant analysis, support vector machine, artificial neural network, naive Bayes and decision tree models by applying different normalization methods. For this purpose, Z-Score Normalization method and min-max normalization method have been used. According to the results of the research, it has been observed that Z-Score Normalization method have provided better outcomes in terms of classification accuracy for all models compared to min-max normalization method.

While determining the normalization method to be used as part of any research, taking the general structure of the data set, sample size and the features of the activation function to be used into account may be considered as the best approach. The fourth factor that should be considered while determining the normalization method to be used is the algorithm that will be used in training stage. In this regard, the selected training function, number of layers, number of iterations and number of nodes have also some importance. For comparing normalization methods, the features belonging to the analysis should be kept constant and the methods should be compared accordingly. After setting the constant parameters, as much as possible normalization method should be tested on the relevant data set and the method providing the best outcome should be selected.

Regarding the wholistic analysis of the contribution of different normalization methods, which were applied on different sample sizes as part of ANN model, on the variance and classification accuracy, it was concluded that the best results were obtained after normalizing via adjusted min-max method. Getting good results at lowest sample size indicates the problem of overfitting. It can be said that the risk of overfitting occurrence is quite high if the developed model works too much on the training set and starts to act by rote or if the training set is too monotonous. Overfitting occurs when the model perceives the noise and random fluctuations of the training data as a concept and learns them. The problem is the noise and fluctuations perceived as concepts will not be valid for a new data, which will affect the generalization ability of the models negatively (Haykin, 1999; Holmstrom & Koistinen, 1992). It is possible to overcome overfitting problem by cross validation method, where data set is divided into pieces to form different training-test pairs and running the model on various data. Overfitting

problem may also be prevented by developing a simpler model and allowing the model to predict. Reducing the number of iterations and removing the nodes that makes least contribution to the prediction power are the other methods that can be used in solving overfitting problem (Haykin, 1999; Holmstrom & Koistinen, 1992; Hua, Lowey, Xiong, & Dougherty, 2006; Zur, Jiang, Pesce, & Drukker, 2009).

Related to the subject, a comparison study, including sigmoid normalization method and other normalization methods that are frequently used in the literature, may be conducted in the future using a data set related to educational sciences. Due to the nature of artificial neural networks outcomes obtained from Matlab software differentiate when the model is rerun. This is due to the fact that the weight values are randomly determined at random, or at a certain interval, according to a given distribution (i.e. Gaussian). As a matter of fact, in case of reconducting the analysis with the same data set, without changing any parameter, some differences may be observed in the outcomes because training, test and validation data sets are randomly determined by the program. This is seen as the other important limitation of the research.

#### 4.1. Limitation of the Research

Sigmoid normalization method could not be tested in the researches since only zero and one type outputs can be generated as a result of sigmoid normalization method. Failure to cover sigmoid normalization method constitutes a limitation of the research.

#### 4.2. Superiority of the Research

In addition to analyze the effect of normalization methods for numeric outputs, the performance of normalization method used in case of categoric output variable was also analyzed as part of the study, which is seen as a superiority of the research. In addition, implementing artificial neural network methods into the education area and performing the analysis by taking different sample sizes into account are considered as the other superiorities of the study.

#### ORCID

Gökhan AKSU  <https://orcid.org/0000-0003-2563-6112>

Cem Oktay GÜZELLER  <https://orcid.org/0000-0002-2700-3565>

Mehmet Taha ESER  <https://orcid.org/0000-0001-7031-1953>

#### 5. REFERENCES


- Aksu, G., & Do an, N. (2018). Veri Madencili inde Kullanılan Ö renme Yöntemlerinin Farklı Ko ullar Altında Kar ıla tırılması, *Ankara Üniversitesi E itim Bilimleri Fakültesi Dergisi*, 51(3), 71-100.
- Ali, A. & Senan, N. (2017). *The Effect of Normalization in VIOLENCE Video Classification Performance*. IOP Conf. Ser.: Mater. Sci. Eng. 226 012082.
- Anderson, J. A. (1990). Data Representation in Neural Networks, *AI Expert*.
- Ayalakshmi, T., & Santhakumaran, A. (2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, 3(1), 1793-8201.
- Azadeh, M., Sheikhalishahi, M., Tabesh, A., & Negahban (2011). The Effects of Pre-Processing Methods on Forecasting Improvement of Artificial Neural Networks, *Australian Journal of Basic and Applied Sciences*, 5(6), 570-580.
- Azimi-Sadjadi, M.R. & Stricker, S.A. (1994). "Detection and Classification of Buried Dielectric Anomalies Using Neural Networks Further Results," *IEEE Trans. Instrumentations and Measurement*, 43, pp. 34-39.
- Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

- Cihan, P., Kalıpsız, O., & Gökçe, E. (2017). Hayvan Hastalığı'nı Te hisinde Normalizasyon Tekniklerinin Yapay Sinir A 1 Performansına Etkisi [Effect of Normalization Techniques on Artificial Neural Network and Feature Selection Performance in Animal Disease Diagnosis]. *e-Turkish Studies (elektronik)*, 12(11), 59-70, 2017.
- Davydov, M.V., Osipov, A.N., Kilin, S.Y. & Kulchitsky, V.A. (2018). Neural Network Structures: Current and Future States. *Open semantic technologies for intelligent systems*, 259-264.
- Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., & Meester, L.E. (2005). *A modern introduction to probability and statistics: Understanding why and how*. United States: Springer-Verlag London Limited.
- Deveci, M. (2012). *Yapay Sinir A ları ve Bekleme Süresinin Tahmininde Kullanılması [Artificial Neural Networks and Used of Waiting Time Estimation]*. Unpublished Master Dissertation, Gazi Üniversitesi Sosyal Bilimleri Enstitüsü, Ankara.
- Elmas, Ç. (2003). *Yapay Sinir A ları*, Birinci Baskı, Ankara: Seçkin Yayıncılık.
- Famili, A., Shen, W., Weber, R., & Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis*, 1, 3-23.
- Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2), 87-107.
- Fraenkel, J.R., & Wallen, N.E. (2006). *How to design and evaluate research in education* (6th ed.). New York, NY: McGraw-Hill.
- Gardner, M. W., & Dorling, S. R. (1998). Artificial Neural Networks (The Multilayer Perceptron) - A Review of Applications in the Atmospheric Sciences. *Atmospheric Environment*, 32, 2627-2636.
- Gerasimovic, M., Stanojevic, L., Bugaric, U., Miljkovic, Z., & Veljovic, A. (2011). Using Artificial Neural Networks for Predictive Modeling of Graduates' Professional Choice. *The New Educational Review*, 23, 175- 188.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Gonzalez, J.M., & DesJardins, S.L. (2002). Artificial neural networks: A new approach to predicting application behaviour. *Research in Higher Education*, 43(2), 235-258
- Gschwind, M. (2007). Predicting Late Payments: A Study in Tenant Behavior Using Data Mining Techniques. *The Journal of Real Estate Portfolio Management*, 13(3), 269-288.
- Hagan, M.T., Demuth, H.B., Beale, M.H., & Jesus, O. (2014). *Neural Network Design*, Boston: PWS Publishing Co.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.
- Hayashi, Y., Hsieh, M-H., & Setiono, R. (2009). Predicting Consumer Preference for Fast-Food Franchises: A Data Mining Approach. *The Journal of the Operational Research Society*, 60(9), 1221-1229.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ.
- Holmstrom, L., & Koistinen, P. (1992). Using additive noise in back-propagation training. *IEEE Trans. Neural Networks*, 3, 24-38
- Hua, J.P., Lowey, J., Xiong, Z., & Dougherty, E.R. (2006). Noise-injected neural networks show promise for use on small-sample expression data. *BMS Bioinform.* 7 (Art. no. 274).
- Hu, X. (2003). DB-H Reduction: A Data Preprocessing Algorithm for Data Mining Applications. *Applied Math. Letters*, 16, 889- 895.
- Hunt, K.J., Sbarbaro, D., Bikowski, R., & Gawthrop, P.J. (1992) "Neural Networks for Control Systems - A Survey. *Automatica*, 28, pp. 1083-1112.

- Karasar, N. (2009). *Bilimsel Ara tırma Yöntemi [Scientific Research Method]*. Ankara: Nobel Yayıncılık.
- Klein, B.D., & Rossin, D.F. (1999). Data Quality in Neural Network Models: Effect of Error Rate and Magnitude of Error on Predictive Accuracy. *OMEGA, The Int. J. Management Science*, 27, pp. 569-582.
- Kriesel, D. (2007). *A Brief Introduction to Neural Networks*. Available at <http://www.dkriesel.com/media/science/neuronalenetze-en-zeta2-2col-dkrieselcom.pdf>
- Krycha, K. A., & Wagner, U. (1999). Applications of Artificial Neural Networks in Management Science: A Survey. *J. Retailing and Consumer Services*, 6, pp. 185-203,
- Lawrance, J. (1991). Data Preparation for a Neural Network, *AI Expert*. 6 (11), 34-41.
- Lou, M. (1993). *Preprocessing Data for Neural Networks*. Technical Analysis of Stocks & Commodities Magazine, Oct.
- Mannila, H. (1996). Data mining: machine learning, statistics, and databases, *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management*, Stockholm, Sweden, June 18–20, 1996.
- Matlab (2002). *Matlab, Version 6.5*. Natick, MA: The Mathworks Inc.,
- Mustaffa, Z., & Yusof, Y. (2011). *A Comparison of Normalization Techniques in Predicting Dengue Outbreak*. International Conference on Business and Economics Research, Vol.1 IACSIT Press, Kuala Lumpur, Malaysia
- Namin, A. H., Leboeuf, K., Wu, H., & Ahmadi, M. (2009). Artificial Neural Networks Activation Function HDL Coder, *Proceedings of IEEE International Conference on Electro/Information Technology*, Ontario, Canada, 7-9 June, 2009.
- Narendra, K. S., & Parthasarathy, K. (1990). Identification and Control of Dynamic Systems Using Neural Networks. *IEEE Trans. Neural Networks*, 1, pp. 4-27.
- Nawi, N. M., Atomi, W. H., Rehman, M. Z. (2013). The Effect of Data Pre-Processing on Optimized Training of Artificial Neural Networks. *Procedia Technology*, 11, 32-39.
- Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in Data mining: An Overview. *International Journal of P2P Network Trends and Technology (IJPTT)*, 4(8), 369-374.
- OECD, (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific and Technical Activities, OECD Publishing, Paris.
- O’Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*, arXiv:1511.08458 [cs. NE], November.
- Özkan, A.O. (2017). Effect of Normalization Techniques on Multilayer Perceptron Neural Network Classification Performance for Rheumatoid Arthritis Disease Diagnosis. *International Journal of Trend Scientific Research and Development*. Volume 1, Issue 6.
- Öztemel, E. (2003), *Yapay Sinir A ları [Artificial Neural Networks]*, stanbul: Papatya Yayıncılık.
- Rafiq, M.Y., Bugmann, G., & Easterbrook, D.J. (2001). Neural Network Design for Engineering Applications. *Computers & Structures*, 79, pp. 1541-1552.
- Ravid, R. (2011). *Practical statistics for educators* (fourth edition). United States: Rowman & Littlefield Publishers.
- Redman, T. C. (1992). *Data Quality: Management and Technology*. New York: Bantam Books.
- Ripley, B.D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.
- Romero, C., Ventura, S. (2011). Educational data mining: a review of the state-of-the-art”, *IEEE Trans. Syst. Man Cybernet. C Appl. Rev.*, 40(6), 601–618.
- Roussas, G. (2007). *Introduction to probability (first edition)*. United States: Elsevier Academic Press.
- Rumelhart, D.E. (1994). The Basic Ideas in Neural Networks. *Comm. ACM*, 37, pp. 87-92.

- Panigrahi, S., & Behera, H. S. (2013). Effect of Normalization Techniques on Univariate Time Series Forecasting using Evolutionary Higher Order Neural Network. *International Journal of Engineering and Advanced Technology*, 3(2), 280-285.
- Sattler, K.U., & Schallehn, E. (2001). A Data Preparation Framework Based on a Multidatabase Language. *Proc. Int'l Symp. Database Eng. & Applications*, pp. 219-228.
- Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61, 85-117.
- Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. (2010). A Comparison of Logistic Regression, Neural Networks, and Classification Trees Predicting Success of Actuarial Students. *Journal of Education for Business*, 85(5), 258-263.
- Silva, C.S. and Fonseca, J.M. (2017). Educational Data Mining: a literature review. *Advances in Intelligent Systems and Computing*, 2-9.
- Stein, R. (1993). Selecting data for neural networks, *AI Expert*.
- Suma, V. R., Renjith, S., Ashok, S., & Judy, M. V. (2016). Analytical Study of Selected Classification Algorithms for Clinical Dataset. *Indian Journal of Science and Technology*, 9(11), 1-9, DOI: 10.17485/ijst/2016/v9i11/67151.
- Upadhyay, N. (2016). Educational Data Mining by Using Neural Network. *International Journal of Computer Applications Technology and Research*, 5(2), 104-109.
- Uslu, M. (2013). *Yapay Sinir A ları ile Siniflandırma [Classification with Artificial Neural Networks]*, İeri statistik Projeleri I [Advanced Statistics Projects I]. Hacettepe Üniversitesi Fen Fakültesi statistik Bölümü, Ankara.
- Vijayabhanu, R. & Radha, V. (2013). Dynamic Score Normalization Technique using Mahalonobis Distance to Predict the Level of COD for an Anaerobic Wastewater Treatment System. *The International Journal of Computer Science & Applications*. 2(3), May 2013, ISSN – 2278-1080.
- Yavuz, S., & Deveci, M. (2012). statiksel Normalizasyon Tekniklerinin Yapay Sinir A ın Performansına Etkisi. [The Effect of Statistical Normalization Techniques on The Performance of Artificial Neural Network], *Erciyes University Journal of Faculty of Economics and Administrative Sciences*, 40, 167-187.
- Yu, L., Wang, S., & Lai, K.K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Trans. Knowl. Data Eng.*, 18, 217–230.
- Wang, F., Devabhaktuni, V.K., Xi, C., & Zhang, Q. (1998). Neural Network Structures and Training Algorithms for RF and Microwave Applications. *John Wiley & Sons, Inc. Int J RF and Microwave CAE*, 9, 216-240.
- Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., Seong, H. Y. (2009). *Predicting NDUM Student's Academic Performance Using Data Mining Techniques*, The Second International Conference on Computer and Electrical Engineering, Dubai, United Arab Emirates, 28-30 December, 2009.
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17, 375-381.
- Zur, R.M., Jiang, Y.L., Pesce, L.L., & Drukker, K. (2009). Noise injection for training artificial neural networks: a comparison with weight decay and early stopping. *Med. Phys.*, 36(10), 4810–4818.

## An Empirical Study for the Statistical Adjustment of Rater Bias

Mustafa İhan \*

<sup>1</sup> Dicle University, Ziya Gokalp Education Faculty, Department of Mathematics and Science Education, Diyarbakir, Turkey

### ARTICLE HISTORY

Received: 28 February 2019

Revised: 24 April 2019

Accepted: 30 April 2019

### KEYWORDS

Bias adjustment,  
Rater bias,  
Many facet Rasch model

**Abstract:** This study investigated the effectiveness of statistical adjustments applied to rater bias in many-facet Rasch analysis. Some changes were first made in the dataset that did not include *rater* × *examinee* bias to cause to have *rater* × *examinee* bias. Later, bias adjustment was applied to rater bias included in the data file, and the effectiveness of the statistical adjustment was further examined. The outcomes pertaining to the datasets with and without bias, and to which the bias adjustment was applied, were compared. It was concluded that diversities created by *rater* × *examinee* bias in examinees' ability estimation, item difficulty indices and measures of rater severity and leniency were, to a large extent, eliminated by bias adjustment. This result indicates that the bias adjustment using many-facet Rasch analysis is a viable way to control rater bias.

## 1. INTRODUCTION

The tests used in education and psychology are categorized as objective tests and subjective tests by the type of scoring (McNamara, Erlandson, & McNamara, 2013). Objective tests consist of the items based on selecting a correct answer from the options provided, such as multiple-choice, true-false, and matching questions (Haladyana, 1997). Scores on objective test do not vary according to the rater, which means that objective tests have higher rater reliability. These tests can be rated easily and quickly, and so they are budget-friendly (Bennett, Ward, Rock, & LaHart, 1990). Subjective tests, on the other hand, use the items that require students to construct their responses, such as open-ended questions. Subjective tests scores tend to vary according to the rater (Bennett, 1991). For this reason, in subjective tests raters are one of the variability sources that affect students' test scores (Eckes, 2005). Rater-based factors are undesired and systematic rater behaviors that lead to the inclusion of variance irrelevant to the construct being measured to students' test scores, and are known as rater effect (Eckes, 2005; Hoyt, 2000). Rater effect includes rater severity and leniency, halo effect, central tendency

**CONTACT:** Mustafa İLHAN ✉ [mustafailhan21@gmail.com](mailto:mustafailhan21@gmail.com) 📍 Dicle University, Ziya Gokalp Education Faculty, Department of Mathematics and Science Education, Diyarbakir, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

effect and range restriction (Saal, Downey, & Lahey, 1980). Bias is also a form of rater effect (Myford & Wolfe, 2004).

### 1.1. Rater Bias (Differential Rater Severity/Leniency)

Bias is raters' unexpectedly severe or lenient scoring regarding an aspect of the assessment process (Knoch, Read & von Randow, 2007). Rater bias can be related to examinees (*rater x examinee*), items (*rater x item*) or both (*rater x examinee x item*). "*Rater x examinee*" bias refers to raters' tendency to give higher or lower scores based on students' prior performances or demographics such as gender, age, and cultural factors (Aubin, St-Onge, & Renaud; 2018; Kumar, 2005). "*Rater x item*" bias refers to whether raters grade all the items on a test with the same severity or leniency (Haiyang, 2010). "*Rater x examinee x item*" bias refers to raters' assignment of lower or higher scores than expected to some students for their performance on some items.

In order to avoid rater bias, rater training (Knoch, Read, & von Randow, 2007; Fahim & Bijani, 2011) and blind scoring using rubrics have been suggested (Hogan & Murphy, 2007). Studies, however, have shown that rater bias can persist despite these precautions. For example, Kondo Brown (2002) investigated whether teachers who received rater training are biased towards some candidates or certain criteria while evaluating university students' Japanese second language writing ability. The performance of 234 university students was graded by three raters using an analytic rubric. The study results showed significant interactions between raters and students, and rater and rating criteria that indicated bias. In a different study by İlhan (2015), 104 students' responses to eight open-ended mathematics questions were graded by seven raters. Despite the training provided to the raters and using a rubric for the scoring, rater bias was not entirely eliminated. Knoch, Read and von Randow (2007) compared the effectiveness of face-to-face and online rater training. They also concluded that rater training cannot completely prevent bias in scoring.

### 1.2. Statistical Adjustment of Rater Bias

The fact that the rater bias persists in spite of using rubrics, blind scoring and rater training brings up the question of whether bias can be statistically adjusted. Indeed, there are studies in the literature on the statistical correction of the rater effects. Raymond and Houston (1990) conducted a study with the purpose of determining and correcting for rater effects in performance assessment. In the research four different procedures; ordinary least squares, weighted least squares, the Rasch model (with a two facets design that includes only raters and examinees and that provides results similar to the Wright and Masters (1982) rating scale model) and data imputation via the E-M algorithm were considered on a simulated data set. The results of the research showed that each of the methods yields more accurate estimates of true levels of performance than the classical approach of summing observed ratings. In the Houston, Raymond and Svec's (1991) study the methods of ordinary least squares, weighted least squares and imputation of the missing data were examined for correcting rater severity and leniency. In the study, simulation data was used and root-mean-squared-error (RMSE) was employed in order to assess the accuracy of the methods in estimating true scores. The research results indicated that the three correction methods used consistently outperformed the procedure of averaging the observed ratings. In another study by Raymond and Viswesvaran (1993), it was aimed to elucidate a simple and flexible method to statistically control for specific types of rating error. In accordance with this purpose, three different models namely ordinary least squares; weighted least squares; and ordinary least squares, subsequent to applying a logistic transformation to observed ratings were performed to data obtained from an oral examination where each of 115 examinees graded by four raters. The study results revealed that the models used for correction of ratings increases reliability. In addition to the methods used in the

researches listed, the literature also includes an approach based on the many-facet Rasch model (MFRM) proposed by Linacre (2018) to adjust rater bias statistically; however, there are no empirical studies of its effectiveness.

### 1.3. Aim of the Study

This study aimed to test the effectiveness of MFRM statistical adjustment of rater bias empirically. It investigates the effects of statistical bias adjustment on estimating the abilities of examinees, on the difficulty indices of items and on measures of rater severity/leniency.

## 2. METHOD

### 2.1. Model of the Study

This study focused on testing a model for the process of bias adjustment, and was therefore designed as basic research. Basic research, rather than seeking answers to real-life problems, addresses issues that offer theoretical contributions to science, build theories and generate new knowledge (Connaway & Powell, 2010). Basic research also formalizes theories and tests hypotheses involving abstract concepts (Bailey, 1994).

### 2.2. Participants

The participants included 95 eighth-grade students, of whom, 49 (51.58%) were female, and 46 (48.42%) were male. Three mathematics teachers graded their responses to open-ended questions.

### 2.3. Data Collection Tools

In the study two data collection tools was used. The first was the Mathematics Achievement Test developed by Ihan (2016). This test contains six open-ended questions. According to results reported by Ihan (2016), the test had a one-dimensional structure. It explained 31.18% of variance ratio, and the factor loads of its items were found to range between .51 and .64.

This study's second data collection tool was a rubric used to grade the students' responses to the open-ended questions. This rubric was also developed by Ihan (2016). The rubric has a holistic structure and four categories: *inadequate* (0), *needs to be developed* (1), *dood* (2) and *very good* (3). Ihan (2016) indicated that these categories were intended to reflect the adequacy of responses on five levels: understanding of the problem, method of solving the problem, the processes carried out to solve the problem, the accuracy of the results obtained and how the solution was obtained.

### 2.4. Data Collection, Psychometric Characteristics, and Analysis

Data were collected in the spring term of 2018. Administering the achievement test to the 95 eighth-grade students was the first stage of data collection. Their responses were graded by three mathematics teachers. The rubric used in the study had been introduced to the raters beforehand. The raters were also told that they should rate all answers to the one question before moving to the next question, and that they should not include variables outside the construct measured, such as appealing handwriting and spatial organization of the responses. After the rating, the data were analyzed using MFRM. FACETS software was used for the analysis.

Statistical indicators of whether the Rasch analysis assumptions were met were investigated firstly. Rasch analysis has three assumptions: unidimensionality, local independence, and model-data fit (DeMars, 2010). However, there is no need to test each assumption one by one since they are all related. That is to say, model-data fit indicates that the unidimensionality assumption has been met (Lee, Peterson, & Dixon, 2010), which indicates that there is no problem with local independence (Nandakumar & Ackerman, 2004). Therefore, the fundamental assumption that needs to be tested is whether there is model-data fit (Güler, Ihan,



Güneyli, & Demir, 2017). This assumption is tested by examining standardized residuals. The number of standardized residuals outside the  $\pm 2$  range should not exceed 5% of the total number of data, and those outside  $\pm 3$  should not exceed 1%, according to Linacre (2018). In this study, the total number of data was 1,710 since it involved 95 students, six items and three raters ( $95 \times 6 \times 3$ ). The number of standardized residuals outside the range of  $\pm 2$  was found to be 76 (4.44%) and the number of standardized residuals outside the range of  $\pm 3$  was found to be 16 (0.94%). This indicated adequate model-data fit, and that the assumptions of Rasch analysis had been met.

After determining that the assumptions were met, the psychometric characteristics of the study data were investigated. The results for reliability and model-data fit in MFRM are shown in Table 1. The infit and outfit indices in all three of the examinee, item and rater facets were within the range of .5 and 1.5, the recommended criteria for their interpretation (Wright & Linacre, 1994). These fit indices indicate model-data fit and the validity of the measurements.

**Table 1.** Results for reliability and model-data fit in MFRM.

Facet	Infit	Outfit	Separation Index	Reliability	df	Chi square
Examinee	.99	1.01	2.19	.83	94	443.00**
Item	.99	1.01	13.20	.99	5	857.20**
Rater	.99	1.01	5.51	.97	2	62.40**

\*\*  $p < .001$

Table 1 shows that the chi-square value for the rater facet was significant, and that the reliability coefficient and separation index were high. This indicated a significant difference between the raters' severity and leniency. Despite this difference, the values reported for the facets of item and examinee indicated that the measures were reliable because the chi-square values for the facets of examinee and item were significant, the reliability coefficients exceeded .80, and the separation indices were higher than 2 (Linacre, 2012). Thus, the students' performances on the different test items can be rated independently, and examinees with different mathematical performances were distinguished with high reliability.

Following the psychometric investigation of the study data, the datasets were prepared for bias adjustment. The comparison of the analysis outcomes obtained from a dataset not involving rater bias and the analysis outcomes reached in case of the inclusion of bias in this dataset and the adjustment of the bias included was thought to be the most convenient way to set forth the effectiveness of the statistical adjustment applied. For this reason; while preparing the dataset for the bias adjustment, the original rater biases were excluded from the dataset to create a dataset with no apparent rater bias—the unbiased dataset. The results of analysis indicated significant relationships between rater 1 and examinee 84 (bias size=1.57,  $t=2.66$ ) and rater 2 and examinee 23 (bias size=1.75,  $t=2.66$ ). These two raters graded two examinees mentioned more leniently than expected. Therefore, the data for examinees 84 and 23 were excluded from the dataset, creating a dataset where three raters graded 93 students' responses to six open-ended mathematics questions and no rater bias. This dataset's measurements of examinees' ability levels, item difficulty indices, rater severity/leniency were used as the criteria for the effectiveness of bias adjustment.

In the second stage of the testing the effectiveness of bias adjustment, some changes were made in the dataset so that it would contain "rater  $\times$  examinee" bias. The grading of rater 1 for examinees 1 to 10 and rater 2 for examinees 11 to 20 were increased by one in some parts of the test and by two in others, creating a dataset where bias was encountered in 20 of the 279 [(93 examinees)  $\times$  (3 raters)] possible interactions between raters and examinees.

In the final stage, the bias adjustment formula was applied to the biases included in the dataset. A fourth facet, bias adjustment, was incorporated in the analysis, along with the three facets of rater, examinee and item. In this facet, *rater* × *examinee* biases in the dataset were listed, and bias adjustment was applied to the grading of rater 1 for examinees 1 to 10 and rater 2 for examinees 11 to 20. No other bias adjustments were done. The *rater* × *examinee* interactions to which the bias adjustment was applied were encoded as 1, and the *rater* × *examinee* interactions where bias adjustment was not necessary were encoded as 2. Thus, syntax containing the four facets of rater, examinee, item and bias adjustment were prepared for many-facet Rasch analysis. At this point, the comparison of the three datasets proceeded.

In this study, the consistency between the ability estimations in the unbiased, biased and adjusted dataset was examined using Pearson's product-moment correlation and the paired samples *t*-test. Correlation analysis and the *t*-test were done using IBM SPSS 20 software. The effect of bias adjustment on item difficulty indices and rater severity/leniency could not be statistically determined since the number of items was limited to six, and the number of raters to three. It was possible only to investigate how close item difficulties and measures regarding raters were to the values in the unbiased dataset.

### 3. RESULT and DISCUSSION

This section includes the study's results. The ability estimations in the unbiased, biased and adjusted datasets are shown in Table 2. As Table 2 shows, there were significant differences between ability estimations in the unbiased and biased datasets. These differences were valid for almost all participants, but were more explicit for the first 20 students who served as a source for the *rater* × *examinee* bias. Table 2 showed that the ability estimations after the application of statistical adjustment to the *rater* × *examinee* bias were significantly closer to the ability scores in the unbiased dataset. This means that the effect of rater bias on the examinee's ability estimations can be controlled by bias adjustment. However, in order to reach a more powerful judgement, the relationship between the ability estimations in the unbiased, biased and bias adjusted datasets needed to be tested statistically. In order to determine statistically how much bias adjustment brings the ability estimations closer to the ability estimations in the unbiased dataset, correlation analysis and the paired samples *t*-test were done. Their outcomes are shown in Table 3.

Table 3 shows that there was a positive, powerful and significant relationship between ability estimations in the unbiased dataset and biased datasets [ $r=.896, p<.001$ ]. However, it should not be overlooked that there was a significant difference between the mean ability scores in these two datasets [ $t_{(92)}=5.03, p<.001$ ]. Better to say, *rater* × *examinee* bias did not have a great impact on the ordering of the examinees' ability levels, but significantly affected their ability estimations. A comparison of the ability estimations in the bias adjusted and unbiased datasets found a perfect positive relationship [ $r=.996, p<.001$ ]. No significant difference was found between the ability estimations in the two datasets [ $t_{(92)}=1.11, p>.05$ ]. This indicated that the effects created by the *rater* × *examinee* bias on the ability estimations can be, to a large extent, eliminated by bias adjustment. The effect of bias adjustment on item difficulty indices and rater severity and leniency measurements are shown in Table 4.

**Table 2.** Ability estimations in the unbiased, biased and adjusted bias datasets.

Examinee Number	No bias	Bias	Adjusted bias	Examinee Number	No bias	Bias	Adjusted bias	Examinee Number	No bias	Bias	Adjusted bias
E1	-0.08	0.58	0.00	E32	-1.04	-0.97	-1.06	E63	0.28	0.25	0.28
E2	1.46	1.73	1.44	E33	-0.44	-0.42	-0.46	E64	-1.26	-1.17	-1.29
E3	-0.08	0.58	0.00	E34	-0.17	-0.17	-0.18	E65	-0.63	-0.60	-0.65
E4	-0.54	0.25	-0.48	E35	-0.08	-0.08	-0.08	E66	-0.83	-0.78	-0.85
E5	-3.15	-1.07	-2.78	E36	-0.73	-0.69	-0.75	E67	-0.35	-0.33	-0.36
E6	-1.50	-0.42	-1.58	E37	-1.04	-0.97	-1.06	E68	-1.50	-1.39	-1.53
E7	-1.38	-0.25	-1.28	E38	-1.26	-1.17	-1.29	E69	0.46	0.41	0.46
E8	-0.83	0.08	-0.73	E39	-0.54	-0.51	-0.55	E70	-0.08	-0.08	-0.08
E9	-0.93	-0.17	-1.13	E40	-2.21	-2.06	-2.26	E71	-0.35	-0.33	-0.36
E10	0.01	0.58	0.00	E41	0.37	0.33	0.37	E72	-0.63	-0.60	-0.65
E11	0.10	0.58	0.09	E42	-1.76	-1.63	-1.79	E73	-0.26	-0.25	-0.27
E12	0.10	0.50	-0.03	E43	0.01	0.00	0.01	E74	-1.04	-0.97	-1.06
E13	-0.63	0.08	-0.65	E44	-0.35	-0.33	-0.36	E75	0.73	0.67	0.74
E14	0.28	0.67	0.21	E45	-0.63	-0.60	-0.65	E76	-0.54	-0.51	-0.55
E15	-1.76	-0.33	-1.35	E46	1.23	1.14	1.26	E77	-0.83	-0.78	-0.85
E16	0.10	0.58	0.09	E47	-1.90	-1.77	-1.94	E78	-0.93	-0.87	-0.96
E17	-0.83	-0.08	-0.91	E48	-1.26	-1.17	-1.29	E79	-0.08	-0.08	-0.08
E18	-0.08	0.41	-0.15	E49	-1.15	-1.07	-1.17	E80	-0.83	-0.78	-0.85
E19	0.28	0.67	0.21	E50	-0.73	-0.69	-0.75	E81	-0.08	-0.08	-0.08
E20	-0.17	0.33	-0.27	E51	-0.54	-0.51	-0.55	E82	-0.26	-0.25	-0.27
E21	-1.15	-1.07	-1.17	E52	-0.54	-0.51	-0.55	E83	-0.17	-0.17	-0.18
E22	-0.44	-0.42	-0.46	E53	-0.35	-0.33	-0.36	E84	-0.54	-0.51	-0.55
E23	0.46	0.41	0.46	E54	-0.35	-0.33	-0.36	E85	-2.21	-2.06	-2.26
E24	-2.21	-2.06	-2.26	E55	-0.83	-0.78	-0.85	E86	0.73	0.67	0.74
E25	-2.60	-2.43	-2.65	E56	-1.62	-1.51	-1.66	E87	0.01	0.00	0.01
E26	0.64	0.58	0.65	E57	-0.63	-0.60	-0.65	E88	-0.08	-0.08	-0.08
E27	0.19	0.16	0.19	E58	-0.93	-0.87	-0.96	E89	0.64	0.58	0.65
E28	-0.17	-0.17	-0.18	E59	-0.73	-0.69	-0.75	E90	0.55	0.50	0.55
E29	0.10	0.08	0.10	E60	-0.54	-0.51	-0.55	E91	-1.04	-0.97	-1.06
E30	-0.26	-0.25	-0.27	E61	-0.93	-0.87	-0.96	E92	-1.50	-1.39	-1.53
E31	-0.93	-0.87	-0.96	E62	-0.35	-0.33	-0.36	E93	-0.44	-0.42	-0.46

**Table 3.** Correlation analysis and paired samples *t*-test results for the comparison of the ability estimations in the unbiased, biased and adjusted datasets.

Comparison	Dataset	Mean (Logit)	Standard Deviation	r	df	t
No bias – Bias	No bias	-.55	.80	.896**	92	5.03**
	Bias	-.36	.75			
No bias – Adjusted bias	No bias	-.55	.80	.996**	92	1.11*
	Adjusted bias	-.56	.79			

\*  $p > .05$ , \*\*  $p < .001$

**Table 4.** Item difficulty indices and rater severity and leniency measurements in the unbiased, biased and adjusted datasets.

	Item Difficulty Indices				Rater Severity/ Leniency Measures		
	No bias	Bias	Adjusted bias		No bias	Bias	Adjusted bias
I1	.86	.74	.85	R1	.00	-.10	-.01
I2	-1.40	-1.21	-1.38				
I3	.53	.48	.53	R2	-.31	-.33	-.30
I4	.14	.13	.13				
I5	-1.29	-1.16	-1.33	R3	.30	.43	.31
I6	1.16	1.02	1.20				

Table 4 shows that the item difficulties and rater measurements in the unbiased and biased datasets were quite different. On the other hand, the item difficulties and rater measurements in the adjusted dataset were extremely close to those of the unbiased dataset. In other words, the differences caused by *rater* × *examinee* bias in the item difficulty indices and rater severity and leniency measurements were largely eliminated by bias adjustment, although not entirely.

#### 4. CONCLUSION

This study investigated the effectiveness of MFRM statistical adjustment of rater biases. Its dataset, which did not include any *rater* × *examinee* bias, was altered to involve *rater* × *examinee* bias. Then, bias adjustment was applied to the rater biases included in the dataset, and the effectiveness of the statistical adjustment was tested. Ability estimations, item difficulties, and rater measurements in the bias adjusted dataset were compared to those in the unbiased dataset. The correlation analysis results failed to indicate complete consistency, despite a strong relationship between ability estimations in the dataset that did not include *rater* × *examinee* bias and the biased dataset. On the other hand, it was determined that there was excellent consistency between the ability estimations calculated after bias adjustment and ability estimations in the unbiased dataset.

A significant difference was also found between the ability estimations in the dataset that did not include *rater* × *examinee* bias and the ability estimations in the biased dataset. No significant differences were found between ability estimations in the bias adjusted dataset and those in the unbiased dataset. All these results reveal that the effects of rater biases on examinees' ability estimations can be eliminated by bias adjustment. This was also the case for item difficulty indices and rater severity and leniency measurements. A comparison of the three datasets determined that differences caused by *rater* × *examinee* bias on item difficulties and rater measurements were almost entirely eliminated.

#### 5. IMPLICATIONS for PRACTICE

This study's results indicate that MFRM bias adjustment can serve as a way to minimize the effects of rater bias. However, it should be underlined that this does not mean that statistical bias adjustment can replace other methods of reducing rater bias such as rater training, blind scoring or using rubrics. The most accurate interpretation based on research results is that statistical adjustment should be performed for observed biases when rater bias occurs despite precautions such as using rubrics or training raters. More clearly, just as statistical controls can be used to support physical controls, but not replace them in scientific researches, bias adjustment should be considered a way to support rater training, blind scoring or the use rubrics, not as an alternative to them.

#### ORCID

Mustafa Ihan  <https://orcid.org/0000-0003-1804-002X>

## 6. REFERENCES

- Aubin, A. S., St-Onge, C., & Renaud, J. S. (2018). Detecting rater bias using a person-fit statistic: A Monte Carlo simulation study. *Perspectives on Medical Education*, 7(2), 83-92. <http://dx.doi.org/10.1007/s40037-017-0391-8>
- Bailey, K. (1994). *Methods of social research*. New York: The Free.
- Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, 2, 1-46. <http://dx.doi.org/10.1002/j.2333-8504.1991.tb01429.x>
- Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). Toward a framework for constructed response items. *ETS Research Report Series*, 1, 1 - 29. <http://dx.doi.org/10.1002/j.2333-8504.1990.tb01348.x>
- Connaway, L. S., & Powell, R. R. (2010). *Basic research methods for librarians*. Santa Barbara, CA: Libraries Unlimited.
- DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221. [http://dx.doi.org/10.1207/s15434311laq0203\\_2](http://dx.doi.org/10.1207/s15434311laq0203_2)
- 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>
- Güler, N., İlhan, M., Güneşli, A., & Demir, S. (2017). An evaluation of the psychometric properties of three different forms of Daly and Miller's writing apprehension test through Rasch analysis. *Educational Sciences: Theory & Practice*, 17(3), 721-744. <http://dx.doi.org/10.12738/estp.2017.3.0051>
- Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, 33(2), 87 - 102. Retrieved from <http://www.celea.org.cn/teic/90/10060807.pdf>
- Haladyana, T. M. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights, MA: Allyn & Bacon.
- Hogan, T. P., & Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427-441. <http://dx.doi.org/10.1080/08957340701580736>
- Houston, W. M., Raymond, M.R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement*, 15(4), 409-421. <http://dx.doi.org/10.1177/014662169101500411>
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, 5(1), 64-86. <http://dx.doi.org/10.1037/1082-989X.5.1.64>
- İlhan, M. (2015). *The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet Rasch model*. Doctoral dissertation, Gaziantep University, Gaziantep, Turkey. Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- İlhan, M. (2016). *Comparison of the ability estimations of classical test theory and the many facet Rasch model in measurements with open-ended questions*. *Hacettepe University Journal of Education*, 31(2), 346-368. <http://dx.doi.org/10.16986/HUJE.2016015182>

- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43. <http://dx.doi.org/10.1016/j.asw.2007.04.001>
- Kondo Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3 - 31. <https://doi.org/10.1191/0265532202lt218oa>
- Kumar, DSP D. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College*, 4, 1 - 15. Retrieved from <http://rmpckl.rmp.gov.my/Journal/BI/performanceappraisal.pdf>
- Lee, M., Peterson, J. J., & Dixon, A. (2010). Rasch calibration of physical activity self-efficacy and social support scale for persons with intellectual disabilities. *Research in Developmental Disabilities*, 31(4), 903-913. <http://dxdoi.org/10.1016/j.ridd.2010.02.010>
- Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial*. Retrieved from <http://www.winsteps.com/a/ftutorial2.pdf>
- Linacre, J. M. (2018). *A user's guide to FACETS Rasch-model computer programs*. Retrieved from <https://www.winsteps.com/manuals.htm>
- McNamara, J. F., Erlandson, D. A., & McNamara, M. (2013). *Measurement and evaluation: Strategies for school improvement*. New York, NY: Routledge.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227. Retrieved from [http://jimelwood.net/students/grips/tables\\_figures/myford\\_wolfe\\_2004.pdf](http://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf)
- Nandakumar, R., & Ackerman, T. A. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 93-105). Thousand Oaks, CA: Sage.
- Raymond, M. R., & Houston, W. M. (1990). *Detecting and correcting for rater effects in performance assessment* (ACT Research Rep. No. 90-14). Iowa City, American College Testing. Retrieved from [http://www.act.org/content/dam/act/unsecured/documents/ACT\\_RR90-14.pdf](http://www.act.org/content/dam/act/unsecured/documents/ACT_RR90-14.pdf)
- Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, 30(3), 253-268. <http://dx.doi.org/10.1111/j.1745-3984.1993.tb00426.x>
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428. <http://dx.doi.org/10.1037/0033-2909.88.2.413>
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370-371. Retrieved from <https://www.rasch.org/rmt/rmt83b.htm>

## Development of Perceived School Counselor Support Scale: Based on the ASCA Mindsets and Behaviors

Mehmet Akif Karaman <sup>1\*</sup>, Cemal Karada <sup>2</sup>, Javier Cavazos Vela<sup>3</sup>

<sup>1</sup> Department of Educational Sciences, Kilis 7 Aralık University, Kilis, Turkey

<sup>2</sup> Department of Educational Sciences, nönü University, Malatya, turkey

<sup>3</sup> Department of Counseling, The University of Texas Rio Grande Valley, Edinburg, TX, USA

### ARTICLE HISTORY

Received: 15 February 2019

Revised: 22 April 2019

Accepted: 04 May 2019

### KEYWORDS

School counselor support scale,  
ASCA national model,  
Developmental school counseling,  
Perceived counselor support,  
Scale development

**Abstract:** This study presents a culturally and psychometrically sound instrument of perceived school counselor support among Turkish high school students. The study has been framed using American School Counseling Association's Mindsets and Behaviors for Students Success Model to create a valuable instrument that measures students' perceptions of their school counselors' support in a different culture, society, and education system. The results of this study supported the theoretical based *Perceived School Counselor Support Scale* long and short forms providing initial and strong evidence based on internal structure and relations to other variables. Internal consistency estimates on subscales ranged from good to strong.

## 1. INTRODUCTION

Puberty is a period in which adolescents need to deal with different issues as a result of psychological and physiological changes. Although most students enter puberty during middle school years, adolescents experience significant issues and make important decisions that can affect their academic, social/emotional, and career development during high school years (Balkin & Schmit, 2016; Ohrt, Limberg, Bordonada, Griffith, & Sherrell, 2016). Hence, a non-familial adult who cares about a student's development as a whole can serve as a protective factor (Karaman, Cavazos Vela, & Lu, 2018; Roe, 2013; Yılmaz & Demir, 2016). These non-familial adults can be teachers, coaches, or school counselors.

Adolescents might need an adult's help during puberty because it is also a period of transition from high school to postsecondary education. Researchers (Ferguson & Lamback, 2014; Suldo & Shaunessy-Dedrick, 2013) stated that students experience stress related to academic performance, career planning, and college admissions. Therefore, school counselors play an important role at this stage helping and supporting students' academic, social/emotional, and career development (Karaman et al., 2018).

**CONTACT:** Mehmet Akif KARAMAN ✉ [makaraman@gmail.com](mailto:makaraman@gmail.com) ☒ Department of Educational Sciences, Kilis 7 Aralık University, Kilis, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

In today's world, school counseling services and school counselors have more value. Hence, contemporary school counseling standards and models are adapted, such as American School Counselor Association's (ASCA) National Model (2003, 2005, 2014a), the international model for school counseling programs (Fezler & Brown, 2011), and the comprehensive school counseling and guidance program in Turkey (Erkan, 2006; National Ministry of Education [NME], 2006). In addition, it is necessary to have assessment tools to measure the efficacy and practicality of programs. One shortcoming in the counseling field is the lack of scales with validity to measure students' perceptions of school counseling services and school counselors (Lapan, Poynton, Marcotte, Marland, & Milam, 2017). Instruments with validity evidence can help researchers, school counselors, and policy makers better understand the nature of school counseling and delivery of efficient services.

### **1.1. School Counseling in Turkey and ASCA National Model**

The school counseling profession has gone through major changes and development since its emergence. The historical corner stones (e.g., industrial revolution, space race) showed the necessity and importance of the profession. Today, many countries have integrated school counseling services into their curriculums. In this respect, ASCA plays an important role creating new visions and models (Fezler & Brown, 2011; Schimmel, 2008). For example, ASCA released the first national model in 2003 for school counseling programs. After this step was taken, we saw a similar development in Turkey (Do an, 2000; Ye ilyaprak, 2005).

The ASCA National Model is "comprehensive in scope preventive in design and developmental in nature" (ASCA, 2012, p. xi). The model aims to promote students' educational and developmental aspects in the academic, career, and personal/social domains with support of school counselors. The ASCA National Model contains three components which are themes, elements, and flow of the model (ASCA, 2012). The four themes, which are leadership, advocacy, collaboration, and systemic change, were designed to achieve maximum program effectiveness via school counselors, parents, and school staff (ASCA, 2012). The elements are accountability, foundation, management, and delivery.

School counseling and guidance programs in Turkey were initiated in the 1950s under the leadership of US education experts invited by Turkish government officials. This step was taken under the Turkish-American cooperation agreement (Ye ilyaprak, 2005). The Turkish school counseling system was changed in parallel to the changes in the US. After the ASCA National Model (2003) was released, the comprehensive school counseling and guidance program, which was prepared based on the developmental perspective, was implemented in the 2006-2007 academic year by school counselors in Turkey (Ergüner-Tekinalp, Leuwerke, & Terzi, 2009; Terzi, Tekinalp, & Leuwerke, 2011). The final version of ASCA National Model (2014a) consists of four components: (a) foundation, (b) management, (c) delivery, and (d) accountability while the comprehensive school counseling and guidance program (Erkan, 2006; NME, 2006) in Turkey consists of five components: (a) group guidance; (b) individual planning; (c) intervention services; (d) program development, research, consultancy and professional development; and (e) other (events that cannot be placed in other program elements). Although the two national models look different, they have many common points as well. Hence, the current study adapted domains of ASCA Mindsets & Behaviors for Students Success (2014b) to create an instrument that can be useful for Turkish school counselors and researchers and adapted and validated by other researchers into different cultures and languages (e.g., English, Arabic, Spanish).

ASCA Mindsets & Behaviors for Students Success (2014b) are organized by domains that "enhance the learning process and create a culture of college and career readiness for all students" (p. 1). These domains are (a) academic development, (b) career development, and (c) social/emotional development. Academic development refers to standards counselors use to



support and maximize students' academic success. The second domain, career development, guides counselors to help students understand the connection between school and work and to support a successful transition from school to higher education or world of work. The last domain, social/emotional development, guides counseling services to help students with social and emotional issues. Specifically, the last domain guides counselors how to help students manage and learn emotions and interpersonal skills. In summary, the current study used a framework based on the aforementioned domains and generated items using the comprehensive counseling and guidance program (Erkan, 2006; NME, 2006).

## **1.2. Counselor Support**

One prominent goal of the school counseling profession is to help all students be successful in schools (ASCA, 2005; Clark & Breman, 2009). In this respect, school counselors and students are the main components of school counseling services. The support students perceive from their counselors can influence their development in academic, career, and social/emotional domains. For example, Poynton and Lapan (2017) stated that students who sought school counselors for assistance when applying to college were more likely to have educational motivation for higher education. In another study, Parker and Ray (2017) found that counseling activities for college and career readiness among Latinx high school students were very important. However, in the same study, it was reported that students indicated personal/social or academic support from their school counselors were less important.

Similar to the current study, Lapan et al. (2017) developed and validated the "College and Career Readiness Counseling Support Scale." Their instrument had five factors but confirmatory factor analysis (CFA) did not confirm the factor structure. In the validation process, these authors found that the frequency and helpfulness of meeting with counselors were correlated with achievement in high school.

Although most of the aforementioned studies focused on counselors' support for college and career readiness, there were noteworthy studies focusing on school counselor support in other areas, such as LGBT youth concerns (Roe, 2013), positive adult role models (Blum, McNeely, & Nonnemaker, 2002), and life skills and individual attention (Ohrt et al., 2016). Roe (2013) used a phenomenological inquiry approach to examine the support gay and bisexual adolescents received from their school counselors. Taking many factors into account (e.g., political beliefs, school counselors' accessibility), students reported that school counselors were helpful when they discussed and listened to students' concerns on LGBT issues. Adolescents stated that it was relaxing when someone listened to their concerns without judging or breaking confidentiality.

The studies mentioned above showed a reality among many important facts: school counselor support has a significant place in students' academic, social/emotional, and career development. This finding is aligned with the ASCA Mindsets & Behaviors for Students Success and the current study aims to create a culturally valuable and psychometrically sound instrument in the counseling field.

## **1.3. Present Study**

The present study has been framed using ASCA Mindsets & Behaviors for Students Success (2014b) to create a valuable instrument which measures students' perceptions of their school counselors' support in a different culture, society, and education system. Previous efforts to create international content standards for school counseling programs (Fezler & Brown, 2011) showed the worthiness of the profession and advocated for school counseling in other countries. After reviewing the literature in English and Turkish, to the best of the authors' knowledge, there is not a theoretically driven instrument developed and validated for K-12 students in both languages. Hence, the purpose of current study was to develop and validate an ASCA National

Model and developmental perspective-based instrument that measures high school students' perceptions of school counseling services and school counselors' support. We aimed to identify long and short forms of the measure that could be useful for Turkish high school students and future adaptation studies (e.g., English version). We utilized the questions listed below to guide the study:

1. Will EFA and CFA identify scales based on the ASCA Mindsets & Behaviors for Students Success that measures high school students' perceptions of school counselors' support in Turkish culture?
2. Will CFA identify the short form of the Perceived School Counselor Support Scale (PSCS)?
3. Will these scales be significantly correlated with mattering and grit?

## **2. METHOD -Study 1: Development of the Perceived School Counselor Support Scale**

### **2.1. Item Generation and Scale Refinement**

The authors followed standards for educational and psychological testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) at the procedure of development and validation of the instrument. First, a literature review of the ASCA National Model (2014a) ASCA Mindsets & Behaviors for Students Success (2014b) and developmental school counseling approach in Turkey (NME, 2006) were conducted to determine factors that influence a student's perception of counselor support. The authors desired to develop an instrument which is efficient and have strong psychometric properties. Hence, we limited the instrument with three factors that reflected three domains of ASCA Mindsets and Behaviors and developmental approach. These three factors (domains) were called a) academic support, b) career support, and c) social /emotional support. Second, 64 items were written based on the domains by the authors and sent to nine expert raters including school counselors who have experience for at least seven years, counselor educators, and measurement evaluation specialists in order to satisfy validity evidence based on test content. Based on their feedback, some items were revised, and seven items were removed from the item pool. The final pool included 57 items. As a final step, a Turkish language and literacy faculty member checked 57 items for grammar and age appropriateness. The language expert grammatically changed a few items and certified that the scale had a 5<sup>th</sup> grade reading level based on the Flesch Reading Ease Formula (Flesch, 1948). A 5-point Likert-type response format, with values 1 *Never*, 2 *Rarely*, 3 *Sometimes*, 4 *Usually*, and 5 *Always*, was used to identify students' perceived support level from counselors.

### **2.2. Data Collection Procedure**

The relevant institutional ethics and research board approved this study. Invitations to school counselors were sent through the e-mail list of the city's national education board. After receiving approval responses from school counselors, we visited schools to meet with principals face-to-face. We initiated the study in schools in which principals admitted us to attend. A family meeting was conducted in each school during the final two weeks of February 2017 and families were informed about the study. A permission form was distributed, and participants whose families gave signed permission forms were included in the study.

When we prepared the measurement package, we inserted bogus items (Moran & Cutler, 1997) to control for response biases (e.g., 1k. Please mark "3" for this question). Data were collected during March 2017. Students attended five high schools across one province of East Anatolia Region. Participants attended a diverse range of high schools (e.g., vocational high schools, general high schools, Anatolian high schools) located in urban and suburban communities. The

data were collected by one of the authors with the help of school counselors during the first 20 minutes of guidance classes. Participation was voluntary, and we distributed measures to only those participants. Incentives were not offered or given to the participants.

### 2.3. Sample

A total of 744 students through 9<sup>th</sup> and 12<sup>th</sup> grade from a vocational, a general and an Anatolian high school participated in Study 1. Eighty-two participants had missing data on five or more items. After removing students with missing data and who did not follow directions on bogus items, the final analysis included 662 participants. The mean age of participants was 16.08 years ( $SD = 1.18$ ; range, 14-19 years). More girls ( $n=353$ , 53.3%) than boys ( $n=309$ , 46.7%) participated. Also, participants reported their grade levels as follows: 9<sup>th</sup> grade ( $n = 147$ , 22.2%), 10<sup>th</sup> grade ( $n= 144$ , 21.8%), 11<sup>th</sup> grade ( $n=189$ , 28.5%), and 12<sup>th</sup> grade ( $n = 182$ , 27.5%).

### 2.4. Measures

**Demographic form:** A demographic form was designed to collect data related to participants' age, gender, and grade levels. The information in the form was included based on feedback from principals and school counselors. We did not include questions related to ethnicity, SES levels, and family background since those could worry or bother some students.

**Perceived School Counselor Support Scale:** The PSCS was developed by the lead author and was based on the ASCA Mindsets & Behaviors for Students Success (2014b) and developmental school counseling approach (Erkan, 2006; NME, 2006; Ye ilyaprak, 2005). Since the core of study is the analysis of this instrument, the following domains describe item development. According to ASCA model and developmental school counseling approach, the first domain is academic development. Hence, when we created possible factors, we named our first factor as academic support inspired by the model and approach. The second domain was career development, and the factor was labeled as career support. Following these, the third domain was social/emotional development, and the third factor was named social/emotional support. A 5-point Likert-type response format with values ranging from one (*never*) to five (*always*) was used. Reliability estimates in the normative sample were evaluated using Cronbach's alpha ( ) to assess internal consistency.

### 2.5. Data Analysis

First, for the purposes of the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), we split the sample with 662 participants into two data sets. We selected a random sample of 50% of 662 students for use in the EFA and the remaining 331 for the CFA. Three analyses were conducted to determine factor structure of the PSCS. The first analysis was parallel analysis (PA), which is a Monte Carlo simulation technique to determine the number of factors to retain in EFA (Ledesma & Valero-Mora, 2007). Parallel Analysis, which was introduced by Horn (1965), compares the observed eigenvalues to account for more variance than the components obtained from random data (Karaman, Balkin, & Juhnke, 2018; O'Connor, 2000). Before we ran a PA, the Kaiser-Meyer-Olkin (KMO) was examined to determine if the data were appropriate for factor analysis. The KMO value of .97 indicated that the data were appropriate for analysis (Leech, Barrett, & Morgan, 2005). ViSta 7.2 program (Young, Valero-Mora, & Friendly, 2006) was used to run PA.

The second analysis conducted was EFA. Based on PA analysis, we used the fixed number of factors in EFA to extract dimensions of the instrument. An EFA using principal axis factoring with a direct oblique rotation was conducted. An oblique rotation was selected since we hypothesized that factors were correlated. The identification of factors was based on factor loadings of .40 or greater. Tabachnick and Fidell (2013) stated that .32 is a good rule of thumb

for the minimum loading of an item. Items that had loadings less than .40 or cross-loaded with no distinct measure of a latent variable were omitted.

The final analysis used was CFA to confirm EFA results and develop PSCS short form. A four-factor model was created based on the PA and EFA results. An essential step was to analyze multivariate normality in this part. The Mardia's statistic indicated that the data had a high value of multivariate kurtosis (9.87; Bentler & Wu, 1993). A Mahalanobis Distance operation was conducted to detect multivariate outliers. Based on the analysis, 8 cases were removed from the data-set, thereby reducing the initial sample of 331 students to 323. The second analysis of Mardia's statistic showed that the multivariate kurtosis decreased dramatically (3.68). We interpreted the chi square statistic ( $\chi^2$ ) and  $p$ -values, as well as comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA) metrics of model fit. When inspecting these values, we used Dimitrov's (2012) standards in which an acceptable model fit is represented in values for the  $\chi^2$  ( $p > .05$ ), CFI  $> .90$ , TLI  $> .90$ , SRMR  $< .06$ , and RMSEA  $< .08$ . When creating the short form, the item selection procedure was based on the statistical methodology conducted by Marteu and Bekker (1992) and Fioravanti-Bastos, Cheniaux, and Fernandez (2011). In this procedure, equal number of items is ranked based on their corrected item-total correlation under subscales. After creating several short forms, multiple CFAs are run. Based on CFA results and internal consistency scores, the best fitted model is chosen as the short form. Hence, we created 12- and 16- item forms to select the best-fitted model based on analyses. Models were compared using Satorra-Bentler chi-square difference test.

## 2.6. Results

### 2.6.1. Factor structure

**Exploratory factor analysis:** Based on PA, four factors were retained. Subsequently, an EFA using principal axis factoring with a direct oblimin rotation was conducted to identify 4 factors. Of the 57 generated items included on the PSCS, 18 were eliminated since they were under the .40 item loading criteria. The fixed number of four factors in EFA explained 70% of the variance across all 39 items. Factor 1 was named as *Career Support* reflecting how school counselor(s) support students in terms of career development. A sample item representing this factor was "My school counselor helps me to learn about careers related to my interests and abilities." The eigenvalue for this factor was 20.71 and explained 53% of the variance across all 39 items. Nine items were retained in the factor, with factor loadings ranging from .51 to .88. Table 1 includes factor loadings of the retained items. Subsequently, Table 2 contains descriptive statistics, intercorrelations of the scores from the respective subscales, and internal consistency ( ) of subscale scores.

Factor 2 was named *Emotional Support*. This factor contained nine items with factor loadings ranging from .61 to .83. The eigenvalue for this scale was 3.22 and explained 8% of the variance. This scale included students' perception of emotional support from school counselor(s). A sample item representing this factor was "My school counselor understands what I am going through."

Factor 3 was named *Social Support* to reflect students' perceived social support from school counselor(s) when interacting with them. This factor contained ten items with factor loadings ranging from .49 to .86. The eigenvalue for this scale was 1.95 and explained 5% of the variance. A sample item representing this factor was "My school counselor encourages me about speaking in the public."

The last scale, factor 4, was named *Academic Support*. This factor contained 11 items with factor loadings ranging from .53 to .81. The eigenvalue for this scale was 1.61 and explained 4% of the variance. This scale reflected students' perceived support from their school

counselor(s). A sample item representing this factor was “My school counselor informs me how to study more efficiently.”

**Table 1.** Instrument Items, Factor Loadings, Corrected Item-Total Correlation Scores, and CFA standardized Parameter Estimates

PSCS Items	CS	ES	SS	AS	CITC	PE
Item 1	<b>.88</b>				.76	.77
Item 2	<b>.76</b>				.71	.62
Item 3	<b>.71</b>				.78	.72
Item 4	<b>.69</b>				.85	.82
Item 5	<b>.66</b>				.78	.75
Item 6	<b>.65</b>				.82	.79
Item 7	<b>.59</b>				.78	.81
Item 8	<b>.54</b>				.77	.83
Item 9	<b>.50</b>				.73	.77
Item 10		<b>.83</b>			.85	.84
Item 11		<b>.80</b>			.84	.79
Item 12		<b>.79</b>			.87	.91
Item 13		<b>.78</b>			.75	.61
Item 14		<b>.78</b>			.78	.82
Item 15		<b>.77</b>			.86	.89
Item 16		<b>.77</b>			.76	.77
Item 17		<b>.66</b>			.68	.72
Item 18		<b>.61</b>			.78	.81
Item 19			<b>-.86</b>		.87	.88
Item 20			<b>-.83</b>		.85	.83
Item 21			<b>-.83</b>		.86	.87
Item 22			<b>-.74</b>		.84	.84
Item 23			<b>-.73</b>		.79	.82
Item 24			<b>-.58</b>		.83	.77
Item 25			<b>-.54</b>		.81	.81
Item 26			<b>-.52</b>		.75	.80
Item 27			<b>-.52</b>		.78	.82
Item 28			<b>-.49</b>		.70	.62
Item 29				<b>.81</b>	.80	.79
Item 30				<b>.75</b>	.70	.74
Item 31				<b>.70</b>	.72	.77
Item 32				<b>.66</b>	.82	.78
Item 33				<b>.66</b>	.80	.81
Item 34				<b>.63</b>	.78	.83
Item 35				<b>.60</b>	.78	.75
Item 36				<b>.60</b>	.81	.81
Item 37				<b>.58</b>	.81	.76
Item 38				<b>.54</b>	.74	.77
Item 39				<b>.53</b>	.57	.65

Note. Factor loadings >.40 are in boldface. PE= Standardized Parameter Estimates. CITC= Corrected Item-Total Correlation Scores. PSCS= Perceived School Counselor Support Scale; CS= Career Support, ES= Emotional Support, SS= Social Support, AS= Academic Support

**Confirmatory factor analysis:** Based on the results of PA and EFA, we hypothesized the 4-factor model would have an appropriate fit. The AMOS version 22 package program was used to compute CFA. We used the second part of the data, which were not included in the EFA, to run CFA. Table 1 presents standardized parameter estimates from the CFA testing the four-factor solution suggested by the EFA. The results,  $\chi^2(696) = 1616.9$ , CFI = .91, TLI = .91, RMSEA [90% CI] = .064 [.060, .068], and SRMR = .048, indicated that four-factor model had an acceptable fit (Dimitrov, 2012).

**Creating PSCS short form (PSCS-S):** Managing and scoring long scales can take a significant amount of time. Moreover, completing a long instrument can be exhausting and lead to measurement errors that can be attributed to incorrect or missed items (Fioravanti-Bastos et al., 2011; Schmidt, Le, & Ilies, 2003). Hence, to select the best items of the PSCS-S scale, items were ranked according to their corrected item-total correlation coefficients (Table 1). The CFA testing for the models was conducted with the same CFA data we used for the extended form.

After examining the corrected item-total correlation coefficients, we created two models. Researchers suggested having at least three items in each factor (MacCallum, Widaman, Preacher, & Hong, 2001; Raubenheimer, 2004). Following this rule, the first model included four factors and 12 items. The second model included four factors and 16 items. Models were compared using Satorra-Bentler chi-square difference test. The testing results for the first model was,  $\chi^2(48) = 105.01$ , CFI = .98, TLI = .97, RMSEA [90% CI] = .061 [.045, .077], and SRMR = .028 indicating a strong fit. A second analysis was run for the second model and results showed that the model had a strong fit,  $\chi^2(98) = 206.03$ , CFI = .97, TLI = .96, RMSEA [90% CI] = .059 [.047, .070], and SRMR = .035. After CFA testing, we used Satorra-Bentler chi-square difference test to choose the best model. Werner and Schermelleh-Engel (2010) stated “if the  $\chi^2_d$  -value is significant, the “larger” model with more freely estimated parameters fits the data better than the “smaller” model” (p. 3). The chi-square difference test was significant,  $\chi^2(50) = 101.01$ ,  $p < .05$ , indicating that Model 2, which was the larger model, had a better fit.

### 3. METHOD - Study 2: Validation of the PSCS Short Form

Following standards for educational and psychological testing (AERA et al., 2014), we collected additional data for validation of PSCS-S. As AERA et al. (2014) reported, the Study 2’s aim was to provide evidence based on internal structure and relations to other variables for the PSCS short form. In this step, two variables were added for convergent validity: (a) mattering and (b) grit. Mattering is an individual’s sense of importance and belonging to family, friends, and society (Sarı & Karaman, 2018). We included mattering because the PSCS has Emotional Support and Social Support subscales which are related to mattering. The second variable, grit, is one’s perseverance and passion toward his/her goals (Cavazos Vela, Hinojosa, & Karaman, 2018). Therefore, we aimed to test if school counselor support was significantly correlated with grit. The following sections include detailed information of the process.

#### 3.1. Participants and Procedures

The data of Study 1 and Study 2 were collected from the same schools within a two-month interval. A total of 760 participants were involved in Study 2. Eight participants were removed from the data set because of high rate of unanswered items. Missing values were replaced by imputed values (EM). The final data set included 752 participants.

Participants’ age ranged from 13 to 18 ( $M = 16.10$  years,  $SD = 1.17$ ). There were 341 boys (45.3%) and 408 girls (54.3%). Three participants preferred not to answer this question. Participants reported their level of classes as follows: 9th grade ( $n = 194$ , 29%), 10th grade ( $n = 228$ , 30%), 11th grade ( $n = 161$ , 21.4%), and 12th grade ( $n = 142$ , 19%). Four participants

failed to respond to this demographic query. To the question of whether they had ever visited the school counselor, 703 of the students (94%) said yes and 45 (6%) said no. Four participants failed to respond to this demographic query. In terms of the reasons to visit their school counselors, 290 students (38.6%) visited due to academic reasons, 198 (26.3%) due to college and career plans, 89 (11.8%) due to emotional issues, 97 (13%) due to social relationships, and 29 participants (3.9%) because of other reasons.

### 3.2. Measures

**Perceived School Counselor Support Scale- short form (PSCS-S):** Based on Study 2's aim, we used the short version of the PSCS. The instrument consists of 16 items under four factors: Academic Support, Career Support, Emotional Support, and Social Support. Cronbach's alpha coefficient scores ranged from .86 to .92.

**General Mattering Scale:** We used the Turkish version of the General Mattering Scale (GMS; Haktanir, Lenz, Can, & Watson, 2016) for the current study. Marcus (1991) developed the original GMS to assess the degree to which individuals believe they are important to others. This 5-point Likert-type assessment yields a single scale score based on participant responses that range from *Very Much* to *Not at All*. Possible scores range from 5 to 20, with higher scores indicative of a greater perception of mattering. Mattering is accounted for by participant responses to items such as "How important do you feel you are to other people?" and "How interested are people generally in what you have to say?" Haktanir et al. (2016) reported an alpha coefficient of .74 for the GMS among first year college students. For the current study, we calculated a Cronbach's alpha of .81.

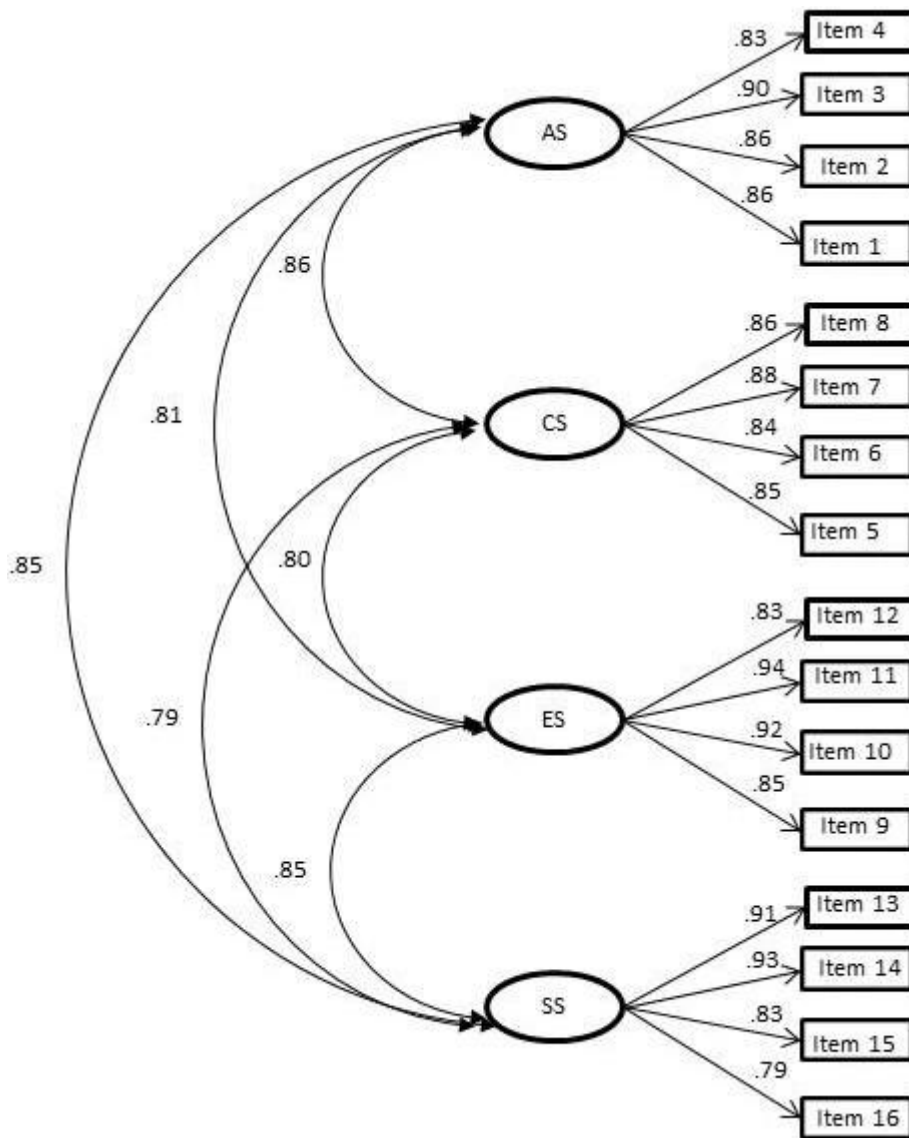
**Short Grit Scale:** We used the Turkish version of the Short Grit Scale (GRIT-S; Sarıçam, Çelik, & Ouz, 2016). The GRIT-S was created by Duckworth and Quinn (2009), measuring grit for long-range goals and trait-level perseverance. The Grit-S is a self-report measure consisting of eight items such as "I am a hard worker" and "I often set a goal but later choose to pursue a different one." A 5-point Likert scale ranging from "Very much like me" to "Not like me at all" is used to indicate the degree to which respondents believe each statement reflects their level of grit. Items 1, 3, 5, and 6 are reverse coded. Duckworth and Quinn (2009) indicated that the scale had adequate test-retest reliability ( $r = .68$ ) after one year and sufficient internal consistency ( $\alpha = .82, .84$ ). Duckworth and Quinn (2009) also determined that self-reporting grit is as reliable as informant reporting. Sarıçam et al. (2016) reported Cronbach alpha coefficient score of .83 for the whole instrument and .68 for test-retest reliability. For the current study, the Grit-S had an internal consistency coefficient of .72.

### 3.3. Results

The analysis showed that all regression coefficients between the latent variables and items were significant. The lowest and highest factor loadings were between latent variables and

Item 16 (.79) and Item 11 (.94), respectively (see [Figure 1](#)). The results of model fit indices showed that the  $\chi^2$  was significant for the hypothesized model,  $\chi^2(98) = 572.83, p < .001; \chi^2/df = 5.84$ . The fit indices indicated a good fit for the data, GFI = .91, TLI = .95, CFI = .96, RMSEA = .08 (90% CI = .074–.087), and SRMR = .03.

Next, to address evidence of relationships to other variables for the PSCS-S, correlational analysis was conducted with the GRIT-S (Duckworth & Quinn, 2009; Sarıçam et al., 2016) and GMS (Haktanir et al., 2016; Marcus, 1991). [Table 3](#) provides descriptive data and correlations. As this table shows, we found evidence for criterion validity. A statistically significant and positive relationship was found between the perceived school counselor total scores and grit ( $r = .08; p < .05$ ) and the general mattering scores ( $r = .17; p < .05$ ). Based on this analysis, higher perceived school counselor scores were correlated with higher grit and general mattering scores.



**Figure 1.** The confirmatory factor analysis model of the Short Perceived School Counselor Support Scale (PSCS-S). The standardized parameter estimates for the PSCS-S are listed. Rectangles indicate the 16 items on the PSCS-S, and ovals represent the 4 latent factors of subscales. Abbreviations represents: CS= Career Support, ES= Emotional Support, SS= Social Support, AS= Academic Support

**Table 2.** Correlations between the Subscales, Means (M), and Standard Deviations (SD) of the PSCS

Scale	M	SD	1	2	3	
1 Career Support	2.43	.94	1.42			
2 Emotional Support	3.27	.95	1.56	.48*		
3 Social Support	2.39	.95	1.45	-.56*	-.52*	
4 Academic Support	2.39	.94	1.39	.63*	.46*	-.63*

Note. PSCS= Perceived School Counselor Support Scale

\* $p < .01$



**Table 3.** Means (*M*), Cronbach's alpha, Correlations between Variables, and Standard Deviations (*SD*) of Variable Scores of the PSCS-S

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	
1. Career Support	3.39	.92	1.23	-	.74*	.74*	.79*	.07	.14*
2. Emotional Support	3.45	.92	1.30		-	.80*	.75*	.09**	.17*
3. Social Support	3.03	.92	1.29			-	.80*	.05	.11*
4. Academic Support	3.11	.88	1.27				-	.08**	.18*
5. Grit	3.30	.72	.69					-	.36*
6. General Mattering	2.91	.81	.71						-

Note. PSCS-S= Perceived School Counselor Support Scale Short Form

\*  $p < .01$

\*\*  $p < .05$

#### 4. DISCUSSION

The purpose of this study was to develop and establish validity evidence for an instrument to assess perceptions of school counselor support among high school students. Because researchers, practitioners, and professional organizations are interested in evaluating students' perceptions of school counselor support (Lapan et al., 2017; Vela, Zamarripa, Balkin, Johnson, & Smith, 2013), having accurate information for measures of perceptions of support advances a school counseling approach by providing researchers and practitioners with information regarding psychometric properties. With the increasing interest in school counseling and different areas of services, there is a need to provide validity evidence for instruments in different languages with high school students.

In today's world, school counselors and the services they provide in terms of academic, career, and social/emotional support have more value. Therefore, contemporary school counseling standards and models are adapted to the needs of the age, such as ASCA's National Model (2003, 2005, 2014a) and the international model for school counseling programs (Fezler & Brown, 2011). Instruments with validity evidence can help researchers, school counselors, and policy makers understand the nature of school counseling and delivering efficient services. We also agree with Lapan et al. (2017) who said that assessments with validity evidence "give students and their families a way to have a voice and advocate for their needs, to know what to expect...to better understand what kinds of college and career services they should be receiving" (p. 85).

Researchers (Lapan et al., 2017; Vela et al., 2013) highlighted the lack of instruments with a theoretical approach and validity evidence that measure perceptions of support from school counselors. The current study used ASCA National Model (2003, 2005, 2014a), ASCA Mindsets & Behaviors for Students Success (2014b), and developmental approach to create items and subscales. The ASCA Mindsets & Behaviors for Students Success highlights three broad domains enhancing learning process and creating a culture of college and career readiness. The results of this study supported the theoretical based PSCS long and short forms providing initial and strong evidence based on internal structure and relations to other variables (AERA et al., 2014). Internal consistency estimates on subscales ranged from good to strong. Also, the PA, EFA, and CFA resulted in a 4-factor model (Factor 1, Factor 2, Factor 3, and Factor 4) with 39 items and accounting for 70% of variance. Factor 1, *Career Support*, contained 9 items reflecting students' perceptions of school counselors' support in career development. Factor 2, *Emotional Support*, contained nine items focusing on students' perceptions of emotional support from school counselors. Factor 3, *Social Support*, contained 10 items reflecting students' perceived social support from school counselor(s). Finally, factor

4, *Academic Support*, contained 10 items to reflect students' perceptions of academic support from school counselor(s).

The short form was also created to be practical and efficient for school counselors and researchers. Therefore, two models were created following previous researchers' suggestions (MacCallum et al., 2001; Raubenheimer, 2004). The complex model including four factors with 16 items had a better fit. In terms of scoring of the subscales, both long and short forms have the same scoring methods. The instrument does not have a total score because each factor's score is calculated separately and evaluated in itself. The logic behind this scoring is related to students' aim to visit counseling services or schools. In other words, one can visit a school counselor for academic reasons but not for emotional or social support. Hence, having a total score will not give an accurate assessment of student's perception on school counselors.

Having a psychometrically sound instrument is also related to evidence of validity based on relations with other variables (AERA et al., 2014). Bivariate correlations analyses provided promising support for convergent validity of the PSCS short form with perceptions of mattering and grit. Evidence based on relations to other variables can influence treatment interventions, program services, or allocation of resources. As an example, if school counselors identify that most students perceive lack of emotional support, they will be able to use this information to create more interventions and services to address students' emotional concerns. Given the sources of validity evidence identified in the current study, exploration of the PSCS-S may provide researchers and school counselors with a meaningful and culturally valuable tool to measure perceptions of support in academic, career, social, and emotional domains.

## **5. IMPLICATIONS for PRACTICE**

First, school counselors may find the PSCS-S useful in identifying the extent to which students met specific career, academic, personal, and social development goals. The PSCS-S offers school counselors with a tool to evaluate students' perceptions of counselor support in various domains. School counselors have an important responsibility to provide direct services to high school students in personal, academic, social, and emotional areas (ASCA, 2016). As a result, the PSCS-S offers school counselors with a mechanism to gather students' perceptions of support and facilitate conversations regarding areas of improvement. Second, the PSCS-S might serve as an outcome tool that can provide evidence to school administrators and policy makers with information regarding the effectiveness of school counseling interventions. As one example, results from a survey with high school students might lead to conclusions that although perceptions of emotional support from counselors are high, perceptions of career support is inadequate. School counselors can gather these types of feedback to determine areas of improvement and inform future services and allocation of resources.

### **5.1. Implications for Future Research**

First, researchers should validate the PSCS-S in different languages with other culturally-diverse populations. These areas of scholarship may assist in determining the degree that some items on other versions of this instrument may be useful and whether items need to be revised. Second, investigations identifying relationships between perceptions of counselor support with other constructs would be useful to demonstrate evidence based on relations to other variables (AERA et al., 2014). If researchers provide convergent and predictive evidences between counselor support and other factors, an important body of literature for the PSCS-S might emerge. Other important factors to investigate include high school test scores, college grade point average, mental health, college self-efficacy, and vocational outcome expectations. Next, researchers can use single group pre-test post-test or between-group designs to examine the impact of school counseling programs and services on students' perceptions of counselor support in career, academic, emotional, and social areas. Potential school counseling methods

that could increase perceptions of development include narrative therapy (White & Epston, 1990), positive psychology (Seligman, 2002), and creative journal arts therapy (Vela et al., 2016). Finally, we created and validated Turkish versions of the counselor support scale with high school students. Factor structure can vary by development levels so researchers should conduct a cross-cultural validation of this instrument with middle school students.

## 5.2. Limitations

Several limitations warrant consideration. First, all data collected in the current investigation came from a non-clinical sample of predominantly Turkish-heritage students from a high school. As a result, validity evidence in the current study is only meaningful if these measures are administered in Turkish with a similar group of high school students. Researchers evaluating factor structure of different versions of the PSCS-S with other populations may provide greater accountability for their perceptions of school counselors. Additionally, findings are not causal (Balkin, 2014) and represent some levels of subjectivity in terms of selecting and developing instrument-items to measure perceptions of school counselor support in academic, social, emotional, and career domains.

## 6. CONCLUSION

In summary, we sought to develop and examine validity evidence of the PSCS-S with a sample of Turkish high school students. The results indicated that the ASCA National Model and Turkey's developmental model of counseling worked in Turkish culture. The items in the instrument, which were written by authors who are from the US and Turkey, reflected a diverse perspective and supported efforts to create an international model of school counseling (Fezler & Brown, 2011). Results from this study also provide promising support for using the PSCS-S to evaluate students' perceptions of counselor support in academic, career, emotional, and social domains. With instruments with strong validity evidence to measure perceptions of counselor support, school counselors and policy makers may be able to evaluate and improve students' perceived feelings of personal, social, academic, and emotional development. The PSCS-S also can help students become self-aware of their perceptions of school counseling services and evaluate interventions, programs, or services provided by school counselors. Furthermore, in the future, this instrument can be adapted and validated in different languages and cultures to measure students' perceptions of support from their school counselors in academic, career, social, and emotional areas.

## ORCID

Mehmet Akif Karaman  <https://orcid.org/0000-0001-7405-5133>

## 7. REFERENCES

- American Counseling Association (2014). *ACA code of ethics*. Alexandria, VA: Author.
- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American School Counselor Association. (2003). *The ASCA national model: A framework for school counseling programs*. Alexandria, VA: Author.
- American School Counselor Association. (2005). *The ASCA national model: A framework for school counseling programs*. Alexandria, VA: Author.
- American School Counselor Association. (2012). *The ASCA National Model: A framework for school counseling programs* (3rd ed.). Alexandria, VA: Author.
- American School Counselor Association. (2014a). *The ASCA national model: A framework for school counseling programs*. Alexandria, VA: Author

- American School Counselor Association. (2014b). *Mindsets and behaviors for student success: K-12 college- and career-readiness standards for every student*. Alexandria, VA: Author.
- Balkin, R. S. (2014). Principles of quantitative research in counseling: A humanistic perspective. *Journal of Humanistic Counseling*, 53, 240 - 248. <https://doi.10.1002/j.2161-1939.2014.00059.x>
- Balkin, R. S., & Schmit, E. L. (2016). Using the crisis stabilization scale to evaluate progress for adolescents in crisis. *Journal of Child and Adolescent Counseling*, 2, 33-41. doi: 10.1080/23727810.2015.1134009
- Blum, R. W., McNeely, C., & Nonnemaker, J. (2002). Vulnerability, risk, and protection. *Journal of Adolescent Health*, 31, 28-39. doi:10.1016/S1054-139X(02)00411-1
- Cavazos Vela, J., Hinojosa, Y., Karaman, M.A. (2018). Evaluation of the Short Grit Scale with Latinx college students. *The Journal of Counseling Research and Practice*, 3(1), 31-42.
- Clark, M. A., & Breman, J. C. (2009). School counselor inclusion: A collaborative model to provide academic and SocialEmotional support in the classroom setting. *Journal of Counseling & Development*, 87, 6-11. doi:10.1002/j.1556-6678.2009.tb00543.x
- Dimitrov, D. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: Wiley.
- Do an, S. (2000). The historical development of counseling in Turkey. *International Journal for the Advancement of Counselling*, 22, 57-67. doi:10.1023/A:1005474126819
- Ergüner-Tekinalp, B., Leuwerke, W., & Terzi, . (2009). Emergence of national school counseling models: Views from the United States and Turkey. *Journal of School Counseling*, 7(33), 1-30.
- Erkan, S. (2006). *Okul psikolojik danı ma ve rehberlik programlarının hazırlanması* [Preparing school counseling and guidance programs]. Ankara, TR: Nobel Yayın Da ıtım.
- Ferguson, R. F., & Lamback, S. (2014). *Creating pathways to prosperity: A blueprint for action*. Report issued by the Pathways to Prosperity Project at the Harvard Graduate School of Education and the Achievement Gap Initiative at Harvard University. Retrieved from <http://www.agi.harvard.edu/pathways/CreatingPathwaystoProsperityReport2014.pdf>
- Fezler, B., & Brown, C. (2011). *The international model for school counseling programs*. Retrieved from [http://www.aassa.com/uploaded/Educational\\_Research/US\\_Department\\_of\\_State/Counseling\\_Standards/International\\_Counseling\\_Model\\_Handbook.pdf](http://www.aassa.com/uploaded/Educational_Research/US_Department_of_State/Counseling_Standards/International_Counseling_Model_Handbook.pdf)
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233. <http://dx.doi.org/10.1037/h0057532>
- Fioravanti-Bastos, A. C. M., Cheniaux, E., & Landeira-Fernandez, J. (2011). Development and validation of a short - form version of the Brazilian state - trait anxiety inventory. *Psicologia: Reflexão e Crítica*, 24, 485 - 494. <https://dx.doi.org/10.1590/S0102-79722011000300009>
- Haktanir, A., Lenz, A. S., Can, N., & Watson, J. C. (2016). Development and evaluation of Turkish language versions of three positive psychology assessments. *International Journal for the Advancement of Counselling*, 38, 286–297. doi:10.1007/s10447-016-9272-9
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. <https://doi.org/10.1007/bf02289447>
- Karaman, M. A., Cavazos Vela, J., & Lu, M. T. P. (2018). Examining the Teacher Support Scale Revised and Counselor Support Scale with Latina/o students. *Adiyaman University Journal of Educational Sciences*, 8, 1-20. doi: 10.17984/adyuebd.357433
- Karaman, M.A., Balkin, R.S., & Juhnke, G. J. (2018). The Turkish adaptation of the Juhnke-Balkin life balance inventory Turkish Form. *Measurement and Evaluation in Counseling and Development*, 51, 141-150. doi: 10.1080/07481756.2017.1308226

- Lapan, R. T., Poynton, T., Marcotte, A., Marland, J., & Milam, C. M. (2017). College and career readiness counseling support scales. *Journal of Counseling & Development, 95*, 77-86. doi:10.1002/jcad.12119
- Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research and Evaluation, 12*, 1–11.
- Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. Mahwah, NJ: Erlbaum.
- MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*, 611-637. doi:10.1207/S15327906MBR3604\_06
- Moran, G. & Cutler, B.L. (1997). Bogus publicity items and the contingency between awareness and media-induced pretrial prejudice. *Law and Human Behavior 21*, 339-344. <https://doi.org/10.1023/A:1024846917038>
- National Ministry of Education. (2006). *İkô retim ve ortaö retim kurumları sınıf rehberli i programı* [K-12 schools class guidance programs]. Retrieved from [https://orgm.meb.gov.tr/alt\\_sayfalar/sinif\\_reh\\_progmm.html](https://orgm.meb.gov.tr/alt_sayfalar/sinif_reh_progmm.html) on July 29, 2018
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*, 396–402. <https://doi.org/10.3758/bf03200807>
- Ohrt, J. H., Limberg, D., Bordonada, T. M., Griffith, C., & Sherrell, R. S. (2016). Adolescents' perceptions of their school counselors' impact. *Journal of Child and Adolescent Counseling, 2*, 1-15. doi:10.1080/23727810.2015.1133996
- Parker, M. M., & Ray, D. C. (2017). School counseling needs of Latino students. *Journal of School Counseling, 15*(16), 1-30.
- Poynton, T. A., & Lapan, R. T. (2017). Aspirations, achievement, and school counselors' impact on the college transition. *Journal of Counseling & Development, 95*, 369-377. doi:10.1002/jcad.12152
- Raubenheimer, J. (2004). An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology, 30*, 59-64. doi: 10.4102/sajip.v30i4.168
- Roe, S. (2014). "Put it out there that you are willing to talk about anything": The role of school counselors in providing support to gay and bisexual youth. *Professional School Counseling, 17*, 153-162.
- Sari, H. . & Karaman, M.A. (2018). Gaining a better understanding of General Mattering Scale: An application of classical test theory and item response theory. *International Journal of Assessment Tools in Education, 5*, 668-681. doi:10.21449/ijate.453337
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206-224.
- Schimmel, C. J. (2008). *School counseling in West Virginia: An examination of school counselors and implementation of WV policy 2315* (Doctoral Dissertation). Retrieved from Theses, Dissertations and Capstones (Paper 320).
- Seligman, M. E. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York, NY: Free Press.
- Suldo, S. M., & Shaunessy-Dedrick, E. (2013). The psychosocial functioning of high school students in academically rigorous programs. *Psychology in the Schools, 50*, 823–843. doi:10.1002/pits.21708
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.)*. Boston, MA: Pearson Education.

- Terzi, ., Tekinalp, B. E., & Leuwerke, W. (2011). Psikolojik danı manların okul psikolojik danı ma ve rehberlik hizmetleri modeline dayalı olarak geli tirilen kapsamlı psikolojik danı ma ve rehberlik programını de erlendirmeleri [The evaluation of comprehensive guidance and counseling programs based on school counseling and guidance services model by school counselors]. *Pegem E itim ve Ö retim Dergisi*, 1(1), 51-60.
- Ünal, A., & Ünal, E. (2010). A case study on the perception of the school counselor by students and teachers. *Journal of Human Sciences*, 7(2), 919-945.
- Vela, J. C., Flamez, B., Sparrow, G. S., & Lerma, E. (2016). Understanding support from school counselors as predictors of Mexican American adolescents' college-going beliefs. *Journal of School Counseling*, 14(7), 1-28.
- Vela, J. C., Zamarrıpa, M., Balkin, R., Johnson, M., & Smith, R. (2013). Understanding Latina/o students' perceptions of high school counselors and acculturation as predictors of enrollment in AP courses. *Professional School Counseling*, 17, 142-152. doi:10.5330/prsc.17.1.a32q312p27351256
- Werner, C., & Schermelleh-Engel, K. (2010). Deciding between competing models: Chi-square difference tests. *Goethe University*. Available online: <https://perma.cc/2RTR-8XPZ> (accessed on 13 July 2018)
- White, M., & Epston, D. (1990). *Narrative means to therapeutic ends*. New York, NY: W. W. Norton.
- Ye ilyaprak, B. (2005). *E itimde rehberlik hizmetleri: Geli imsel bir yakla ım* [Guidance services in education: A developmental perspective]. Ankara, TR: Nobel Yayın Da ıtım.
- Yılmaz, F & Demir, S. (2016). The validity and reliability study of revised School Climate Teacher Survey's Turkish version. *International Journal of Assessment Tools in Education*, 3(1), 85-100.
- Young, F.W., Valero-Mora, P., & Friendly, M. (2006). *Visual statistics seeing data with dynamic interactive graphics*. Hoboken, NJ: Wiley.

## Investigating a new method for standardising essay marking using levels-based mark schemes

Jackie Greateorex <sup>1\*</sup>, Tom Sutch <sup>1</sup>, Magda Werno <sup>1</sup>, Jess Bowyer <sup>2</sup>, Karen Dunn <sup>3</sup>

<sup>1</sup> Cambridge Assessment, Triangle Building, Shaftesbury Road, Cambridge, UK, CB2 8EA

<sup>2</sup> University of Exeter, St Luke's Campus, Heavitree Road, Exeter, UK, EX1 2LU

<sup>3</sup> British Council, 10 Spring Gardens, London SW1A 2BN, UK

### ARTICLE HISTORY

Received: 18 January 2019

Revised: 11 April 2019

Accepted: 23 April 2019

### KEYWORDS

Comparative judgement,  
Marking,  
Standardisation,  
Reliability,  
Essay

**Abstract:** Standardisation is a procedure used by Awarding Organisations to maximise marking reliability, by teaching examiners to consistently judge scripts using a mark scheme. However, research shows that people are better at comparing two objects than judging each object individually. Consequently, Oxford, Cambridge and RSA (OCR, a UK awarding organisation) proposed investigating a new procedure, involving ranking essays, where essay quality is judged in comparison to other essays. This study investigated the marking reliability yielded by traditional standardisation and ranking standardisation. The study entailed a marking experiment followed by examiners completing a questionnaire. In the control condition live procedures were emulated as authentically as possible within the confines of a study. The experimental condition involved ranking the quality of essays from the best to the worst and then assigning marks. After each standardisation procedure the examiners marked 50 essays from an AS History unit. All participants experienced both procedures, and marking reliability was measured. Additionally, the participants' questionnaire responses were analysed to gain an insight into examiners' experience. It is concluded that the Ranking Procedure is unsuitable for use in public examinations in its current form. The Traditional Procedure produced statistically significantly more reliable marking, whilst the Ranking Procedure involved a complex decision-making process. However, the Ranking Procedure produced slightly more reliable marking at the extremities of the mark range, where previous research has shown that marking tends to be less reliable.

## 1. INTRODUCTION

General Certificate of Secondary Education (GCSE), Advanced Subsidiary (AS) and Advanced Level (A Level) are school examinations taken in the UK. Given the high-stakes nature of these examinations, it is essential that marking reliability is high and that standardisation (examiner training to accomplish uniform use of the mark scheme) is effectual, so that marks and grades are dependable. Generally, marking reliability is greater for short

CONTACT: Jackie Greateorex ✉ [greateorex.j@cambridgeassessment.org.uk](mailto:greateorex.j@cambridgeassessment.org.uk) 📧 Research Division, Cambridge Assessment, Triangle Building, Shaftesbury Road, Cambridge, UK, CB2 8EA

ISSN-e: 2148-7456 / © IJATE 2019

answer questions than for questions requiring a long response which are marked with levels-based mark schemes<sup>†</sup>. Consequently, effective procedures for standardising examiners' essay marking are crucial.

It may be feasible to improve the current approach to standardisation and maximise essay marking reliability. The purpose of standardisation is to ensure that all examiners apply the mark scheme fairly and consistently, and procedures vary between Awarding Organisations, subjects and units. Traditionally, standardisation consists of practice marking, a meeting, where examiners are trained to apply the mark scheme, typically by marking a number of scripts as a group. Examiners then individually mark a sample of scripts at home, which are checked by their Team Leader, or the Principal Examiner (PE: the lead marker in charge of the examination or qualification) in smaller subjects. The Team Leader or PE may require further samples of marking to be checked.

In addition to standardisation there are several procedures for maximising marking quality including scaling (correcting consistently lenient or severe marking), and marker monitoring (observing marking post standardisation). The focus of the research is standardisation and it is beyond the scope of the research to account for these additional procedures.

Laming (2004) concludes from extensive research that people are better at comparing two objects than making absolute judgements about an object. Subsequently, a new approach to standardisation, forthwith called the Ranking Procedure, was proposed. This procedure focuses on comparing essays with one another and ranking them from the best to the worst.

The present research had two aims:

- to investigate whether the Ranking Procedure and Traditional procedure resulted in equivalent levels of marking reliability or whether the reliability from one was demonstrably better;
- to evaluate whether examiners considered the Ranking Procedure to be useful, how they conducted their marking, and whether the procedure was efficient.

### 1.1. Literature Review

Extended response questions are widely regarded as the most difficult questions to mark and are associated with the lowest levels of marking reliability (Black, Suto, & Bramley, 2011; Suto, Nádas, & Bell, 2011b). Consequently, there has been much research investigating why extended response questions have lower reliability, and how this can be improved.

One suggestion is that marking extended response questions entails a high cognitive load for the examiner, which could in turn lead to reduced marking reliability. Suto and Greatorex (2008) found that more complex cognitive strategies, such as evaluating and scrutinising, were used significantly more in the marking of GCSE Business Studies, which uses a levels-based mark scheme, than in GCSE Mathematics, which uses a more objective points-based mark scheme. Senior examiners in the same study suggested that it might be useful to train examiners in the use of these cognitive strategies. This may particularly benefit new examiners, as research indicates that examiners with lower subject expertise, marking and teaching experience mark extended response questions less accurately than others (Suto, Nádas, & Bell, 2011a).

---

<sup>†</sup> Levels-based mark schemes (levels-of-response mark schemes) are generally used for marking extended written responses. Such mark schemes often divide the available marks into smaller mark bands, each mark band is associated with a level and a description of the type of answer that will obtain a mark from within a given mark band. The examiner classifies a candidate's response into a level and then decides which mark from the associated mark band is most appropriate. For more detailed descriptions see Pinot de Moira (2013) and Greatorex and Bell (2008a).



Attempts to increase marking reliability in extended response questions have tended to focus on two key aspects of the marking process: standardisation (or examiner training), and mark schemes. Both are particularly pertinent to the current study.

## **1.2. Improving Standardisation**

The greatest recent change to standardisation procedures (also called examiner or rater training in the literature) has been the transition from face-to-face to online standardisation. Some research indicates that online standardisation may be very slightly more effective in increasing marking accuracy, although in the context of English as a Second Language (ESL) assessment (Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010; Wolfe & McVay, 2010). Research into the use of online standardisation in UK high-stakes examinations indicates that online standardisation is equally as effective as face-to-face standardisation (Billington & Davenport, 2011; Chamberlain & Taylor, 2010).

There is evidence to suggest that face-to-face or online meetings are not particularly effective for increasing marking accuracy on their own (Greatorex & Bell, 2008b; Raikes, Fidler, & Gill, 2009), and should be combined with additional feedback (Greatorex & Bell, 2008b; Johnson & Black, 2012). The type of feedback received does not affect marking reliability, whether it is iterative or immediate, personalised or prewritten, or targeted at improving accuracy or internal consistency (Greatorex & Bell, 2008b; Sykes et al., 2009). However, Johnson and Black (2012) found that examiners find feedback most helpful when it is immediate, refers to the mark scheme and focuses on specific problems with scripts.

Standardisation is of greatest benefit for new or less experienced examiners, whilst having little or no effect on the marking accuracy of experienced examiners (Meadows & Billington, 2007; Meadows & Billington, 2010; Raikes et al., 2009; Suto, Greatorex, & Nádas, 2009). Additional background effects such as subject knowledge and expertise also affect marking reliability (Suto & Nádas, 2008). Despite this, with adequate training, some examiners with no or little teaching and marking experience can become as reliable as the most experienced examiners, although these individuals may be difficult to identify before the standardisation process begins (Meadows & Billington, 2010; Suto & Nádas, 2008). It is noteworthy that research by Meadows and Billington (2010) and Suto and Nádas (2008) related to questions requiring short or medium length responses rather than essays, and therefore the findings may not generalise to examinations marked with levels-based mark schemes.

## **1.3. Mark Schemes**

Alternative studies have investigated whether changes to levels-based mark schemes could improve marking reliability. A key factor is that levels-based mark schemes are often lengthy and contain a lot of information. Whilst more constrained mark schemes are associated with higher levels of reliability (Bramley, 2009; Pinot de Moira, 2013; Suto & Nádas, 2009), they do so by restricting the number of creditable responses and thus would compromise the validity of extended-response assessment (Ahmed & Pollitt, 2011; O'Donovan, 2005; Pinot de Moira, 2011a; Pinot de Moira, 2011b).

An alternative is to use a holistic, rather than analytic, mark scheme. Holistic mark schemes are where one overall mark is given to a response. The mark scheme may specify different elements of performance, but the examiner attaches their own weighting to each feature. In analytic levels-based mark schemes, the examiner awards separate marks for individual elements of a response. There is no clear consensus as to which is more valid and reliable.

The evidence suggests that inter-rater reliability is higher in holistic scoring (Çetin, 2011; Harsch & Martin, 2013; Lai, Wolfe, & Vickers, 2012). However, analytic scoring is particularly helpful in diagnostic English as a Second Language (ESL) assessment as features

of a student's writing can be assessed individually and that information then be fed back to the candidate to guide future learning (Barkaoui, 2011; Knoch, 2007; Lai et al., 2012; Michieka, 2010). Holistic mark schemes, on the other hand, can obscure differences in individual traits of a student's response, as well as how examiners weigh and apply different assessment criteria (Harsch & Martin, 2013).

A style of holistic marking that has particular relevance for the current study is Comparative Judgement (CJ). This method entails deciding which is the better of two scripts, thus making holistic but also *relative* judgements about script quality. Examiners make a series of these judgements, until each script has been judged a number of times. A rank order of all scripts is then statistically compiled, usually by fitting a Bradley-Terry model to the paired comparison data.

If the pairs are presented online and the data can be analysed in 'real time' it is possible to make the presentation of pairs 'adaptive'. This means that as more judgements are made, examiners are given scripts that appear to be closer together in quality, in order to make more nuanced distinctions between scripts and reduce the overall number of comparisons that need to be made. This process is known as 'adaptive comparative judgement' (ACJ), Pollitt (2012a) and Pollitt (2012b).

Whilst CJ is most often used for comparability studies, it is argued that it is more valid and reliable than traditional marking, as examiners are simply making overall judgements about script quality (Kimbell, 2007; Kimbell, Wheeler, Miller, & Pollitt, 2007; Pollitt, 2009, 2012a, 2012b; Pollitt, Elliott, & Ahmed, 2004). A project, which used ACJ to assess Design and Technology portfolios, found a reliability coefficient of 0.93 (Kimbell, 2007), whilst Whitehouse and Pollitt (2012) found a reliability coefficient of 0.97 when using ACJ in AS level Geography papers. However, adaptivity can inflate the reliability coefficient, so the high reliability found in these studies is disputed (Bramley, 2015; Bramley & Vitello, 2018). Moreover, there is empirical evidence that the strength of CJ lies in multiple judgements and a strong statistical model, rather than comparing one script directly with another (Benton & Gallagher, 2018). Also the process is very time-consuming (Whitehouse & Pollitt, 2012). Consequently, there are serious doubts as to whether it is practicable in large-scale, high-stakes assessment.

#### 1.4. Research Questions

The experiment tested the following hypotheses:

H0: The Traditional and Ranking Standardisation result in equivalent levels of marking reliability.

H1: The Traditional or Ranking Standardisation result in more reliable marking.

Examiners' perspectives were collected to answer the following questions:

- How did the examiners undertake the Ranking Procedure?
- Was the Ranking Procedure (in)efficient and (un)suitable for upsampling or digitising?

## 2. METHOD

The project utilised candidates' scripts from an OCR AS level History examination. This examination was chosen for several reasons: firstly, the entry was large enough to select a wide range of scripts. Secondly, the questions required essay responses and were marked using a levels-based mark scheme.

The Principal Examiner from live examining was used as the Principal Examiner for this study. Ten Assistant Examiners (examiners) participated in the study. They had not marked this paper

in live examining, but they had either been eligible to mark it or had marked a similar examination (e.g. another A level History paper).

## **2.1. Design**

The experiment had two conditions.

- The control condition was a simulation of a traditional standardisation process. This is used within some current awarding organisations. Each examiner marked a Provisional Sample. They attended a standardisation meeting where their marking was standardised using the Traditional Mark Scheme. After the meeting, the examiners marked the Standardisation Essays and received feedback on their marking from the PE. The PE decided whether each examiner could proceed to the next stage of the experiment, re-mark the Standardisation Sample or mark a further Standardisation Sample. Finally, the examiners marked the Allocation.
- The experimental condition, called Ranking Standardisation, broadly followed the same process as the control condition. Each examiner ranked a Provisional Sample. They attended a standardisation meeting at which they learnt how to rank responses (with no marks) and then learnt how to mark the ranked essays. After the meeting each examiner rank ordered the Standardisation Sample from the best to the worst response. They received feedback on their ranking from the PE. The PE decided whether each examiner could proceed to the next stage of the experiment or re-rank the Standardisation Sample. Subsequently, each examiner marked the Standardisation Sample. The PE decided whether each examiner could proceed to the next stage of the experiment, re-mark the Standardisation Sample or rank and mark a further Standardisation Sample. Finally, the examiners marked the Allocation.

The design was within subjects. Marking reliability was measured at the end of the experiment. Counterbalancing was achieved by:

- Allocating examiners to groups based on which date they were available to attend a meeting<sup>‡</sup>
- Conducting conditions in the order determined by a 2x2 Latin Square:
  - group 1 control condition then experimental condition
  - group 2 experimental condition then control condition.

This was to guard against the order of conditions affecting the results.

## **2.2. Materials**

### **2.2.1. Scripts**

All the essays involved were responses to the two most popular questions for that examination paper (referred to here as questions A and B). Each had a maximum of 50 marks available. Scripts were anonymised for use by examiners.

The experiment involved four samples of essays: The Provisional, Meeting and Standardisation samples, as well as the examiners' final marking allocation. This was intended to mirror the real standardisation process. Participants marked the Provisional Sample before the standardisation meeting; the Meeting Sample was used in the standardisation meeting to teach participants about applying the mark scheme correctly; and the Standardisation Sample was marked after the meeting to ensure participants were applying the mark scheme correctly and to gain a measure of inter-rater reliability. After the standardisation process was completed, participants were asked to mark an allocation of 50 essays. The essays covered a range of quality of performance.

---

<sup>‡</sup> It was assumed that availability would be as random as any other way of allocating examiners to groups.

### **2.2.2. Mark Schemes**

#### *Traditional Mark Scheme*

The Traditional Mark Scheme was the live mark scheme for question A or B. The live mark scheme included level descriptors and a description of content for each question.

#### *Ranking Mark Scheme*

In the Ranking Mark Scheme the level descriptors from the live mark scheme were replaced with a brief description of the characteristics of quality of performance. The Ranking Mark Scheme was written by the PE and reviewed by OCR. After the standardisation meeting, an indication of the marks given to each Meeting Essay was added. For an example of the Ranking Mark Scheme see [Figure 1](#).

### **2.2.3. Examiner Questionnaire**

A questionnaire was developed that included open and closed questions. The questionnaire focused on the usefulness of aspects of standardisation, how the Ranking Procedure might be upscaled or conducted on-screen, and related merits and limitations.

## **2.3. Controls**

Control mechanisms were in place. First, examiners were allocated to groups based on availability. Secondly, group 1 completed each stage of the control condition using the Traditional Mark Scheme on question A before completing the parallel stage of the experimental condition using the Ranking Mark Scheme for question B. Group 2 completed each stage of the experimental condition with the Ranking Mark Scheme on question A before completing the parallel stage of the control condition with the Traditional Mark Scheme for question B. Thirdly, all examiners marked the same essays at each stage of the experiment. Fourthly, none of the examiners marked the question paper in live marking, as such participants would have violated the crossover design by experiencing the control condition before the experimental condition.

These controls and the within subjects design enabled a direct comparison between the reliability of marking generated by the two experimental conditions.

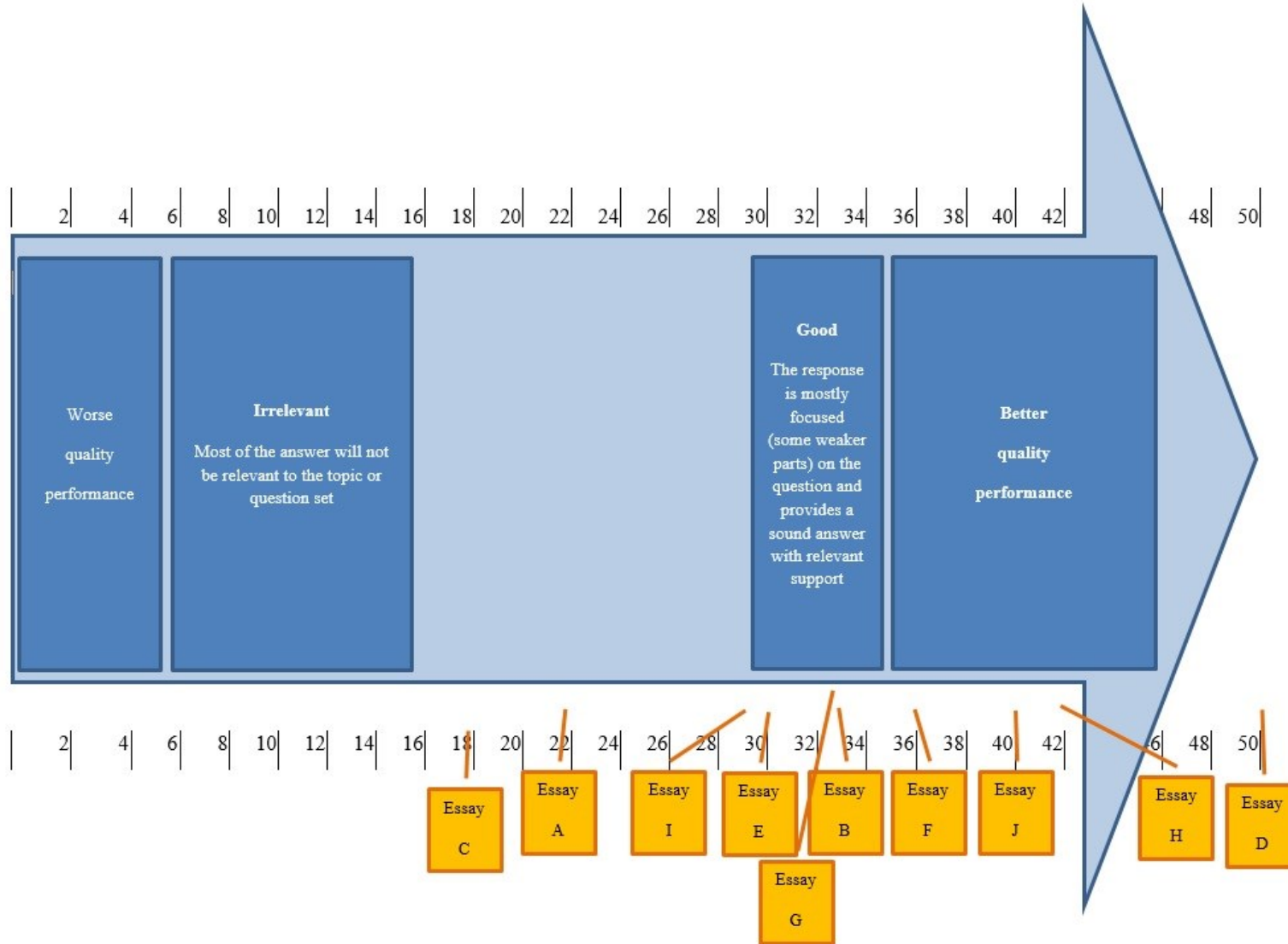


Figure 1. Ranking Mark Scheme Question A

## 2.5. Analysis

### 2.5.1. Quantitative

The within-subjects aspect of the design was statistically efficient because each examiner served as their own control, so examiner-level variation was isolated.

The experiment was more complex than a standard crossover design, because the measurement taken for each examiner for each condition also had a structure; an average of 50 essays. This was exploited in the analysis by modelling for candidate level effects.

The difference from the definitive mark<sup>§</sup> for each marked essay was computed. The appropriate definitive mark was used depending on whether the essay was marked with the Traditional or Ranking Mark Scheme.

Different mark schemes were used for the two procedures, therefore the definitive marks for a given response could vary depending on procedure. In turn, this could mean that the effective mark range may vary and thus differences between examiners could be affected. To take an extreme example, if a mark scheme strongly fostered marking towards the centre of the mark range, the examiners would comply and the differences between examiners would be small, giving a false impression of high reliability. In order to address this, standardised differences were computed, by dividing the difference between the examiner and PE marks (definitive marks) by the standard deviation of the definitive mark for the given question under the appropriate procedure.

Analyses of variance were calculated using a variety of dependent variables. A standard analysis for a 2×2 crossover design, as described by Senn (2002) for example, would be possible at the examiner level (using the mean difference under each procedure), but this would not exploit the multiple measurements for each examiner and the variation at a candidate and response level. As a result, a more complex model was applied:

$$Y_{ijk} = \mu + m_{(i)k} + q_j + \tau_{d[i,j]} + c_l + cr_{jl} + e_{ijkl}$$

where the terms are as follows, notation based on Jones and Kenward (1989):

- $Y_{ijk}$  : Random variable representing marking difference (with observed values  $y_{ijk}$ ) – either actual difference, or absolute difference as appropriate
- $\mu$  : General mean
- $m_{(i)k}$  : The effect of examiner  $k$  in group  $i$
- $q_j$  : The effect of question (and also period)  $j$
- $\tau_{d[i,j]}$  : The direct effect of the treatment (procedure) used in period  $j$  for group  $i$
- $c_l$  : The effect of candidate  $l$
- $cr_{jl}$  : The effect of response by candidate  $l$  to question  $j$
- $e_{ijkl}$  : A random error for candidate  $l$ , examiner  $k$ , period/ question  $j$  and group  $i$ , assumed to be independently and identically normally distributed with mean 0 and variance  $\sigma^2$ .

Note that no carry-over effect (denoted by Jones and Kenward (1989) as  $\lambda$ , and capturing any effect of the method used for the first period on the results in the second period) was included in the model. We followed the advice of Senn (2002) in not testing for this and carrying out a two-stage analysis, as was once common<sup>\*\*</sup>.

<sup>§</sup> There are several legitimate ways to calculate the definitive mark. For the purposes of this study, the Principal Examiner's marks were used as the definitive marks.

<sup>\*\*</sup> A two-stage analysis first tests for the presence of a carry-over effect, then if such an effect is found, only data from the first period are used to test for the treatment (in our case, procedure) effect. As Senn (2002) explains, this

In our experiment it was not possible to separate any effect of period (that is, whether examiners' reliability changed between the first and second sets of responses marked) from the question marked (any effect on reliability due to the essay question, or the History topic) because all examiners in both groups were standardised on and marked question A first, followed by question B.

### 2.5.2 Qualitative

A thematic analysis of the qualitative responses to the questionnaire was guided by advice from Braun and Clarke (2006). There were four themes in the data, however, our focus is:

- Decision Process for the Ranking Procedure (including an Initial Sorting stage)
- Ranking is Time-consuming

Regarding the Decision Process a diagram was drawn to represent the data. A second researcher checked the diagram against their reading of the data.

## 3. FINDINGS

### 3.1. Quantitative

In the six analysis of variance models most effects were strongly statistically significant, reflecting the size of the sample. For the purposes of brevity these figures are not included.

The focus of the research was the procedure effect, which was significant at the 5% level for all analyses except standardised absolute difference from the average examiner mark (Table 1). The unstandardised differences were more easily interpreted as they are articulated in terms of raw marks, while the standardised differences are the number of standard deviations (of the definitive mark distribution). Using the measure of actual difference, the examiners were more severe (by 0.9 of a mark), and further away from the definitive mark, under the Ranking Procedure. When absolute difference from the definitive mark is considered, there was greater marking error (by 0.5 of a mark on average) using the Ranking Procedure than the Traditional Procedure; the difference was somewhat smaller (0.3 marks) when the average examiner mark was used as the comparator.

The standardisation of the differences had a small influence on the direction of the results, but did reduce the significance of the procedure effect when considering the absolute difference. When focusing on the absolute difference with respect to the average examiner mark the difference became statistically insignificant (at the 5% level) when standardised differences were used.

In short, the results supported the hypothesis (H1: The Traditional or Ranking Standardisation result in more reliable marking) and the Traditional Standardisation procedure yielded greater marking reliability.

Figure 2 and Figure 3 show the mean error for each of the responses to question A and B (the results for each procedure originated from a different group of examiners). The x-axis shows the responses arranged by definitive mark, and the 10 Meeting Essays from the experimental condition are shown as vertical lines<sup>††</sup>. The three panels for each question show the same results with different y-axes:

- actual difference from definitive mark
- absolute difference from definitive mark
- absolute difference from average examiner mark<sup>‡‡</sup>.

---

approach is flawed because the test based on the first period only is highly correlated with the pre-test for carry-over, and is thus heavily biased.

<sup>††</sup> Note that these vertical lines are not necessarily the definitive marks for the control condition, but they are retained to enable comparison between the two halves of each graph.

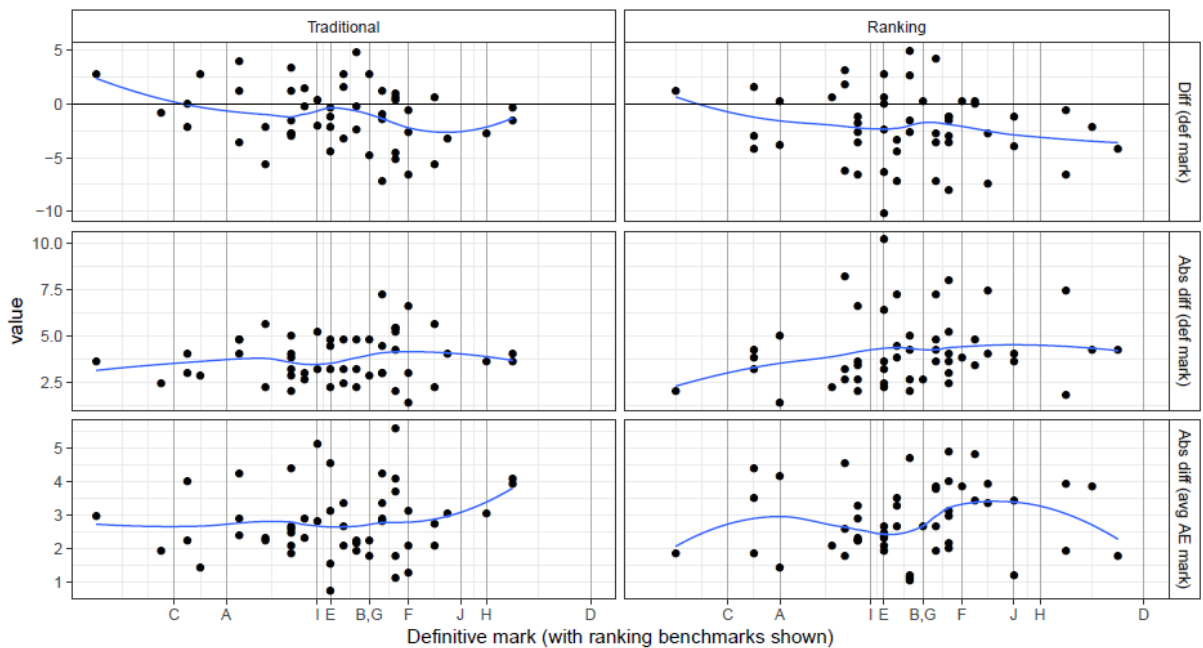
<sup>‡‡</sup> Average actual difference from average examiner mark is not shown, as it is zero for each response.

**Table 1.** Effect sizes for procedure, and estimates of mean

Response	Estimates of means under each procedure		Effect of procedure			
	Traditional	Ranking	Estimate	Standard Error	t value	Pr >  t
Actual difference	-1.628	-2.518	-0.890	0.247	-3.60	0.0003
Absolute difference	3.780	4.326	0.546	0.178	3.07	0.0022
Actual difference (standardised)	-0.2449	-0.3622	-0.1173	0.0363	-3.23	0.0013
Absolute difference (standardised)	0.5688	0.6237	0.0550	0.0261	2.11	0.0353
Absolute difference (average examiner mark)	2.68	3.01	0.332	0.135	2.47	0.0138
Absolute difference (average examiner mark) (standardised)	0.4084	0.4179	0.00949	0.01949	0.49	0.6266

There were few discernable trends. The proximity between the marks for Meeting Essays and the definitive mark of the target essay had no clear effect on the marking reliability of question A or question B. For actual difference in question A, the negative gradient suggests a slight tendency for examiners to be harsher for higher marks (that is, they mark closer to the middle of the mark scale than the PE) in both conditions.

For the bottom panel (absolute difference from average examiner mark) there were a few indications that the Ranking Procedure yielded greater marking reliability at the extremes of the mark range. For question A, there was more consensus among examiners using the Ranking than the Traditional Procedure at the upper end of the mark range. For question B a similar effect was observed at the lower end of the mark range, although, the responses in the allocation had definitive marks lower than the lowest Meeting Essay, G.



**Figure 2.** Mean error for each response along with definitive mark and Meeting Essays: question A



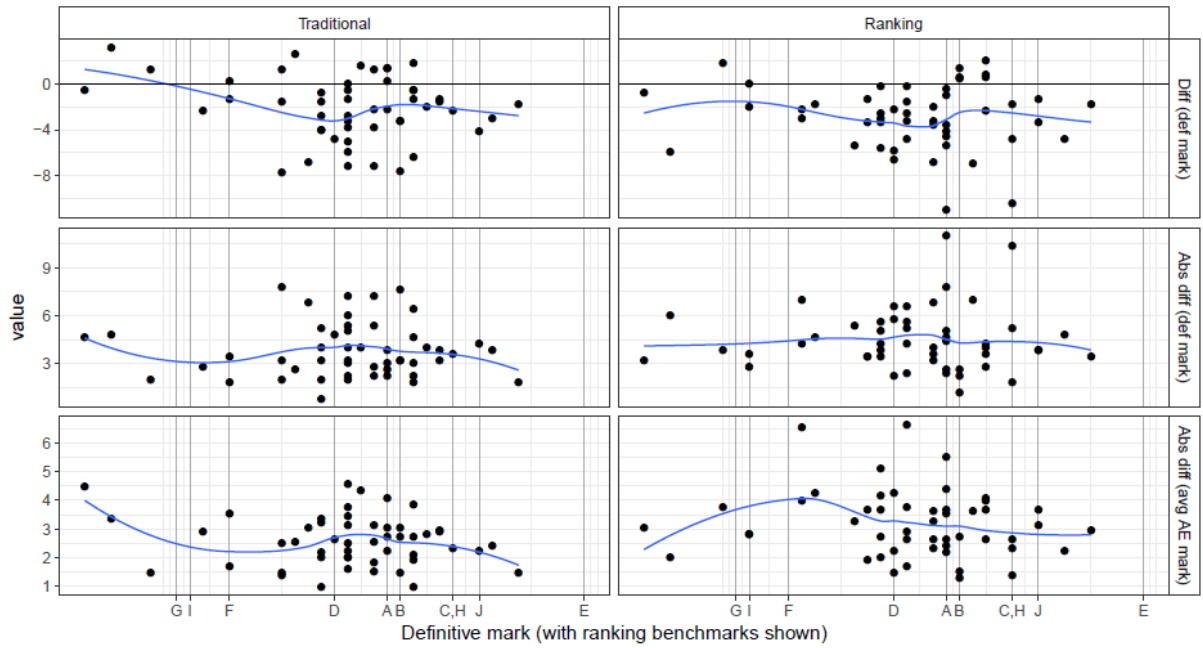


Figure 3. Mean error for each response along with definitive mark and Meeting Essays: question B

### 3.1. Qualitative

The Decision Process for the Ranking Procedure was a complex, multi stage process comprising several paired comparisons of essays (Figure 4). The core Decision Process was preceded in some instances by the Initial Sorting of the essays.

Nine examiners said that the Ranking Procedure was more time-consuming than the Traditional procedure. Reasons cited by examiners included:

- the difficulty of judging how essays compared to one another
- re-reading essays
- dealing with the accumulation of essays available to compare with the target essay
- lack of familiarity.

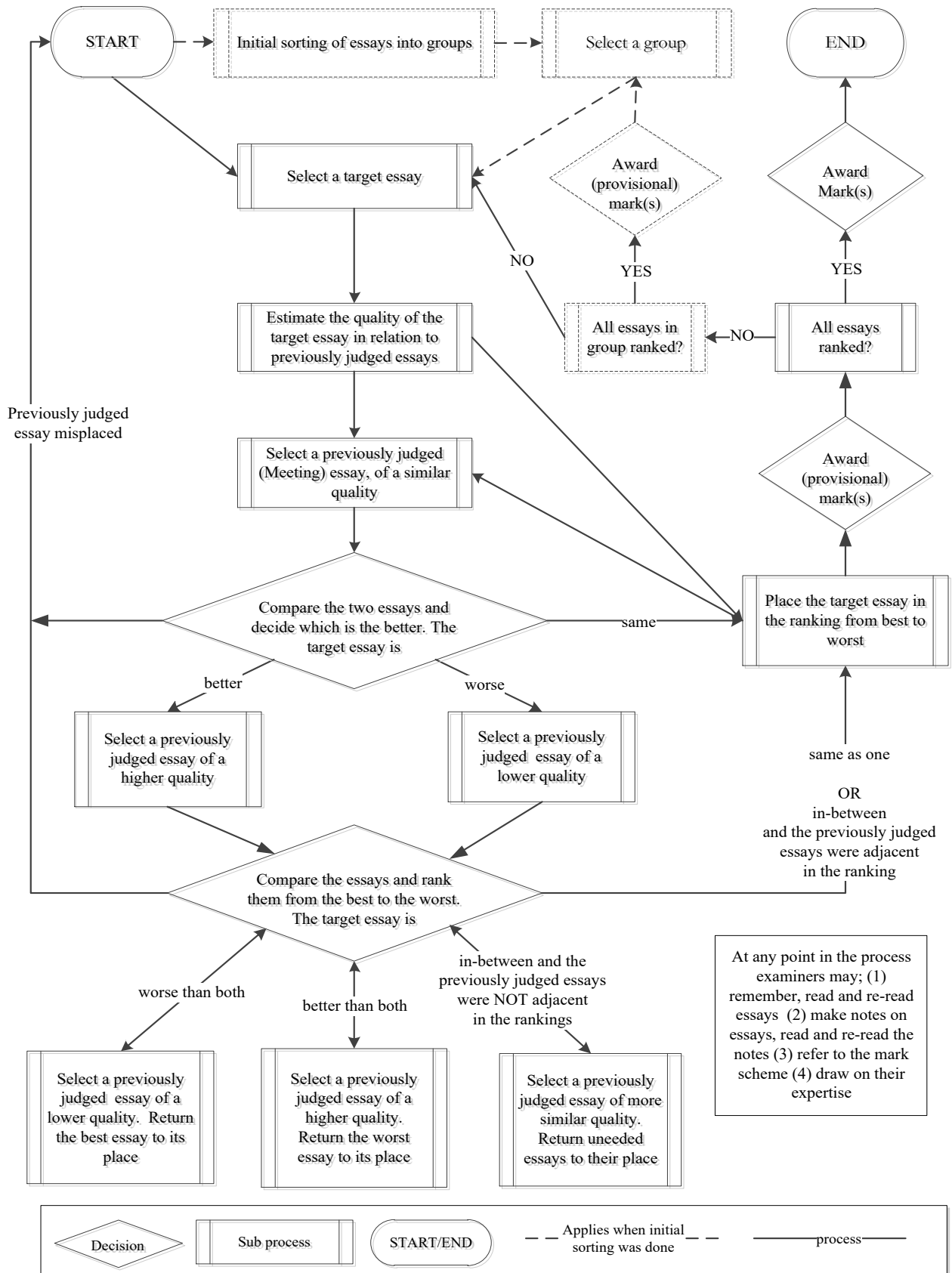


Figure 4. Decision Process for the Ranking Procedure

#### **4. DISCUSSION**

Prior studies illustrate that people's ability to compare two objects surpasses our ability to make absolute judgements about an object (Laming, 2004). Therefore, OCR proposed investigating a new procedure, the Ranking Procedure focusing on comparing the quality of essays and ranking them according to their quality before assigning marks. This research evaluated the marking reliability resulting from both Traditional and Ranking Standardisation, and examiners' experiences of the procedures. The research has limitations, which are outlined below. However, the research generated important findings.

The experiment was designed to simulate standardisation processes. It was beyond the scope of the research to incorporate the many checks and balances that are used to achieve reliable marking in addition to standardisation, for instance scaling. Therefore, the marks and statistics from the experiment were not directly comparable to marking reliability in live marking. However, the experimental data were suitable for testing the hypotheses.

The examiners were likely to be more familiar with the Traditional Procedure than the Ranking procedure, which may result in the former yielding more reliable marking. Arguably, examiners new to both procedures should have been recruited to ensure the experiment was a fair evaluation. However, if an Awarding Organisation were to switch from the Traditional Procedure to another procedure then many examiners would, in the short term, be more familiar with the Traditional Procedure. Consequently, the experiment is an authentic comparison of reliability delivered from the Traditional Procedure and a potential new procedure.

There was insufficient time between operational activities for examiners to complete one condition after another. Therefore, a departure from a crossover design was invoked. Group 1 undertook each stage of the experiment using the Traditional Mark Scheme on question A before completing the parallel stage of the experiment using the Ranking Mark Scheme for question B. Group 2 undertook each stage of the experiment using the Ranking Mark Scheme on question A before completing the parallel stage of the experiment using the Traditional Mark Scheme for question B. Each examiner marked the same essays as the other examiners at each stage of the experiment. The interweaving of stages may have had a confounding effect on each condition. Ideally there would be a 'wash-out' period between the two conditions, to allow any effects from the first half of the experiment to dissipate before commencing the second. The lack of a wash out period was a major practical constraint to the design. It was hoped that the effect of the conditions would be large enough to overpower any confounding variables, particularly as the interweaving of stages was common to the groups (with the two procedures reversed).

Several findings emphasised the limitations of the Ranking Procedure. First, the Traditional Procedure produced greater marking reliability than the Ranking Procedure. The Traditional Standardisation resulted in smaller mean differences for all measures. Second, the procedure effect was statistically significant at the 5% level for all measures of reliability, with the exception of standardised absolute difference from the average examiner mark. This concurred with previous research. When alternatives to Traditional Standardisation were investigated they did not consistently lead to better marking reliability than Traditional Standardisation (Greatorex & Bell, 2008b). Additionally, the mark scheme and feedback to examiners improve marking reliability, but exemplar scripts do not improve marking reliability in terms of absolute difference between the PE's and examiners' marking (Baird, Greatorex, & Bell, 2004). Finally, the high reliability of using paired comparisons in marking (ACJ) has been disputed (Bramley, 2015; Bramley & Vitello, 2018). Based on the reliability measures Ranking Standardisation was not as effective as Traditional Standardisation.

There were additional limitations of the Ranking Procedure. First, it was time-consuming, due to the need to re-read essays. Both using ranking of more than two objects and involving

adaptivity have the potential to reduce the time taken. Whitehouse and Pollitt (2012) maintained that ACJ with pairs was not viable as a form of summative assessment for large scale public examinations in England in its current form as it was too time consuming. This study suggested that the Ranking procedure is also likely to be too time-consuming for these purposes. Secondly, for examiners the Decision Process for the Ranking Procedure is complex, indeed more complex than the decision process in CJ with pairs. Together the thematic analysis and the literature suggested that in its current form the Ranking Procedure was too complex and too time-consuming to be used for summative assessments for large scale public examinations in England.

However, the Ranking Procedure had merits which suggest it is worth further consideration. The Ranking Procedure overcame one drawback of levels-based mark schemes: that marking can be less reliable at the extremities of the mark range. Ideally marking is reliable throughout the mark range. However, prior research showed that marking reliability was better towards the centre of the mark range and not as good towards the top and bottom of the mark scale (Pinot de Moira, 2013) and a remedy is sought. Our qualitative data included comments from examiners that the Ranking Procedure gave greater discrimination and a lower prospect of inaccurate marking. The qualitative findings aligned with the quantitative evidence that at the extremities of the mark scale the Ranking Procedure performed better than the Traditional Procedure. In question A, there was higher reliability among examiners using the Ranking Procedure than the Traditional Procedure at the upper end of the mark range, and in question B a similar effect was noted at the lower end of the mark range. There is no clear cause for this result. It is possible that the holistic marking of the Ranking Procedure outperformed the analytical marking of the Traditional procedure in supporting reliability at the extremes of the mark range. Further consideration may be given to how (features of) the Ranking Procedure can be applied to mark essays at the extremes of the mark range.

## **5. CONCLUSION**

This study investigated the merits and limitations of the Ranking Procedure compared with the Traditional Procedure. The limitations of the Ranking Procedure were that it yielded less reliable marking than the Traditional Procedure, was time consuming and entailed an inefficient Decision Process. However, the Ranking Procedure had several merits. Examiners noted that the Ranking Procedure gave greater discrimination and a lower prospect of inaccurate marking. The quantitative results indicate that the Ranking Procedure produced reliable marking at the extremities of the mark range, whereas traditional levels-based mark schemes tend to generate more reliable marking in the middle of the mark range and less reliable marking at the extremities. The findings suggest that the Ranking Procedure is unsuitable for implementation in public examinations in its current form. However, it may be advantageous to explore techniques for refining the Ranking Procedure so that its merits may be realised, for example in awarding or research studies.

### **ORCID**

Jackie Greatorex  <https://orcid.org/0000-0002-2303-0638>

Tom Sutch  <https://orcid.org/0000-0001-8157-277X>

Karen Dunn  <https://orcid.org/0000-0002-7499-9895>

### **Conflicts of Interest**

The authors declared no conflict of interest.

### **Acknowledgements**

We would like to thank colleagues in OCR and the Research Division at Cambridge Assessment for their help and advice with the study, particularly Beth Black who had the original ideas for

the Ranking Procedure. We would also like to thank the examiners who engaged with the research, especially the Principal Examiner who played a pivotal role in the project.

## 6. REFERENCES

- Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, 18(3), 259-278. doi: <http://dx.doi.org/10.1080/0969594X.2010.546775>
- Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, 11(3), 331-348.
- Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279-293.
- Benton, T., & Gallagher, T. (2018). Is comparative judgement just a quick form of multiple marking. *Research Matters: A Cambridge Assessment Publication* (26), 22-28.
- Billington, L., & Davenport, C. (2011). On line standardisation trial, Winter 2008: Evaluation of examiner performance and examiner satisfaction. Manchester: AQA Centre for Education Research Policy.
- Black, B., Suto, W. M. I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice*, 18(3), 295-318.
- Bramley, T. (2009). Mark scheme features associated with different levels of marker agreement. *Research Matters: A Cambridge Assessment Publication* (8), 16-23.
- Bramley, T. (2015). Investigating the reliability of Adaptive Comparative Judgment *Cambridge Assessment Research Report*. Cambridge, UK: Cambridge Assessment.
- Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 1-16. doi: 10.1080/0969594X.2017.1418734
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101.
- Çetin, Y. (2011). Reliability of raters for writing assessment: analytic - holistic, analytic-analytic, holistic-holistic. *Mustafa Kemal University Journal of Social Sciences Institute*, 8(16), 471-486.
- Chamberlain, S., & Taylor, R. (2010). Online or face to face? An experimental study of examiner training. *British Journal of Educational Technology*, 42(4), 665-675.
- Greatorex, J., & Bell, J. F. (2008a). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education*, 23(3), 333-355.
- Greatorex, J., & Bell, J. F. (2008b). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education*, 23(3), 333-355.
- Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20(3), 281-307.
- Johnson, M., & Black, B. (2012). Feedback as scaffolding: senior examiner monitoring processes and their effects on examiner marking. *Research in Post-Compulsory Education*, 17(4), 391-407.
- Jones, B., & Kenward, M. G. (1989). *Design and Analysis of Cross-Over Trials*. London: Chapman and Hall.
- Kimbell, R. (2007). e-assessment in project e-scape. *Design and Technology Education: an International Journal*, 12(2), 66-76.

- Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007). E-scape portfolio assessment. Phase 2 report. London: Department for Education and Skills.
- Knoch, U. (2007). ‘Little coherence, considerable strain for reader’: A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, 12(2), 108-128. doi: [10.1016/j.asw.2007.07.002](https://doi.org/10.1016/j.asw.2007.07.002)
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12, 26-43. doi: [10.1016/j.asw.2007.04.001](https://doi.org/10.1016/j.asw.2007.04.001)
- Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2012). Halo Effects and Analytic Scoring: A Summary of Two Empirical Studies *Research Report*. New York: Pearson Research and Innovation Network.
- Laming, D. (2004). *Human judgment: the eye of the beholder*. Hong Kong: Thomson Learning.
- Meadows, M., & Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection*. Manchester: National Assessment Agency.
- Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: Centre for Education Research and Policy.
- Michieka, M. (2010). Holistic or Analytic Scoring? Issues in Grading ESL Writing. *TNTESOL Journal*.
- O'Donovan, N. (2005). There are no wrong answers: an investigation into the assessment of candidates' responses to essay-based examinations. *Oxford Review of Education*, 31, 395-422.
- Pinot de Moira, A. (2011a). Effective discrimination in mark schemes. Manchester: AQA.
- Pinot de Moira, A. (2011b). Levels-based mark schemes and marking bias. Manchester: AQA.
- Pinot de Moira, A. (2013). Features of a levels-based mark scheme and their effect on marking reliability. Manchester: AQA.
- Pollitt, A. (2009). *Abolishing marksism and rescuing validity*. Paper presented at the International Association for Educational Assessment, Brisbane, Australia. [http://www.iaea.info/documents/paper\\_4d527d4e.pdf](http://www.iaea.info/documents/paper_4d527d4e.pdf)
- Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157-170.
- Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281 - 300. doi: <http://dx.doi.org/10.1080/0969594X.2012.665354>
- Pollitt, A., Elliott, G., & Ahmed, A. (2004). *Let's stop marking exams*. Paper presented at the International Association for Educational Assessment, Philadelphia, USA.
- Raikes, N., Fidler, J., & Gill, T. (2009). *Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology* Paper presented at the British Educational Research Association, University of Manchester, UK.
- Senn, S. (2002). *Cross-Over Trials in Clinical Research*. Chichester: Wiley.
- Suto, I., Nádas, R., & Bell, J. (2011a). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26(1), 21-51.
- Suto, W. M. I., & Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice*, 15(1), 73-89.
- Suto, W. M. I., Greatorex, J., & Nádas, R. (2009). Thinking about making the right mark: Using cognitive strategy research to explore examiner training. *Research Matters: A Cambridge Assessment Publication*(8), 23-32.

- Suto, W. M. I., & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, 23(4), 477-497. doi: 10.1080/02671520701755499
- Suto, W. M. I., & Nádas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*, 24(3), 335-377. doi: <http://dx.doi.org/10.1080/02671520801945925>
- Suto, W. M. I., Nádas, R., & Bell, J. (2011b). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26(1), 21-51.
- Sykes, E., Novakovic, N., Greatorex, J., Bell, J., Nádas, R., & Gill, T. (2009). How effective is fast and automated feedback to examiners in tackling the size of marking errors? *Research Matters: A Cambridge Assessment Publication* (8), 8-15.
- Whitehouse, C., & Pollitt, A. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. Manchester: AQA Centre for Education Research and Policy.
- Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment*, 10(1). <http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1601/1457>
- Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context*. New York: Pearson Research and Innovation Network.

## Teaching Game and Simulation Based Probability

Timur Koparan <sup>1\*</sup>

<sup>1</sup> Zonguldak Bülent Ecevit University, Ere li Faculty of Education, Department of Mathematics and Science Education, Zonguldak, Turkey

### ARTICLE HISTORY

Received: 13 November 2018

Revised: 22 April 2019

Accepted: 04 May 2019

### KEYWORDS

Probability teaching,  
Educational game,  
Simulation,  
Experimental probability,  
Theoretical probability

**Abstract:** Technology and games are the areas where learners are most interested in today's world. If these two can be brought together within the framework of learning objectives, they can be an advantage for teachers and students. This study aims to investigate the learning environment supported by game and simulation. The games were used to evaluate the basic probability knowledge of the prospective teachers, to demonstrate the role of problem solving in the formation of the mathematical knowledge, and to enable discussing mathematical ideas in a worksheet. Simulations were used for visualization and a large number of experiments. The sampling of the study, by which case study research is adopted, is comprised of 40 prospective teachers at a state university in Turkey. The data were collected by introducing nine open-ended questions by means of games, worksheets and simulation activities. The questions asked relevant to the games include making predictions about the fairness of the games, playing the games in small numbers and in big numbers and the observation of the scores, calculation of the winning probabilities of the gamers both experimentally and theoretically, and their comparisons. The process of finding out the probability information underlying the games by the prospective teachers was analyzed qualitatively by means of worksheets, simulations and in-class observation, and the ways of thinking, intuitions, estimations, strategies, and opinions about the learning situation of the participants were tried to be determined. The results obtained put forward that the learning situation that was set up simultaneously contributed to the knowledge of probability and probability teaching of the prospective teachers; and that the candidates' opinions about the learning situation are positive.

## 1. INTRODUCTION

Probability includes a lot of disciplines (physics, economics, meteorology, genetics, insurance) due to having a wide range of application. Furthermore, the language of probability has come into most part of our daily lives. For instance, the probabilities of the side effects of a medicine, home accidents, raining, chance games, sports competitions, Gal (2005) claim that needs in the real world are supposed to be part of the thought inclining, the things thought at school, assessed

---

CONTACT: Timur KOPARAN ✉ [timurkoparan@gmail.com](mailto:timurkoparan@gmail.com) 📧 Zonguldak Bülent Ecevit University, Ere li Faculty of Education, Department of Mathematics and Science Education, Zonguldak, Turkey

ISSN-e: 2148-7456 /© IJATE 2019



and valued. In this point of view, since individuals come across situations of uncertainty a lot of times and in a lot of places in everyday life, probability has gained importance as a content area in which students should have experience so as to become knowledgeable citizens from the school years. Probability teaching can increase people's levels of interpreting what they see. That's why probability has reached a more important status in the teaching programs of many countries recently (National Council of Teachers of Mathematics, 2000; Koparan & Kaleli Yılmaz, 2015). However, probability is a subject which students have difficulty in learning (Ben Zvi & Garfield, 2004; Koparan & Kaleli Yılmaz, 2015; Koparan & Taylan Koparan, 2019). Another difficulty for teaching probability is the fact that intuitions and truths aren't concordant. In probability teaching, sufficient education and support are needed in order to develop the intuitions of both teachers and students (Batanero, Contreras, Fernandez & Ojeda, 2010). However, such deficiencies as the fact that subjects are generally discussed in teacher centered class situations; that suitable teaching material is missing (Gürbüz, 2006); and that most of mathematics teachers are unqualified to teach probability actively (Bulut, Yetkin, & Kazak, 2002) require studies on developing and applying teaching material and methods, and evaluating the applications in this respect. Since visuality needed for probability problems can't be provided in traditional situations, there is need for alternative learning situations (Koparan, 2019). Using games and simulations are some of these alternative situations.

### **1.1. Game-Based Learning**

Games have always played an important role in learning mathematics as they encourage mathematical thinking (Kamii & Rummelsburg, 2008). Teachers know that the situations in which students learn by doing and experiencing are more valuable. This is closely connected with learning by discovering. However, this is a hard educational problem and teachers usually fail to fulfill it. Games are good examples of learning together in the learning situation

(Bragg, 2007). Learning takes place in a context. For the students, games immediately become useful in learning. Because, there is an attempt to play the game and to contribute to play it better. There are some studies revealing that game-based learning enriches the learning environment, improves the students' performance, increases the students' motivation, provides the opportunity to work with the group and provides a fun learning environment (Hamalainen, 2008; Nisbet ve Williams, 2009; Burguillo, 2010; Ahmad et al., 2010; Gürbüz et al., 2014)

In this study, the games were used to evaluate the basic probability knowledge of the prospective teachers, to demonstrate the role of problem solving in the formation of the mathematical knowledge, and to enable discussing mathematical ideas.

### **1.2. Using Simulations in Probability Teaching**

The potential which dynamic statistical software has can be used to establish learning situations suitable for both teachers and students (Koparan, 2015; Koparan & Kaleli Yılmaz, 2015; Koparan & Taylan Koparan, 2019). Upon the increase in the importance of the subject of probability in teaching programs and access to technology at schools, and on the purpose for teachers to have students to experience repeated trials of the same event, using concrete material and experiments through computer simulations should be encouraged (Batanero, Henry, & Parzysz, 2005). Therefore, students need experimental research to understand the theoretical bases of probability. The experimental research is an opportunity to improve their stochastic intuitions, to help them establish a sound understanding of probability, and to motivate them (Borovcnik & Kapadia, 2009). The studies carried out on probability teaching have recommended using computers as a way to understand abstract or difficult concepts and to increase students' talents (Mills, 2002; GAISE, 2005; Gürbüz, 2008; Koparan, 2015; Koparan, 2016). Batanero, Henry and Parzysz (2005) emphasize that students should execute the simulations to help them solve simple probability problems which are impossible through

physical experiments in computer courses at schools. Simulation is the most suitable strategy in focusing better on concepts and in decreasing technical computations (Borovcnik & Kapadia, 2009). Simulations provide an opportunity to strengthen understanding statistical ideas (Konold, Harradine, & Kazak, 2007) and to support the students' process of learning while studying experiments of chance (Maxara & Biehler, 2007).

Along with the development in technology, thanks to some software, modeling mathematical situations has become easier. So, the students, who have been motivated by means of experiments, can research and discover theoretical solutions. Furthermore, they can be convinced about the theoretical solution model by comparing the theoretical and experimental solutions of a problem. Some researchers have stated that modeling strengthens the applicability of mathematical ideas in life, learning new mathematical concepts and stochastic intuitions, and contribute to understanding mathematical concepts (Maxara & Biehler, 2007; Koparan, 2016). Moreover, experimental and theoretical probabilities should be tied up. That is, modeling should offer the advantage of understanding how algebraic facts influence the observed situations.

The simulation based approach requires a special learning attempt for both teachers and students. Since they require not only a statistical perspective but also modeling skills. But most teachers have very little experience about carrying out experiments of probability and using instruments of simulation, and they might have difficulty in the application of the experimental approach (Stohl, 2005). That's why, there is need to put forward new approaches relevant to probability teaching. The modeling of probability may enable the connection between the real world and mathematics (Greer & Mukhopadhyay, 2005). Beg (1995) points out that modeling in statistics provides an ideal teaching platform for mathematical discoveries due to execution with pictorial and concrete material instead of equations and graphics.

Hawkins (1990) points out that probability teaching can't be reduced to conceptual structures and tools of problem solving only, also, there is obligation to generate ways of logical inference and right intuitions in students. Not only does probability teaching offer different models but also it ensures thinking about such questions in deep as how to get information from sources, or why a model is suitable. Prospective teachers don't take the courses of probability and statistics before the third grade at the university. Unfortunately, teachers aren't always able to have a good preparation period to teach probability during their initial training because of time pressure as well. More research is needed so as to clarify the basic components in probability teaching at every level and the preparation of teachers. In recent years, important research that focused on the education of teaching of mathematics and professional development (Ponte & Chapman, 2006; Hill, Sleep, Lewis, & Ball, 2007) hasn't been redounded on teaching of statistics and probability. This denotes that teaching of statistics and probability is an important research area needing to be developed at school level.

There have been different viewpoints about what the best probability teaching should be so that they can be interpreted by everyone in different situations (Jones, Langrall, & Mooney, 2007). These viewpoints depend on different interpretations of probability. People think of probability in at least three different ways (classical, frequent and subjective) and these viewpoints may appear in the process of teaching and learning. Each of these interpretations has advantages and disadvantages (Batanero, Henry, & Parzysz, 2005). If students need to develop a significant sense about probability, it is of importance for them to accept these different interpretations, and to discover the connections among them and different contexts through which one of them might be useful.

By this study, it was aimed to offer games containing contingent situations to prospective teachers; to get predictions made about the games; to observe the games for few trials and to make experimental probability calculations using simulations for multiple trials; to design such

a learning situation as to enable to put forward relations and models to help to understand theoretical solutions; and to evaluate that learning situation. In accordance with this aim, games based activities were developed by the researcher about probability teaching and those activities were supported by worksheets and simulations. It was thought that the activities, which contained games, predictions, experiments, observation and discussion, would reveal the knowledge of the prospective teachers about the probability buried in the games; and that, besides, they would offer samples of the contents of teaching needed in probability teaching and the use of technology in probability teaching. This study is also going to try to find answers to the questions about how to integrate the use of games and simulations with probability teaching. In accordance with these aims, the problem of research was specified as “What are the efficacy of the use of games and simulations in probability teaching and the opinions of the prospective teachers about it?”.

### 1.3. Conceptual framework

The games based activities executed in this study are based on the Strategy of Predict, Observe and Explain (POE). The POE strategy was developed by White and Gunstone (1992) to expose students’ predictions and reasons for what they did for a particular event. There are three steps in POE strategy.

**Step 1. Predict:** Students are asked to write their predictions about what will happen.

**Step 2. Observe:** Time is given for experiment and observation. Students are asked to write down what they observe.

**Step 3. Explain:** They are asked to think again and take into account the observation. After the students write their explanations on their papers, they discuss their ideas together.

In this study, the prospective teachers were asked to make predictions about the game before playing it. Later, at the observation stage, the games are played with few trials and the gamers think of what to do to win the game. Then, observation was carried out to review the results of multiple trials using simulations. And at the explain stage, which is the final stage, it was aimed to focus on the theoretical solutions and compare the experimental results to the theoretical ones.

Understanding the initial ideas of learners in probability teaching may be used to inform teachers about the ways of thinking of learners; to produce discussions and to motivate students to learn (Joyce, 2006). Surprising cases create situations in which students can be ready to revise their personal theories. The strategy is based on the following principles:

1. If students aren’t asked for predictions about what to happen in the problems, it may not be possible to observe them carefully.
2. Writing their predictions motivates students to learn the answer.
3. Asking students to explain the reasons of their predictions gives teachers their understanding. This may be useful in case of misunderstandings to happen and to improve the understanding they have.
4. Explaining and evaluating their predictions, and hearing others’ predictions help students start to evaluate their learning and build new meanings.

In this study, it was thought that it would be appropriate to use POE strategy, which provides the opportunity of re-examining their own theories, revealing the current understanding of prospective teachers. At the same time the POE is a strategy compatible with the constructivist learning theory (Küçüközer, 2013).

## **2. MATERIAL and METHOD**

In this study, case study research was used. The sampling of the study is comprised of 40 (28 girls and 12 boys) volunteering prospective mathematics teachers studying in a class of totally 60 (45 girls and 15 boys) prospective teachers at a state university. The prospective teachers did those activities in the scope of the statistics and probability course for the first time along with their university life. They have encountered “TinkerPlots Sampler Tool” (Konold & Miller, 2004) last year in Teaching Technology and Material Development Course. However, they only applied data analysis and basic statistical concepts. The prospective teachers had had no experience of probabilistic problems and modeling studies until this study.

### **2.1. Data collection tools**

In this study, games supported by worksheets, in class observations carried out by the researcher, simulation sections made up by means of dynamic software, and an evaluation form to determine the prospective teachers’ opinions were used as means of data collection. The opinions of a specialist who gave the probability course in the selection of the games and formation of the evaluation form were obtained. The six games played by the prospective teachers and the questions about each game are presented in tables under the title of findings. The nine open ended questions in the evaluation form aiming to find out the prospective teachers’ opinions are also presented in the findings part. During the activities, the researcher made participant observation.

### **2.2. Procedure**

In this study, the worksheets developed by the researcher and the simulations developed using TinkerPlots, a software of dynamic statistics, were used as the data collection instruments. The researcher is also a faculty member teaching the prospective teachers probability statistics. They were informed about TinkerPlots 2.3.2 Sampler and how to make simulations for 4 class hours prior to the study. The activities of probability teaching through games lasted 6 six weeks as to 2 hours per week in extra-curricular times. The groups of two were selected randomly for the games. The worksheets were handed out to the groups and the games were introduced. The gamers were asked to write down their predictions about the game. Later, the game was played by the gamers using few dices or less money (20 or 50). After they had played the game, the gamers were asked to make predictions again. And after that, the prospective teachers were asked to prepare a simulation about the game, to use the simulation for multiple trials, and to compute the experimental probability. And finally, they were asked to make a theoretical probability calculation and compare the experimental and theoretical probability results. At the end of the six-week period, the prospective teachers were asked their opinions about the use of simulation supported games in probability teaching.

### **2.3. Data analysis**

The ways of thinking of the prospective teachers about the games were tried to put forward by qualitatively analyzing the data obtained from the worksheets and observations. Some opinions were supported by the simulations made by the prospective teachers. The opinions about the games were also qualitatively analyzed and the outstanding opinions were presented directly with the quotations. The researcher observed the processes of making predictions, playing games, making experimental probability calculations and creating simulation models, using them and passing them to the theoretical calculations. The answers given to the nine open ended questions were analyzed in determining the opinions of the prospective teachers for the game and simulation based learning environment. The answers given to each question are examined and classified into similar and different ideas.

### 3. RESULTS

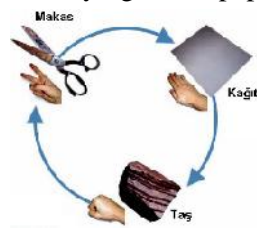
In this research, the results were presented under two titles as *those obtained from the games and simulations* and *from the prospective teachers' opinions about the use of games and simulations*. The findings obtained from the games and simulations contain the prospective teachers' predictions, observations and explanations, and the simulation sections. And the findings obtained from the prospective teachers' opinions about the use of games and simulations contain the qualitative analysis of the answers to the nine open ended questions in the evaluation form.

#### 3.1. The findings obtained from the games and simulations

In this part, the findings obtained from the worksheets and the simulation sections about each game are presented altogether. On the worksheets, the predictions by the prospective teachers about the games, the observations and experimental probability calculations for the few and multiple trials, and in the explanations part, the results about the theoretical probability calculations and the ways of thinking were focused on. And in the simulation sections, the process of the simulation made by the prospective teachers was tried to be revealed.

**Table 1.** The rock paper-scissors-game

The rule of the game	The questions about the game
<p>The game is named as "hands game". Because, both gamers use hand signs while playing the game. The hand signs are as follows:            Fist: Stone            Palm: Paper            Index &amp; Middle Fingers: Scissors            While playing, hands are moved saying stone, paper, scissors and one of them is chosen at the 4<sup>th</sup> motion. The meanings of the probable 3 scores are:            Stone breaks scissors.            Paper wraps stone.            Scissors cut paper.</p>	<p>Think whether the game Stone-Paper-Scissors is fair after playing it. That is, is the chance to win for every gamer equal? In order to understand whether the game is fair, make a tree diagram or a list of the probable scores, and also the list of the winners. How many probable scores are there?            How many choices can the 1<sup>st</sup> gamer make?            How many choices can the 2<sup>nd</sup> gamer make?            What is the multiplication of these two scores?            Compare the multiplication and the number of the probable scores. Compare the winning numbers of the Stone-Paper-Scissors game.            Decide whether the game is fair.</p>



**Predict:** When the predictions were examined before beginning the game in Table 1, it was seen that, of the prospective teachers, 35 said that the game was fair and 15 not. Some statements about these opinions are presented below.

*"I think it isn't fair. Winning depends on the move you do at the last moment."*

*"It isn't fair. You can't know what the opponent is doing. It depends on chance."*

*"I think it isn't fair. Because there will be many cases in which the game ends in a draw."*

*"The cases in which the game ends in a draw are many, but the number of the cases when stone, paper or scissors wins will be equal. That's why the game is fair."*

*"Stone, paper or scissors won't have the upper hand against one another. Their theoretical probabilities are equal. The probability that the game ends in a draw is a little bigger than these."*

**Observe:** The prospective teachers played the game twenty times and made experimental probability calculations. Later, they were asked to make models for multiple trials. In Figure 1, the model made can be seen.

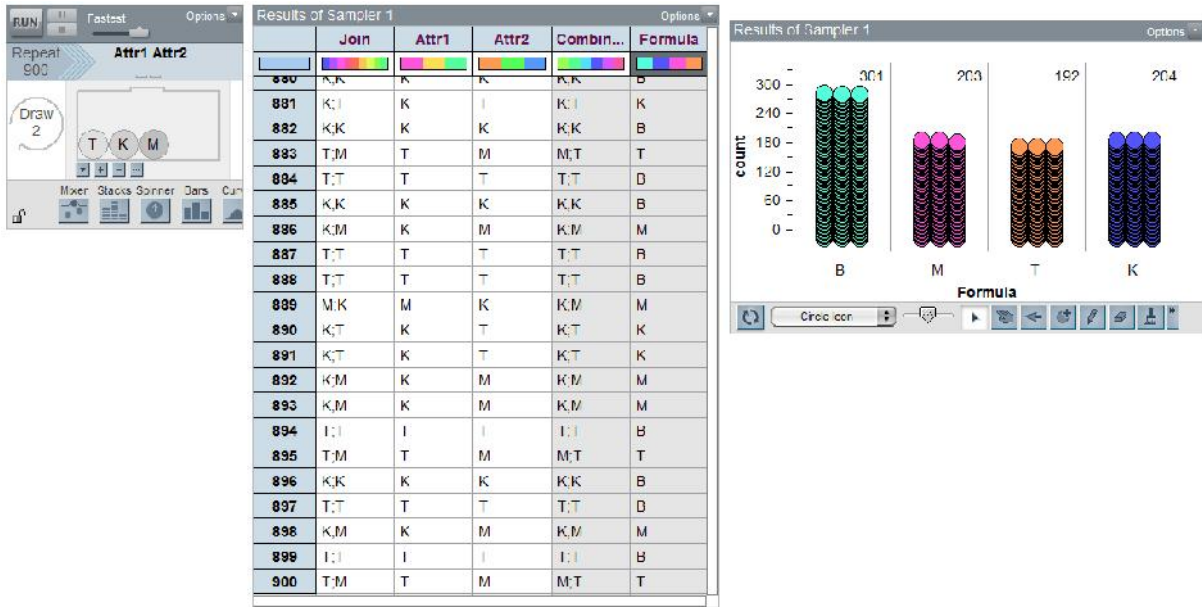


Figure 1. The simulation model made for the stone paper scissors game.

The sampler on the left in Figure 1 is the simulation model made for the stone paper scissors game. T stands for stone, K for paper and M for scissors. It is displayed that, by Draw 2, two of the stone, paper and scissors can be selected randomly, and by Repeat 900, the game can be repeated 900 times. In the figure in the middle, Attr1 displays the moves of the 1<sup>st</sup> gamer and Attr2 displays the moves of the 2<sup>nd</sup> gamer, and Formula displays the winning move. Finally, result of Sampler 1 displays the winning numbers of stone, paper and scissors and the draws.

**Explain:** The prospective teachers filled in the tree diagram in the worksheet after the predictions and calculated the theoretical probabilities. In Figure 2, the tree diagram and the calculations of theoretical probabilities can be seen.

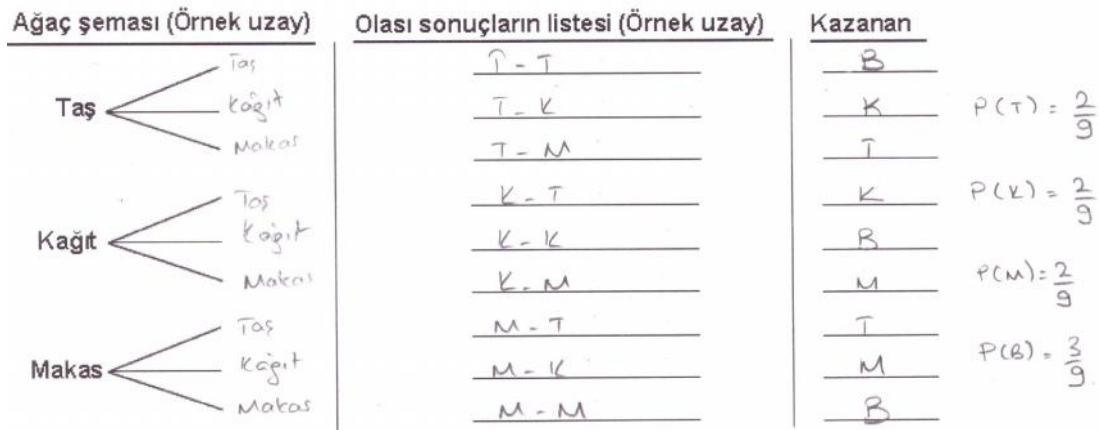


Figure 2. The tree diagram completed by the prospective teachers in the worksheet of the stone paper scissors game.

As can be seen in Figure 2, the prospective teachers fulfilled the directive on the worksheet and filled in the gaps. The worksheet enabled the prospective teachers to see more clearly all of the probable scores and the number of the desired situations. Thus, the theoretical probabilities were calculated correctly. The mathematical facts underlying the experimental calculations were understood much better.

**Table 2.** The difference of the dices game

The rule of the game	The questions about the game
Two friends decide to play a dice tossing game. They throw two dices and find the difference between the dices by extracting the smaller number from the bigger one. If the difference is 0,1 or 2, the first gamer, or 3,4,5, the 2 <sup>nd</sup> gamer wins.	Do you think this game is fair? Explain your predict. Play the game with your friend next to you 50 times. Record the scores in the table. Has your opinion about the equity of the game changed after playing it? If it has, how? What have you noticed after playing the game and seeing the scores which you didn't before? Please, calculate the winning probabilities of the gamers theoretically.



**Predict:** When the predictions were examined before beginning the game in Table 2, whose rules are given, it was seen that, of the prospective teachers, 24 said that the game was fair and 26 not. Some statements about these opinions are presented below.

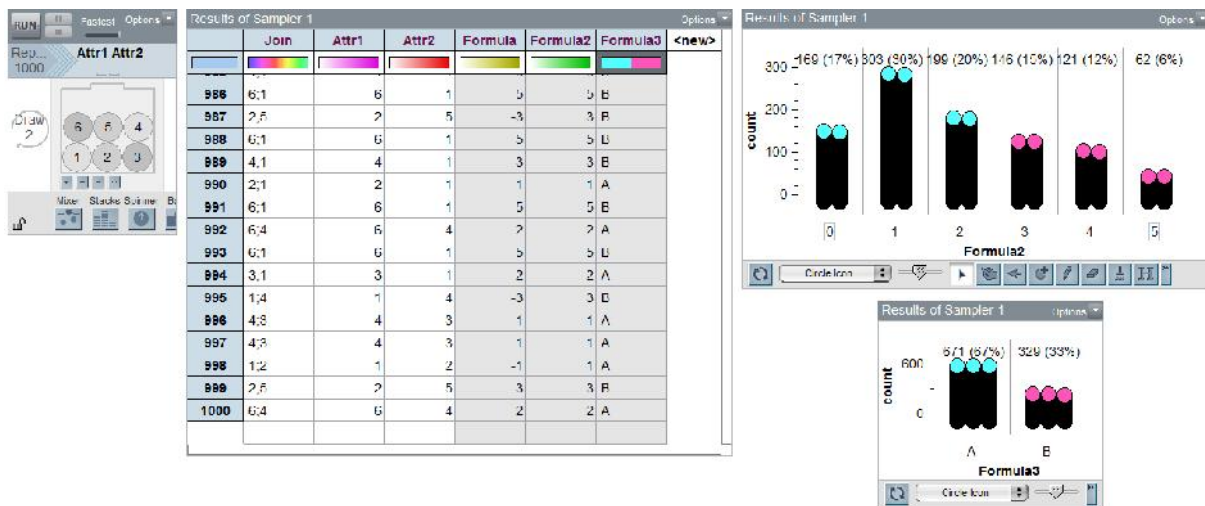
*“I think the game is fair. Because, there are 3 situations for both gamers – 0,1,2 for the first and 3,4,5 for the second. So, I think it is fair.”*

*“It is fair, because, I think the difference will be equal in probability.”*

*“It is fair, because, the difference between the numbers will be the same.”*

*“It isn't, because the probability for the difference to be small is less. E.g. there are two situations for 5 to come but ten for 1.”*

**Observe:** The prospective teachers played the game 50 times and made experimental probability calculations. Later, they were asked to make models for multiple trials. In Figure 3, the model made can be seen.



**Figure 3.** The simulation model made for the difference of the dices game.

The sampler on the left in Figure 3 is the simulation model made for the difference of the dices game. It is displayed that, by Draw 2, two of the numbers 1,2,3,4,5 can be selected randomly, and by Repeat 1000, the trial can be repeated 1000 times. In the figure in the middle, Attr1 displays the 1<sup>st</sup> dice and Attr2 displays the 2<sup>nd</sup> one, Formula 1 displays the difference of the dices, Formula 2 displays the absolute value of the difference of the dices, and Formula 3 displays the winning status of the gamers A and B. While Result of Sampler 1 displays the

distribution of the difference of the dices in consequence with 1000 trials, the graphic lower right hand displays the winning numbers and percentages of the gamers A and B.

**Explain:** The probability information underlying the game was asked at the end of the few and multiple trials, and they were asked to calculate the theoretical probability. Some situations and ways of thinking are presented below.

Some students calculated the probabilities wrongly because they counted the situations wrongly, e.g.  $23/36$  and  $13/36$ ,  $3/5$  and  $2/5$  etc. It was seen that some of them got the wrong result because they didn't take the permutation of the dices into consideration. E.g.,

*In order for the difference of the dices to be 0, they must come (6, 6), (5, 5), (4, 4), (3, 3) (2, 2), (1, 1), to be 1, (6, 5), (5, 4), (4, 3), (3, 2), (2, 1), to be 2, (6, 4), (5, 3), (4, 2), (3, 1), to be 3, (6, 3), (5, 2), (4, 1), to be 4, (6, 2), (5, 1), and to be 5, (6, 1). There are totally 15 conditions for 0,1,2 and 6 conditions for 3,4,5. The probabilities are found to be  $15/21$ .*

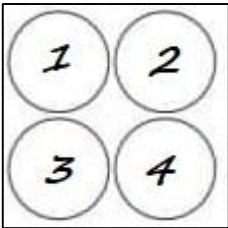

Some of the prospective teachers were seen to have evaluated all the situations and calculated the probabilities correctly.

*In order for the difference of the dices to be 0, they must come (6, 6), (5, 5), (4, 4), (3, 3) (2, 2), (1, 1), to be 1, (6, 5), (5, 6), (5, 4), (4, 5), (4, 3), (3, 4), (3, 2), (2, 3), (2, 1), (1, 2), to be 2, (6, 4), (4, 6), (5, 3), (3, 5), (4, 2), (2, 4), (3, 1), (1, 3), to be 3, (6, 3), (3, 6), (5, 2), (2, 5), (4, 1), (1, 4), to be 4, (6, 2), (2, 6), (5, 1), (1, 5), and to be 5, (6, 1), (1, 6). There are totally 24 conditions for 0,1,2 and 12 conditions for 3,4,5. The probabilities are found to be  $24/36$  and  $12/36$ .*

It was seen that a great majority of the prospective teachers changed their minds after playing the game, and very few were convinced that the game wasn't fair following the theoretical probability calculations.

*“Yes, I've changed my opinion. The game isn't fair. My friend was right.”*

**Table 3.** Dart game

The rule of the game	The questions about the game
Think that the game in the figure is played by rolling the darts three times. It is thought to make use of a dice so as to simulate the dart scores. It is required 1 point to be won when the dice comes 1, no point when 2,3 and 4, and the game to be disregarded when 5 and 6.	Is the suggestion suitable for the simulation of the game? Is it useful? Please, explain the reason. Please, predict the probability for a gamer to win 2 points at 3 throws in a dart game. Please, predict how many times a gamer can win 2 points assuming they play the game 50 times. With real trials, please, play the game 50 times. Please, record the scores you have got. Compare your predictions and the scores in the trials. Please, calculate the theoretical probability and compare it to the experimental probability results.
	

**Predict:** It was asked whether the suggestion given was suitable before beginning to play the game given in Table 3 in order to simulate the game. Some of the opinions about it are as follows;

*“I think it isn't suitable to use a dice. The dice has 6 sides but there are 4 zones in the dart game.”*

*“I think it isn't suitable, because we can hit in the dart game. The dice is a matter of luck, but the dart game is a matter of skill.”*



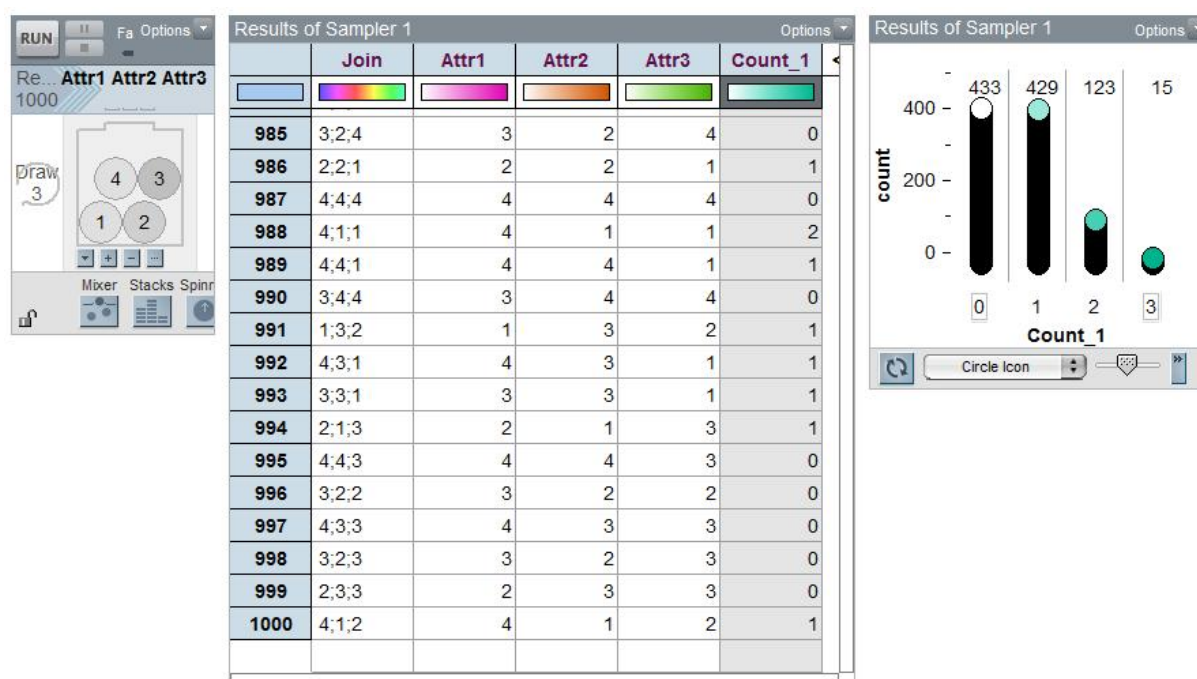
*“I think it isn’t suitable, because the dice has 6 sides but there are 4 zones on the dart board. That’s why it isn’t suitable.”*

*“I think it isn’t suitable, because we disregard when it comes 5 or 6 and throw it over. We roll it 3 times in some trials and more in some others.”*

*“I think it is suitable, because 1 is a point and 2,3 and mean the missed point, and if it comes 5 or 6 it means to roll it over. There are 4 outputs as a result. It is suitable to use a dice.”*

*“I think it is suitable. If the dart game is played one by one, it takes longer. It has been thought to make use of the dice so that everyone can play simultaneously and the scores can be recorded. It is also suitable to disregard the 5 and 6 sides of the dice and roll it over.”*

**Observe:** The prospective teachers played the game 50 times and made experimental probability calculations. Later, they were asked to make models for multiple trials. In Figure 4, the model made can be seen.



**Figure 4.** The simulation model made for the difference of the dart game.

The sampler on the left in Figure 4 is the simulation model made for the difference of the dart game. It is displayed that, by Draw 3, three of the numbers 1, 2, 3, 4 can be selected randomly, and by Repeat 1000, the trial can be repeated 1000 times. In the figure in the middle, Attr1, Attr2 and Attr3 displays the numbers drawn respectively, and count displays the points gained at a trial. Result of Sampler 1 displays the distribution of the points 0, 1, 2 and 3 gained at all the trials.

**Explain:** Following the observation stage, the theoretical probability calculations were performed. Quotes from some of the situations obtained are as follows.

Some of the prospective teachers calculated the probability to win 1 point correctly.

$$“SSD+SDS+DSS = (1/4.1/4.3/4). 3=9/64=0.14”$$

Several prospective teachers calculated the theoretical probabilities of all situations correctly.

“While calculating the probability to win 0 point out of 3 rolls, 1 should never come. That’s to say, it is found that the number of the trios with 2,3 and 4 is  $3 \times 3 \times 3 = 27$ , all situations is  $4 \times 4 \times 4 = 64$ , the probability is  $27/64 = 0.42$ . While calculating the probability to win 1 point out of 3 rolls, 1 should come only once. That’s to say, there should be such trios as (1,2,2), (1,3,3), (1,4,4), (1,2,3), (1,2,4), (1,3,4). While calculating the first three, a calculation of repeating permutation is performed. Accordingly, the number of all the situations fulfilling the requirement of 1 point is found to be  $3+3+3+6+6+6=27$ . And the probability is found to be  $27/64=0.42$ . And while calculating the probability to win 2 points out of 3 rolls, (1,1,2), (1,1,3) (1,1,4) situations are in point. There are totally  $3+3+3=9$  situations by a calculation of repeating permutation. The probability is found to be  $9/64=0.14$ . The probability to win 3 points out of 3 rolls is a single situation as (1,1,1). The probability is found to be  $1/64=0.02$ .”


In Table 4, the experimental and theoretical probability results of a prospective teacher about the points to gain in a dart game can be seen.

**Table 4.** The experimental and theoretical probability results of a prospective teacher about the points to gain in a dart game.

Approaches	0 point	1 point	2 points	3 points
Experimental probability	0.30	0.56	0.10	0.06
Theoretical probability	0.42	0.42	0.14	0.02

When the worksheets were reviewed at the end of the study, it was seen that 40 prospective teachers had calculated the theoretical probability correctly, that six of them had calculated it incorrectly, and that 4 of them had only calculated the experimental probability, not the theoretical one.

**Table 5.** Basketball game

The rule of the game	The questions about the game
<p>A basketball player scores averagely 3 of every 4 shots. In other words, he has a chance of 75% at every shot.</p> 	<p>Is it suitable to use a coin to simulate this problem?                      Would it be suitable to use a dice assuming that 1,2 and 3 are shots on target, 4 is the missed shot, and 5 and 6 are the disregarded shots?                      Please, calculate the probability for the basketball player to score all 6 or 5 of the 6 shots he throws according to the results you will obtain for 50 trials. Calculate it theoretically. Compare it to the experimental results.</p>

**Predict:** It was seen that there were both positive and negative opinions about using a coin or a dice in order to simulate the game in Table 5. Some of the quotes about these opinions are as follows:

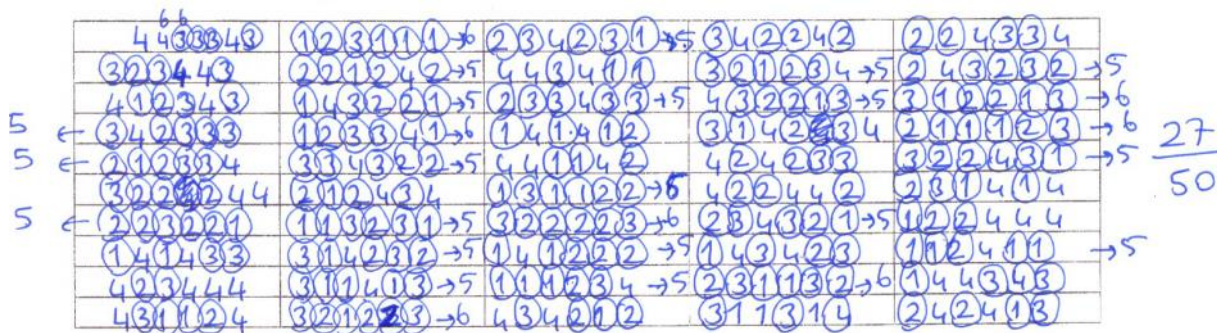
“A coin is suitable, because he either can or can’t score every shot. There are two situations.”

“A coin produces outputs with equal probabilities. It is either heads or tails. Since 3 of the 4 shots in this game will be scores, there won’t be outputs with equal probabilities. That’s why a coin can’t be used in the simulation of this problem. I think a dice is suitable.”

“What can be the connection between a dice and a basketball player? I think it isn’t suitable.”

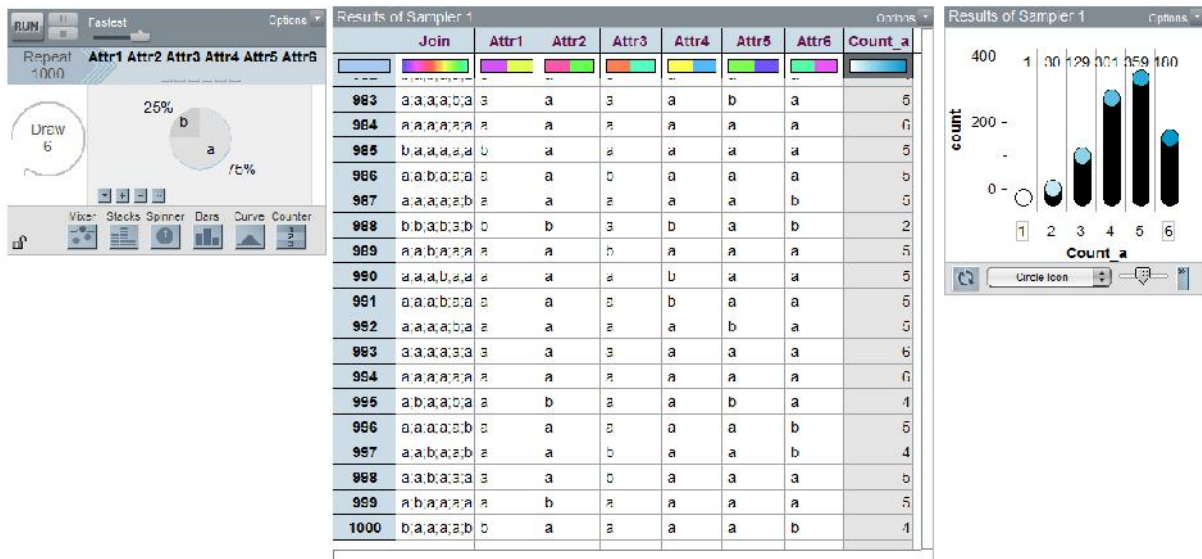
“Since the sample space is different, I think it isn’t suitable.”

**Observe:** The prospective teachers played the game 50 times with their partners and calculated the experimental probability according to the results they got. One of the experimental results is shown in Figure 5.



**Figure 5.** The sample experimental study by the prospective teachers for the basketball game (50 trials)

While the experimental results change between 23/50 and 35/50, it was observed that the values 26/50 and 27/50 were more. In Figure 5, the results of 50 trials are shown. As can be seen in Figure 5, 27 out of 50 trials with a single dice resulted in 5 or 6 scores. It was obtained that the probability for the basketball player to score 5 or all 6 of 6 shots was experimentally  $27/50=0,54$ . The simulation model made to do more trials is shown in Figure 6.



**Figure 6.** The simulation model made for the basketball game

The sampler on the left in Figure 6 is the simulation model made for the basketball game. It is displayed that, by Draw 6, 6 figures should be produced from the spinner randomly, and by Repeat 1000, the trial can be repeated 1000 times. The figure in the middle indicates the figures produced for each trial and the number of the successful shots at a trial.

Result of Sampler 1 displays the distribution of the successful shots in consequence with all trials. The experimental probability was obtained to be  $539/1000=0,539$  according to the simulation results.

**Explain:** Following the observation step, it was proceeded to the theoretical probability calculations. The prospective teachers proceeded to note down all the situations in the

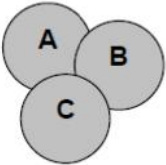
probability for at least 5 out of 6 shots to be scores and add their probabilities. Below is the section about this solution.

*“If the scores are represented by B and the missed shots by K, the possible outputs for the theoretical probability will take place as all 6 scores (BBBBBB) and 5 scores (KBBBBB, BKBBBB, BBKBBB, BBBKBB, BBBBKB and BBBBBK). The probability for all 6 scores is  $(0,75)^6$*

*The probability for 5 scores and a missed shot is  $(0,75)^5 \cdot (0,25)$*

*The theoretical probability according to the Binomial probability distribution is found to be  $(0,75)^6 + 6 \cdot (0,75)^5 \cdot (0,25) = 0,5339$ ”*

**Table 6.** Token flipping game

The rule of the game		The questions about the game
<p>On the side, you see three tokens one of which has an A and a B, the second an A and a C, and the third a B and a C side. The rules of the game played with these tokens are as follows:</p> <p>First, who is to be Player 1 and who 2 is decided. Player 1 flips up all the three tokens. When they fall on the ground, if there is any matching, Player 1 wins 1 point. In case of no matching (all three sides of tokens different), Player 2 wins 1 point. The first player to win 20 points wins the game.</p>		<p>Please, write down your predictions about whether this game is fair. Play this game with your friend until one of you wins 20 points. At every trout, put a line in the winning player’s side in the tally. Please, calculate the winning probabilities of players in this game theoretically. Then, compare the experimental and theoretical probability results.</p>

**Predict:** When the predictions were reviewed before beginning to play the game seen in Table 6, of the prospective teachers, 23 stated that the game wasn’t fair, 8 it wasn’t fair and in favor of Player, and 9 it was fair. Some of the quotes about those opinions are as follows:

*“I don’t think it is fair, because the tokens aren’t shuffled.”*

*“I think it is fair, because the probabilities are equal as there are two of every letter.”*

*“The game is very fair.”*

*“The game isn’t fair at all.”*

*“It isn’t fair. The probability that the token won’t match is higher.”*

*“It isn’t fair. Player 2 wins.”*

*“I think it isn’t fair. Player 1 has more chance.”*

*“The probability for all three to be different is less than two to be the same.”*

**Observe:** the prospective teachers played the game with their partners 50 times and calculated the experimental probability according to the results they obtained. Later, they made a model suitable for the problem to make an observation for multiple trials. In Figure 7, the model made is shown.

In Figure 7, the Sampler on the left is the simulation model made for the token flipping game. The first box represents the token with sides A and B, the 2<sup>nd</sup> the one with A and C, and the 3<sup>rd</sup> the one with B and C. Draw 3, displays that one for each from the boxes should be chosen randomly, and Repeat 1000 displays that the trial should be repeated 1000 times. In the figure in the middle, Attr1, Attr2 and Attr3 display the tokens drawn from the boxes, and in the Formula1 column, 1 indicates that there isn’t a match and 2 indicates that there is a match. Result of Sampler 1 displays the numbers of the situations with and without matches in consequence with 1000 trials, that is to say, the winning numbers and percentages of Player 2 and 1 respectively.

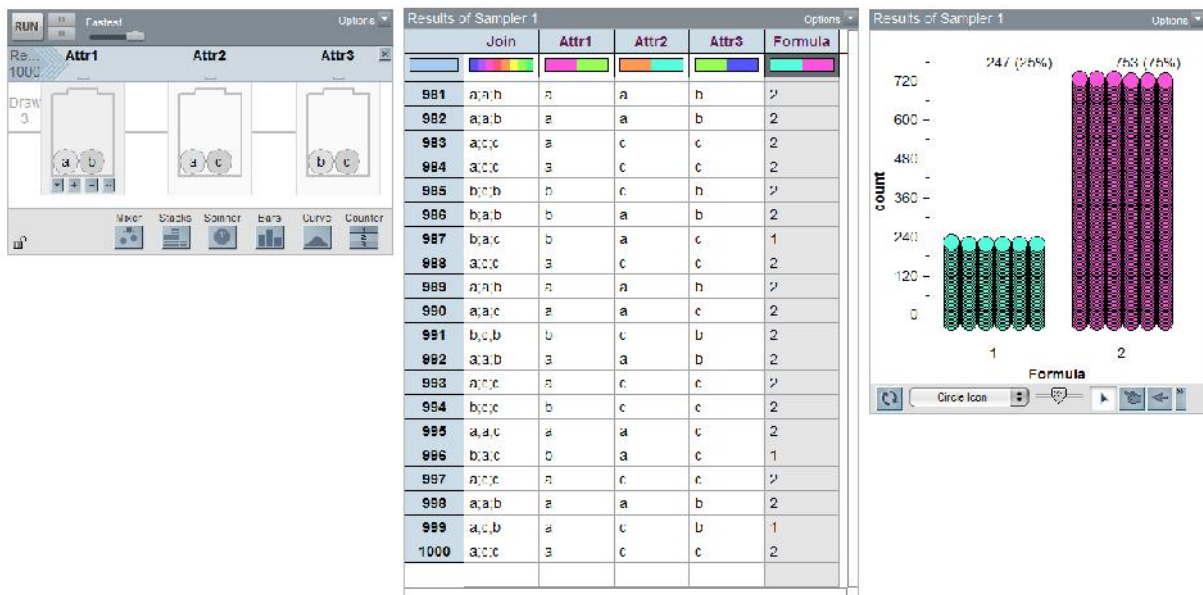
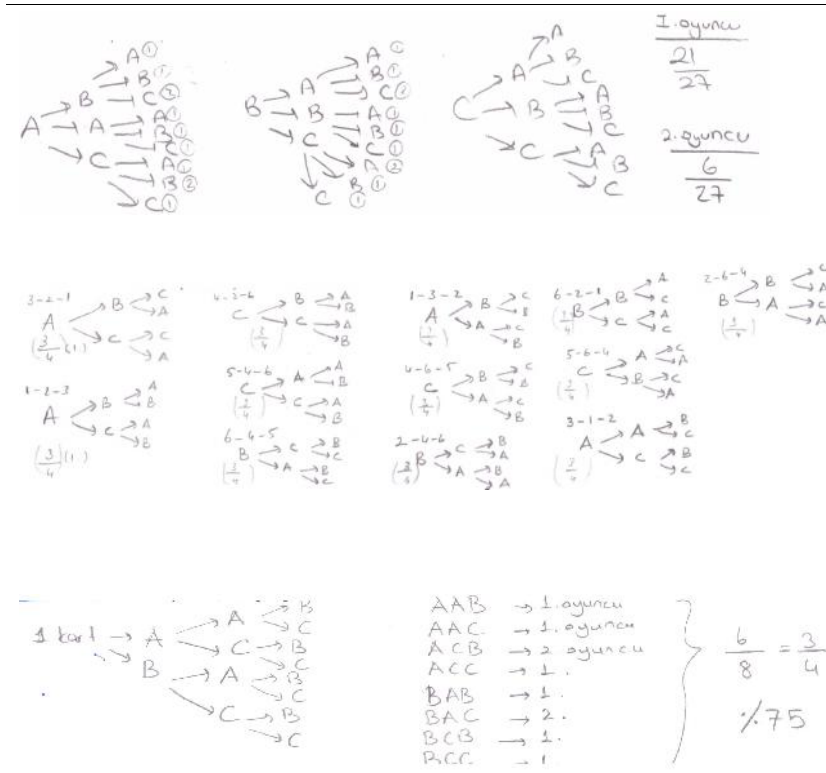


Figure 7. The simulation model made for the token flipping game

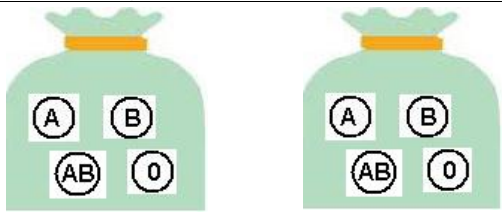
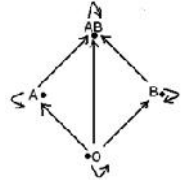
**Explain:** following the observation stage, it was proceeded to the theoretical probability calculations. It was observed that the prospective teachers tried to make up tree diagrams so as to be able to calculate the theoretical probability for the token flipping game. In Table 7, sections from the worksheets are shown.

Table 7. Samples from the tree diagrams made for the token flipping game



It was seen that some of the prospective teachers made wrong tree diagrams, and thus found wrong theoretical probabilities. Some of the prospective teachers found out the theoretical probability by enumerating the cards as 1-2-3 and writing down the situations of permutations. It was seen that some of the prospective teachers found out the result by a single tree diagram.

**Table 8.** Blood donation game

The rule of the game	The questions about the game
<div style="text-align: center;">  </div> <p>In each of two bags, there are 4 cards with the blood types on. The rule of the game is that those in the 1<sup>st</sup> bag are donors and those in the 2<sup>nd</sup> are receptors. When a random card from each bag is drawn, if it is suitable for the first card to donate to the second, player 1, otherwise player 2 wins.</p>	<p>Please, write down whether this game is fair, and your predictions.</p> <div style="text-align: center;">  </div> <p>Make use of the correlation of blood donation in the figure and play the game 32 times with your friend. What kind of a situation do you expect as the number of trials increases? Please, write down your opinions. Calculate the theoretical probability. Compare the experimental and theoretical probability results.</p>

**Predict:** Before beginning to play the blood donation game shown in Table 8, the predictions were reviewed. Of the prospective teachers, 27 said that the game wasn't fair, and 13 it was. Of the participants thinking the game wasn't fair, 14 said the game was in favor of player 1, and 13 in favor of player 2. The quotes about some of the opinions are as follows:

- "It's unfair. The probabilities to win will change as to what player 1 draws."*
- "It's unfair, because the 4 blood types aren't distributed equally."*
- "It's fair, because there are the same cards in both bags."*
- "It's fair, ability or inability to donate will be equal."*
- "I think this game is unfair. Player 1 has more advantages."*
- "I think it's unfair. The probability for player 2 is bigger. In case of inability to donate, always player 2 will win."*
- "The probability for player 1 is bigger, because ability to donate is more in all situations."*

**Observe:** The prospective teachers played the game with their partners 32 times and calculated the experimental probability in accordance with the results they got. Later, they made a model suitable to the problem to observe multiple trials. In Figure 8, the model made is shown.

The Sampler on the left in Figure 8 is the simulation model made for the blood donation game. It represents the blood types A, B, O and AB. Draw 2, displays that two of them can be chosen randomly, and Repeat 1000 displays the trials will be repeated a 1000 times. In the shape in the middle, Attr 1 and Attr2 display the names of the randomly chosen blood types, in the Formula 1 column, D displays the suitable matchings for donation, and Y unsuitable matchings. Result of Sampler 1 displays the number of incorrect and unsuitable situations according to the result of a 1000 trials, that is to say, the numbers and percentages of player 1 and 2.

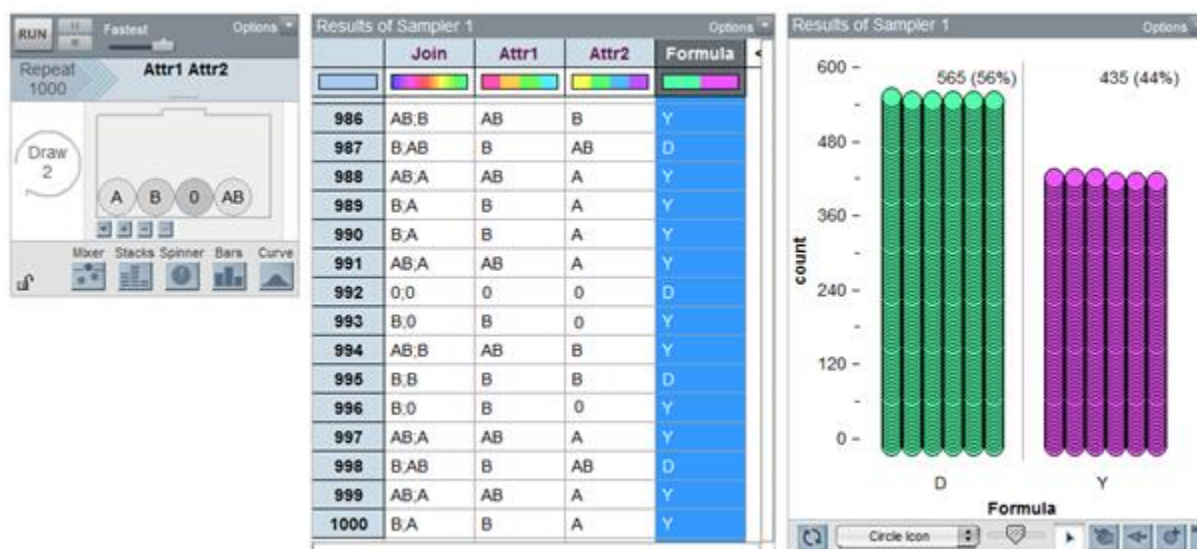


Figure 8. The simulation model made for the blood donation game.

**Explain:** It was seen that the prospective teachers used different methods to calculate the theoretical probability for the blood donation game. In Table 9, sections from the worksheets are shown.

Table 9. Samples from the diagrams made for the blood donation game

<p>1. oyuncu için; <math>\frac{9}{16}</math> kazanma olasılığı                  2. oyuncu için; <math>1 - \frac{9}{16} \rightarrow \frac{7}{16}</math></p>	<p>Some of the prospective teachers used the tree diagram to find out all situations.</p>
<p>1. oyuncunun kazanma olasılığı = <math>\frac{9}{16} = 0,56</math>                  2. oyuncunun kazanma olasılığı = <math>\frac{7}{16} = 0,44</math></p>	<p>Some of the prospective teachers proportioned the number of the suitable situations to the number of all situations by listing all situations.</p>
<p>1. oyuncunun kazanma olasılığı = <math>\frac{9}{16}</math>                  2. oyuncunun kazanma olasılığı = <math>\frac{7}{16}</math></p>	<p>Some of the prospective teachers got the result by finding out the probabilities for suitable blood types in bag 2 to match every blood type in bag 1 and adding them.</p>

### 3.2. The Findings Obtained from the Opinions of the Prospective Teachers

The first question aiming to determine the opinions of the prospective teachers was “How did you like the games you have played in probability courses so far? Did you enjoy them? Please, write down your opinions.” Nearly all of the prospective teachers delivered positive opinions. Some of the quotes from those opinions are as follows:

- ) “Yes, it made the course more fun. I enjoyed playing the games.”
- ) “The games were definitely fun, good and interesting.”
- ) “The games were related to the subjects we studied. They were both useful and entertaining. They made me reframe probability.”

- ) *“I found them not only entertaining but also educative. They made me participate in the course more willingly.”*
- ) *“We played and learned by having fun. We have understood the concepts of probability better. We have noticed our wrong opinions.”*
- ) *“I think such different activities are important if we want to do student centered classes and attract attention to subjects. I really enjoyed playing the games.*
- ) *“I enjoyed playing the games. The fact that my guesses conflicted with the theoretical results increased my curiosity about the games and made me think about the subject a little more.”*
- ) *“I didn’t like the subject of probability much when I was at high school. The games and simulations affected me positively to take the course of probability into consideration and encouraged me to study that subject.”*
- ) *“It sometimes took us long to get the result.”*
- ) *“Studying probability with games becomes more enjoyable. The activities we did gave us useful information to plan student centered courses in our career.”*
- ) *“The coins flipped and dices rolled, card drawing games, pouches blown up by all the class, etc. Shortly, I enjoyed all the games we played a lot. I can say the fact that I thought I was incompetent at the subject of probability before, and that therefore I thought the course was boring has decreased thanks to this method of teaching. Especially when I noticed how our intuitions misled us, I realized the interesting aspect of probability rather than its boring side.”*

Unlike all these opinions, a student said

- ) *“The games were boring, I didn’t enjoy.”*

The second question aiming to determine the opinions of the prospective teachers was “What kind of difficulty did you have while you played the games in the classes of probability?” In this question, more than half of the prospective teachers answered that they didn’t have much difficulty. Some quotes from those who said they had difficulty are as follows:

- ) *“Everytime it came 5 or 6 when we rolled the dice, we disregarded it and rolled again, I always rolled 5 or 6 all the time. It caused the game to take long unnecessarily.”*
- ) *“We were rolling the dice in the games. Because we were at the desks, the dices fell down and I had difficulty finding them. I didn’t have any other difficulty.”*
- ) *“I had difficulty finding theoretical probabilities.”*
- ) *“At the beginning, I had difficulty making guesses and find out the theoretical probabilities. But, as I got used to the games, I had less difficulty.”*
- ) *“Because we played the games in pairs, we couldn’t exchange ideas enough in some games.”*

The third question aiming to determine the opinions of the prospective teachers was “Do you think educative games should be used in probability teaching? Why (not)?” It was seen that all of the prospective teachers had answered the question positively. Some of the quotes from those opinions are as follows:

- ) *“Yes, most of students almost hate the subject of probability. By the help of these games, students can be enabled to learn while having fun and love the course.”*
- ) *“I think they should definitely be used, because many people’s experience and knowledge about probabilities in chance games is rather insufficient. That is the reason why it may be useful to learn some things by putting them into practice and have experience.*
- ) *“I definitely think so. Yes, because it is catchier to learn by seeing and practice instead of memorizing formulas.”*



- ) *“I certainly think they should be used. I think they are necessary in respect of the fact that education should be more efficient and that it be brought to students’ attention.”*
- ) *“Absolutely, educative games should be used in probability teaching. In this way, I find them useful in respect of the fact that, for students, they are preparation for classes, preliminary information, and remarkable.”*
- ) *“Yes, I think so. Because, students may get bored if we always use the direct instruction method. We should try different ways and methods in teaching.”*
- ) *“I think they should be used. Because, they give information about all the probable situations and their distribution.”*
- ) *“Probability is such a subject that needs thinking extremely over it. As a matter of fact, it is also hard to materialize it. That’s why I find these games useful for probability teaching.”*
- ) *“Difficulty in probability is due to inability to think multilaterally. Games and simulations help get rid of that difficulty.”*
- ) *“Yes, being able to see and compare the rates of the experimental and theoretical probabilities satisfies me more than doing classical calculations on paper. In addition, teaching a subject which students are afraid of in this way makes it more enjoyable.”*

The fourth question aiming to determine the opinions of the prospective teachers was “How could using educative games in probability teaching contribute to students?” Some of the quotes from those opinions are as follows:

- ) *“Students get it in return for probability in everyday life.”*
- ) *“By playing games of probability, students’ viewpoints can change and they may comprehend it better.”*
- ) *“It enables them to have more ideas about probability, and, also, it increases permanence in what they have learnt.”*
- ) *“First of all, I think students can focus their attention and give the information to be wanted from them easier. Since it is an occasion on which they can realize that the subject isn’t only comprised of formulas, I find it useful.”*
- ) *“The students’ attitude towards the subject of probability and mathematics may change positively.”*
- ) *“Their intuitions towards probability may develop.”*
- ) *“They could find a chance to see all the probable situations.”*
- ) *“I think games could help students gain communication, reasoning and association skills.”*
- ) *“It enables producing rational solutions instead of memorizing the solutions of problems. Thus, they can easily cope with the probability problems they may come across in everyday life.”*

The fifth question aiming to determine the opinions of the prospective teachers was “Might using games in probability teaching have any disadvantages? What, if any?” most of the prospective teachers said it mightn’t have any. Very few of them said it might. Some of the quotes from those opinions and their frequencies are as follows:

- ) *“The games may be time-consuming.” (7)*
- ) *“During playing the games it may be hard to manage the classroom.” (5)*
- ) *“The teacher needs to master the subject and have experience with the game.” (3)*
- ) *“It may cause a noise.” (3)*

The sixth question aiming to determine the opinions of the prospective teachers was “What is your opinion about using material (coins, dices, cards, simulation, etc.) in probability teaching?” Some of the quotes from the opinions about the question are as follows:

- ) *“I liked using especially simulation, and I’m thinking of using it in my teaching career. As the number of experiments done increases, the theoretical value is approximated more and more, and learning doesn’t take place by memorization; they are more convincing and significant.”*
- ) *“Coins, dice and cards are useful, but simulation is much better material, because either few or multiple trials can be done.”*
- ) *“I think these materials are very important and necessary. I understood it better after playing the games.”*
- ) *“It will be appropriate to use them in classes because learning visually and by practice is permanent.”*
- ) *“It would be better to use different material, particularly those we can see every day and have with us, such as coins. It is actually a sign that it is easy to get information easily if wished. Just a viewpoint.”*
- ) *“Learning visually and by practice is always permanent. The games we can make with coins, dices and cards always be remembered easily. So, it is useful to learn something using plenty of material.”*
- ) *“It calls attention more, and we can realize that probability doesn’t just consist of formulas because the fact that we practice the experimental probability recurrently approximates us to the theoretical probability.”*
- ) *“Yes, I’m in favor of using such material, because learning will be much easier if there are something concrete for students, and as these materials are exactly relevant to the subject of probability, they would contribute a lot to learning.”*
- ) *“Probability is more incomprehensible and difficult without material and games. To speak for myself, probability was more theoretical and incomprehensible in previous years. It is more comprehensible by the help of material.”*

The seventh question aiming to determine the opinions of the prospective teachers was “What should be taken into consideration when choosing educative games to use in classes?” Some of the quotes from those opinions are as follows:

- ) *“Experimental material of is probability should be used without confusing students.”*
- ) *“Games should be associated with the subject.”*
- ) *“Using time in games should be taken into consideration.”*
- ) *“The explanations about games on worksheets should be clear and intelligible.”*
- ) *“I think it would be better to play the games with easy to use material which are used in everyday life and available.”*
- ) *“We should be careful about whether the games can achieve their purpose.”*
- ) *“Games to call students’ interest can be selected.”*
- ) *“They should be suitable for students’ levels.”*
- ) *“Games should be clear, simple and educative.”*
- ) *“They should enable discussing.”*
- ) *“They should be economic, portable and strong.”*
- ) *“The classroom and students should be organized in accordance with games.”*
- ) *“Students should be encouraged to participate actively.”*

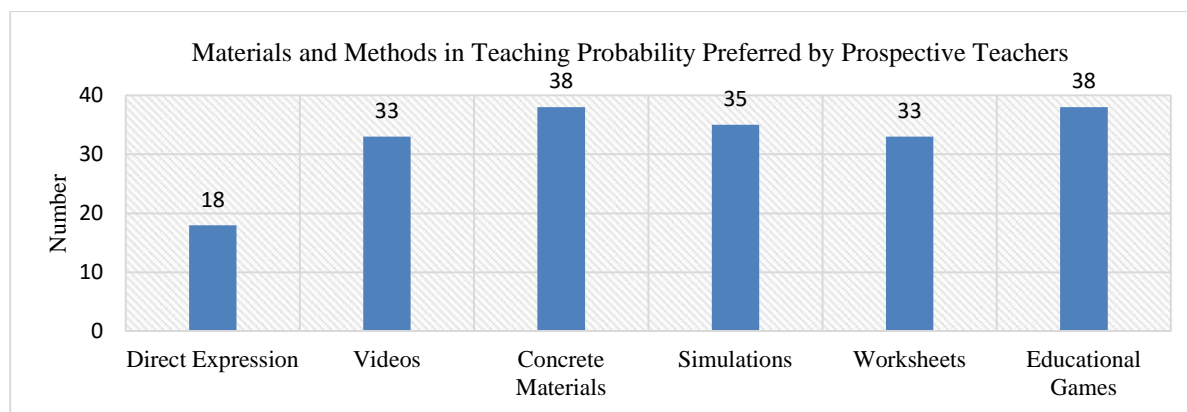
The eighth question aiming to determine the opinions of the prospective teachers was “Would you think of using games in teaching probability in your teaching career? Why (not)?” Some of the quotes from those opinions are as follows:

- ) *“It would be the method for me to use, but I think I would concentrate on simulation, because anything involving technology is more interesting for me.”*

- ) *“I’m thinking of using the games played with dices and coins, because students should sometimes see the probability results with their own eyes by practicing themselves.”*
- ) *“I’m absolutely thinking of it. Now that I had fun at this age, I think it would call the interest of the children of an age level that are keen on games more, and they could understand without having much difficulty.”*
- ) *“I’m certainly thinking of it. Children learn what they need to; and they need games more at those ages. If we can make use of games in teaching the subject of probability, this is an advantage and it should be used.”*
- ) *“Yes, because even I have changed my viewpoint towards the probability course at this age, after playing the games. I think it would have a much more positive effect if we apply it to students of a low age level.”*
- ) *“Yes, I’m thinking of using them, because probability is an abstract subject. I think different approaches are necessary in order to understand some concepts more clearly. I also think that it contributes to students to improve their intuitions.”*
- ) *“Yes, I would think of it. Probability is such a subject that can be associated with real life as a whole. I would try to teach by practice and experience, and make use of games.”*
- ) *“I wouldn’t think of using games frequently, because it would take long. I might sometimes use them in order to attract students’ attention, consolidate what they have learnt and make the class hours’ fun.”*

As can be seen in the quotes, the prospective teachers want to use educative games in probability teaching for various reasons when they become teachers.

And in the ninth question aiming to determine the prospective teachers’ opinions, they were asked about the options they prefer to be used in probability teaching. They could make more than one choice. The findings obtained are shown in Figure 9.



**Figure 9.** Materials and methods in probability teaching preferred by prospective teachers

As can be seen in Figure 9, while most of the prospective teachers stated that videos, concrete material, worksheets and educative games should be used in probability teaching, more than half of them pointed out that the direct instruction method shouldn’t be used. This can be interpreted that the traditional probability teaching isn’t preferable.

#### 4. DISCUSSION

It was seen that the prospective teachers had difficulty in making predictions before playing the games or they made incoherent predictions. This proves that the prospective teachers couldn’t develop proper strategies while making predictions and had wrong intuitions. Hence, there have been researchers who point out that there is need for enough education and support to improve

the intuitions of both teachers and students in probability teaching (Hawkins, 1990; Batanero, Contreras, Fernandes, & Ojeda, 2010).

On the other hand, it can be said that the explanations made after playing the games with concrete material and simulations were generally right. This case suggests that it is useful and necessary not to use an expression only dependent on theoretical knowledge but to do scientific studies enabling discussing situations in which real or realistic life conditions containing probability are supported by games and simulations. Indeed, Koparan (2016) stated that modern approaches should be used in probability teaching.

In this study, using games and simulations contributed to linking the probability knowledge of the prospective teachers to real life situations and to the coming up of the mathematical knowledge underlying the games. Indeed, Koparan (2016) stated that the visualization needed concerning the probability problems hadn't been provided in traditional environments, and that alternative learning environments were needed. However, it enabled the prospective teachers to understand the relation between the experimental and theoretical probabilities. Indeed, Batanero, Henry, & Parzysz, (2005) pointed out that students should accept different comments in order to develop a significant perception and discover the connections between them and different contexts which any of them might be useful for.

In this study, a POE strategy-based educational gaming approach is presented in probability teaching. Researcher observations revealed that the POE strategy-based game approach could significantly increase the interest of students and their retention. That is, the POE can help prospective teachers clarify their own individual opinions and effective in promoting a durable conceptual change, as indicated by some researchers (Küçüközer, 2013; Akpınar, 2014). This means that the POE strategy is an effective way of teaching probability and is fully integrated into the game.

Prospective teachers stated that the use of play, study material, concrete material and simulation in probability teaching makes the teaching of probability more enjoyable and fun. These opinions of prospective teachers are in line with the findings obtained in the study of Kaya and Elgün (2015). Prospective teachers stated that the use of play, study material, concrete material and simulation in probability teaching makes the teaching of probability more enjoyable and fun. Prospective teachers think that the anxiety and fear of the student about the probability lesson can be reduced by the presentation of probability course with games, concrete materials and simulations in a comprehensible way in amusing learning environments. There are studies (Katmada, Mavridis, & Tsiatso, 2014) supporting this opinion of preservice teachers and indicating that games are used in mathematics courses and thus positive changes occur in students' attitudes in courses.

It was seen that the prospective teachers' opinions about the learning situation made up are positive in general. The prospective teachers stated that the probability classes became more attractive and enjoyable with such materials as games, worksheets, simulations, videos, dices and coins. Indeed, Gürbüz (2006) stated that the fact that the subjects are generally taught in teacher centered classrooms and suitable teaching materials are missing or not used affect probability teaching negatively.

## **5. CONCLUSION**

It shouldn't be forgotten that prospective teachers tend to teach however they have learnt. The role of prospective teachers is big in giving individuals basic skills associated with probability. For this reason, it is important that prospective teachers should gain different experiences associated with probability teaching at university. It is necessary that prospective teachers should know the ways to get to the probable answer using certain strategies whenever they come across any situation with probabilities, and bring this way of thinking to students as well.

By this study, it was demonstrated how to present a subject in different ways in teacher training, how to do a didactic analysis in a similar situation, what the basic probability ideas are, and at what level to use these formulated similar situations with secondary school students. Such analyses should make an essence in teacher training with respect to mathematical and didactic viewpoint. Prospective teachers expressed that they wanted to use new approaches in this study when they would teach probability when they become teachers. However, the opinions obtained from the prospective teachers suggest that the activities analyzed in this study contributed to the knowledge of probability and probability teaching of the prospective teachers synchronously.

## ORCID

Timur Koparan  <https://orcid.org/0000-0002-3174-2387>



## 6. REFERENCES

- Ahmad, W., Shafie, A., & Latif M. (2010). Role-playing game-based learning in mathematics. *Electronic Journal of Mathematics & Technology*, 4(2), 184-196.
- Akpınar, E. (2014). The Use of Interactive Computer Animations Based on POE as a Presentation Tool in Primary Science Teaching. *Journal of Science Education and Technology*, 23(4), 527-537.
- Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Eds.), *Exploring probability in school: Challenges for teaching and learning*, (pp. 15-37). Netherlands: Kluwer.
- Batanero, C., Contreras, J. M. Fernández, J. A. & Ojeda, M. M. (2010). Paradoxical games as a didactic tool to train teachers in probability. Publicación en C, Reading (Eds.), Proceedings of the Eight International Conference on Teaching Statistics [CD-ROM]. Lubjana: International Association for Statistical Education. ISBN: 978-90-77713-54-9. Tipo de contribución: Trabajo referido. 4 -6 Julio 2010.
- Begg, A. (1995). Statistics and the mathematical processes. *Teaching Statistics*, 17(2), 40–45.
- Ben-Zvi, D., & Garfield, J. (2004). *The Challenge of developing statistical literacy, reasoning, and thinking*, Kluwer Academic Publishers.
- Borovcnik, M., & Kapadia, R (2009). Research and developments in probability education. *International Electronic Journal of Mathematics*, 4(3), 111-130.
- Bragg, L. (2007). Students' conflicting attitudes towards games as a vehicle for learning mathematics: A methodological dilemma. *Mathematics Education Research Journal*, 19(1), 29–44.
- Bulut, S., Yetkin . E., & Kazak S. (2002). Investigation of prospective mathematics teachers'probability Achievement, Attitudes Toward Probability and Mathematics with Respect to Gender. *Hacettepe University Journal of Education*. 22, 21-28.
- Burguillo, J. C. (2010). Using game theory and competition-based learning to stimulate student motivation and performance. *Computers and Education*, 55, 566–575.
- Gaise (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report: A curriculum framework for PreK-12 statistics education*. The American Statistical Association (ASA). <http://www.amstat.org/education/gaise/>
- Gal, I. (2005). Towards “probability literacy” for all citizens: building blocks and instructional dilemmas. In G.A. Jones (Eds.) *Exploring probability in school: Challenges for teaching and learning*, (pp. 39–63). New York: Springer.
- Gürbüz, R. (2006). Olasılık kavramlarının ö retimi için örnek çalı ma yapraklarının geli tirilmesi [Development of study sheets for the teaching of probability concepts]. *Çukurova University Journal of Faculty of Education*, 3(1), 111–123.

- Gürbüz, R. (2008). Olasılık konusunun ö retiminde kullanılabilecek bilgisayar destekli bir materyal [A computer aided material for teaching ‘probability’ topic]. *Mehmet Akif Ersoy University Journal of Faculty of Education*, 8(15), 41-52.
- Gürbüz, R., Erdem, E., & Uluat B. (2014). Reflections from the Process of Game-Based Teaching of Probability. *Croatian Journal of Education*, 16(3), 109-131.
- Greer, G., & Mukhopadhyay, S. (2005). Teaching and learning the mathematization of uncertainty: historical, cultural, social and political contexts. In: G.A. Jones (Eds.) *Exploring probability in school: Challenges for teaching and learning*, (pp. 297–324). New York: Springer.
- Hamalainen, R. (2008). Designing and evaluating collaboration in a virtual game environment for vocational learning. *Computers & Education*, 50, 98–109.
- Hawkins, A. (1990). Training teachers to teach statistics. Voorburg: International Statistical Institute.
- Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers’ mathematical knowledge: What knowledge matters and what evidence counts? In F. Lester (Eds.), *Second Handbook of Research on Mathematics Teaching and Learning*, (pp. 111-155). Charlotte NC: Information Age Publishing.
- Jones, G.A., Langrall, C.W., & Mooney, E.S. (2007). Research in probability: responding to classroom realities. In: F.K. Lester Jr (Eds.) *Second Handbook of Research on Mathematics Teaching and Learning*, (pp. 909–955). Reston: The National Council of Teachers of Mathematics.
- Joyce, C. (2006). Predict, observe, explain (POE). <http://arb.nzcer.org.nz/strategies/poe.php> (accessed on 10 June 2017)
- Kamii, C., & Rummelsburg, J. (2008). Arithmetic for first graders lacking number concepts. *Teaching Children Mathematics*, 14(7), 389–394.
- Katmada, A., Mavridis, A., & Tsiatsos, T. (2014). Implementing a gam efor supporting learning in mathe-maticss. *The Electronic Journal of e-Learning*, 12(3), 230-242.
- Kaya, S., & Elgün, A. (2015). E itsel oyunlar ile desteklenmi fen ö retiminin ilkokul ö rencilerinin akademik ba arısına etkisi [The influence of instructional games in science teaching on primary students’ achievement]. *Kastamonu Education Journal*, 23(1), 329-342.
- Konold, C. & Miller, C. (2004). TinkerPlots™ Dynamic Data Exploration 1.0. Emeryville, CA.: Key Curriculum Press.
- Konold, C, Harradine A, & Kazak S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, 12(3), 217-230.
- Koparan, T., & Kaleli Yılmaz, G. (2015). The effect of simulation-based learning on prospective teachers’ inference skills in teaching probability. *Universal Journal of Educational Research*, 3(11), 775-786.
- Koparan, T. (2015). Olasılık Ö retiminde Simülasyon Kullanımı [Using similation in teaching probability]. *Ondokuz Mayıs University Journal of Faculty of Education*, 34(2), 22-36.
- Koparan, T. (2016). Using simulation as a problem solving method in dice problems. *British Journal of Education, Society & Behavioural Science*, 18(1), 1-16.
- Koparan, T. (2019). Examination of the dynamic software-supported learning environment in data analysis, *International Journal of Mathematical Education in Science and Technology*, 50(2), 277-291.
- Koparan, T., & Taylan Koparan, E. (2019). Empirical Approaches to Probability Problems: An Action Research. *European Journal of Education Studies*, 5(10), 100-117.
- Küçüközer, H. (2013). Designing a powerful learning environment to promote durable conceptual change. *Computers & Education*, 68, 482-491.

- 
- Maxara C, & Biehler R. (2007). Constructing stochastic simulations with a computer tool students' competencies and difficulties. In D. Pitta, Pantazi, & P. G. Philippou (Eds.), *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education*. Larnaca, Cyprus.
- Mills, J. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1), 1-20.
- National Council of Teachers of Mathematics (NCTM), (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- Nisbet, S., & Williams, A. (2009). Improving students' attitudes to chance with games and activities. *Australian Mathematics Teacher*, 65(3), 25–37.
- Ponte, J. P., & Chapman, O. (2006). Mathematics teachers' knowledge and practices. In A. Gutierrez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education: Past, present and future*, (pp. 461-494). Roterdham: Sense.
- Stohl, H. (2005). Probability in teacher education and development. In G. Jones (Ed.). *Exploring probability in schools: Challenges for teaching and learning* (345-366). New York: Springer
- White, R., & Gunstone, R. F. (1992). *Prediction-Observation-Explanation*. In R. White & R. F. Gunstone, *Probing understanding* (pp. 44-46). London, England: The Falmer Press.

## Explanatory Item Response Models for Polytomous Item Responses

Luke Stanke <sup>1\*</sup>, Okan Bulut <sup>2</sup>

<sup>1</sup>Tessellation Minneapolis, MN, USA

<sup>2</sup>Centre for Research in Applied Measurement and Evaluation University of Alberta

### ARTICLE HISTORY

Received: 19 January 2019

Revised: 08 May 2019

Accepted: 15 May 2019

### KEYWORDS

Polytomous IRT,  
Explanatory item response  
modeling,  
Assessment,  
Partial Credit Model

**Abstract:** Item response theory is a widely used framework for the design, scoring, and scaling of measurement instruments. Item response models are typically used for dichotomously scored questions that have only two score points (e.g., multiple-choice items). However, given the increasing use of instruments that include questions with multiple response categories, such as surveys, questionnaires, and psychological scales, polytomous item response models are becoming more utilized in education and psychology. This study aims to demonstrate the application of explanatory item response theory (IRT) models to polytomous item responses in order to explain common variability in item clusters, person groups, and interactions between item clusters and person groups. Explanatory forms of several IRT models – such as Partial Credit Model and Rating Scale Model – are demonstrated and the estimation procedures of these models are explained. Findings of this study suggest that explanatory IRT models can be more parsimonious than traditional IRT models for polytomous data when items and persons share common characteristics. Explanatory forms of the polytomous IRT models can provide more information about response patterns in item responses by estimating fewer item parameters.

## 1. INTRODUCTION

Item response theory (IRT) models have been widely used for the design, scoring, and scaling of educational and psychological assessments during the past three decades (Bond & Fox, 2001; Embretson & Reise, 2000; Lord, 1980; van der Linden & Hambleton, 1997; Wright & Masters, 1982). Dichotomous IRT models, such as the Rasch model (RM; Rasch, 1960/1980) and two-parameter logistic model (2PL; Birnbaum, 1968), have been more common in practice due to the popularity of standardized assessments with dichotomously scored multiple-choice items. However, today's educators desire to differentiate their students with more innovative assessment tools that consist of not only dichotomous items but also items with more than one score level (i.e., polytomous items). Similarly, many researchers prefer to use surveys, questionnaires, and scales with Likert-type items that often consist of multiple, ordered response categories (e.g., strongly disagree, disagree, agree, and strongly agree). To

**CONTACT:** Okan BULUT ✉ [bulut@ualberta.ca](mailto:bulut@ualberta.ca) ☒ Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB, T6G 2G5 Canada

ISSN-e: 2148-7456 /© IJATE 2019



accommodate such items, polytomous IRT models need to be utilized. Polytomous IRT models, including the Nominal Response Model (NRM; Bock, 1972), the Graded Response Model (GRM; Samejima, 1969), the Sequential Response Model (SRM; Tutz, 1990, 1991), the Rating Scale Model (RSM; Andrich, 1978), and the Partial Credit Model (PCM; Masters, 1982), can be used for items with either nominal or ordered response categories.

The traditional IRT models – regardless of number of response categories – can only provide direct information regarding respondents’ trait levels in aptitude, achievement, cognitive abilities, and so on, as well as item information concerning the difficulty, discrimination, and fit of an item to the selected IRT model. Although researchers and practitioners often use these descriptive measures for making decisions regarding respondents and items, traditional IRT models are not able to identify any systematic effects that result from the design of a measurement instrument. That is, these models do not explain common variability across items or across respondents based on the design or theory behind the instrument. Measuring the commonality of responses is an important step in test development because it allows test developers to assess the degree which construct-relevant or construct-irrelevant features – including linguistic, communicative, cognitive, cultural, or physical features – are related to the construct being measured (AERA, APA, NCME, 2014). For example, consider a test that is taken by examinees who either are native speakers of English or speak English as a secondary language, but understanding the English language is not relevant to the target construct being measured. Under traditional IRT models, there would be no way of directly estimating the mean difference between primary and secondary English speakers’ performances.

Expanding the same example, assume that this test assesses mathematical knowledge for middle-school students and researchers are interested in examining the potential effects of including graphics on test items to assist students in answering the items. The researchers can create two equivalent test forms where one test form contains images on half of the items, while the remaining items have no images. The second form contains the same items as the first form, but the presence or absence of images on items is the opposite of the first form. Using traditional IRT models, there would be no way of directly estimating the impact of using images on the difficulty levels of these test forms.

Information about the mean differences between primary and secondary English speakers or the impact of images in test items can be directly estimated using Explanatory Item Response Modeling (EIRM). De Boeck and Wilson (2004) introduced the EIRM framework for measuring common variability in item clusters, respondent groups, or the interactions between item clusters and respondent groups. Instead of estimating the descriptive effects of respondents’ trait level or item difficulty, the explanatory item response models extract information from responses by including explanatory variables. Under the EIRM framework, traditional IRT models can be formulated as a subset of models that belong to a larger class of models – generalized linear mixed models (GLMMs). GLMMs can function as explanatory IRT models when the model includes an item covariate, a person covariate, or a person-by-item covariate (De Boeck & Wilson, 2004; Wilson, De Boeck, & Carstensen, 2008). The EIRM approach defines responses to items as repeated measures nested within each respondent in a multilevel framework. Within a multilevel model, the effects of explanatory variables can be estimated either as fixed or random effects. The linear logistic test model (LLTM; Fischer, 1973; De Boeck, 2008), the latent regression Rasch model (Zwinderman, 1991), and the latent regression LLTM are widely-used forms of explanatory IRT models (Desjardins & Bulut, 2018).

With EIRM, the object of measurement is typically not at the item or respondent levels, but a higher level to explain the relationship among the items or respondents. In the earlier example, explanatory IRT models can provide information to explain the mean differences between

primary and secondary speakers of English, and help determine whether there is any impact of including images on the difficulty of items within the same model. Therefore, researchers can analyze item response data from tests using a perspective that goes beyond common practices in psychology and educational measurement (De Boeck & Wilson, 2004, p.7). The inclusion of explanatory variables in IRT models is typically based on a pre-defined theory. In the case of the example above, explanatory variables indicating English as a primary language (a person covariate) and presence of an image in an item (an item covariate) could be included as predictors in the same model.

### **1.1. Significance of Study**

To date, EIRM has been mostly applied to either dichotomous data or pseudo-dichotomous data where polytomous response categories have been collapsed into binary categories through the selective grouping of ordered or nominal response categories (e.g., Bulut, Palma, Rodriguez, & Stanke, 2015; De Boeck & Partchev, 2012; Plieninger & Meiser, 2014; Prowker & Camilli, 2007; Scheiblechner, 2009; Verhelst & Verstralen, 2008). Despite more recent attempts that described how to estimate explanatory IRT models for items with ordered or nominal response categories (e.g., Jiao & Zhang, 2014; Wang & Liu, 2007; Tuerlinckx & Wang, 2004), the proposed models have been limited in terms of utilizing a familiar polytomous IRT model (e.g., GRM, PCM, and RM) within the EIRM framework. Also, these models mostly focused on the first threshold between item response categories as it is often interpreted as the difficulty of polytomous items. In this study, we aim to establish a basis for explanatory IRT models for polytomous item response data, not by formulating a new model, but elucidating the flexibility and usefulness of the existing polytomous explanatory IRT models. We used a real dataset to demonstrate the utility of the explanatory IRT models by examining the threshold parameters and model fit statistics. In addition, we described a new parameterization of the explanatory IRT models for polytomous response data that allows a straightforward estimation of these models in R (R Core Team, 2018).

### **1.2. Theoretical Background**

#### **1.2.1. Explanatory Item Response Modeling**

Explanatory item response models can utilize IRT for both measurement and explanation purposes (De Boeck & Wilson, 2004). The main advantage of these models is the flexibility to analyze items and respondents, while simultaneously decomposing common variability across item- and respondent groups (Briggs, 2008). In addition, EIRM allows a theory to be directly imputed into IRT models. EIRM has been applied to a wide array of psychometric and measurement studies, including construct validity studies aiming to explain common variability in item parameters (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Embretson, 2006), latent growth modeling (Wilson, Zheng, & McGuire, 2012), local item dependence studies (Wang & Wilson, 2005), differential functioning (Luppescu, 2002; Williams & Beretvas, 2006; French & Finch, 2010), item parameter drift (Bulut et al., 2015), and contextual effect studies (Albano, 2013; Kan & Bulut, 2014; Kubinger, 2008).

Despite the increasing popularity of EIRM in educational and psychological settings, there are only a few instances of EIRM where researchers used explanatory variables to explain the item-level or person-level variation in polytomous response data. One of the very first attempts to study EIRM with polytomous data was Tuerlinckx and Wang's (2004) study where the authors fit a series of models to a verbal aggression dataset that consisted of 24 items with three response categories. The study examined model fit of five models: a RSM, an explanatory RSM with two person covariates, a PCM, an explanatory PCM with two person covariates, and an explanatory PCM with using five item covariates crossed with threshold parameters and two person covariates. The two polytomous explanatory models with only person covariates estimated two more parameters than their traditional counterparts. In both cases, the explanatory models with person characteristics fit better than the

traditional polytomous IRT models. However, neither of these explanatory models could explain the location of thresholds or the distance between thresholds in the items.

Tuerlinckx and Wang’s explanatory IRT model with both item and person covariates is the most interesting of the three polytomous explanatory models. This model estimates 13 parameters in total – 5 item explanatory variables for the first threshold, 5 item explanatory variables for the second threshold parameter, the two person covariates, and the variance component for the person trait level. Compared to the traditional PCM, the model estimates 36 fewer parameters. Unlike the first two explanatory models, this model uses covariates to explain the location of thresholds. Threshold locations for each item can be approximated by summing the coefficients where relevant item characteristics are present. Using AIC and BIC, this model did not fit as well compared to the other two explanatory models. While both the first and second threshold locations are estimated using item covariates, the parameterization of this model does not make it easy to explain the distance between thresholds. Furthermore, although this model uses explanatory variables to estimate the location of all thresholds, the way the model is parameterized, without reference to prior thresholds, makes the estimated coefficients more difficult to interpret. Like traditional IRT models, it is important to refer to the prior threshold locations when developing polytomous explanatory models.

### 1.2.2. Modeling Polytomous Data

The notation for the EIRM framework is similar to traditional IRT models. Under the PCM, with adjacent item response categories indexed by  $j$  and possible item scores from 0 to  $J$ , the log-odds of selecting response category  $j$  over  $j - 1$  on item  $i$  ( $i = 1, 2, 3, \dots, K$ ) for person  $n$  can be written as:

$$\log\left(\frac{P_n}{P_n(j-1)}\right) = \theta_n - (\delta_i + \tau_{ij}), \tag{1}$$

where  $\theta_n$  represents the latent trait of person  $n$  and it is normally distributed as  $N(\mu_n, \sigma_n^2)$ . Traditionally, the  $\delta_i$  is considered an overall index of item difficulty; however, this is actually the location of the threshold between the first ( $j = 0$ ) and second ( $j = 1$ ) response categories for item  $i$ . The first threshold is often treated as the item difficulty because the first threshold represents the first step to obtain at least a partial credit instead of the lowest possible score on the item. When item response data are dichotomous, there are no estimates for  $\tau_{ij}$ . Therefore, a single threshold parameter,  $\delta_i$ , becomes the item difficulty parameter. When three or more response categories exist,  $\tau_{ij}$  represents the distance between the  $(j - 2)/(j - 1)$  threshold and the  $(j - 1)/j$  threshold for item  $i$ .

In the explanatory form of PCM (EPCM; Tuerlinckx & Wang, 2004), the log-odds of selecting response  $j$  over  $j - 1$  on item  $i$  for person  $n$  can be written as:

$$\log\left(\frac{P_n}{P_n(j-1)}\right) = \mathbf{Z}_{nij} \boldsymbol{\theta}_n - \mathbf{X}'_{ni} \boldsymbol{\delta}_i + \epsilon_{nij}, \tag{2}$$

where  $\mathbf{Z}_{nij}$  is a matrix that can be used to estimate both fixed- and random-effects related to the person traits. When fitting a traditional IRT model,  $\mathbf{Z}_{nij}$  would be a vector of ones. For the earlier example with examinees who are either native speakers of English or speak English as a second language, the  $\mathbf{Z}_{nij}$  matrix could include an additional column of ones (for native speakers) and zeros (for non-native speakers) to estimate a fixed effect for English as a primary language as well as a column of ones to estimate a residual person effect. Similar to  $\mathbf{Z}_{nij}$ ,  $\mathbf{X}_{nij}$  is a matrix of item-related information that describes the characteristics of individual items. With traditional IRT models, a matrix with  $K - 1$  columns indicating the item would be used to

estimate item difficulties for individual items. In the case of the example above, a vector of ones (for items with images) and zeros (for items without images) can be included as an additional column in  $\mathbf{X}_{nij}$  to estimate the impact of the absence or presence of images for person  $n$  on item  $i$ . Finally,  $\delta_{ij}$  in Equation 2 represents the distance between the  $(j-2)/(j-1)$  threshold and the  $(j-1)/j$  threshold for item  $i$ , as in the traditional PCM.

Equation 2 illustrates one of the main issues that often occur when utilizing EPCM. While explanatory variables are used to describe the traits of respondents and the difficulty of initial thresholds, the model contains no parameters to describe the common variation beyond the initial threshold parameter. In Equation 2, the other threshold parameters are an afterthought; and thus the model allows for parameters to explain the common variability between initial thresholds, respondents, and their interactions, but fails to do the same for subsequent thresholds. RSM forces thresholds between  $J$  and  $J-1$  to be equidistant for all items, whereas PCM allows for unique estimates of all thresholds across the items. Extending the EIRM framework to all thresholds can allow distances between thresholds to be explained using available covariates.

Natesan, Limbers, and Varni (2010) extended the polytomous EIRM research by applying an explanatory form of GRM to polytomous response data. The study combined the polytomous model with cumulative logits (Tuerlinckx & Wang, 2004) and a 2-level latent regression model (Van den Noortgate & Paek, 2004). The authors compared the fit of two models, a GRM and an explanatory GRM model that was the combination of the 2-level latent regression model and the polytomous model with cumulative logits. This explanatory model contained a person covariate related to emotional quality of life and no additional item covariates. The two models were estimated using both Bayesian likelihood estimation in WINBUGS (Lunn, Thomas, & Spiegelhalter, 2000) and a standard IRT model fitting approach in MULTILOG 7 (Thissen, Chen, & Bock, 2003) on data from a five-item emotional functioning scale. The authors argued that the model with the explanatory person covariate was better because a clearer picture of emotional functioning was obtained when a measure of emotional quality of life was included in the model (Natesan et al., 2010). Although the authors used a person covariate in their explanatory IRT model, their study did not focus on explaining the distance between item thresholds.

The most common application of explanatory IRT models to polytomous response data is the identification of differential item functioning (DIF) in items with three or more response options (Williams & Beretvas, 2006; Vaughn, 2006). These models include an interaction term between a person covariate (e.g., gender) and the initial threshold parameter for a given item. The distance between additional thresholds is estimated with a single threshold parameter when using the explanatory RSM or multiple threshold parameters when using the explanatory PCM (i.e., EPCM). With very little implementation of these models, one of the goals of this study is to display the flexibility of polytomous explanatory IRT models that can help researchers understand the context of the distance between thresholds.

### **1.2.3. Alternative Parameterizations**

To recap, there have been very few studies that implemented explanatory IRT models with polytomous response data. There are two major reasons for the scarcity of such studies. First, explanatory IRT models for polytomous response data often require a large number of parameters to be estimated, which may be computationally intensive, especially if the data include many items and respondents. Also, as the number of parameters that the model yields increases, the interpretation of model results becomes more difficult (Bulut et al., 2015). Second, the number of software programs for estimating explanatory models with polytomous data has been limited due to the parameterization of these models. Previous research utilized different software programs for the estimation of explanatory IRT models with polytomous response data, such as WINBUGS (Jiao & Zhang, 2014;

Natesan et al., 2010), HLM (Williams & Beretvas, 2006), and the PROC NL MIXED procedure in SAS (Tuerlinckx & Wang, 2004). However, some of these programs (e.g., HLM and SAS) are only commercially available and the others (e.g., WINBUGS) require a strong understanding of the Bayesian modeling.

To avoid the problems described above, some researchers restructured polytomous response data into dichotomous response data and utilized free software programs that are capable of estimating GLMMs with dichotomous data (e.g., Bulut et al., 2015; De Boeck & Partchev, 2012; Plieninger & Meiser, 2014; Prowker & Camilli, 2007; Scheiblechner, 2009; Verhelst & Verstralen, 2008). However, changing the original structure of data often results in information loss and thus adds additional bias to the inferences made from the estimated models. Alternatively, some researchers maintained the original structure of polytomous data but only focused on explaining the initial threshold – or item location – in the estimation process and ignored subsequent thresholds (e.g., Tuerlinckx & Wang, 2004). However, if explanatory IRT models only examine the initial threshold in the items, potential relationships that may exist across all thresholds could be missed when interpreting the model results.

To solve some of these technical problems, a different parameterization of polytomous IRT models is necessary. Consider the number of thresholds estimated using the traditional RSM and PCM models for a dataset that has  $J$  response categories for each item ( $i = 1, 2, 3, \dots, K$ ). Under the traditional RSM, a total of  $J-1$  thresholds need to be estimated beyond the initial thresholds for each item. The estimation of only  $J-1$  parameters is quite restrictive because it assumes that the distance between two thresholds is the same across all items. When researchers assume and thus fit such a model, they have strong a theory about responses. If the researchers fit a polytomous explanatory IRT model with predictors related to their theory, then the model can produce conclusions with greater fidelity.

Under the traditional PCM, a total of  $(J-1) \times K$  threshold parameters will be estimated beyond the initial threshold parameters for each item. The estimation of  $(J-1) \times K$  parameters assumes that the distance between any two thresholds on any item is unique. When researchers fit the PCM, there is no *a priori* theory regarding their models, although the distance between thresholds might be related to item characteristics, respondent characteristics, or an interaction between the two. If a researcher chooses an explanatory IRT model with covariates that explains the distances between thresholds over traditional approaches such as the RSM or PCM, then the researcher is potentially choosing a model that restricts the number of threshold estimates and ties those estimates to an underlying theory. The following section elaborates on these potential models, using a new parameterization.

#### 1.2.4. Strictly Threshold Explanatory Models

The Strictly Threshold Explanatory Model (STEM) is a compromise of the RSM and the PCM that utilizes EIRM. Rather than estimating a single distance between thresholds as in the RSM, or the unique distances between item-by-step threshold combinations as in the PCM, the STEM constrains the estimation of the distances between threshold locations based on common item and/or person characteristics. In the STEM model, the initial item thresholds (i.e., item difficulties) are estimated without the use of explanatory variables; however, subsequent distances are estimated using explanatory variables. Consider the earlier example where including graphics on mathematics items and respondents’ primary language are likely to affect responses to items and particularly the thresholds. For this example, items are scored in one of three categories *incorrect*, *partially correct*, and *correct*. The STEM can be used to estimate this model as follows:

$$\log\left(\frac{P_n}{P_n^{(j-1)}}\right) = \beta_{0i} + \beta_{1i}(\text{Primary English})_n + \beta_{2i}(\text{Other Language})_n + \beta_{3i}(\text{Images})_i. \tag{3}$$

In this model,  $\theta_n$  still represents the trait level of person  $n$  (i.e., mathematical ability),  $\tau_{ij}$  represents the initial threshold location (i.e., the *incorrect/partially correct* threshold) for item  $i$ . The distance between the *incorrect/partially correct* and the *partially correct/correct* threshold is estimated using three parameters,  $\delta_1$ ,  $\delta_2$ , and,  $\delta_3$ , which represent the distance between thresholds for primary English speakers controlling for images on items, secondary English speakers controlling for images on items, and the presence of images on items controlling for English language status.

While the STEM is a compromise between the RSM and PCM, the model constraints are beneficial for this particular model. Like the PCM, the STEM does not have fixed step threshold parameters that are equal across all items, rather several step parameters based on features embedded within the items. Like the RSM, the STEM is easier to interpret than the PCM due to fewer numbers of estimated parameters. This interpretation is aided by the use of EIRM and distances between item thresholds are now due to an interaction between the threshold and the item features.

### 1.2.5. Explanatory Partial Credit Model

The STEM utilizes EIRM, but only for restricting, explaining, and measuring thresholds beyond the first threshold and does not use explanatory variables for the location of the initial threshold locations (i.e., the *incorrect/partially correct* thresholds). If the STEM seems appropriate, then using explanatory variables for locating the initial threshold and subsequent thresholds for items is a logical extension. To illustrate the Explanatory Partial Credit Model (EPCM) model, consider the same example for the STEM. For the EPCM, both initial thresholds and distances between thresholds are estimated using explanatory variables. Thus, the EPCM can be written as:

$$\log\left(\frac{P_n}{P_{ni(j-1)}}\right) = \theta_n - (\delta_1(\text{Primary English})_n + \delta_2(\text{Other Language})_n + \delta_3(\text{Images})_i + \delta_1(\text{Primary English})_n + \delta_2(\text{Other Language})_n + \delta_3(\text{Images})_i) \quad (4)$$

In this model,  $\theta_n$ ,  $\delta_1$ ,  $\delta_2$ , and,  $\delta_3$  have the same meaning as the STEM, and the estimates of  $\delta_1$ ,  $\delta_2$ , and  $\delta_3$  represent the effects of primary English speakers, secondary English speakers, and images in items on the initial threshold locations, respectively. Components present on a particular item and respondent in Equation 4 are additive. Thus, the initial threshold for an item with an image from an examinee whose primary language is English would be located at  $\delta_1 + \delta_3$ . For the same item, the distance between the *incorrect/partially correct* and *partially correct/correct* thresholds is equal to  $\delta_1 + \delta_3$ . For this example, a total of seven parameters will be estimated regardless of the number of items on the assessment: three parameters for the initial threshold, three parameters for the distances between the thresholds, and a variance component for respondent trait level. A more generalized formula for the EPCM can be written as:

$$\log\left(\frac{P_n}{P_{n(j-1)}}\right) = \mathbf{Z}_n \boldsymbol{\theta}_n - \mathbf{X}'_n \boldsymbol{\delta}_i + \mathbf{W}'_n \boldsymbol{\tau}_{ii}, \quad (5)$$

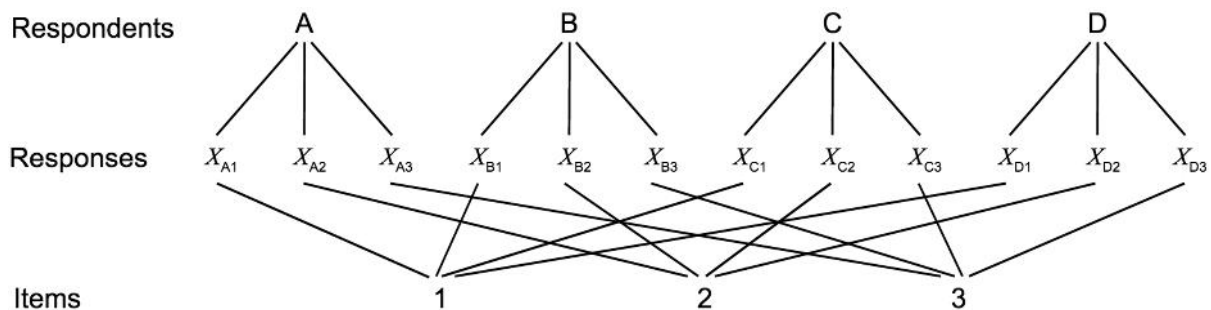
where  $\mathbf{Z}_n \boldsymbol{\theta}_n - \mathbf{X}'_n \boldsymbol{\delta}_i$  has the same meaning as the adjacent categories logit model introduced in Equation 2, and  $\mathbf{W}_n$  is a matrix of indicator variables used to estimate both fixed- and random-effects related to the distances between thresholds,  $\boldsymbol{\tau}_{ii}$ .

There is an important caveat when fitting explanatory IRT models to polytomous data. In the GLMM framework, the number of item-related parameters that can be estimated for each person group is limited to the total number of item-by-step threshold combinations. For example, if a test has 10 items with four response options, thus three thresholds, then a maximum of 30 parameters (10 items x 3 thresholds) per person group could be estimated for the test (i.e., ten for each step level). Therefore, researchers who utilize explanatory models need to ensure that the explanatory IRT model of interest

with item and person covariates is capable of estimating all the parameters given the constraint on the number of parameters to be estimated.

### 1.2.6. Cross-Classified Explanatory Partial Credit Model

The cross-classified EPCM extends the EPCM by including an additional variance component for item difficulty. The result is a model that contains two random-effects, a random-effect for person trait level and a random-effect for item thresholds. Figure 1 displays a network graph describing the cross-classified nature of random items and random persons within the GLMM framework (Beretvas, 2008). By including the random item effect in the cross-classified EPCM, the model acknowledges that additional unaccounted item-related variability exists in the data. The random effect is described as a residual item difficulty because the model already includes item-level predictors. The difficulty of each item can be found by extracting item difficulties from a posterior distribution and combining the values with the relevant item-level predictors. Since these models include fixed-effect predictors for items and an additional random effect that accounts for residual item variability, items can be considered partly random or mixed effects (Van Den Noortgate, De Boeck, & Meulders, 2003).



**Figure 1.** A network graph depicting the cross-classified nature of items and examinees.

As explained earlier, polytomous explanatory IRT models can recover item thresholds either as a fixed effect or as a random effect (Wang, Wilson, & Shih, 2006; Wang & Wu, 2011). From a theoretical standpoint, fitting an IRT model with explanatory item covariates assumes that the researcher has a conceptual understanding of the response process. It is unlikely that the researcher would be able to identify and include all the item-related covariates that can affect the difficulty of an item. The inclusion of the random-effect for item thresholds represents an effect for all of the unexplained components that are not included as fixed explanatory item threshold predictors. Including a random effect suggests that not all features that affect item difficulty are included in the model, but their net effect is a normal distribution of item difficulties with some known mean and variance.

Random item models have been extended to EIRM in several different contexts including, but not limited to, explaining a construct (De Boeck, 2008; Janssen, 2010; Janssen, Schepers, & Peres, 2004), understanding the components of item sets created using automatic item generation (Holling, Bertling, & Zeuch, 2009), predicting item difficulty (Hartig, Frey, Nold, & Klieme, 2012), understanding the impact of cognitive supports on alternative assessments (Ferster, 2013), investigating differential facet functioning (Cawthon, Kaye, Lockhart, & Beretvas, 2012), and modeling item position effects (Albano, 2013). Extending the EPCM in Equation 4, which parameterizes the model for the example considering the role of images in item difficulty for primary and secondary English-speaking students on a mathematics test, the cross-classified EPCM can be written as:

$$\log\left(\frac{P_n}{P_n(j-1)}\right) = \eta - (\beta_1(\text{Primary English})_n + \beta_2(\text{Other Language})_n + \beta_3(\text{Images})_i + \beta_1(\text{Primary English})_n + \beta_2(\text{Other Language})_n + \beta_3(\text{Images})_i + \epsilon_i) \quad (6)$$

Compared to Equation 4, the only difference in Equation 6 is the additional parameter of  $\epsilon_i$ , where  $\epsilon_i \sim N(0, \sigma_i)$ , representing the random effect for residual item difficulty. If the estimate of  $\epsilon_i$  is zero, then the model in Equation 6 is equivalent to the EPCM in Equation 4. By including the random item effect in the cross-classified EPCM, the model acknowledges that additional unaccounted item-related variability exists within the data. Since these models include fixed-effect predictors for items and an additional random effect that accounts for residual item variability, items can be considered partly random (Van Den Noortgate et al., 2003).

The explanatory IRT models outlined in this section can be estimated in several ways, most typically marginal maximum likelihood estimation in conjunction with the EM algorithm (Bock & Aitkin, 1981) or restricted maximum likelihood in conjunction with the Laplace estimation. These models can also be estimated using Bayesian methods such as the Markov Chain Monte Carlo estimation method (Gelman, Carlin, Stern, & Rubin, 2013). The aforementioned methods are available through a wide variety of statistical software programs. In this study, we use the *firm* package (Bulut, 2019) in R (R Core Team, 2018) for estimating traditional and explanatory IRT models for polytomous response data. The *firm* package is essentially a wrapper for the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), which is capable of estimating various GLMMs using a restricted maximum likelihood method. In the following sections, we demonstrate how to estimate each of the explanatory IRT models discussed earlier as well as their traditional counterparts using a real dataset with polytomous item responses.

## 2. METHOD

### 2.1. Data

For this study, we used a verbal aggression dataset (Vansteelandt, 2000) to demonstrate the estimation of explanatory IRT models with polytomous response data. The verbal aggression dataset consists of the responses to a verbal aggression measure based on potentially frustrating situations. A total of 316 first-year psychology students from a Belgian university (243 women and 73 men) responded to the items about four situations: a bus failing to stop, missing a train, the grocery store closing immediately prior to entering, and an operator disconnecting a call because the respondent can no longer pay. For each situation, the students were asked to decide whether they would curse, shout, or scold; whether they would either do the chosen behavior or just want to do it; and whether they would blame themselves or others for the situation. The situations did not follow a factorial design, but each situation type prompt occurred 6 times resulting in 24 items in total. The response options were *no*, *perhaps*, or *yes* for each item. Table 1 shows a descriptive summary of the items and item covariates in the verbal aggression dataset.

We selected this particular dataset because (1) it is a well-known dataset since it has been used as an illustrative example in previous demonstrations of explanatory IRT models (De Boeck, 2008, 2011; De Boeck & Wilson, 2004, Tuerlinckx & Wang, 2004); (2) it is publicly available through many packages in R (e.g., *lme4* and *difR*) – which would allow readers to replicate the analyses presented in this study (see the Appendix for the R codes); and (3) the small sample size of the verbal aggression dataset justifies the need for estimating a parsimonious model for exploratory purposes rather than a traditional IRT model with many item parameters. Previous research suggested that when polytomous items have three response categories, sample sizes of 300 (or possibly more, if the number of response categories is larger) might be necessary to obtain robust estimates of item threshold parameters (Linacre, 2002; Reise & Yu, 1990). The verbal aggression dataset narrowly exceeds the suggested



sample size for polytomous IRT modeling. Therefore, we highlight the trends in the results from traditional IRT models (e.g., RSM and PCM) but intentionally avoid any further interpretation.

**Table 1.** Explanatory Variables and Response Frequencies in the Verbal Aggression Dataset

Item	Situation	Explanatory Variables			Response Options		
		Behavior	Mode	Blame	No	Perhaps	Yes
1	1	Curse	Want	Other	91	95	130
2	1	Scold	Want	Other	126	86	104
3	1	Shout	Want	Other	154	99	63
4	2	Curse	Want	Other	67	112	137
5	2	Scold	Want	Other	118	93	105
6	2	Shout	Want	Other	158	84	74
7	3	Curse	Want	Self	128	120	68
8	3	Scold	Want	Self	198	90	28
9	3	Shout	Want	Self	240	63	13
10	4	Curse	Want	Self	98	127	91
11	4	Scold	Want	Self	179	88	49
12	4	Shout	Want	Self	217	64	35
13	1	Curse	Do	Other	91	108	117
14	1	Scold	Do	Other	136	97	83
15	1	Shout	Do	Other	208	68	40
16	2	Curse	Do	Other	109	97	110
17	2	Scold	Do	Other	162	92	62
18	2	Shout	Do	Other	238	53	25
19	3	Curse	Do	Self	171	108	37
20	3	Scold	Do	Self	239	61	16
21	3	Shout	Do	Self	287	25	4
22	4	Curse	Do	Self	118	117	81
23	4	Scold	Do	Self	181	91	44
24	4	Shout	Do	Self	259	43	14

### 2.2. Model Overview

The following IRT models were fit the verbal aggression dataset: the RSM, the PCM, the EPCM, and the cross-classified EPCM. All of the models focused on the estimation of the first threshold (i.e., *no/perhaps* step) and the second threshold (i.e., *perhaps/yes* step) for each item. As explained earlier, the RSM and the PCM are traditional IRT models and thus do not include any item-level or person-level covariates. Note that we included the RSM and PCM for illustrative purposes only; we do not intend to make any inferences from the estimated threshold parameters due to having a small sample size in the verbal aggression dataset. The primary focus of this study was the two explanatory IRT models: the EPCM and the cross-classified EPCM. These models aimed to explain the variability between the step thresholds using item covariates.

Equation 7 shows the RSM and the PCM for the verbal aggression dataset.  $\theta_n$  represents the overall verbal aggression level of person  $n$ ,  $\beta_i$  is the initial threshold between the *no* and *perhaps* response categories for item  $i$ , and  $\delta_{ii}$  represents the distance between the *no/perhaps* threshold and the *perhaps/yes* threshold for item  $i$ . The only difference between the RSM and the PCM is that  $\delta_{ii}$  is the same across all items in the RSM. That is, the distance between the *no/perhaps* threshold and the *perhaps/yes* threshold is constant across all of the items:

$$\log\left(\frac{P_n(n)}{P_n(p\ na)}\right) \text{ or } \log\left(\frac{P_n(p\ na)}{P_n(y)}\right) = \theta_n - (\beta_i + \delta_{ii}). \tag{7}$$

Equation 8 demonstrates the EPCM with the item-related covariates. In addition to behavior type (i.e., curse, scold, or shout), blame type (others or self) and blame mode (want or do) were used as explanatory covariates in the model. Because the items follow a within-group membership and not between-group membership, all test characteristics cannot be estimated simultaneously because of over-specification. As a result, only a single parameter is needed to estimate the effect of blaming self over blaming others. Similarly, a single parameter is needed to estimate the effect of wanting versus doing an act of verbal aggression.

$$\log\left(\frac{P_n(n)}{P_n(p\ na)}\right) \text{ or } \log\left(\frac{P_n(p\ na)}{P_n(y)}\right) = \ln\left(\beta_1(\text{Curse})_i + \beta_2(\text{Scold})_i + \beta_3(\text{Shout})_i + \beta_4(\text{Do})_i + \beta_5(\text{Self})_i + \beta_1(\text{Curse})_i + \beta_2(\text{Scold})_i + \beta_3(\text{Shout})_i\right) \quad (8)$$

The parameters  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  in Equation 8 indicate the distances between the *no/perhaps* step thresholds and the *perhaps/yes* step thresholds. Behavior type (i.e.,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ ) also explains the initial step thresholds for the *no/perhaps* thresholds. The parameter  $\beta_4$  is the difficulty associated with going from wanting to complete a behavior to doing a behavior and  $\beta_5$  represents the difficulty associated with going from blaming others to blaming oneself.

Equation 9 demonstrates the cross-classified EPCM with the item-related covariates. This model includes all of the elements in the EPCM in Equation 8. In addition, there is an error term,  $\epsilon_i$ , which presents the random effect for residual item difficulty.

$$\log\left(\frac{P_n(no)}{P_n(p\ na)}\right) \text{ or } \log\left(\frac{P_n(p\ na)}{P_n(y)}\right) = \ln\left(\beta_1(\text{Curse})_i + \beta_2(\text{Scold})_i + \beta_3(\text{Shout})_i + \beta_4(\text{Do})_i + \beta_5(\text{Self})_i + \beta_1(\text{Curse})_i + \beta_2(\text{Scold})_i + \beta_3(\text{Shout})_i + \epsilon_i\right) \quad (9)$$

The models summarized in Equations 7 through 9 were fit to the verbal aggression dataset using the *erm* package (Bulut, 2019). The *erm* package controls the *glmer* function from the *lme4* package (Bates et al., 2015) and prints model results in a simpler output. The *glmer* function is capable of fitting a GLMM to a dependent variable that follows a binominal distribution within a multilevel structure. Therefore, regular response data in a wide format (persons as rows and items as columns) need to be reformatted into a long format (items nested within persons) and contained indicator codes for items and responses. In addition, polytomous item responses must be transformed into a dichotomous form. The *polyreformat* function from the *erm* package can transform polytomous items into multiple dichotomous items, without distorting the original response structure. In this study, the response categories of no, perhaps, and yes were dichotomized by creating new labels for each response category. Table 2 shows the reformatted response categories for the verbal aggression dataset.

**Table 2.** Reformating Polytomous Responses into Multiple Dichotomous Responses

Original Response	Category “perhaps”	Category “yes”
No	0	NA
Perhaps	1	0
Yes	0	1

Because the five IRT models in this study were not nested within each other, a direct comparison between the models using a chi-square test was not possible. Instead, we compared the models using the relative fit indices of the Akaike Information Criterion (AIC; Akaike, 1974)

and Bayesian Information Criterion (BIC; Schwarz, 1978). The AIC and BIC indices can be calculated using deviance statistics from each model, where deviance is

$$\text{Deviance} = -2(\log\text{likelihood}) \quad (10)$$

and AIC and BIC fit indices can be computed as

$$\text{AIC} = \text{Deviance} + (2 \times k), \text{ and} \quad (11)$$

$$\text{BIC} = \text{Deviance} + (df \times \log(n)) \quad (12)$$

where  $k$  is the number of estimated parameters and  $n$  is sample size. AIC and BIC were chosen for several reasons. First, while AIC and BIC answer two different questions, when the criteria agree on the best model, this provides reassurance on the robustness on the model choice (Kuha, 2004). Second, regardless of the criteria of use for both AIC and BIC, readers have a preferred relative fit index.

### 3. RESULTS

Table 3 displays the estimated locations of item difficulties and step distances for the RSM and PCM. The RSM is a more restrictive model than the PCM. Item difficulty was estimated for each item individually, while the step distance for *perhaps* to *yes* was fixed (0.54 logits) across the items. The most difficult item based on the location of the *no/perhaps* threshold (2.69 logits) was the item S3DoShout, which is about whether the respondent would do a shouting behavior when the grocery store closes just as he or she is about to enter. Also, the RSM indicates that selecting the *yes* option is  $\exp(0.54) = 1.72$  times more difficult than selecting *no* and *perhaps* options in the items, after controlling for the latent trait (i.e., verbal aggression) level.

Unlike the RSM, the PCM allows each item to have a unique item difficulty (i.e., the threshold for *no* and *perhaps* options) and a unique step parameter for the distance between the *perhaps* and *yes* option. Based on the estimated item difficulties from the PCM, the item S3DoShout was still the most difficult item (2.71 logits for the *no/perhaps* threshold) among the 24 items – which is not surprising given the very low frequency of response option “yes” for this particular item (see Table 1). Unlike the fixed step parameters in the RSM, the PCM had unique step parameters for the distance from *perhaps* to *yes*, ranging from -0.10 to 1.13 across the 24 items. This result suggests that the items in the verbal aggression dataset did not have similar distances from *perhaps* to *yes*, and thus unconstrained step parameters from the PCM can possibly explain more variation among the items.

The PCM results in Table 3 show that four items related to the shouting behavior (S1DoShout, S2DoShout, S4WantShout, and S4DoShout) have a negative distance parameter for *perhaps/yes*, indicating that the thresholds are not ordered in the same order as the response categories (no, perhaps, and yes). That is, selecting the option “yes” over “perhaps” in these four items was easier for the respondents. This psychometric phenomenon is often called *disordered thresholds* or *reversed deltas* in the literature (Adams, Wu, & Wilson, 2012). In the current study, disordered thresholds may be an indicator of some response processes where respondents prefer to manifest their verbal aggression more explicitly by selecting “yes” rather than “perhaps”. Because the items with disordered thresholds seem to be related to different behavior types (i.e., shouting vs. others), using this characteristic within the EIRM framework can help elucidate the disordered threshold problem.

**Table 3.** Locations of No/Perhaps and Perhaps/Yes Thresholds for the Rating Scale Model (RSM) and Partial Credit Model (PCM)

Items	RSM				PCM			
	Location of no/perhaps		Distance to perhaps/yes		Location of no/perhaps		Distance to perhaps/yes	
	<i>Est.</i>	<i>SE</i>	<i>Est.</i>	<i>SE</i>	<i>Est.</i>	<i>SE</i>	<i>Est.</i>	<i>SE</i>
S1WantCurse	-0.53	0.12	0.54	0.05	-0.38	0.16	0.26	0.21
S1WantScold	-0.13	0.12	0.54	0.05	0.12	0.16	0.00	0.21
S1WantShout	0.35	0.12	0.54	0.05	0.35	0.15	0.56	0.22
S2WantCurse	-0.73	0.12	0.54	0.05	-0.95	0.17	0.89	0.21
S2WantScold	-0.17	0.12	0.54	0.05	0.01	0.16	0.17	0.21
S2WantShout	0.30	0.12	0.54	0.05	0.51	0.15	0.02	0.22
S3WantCurse	0.12	0.12	0.54	0.05	-0.09	0.14	1.02	0.21
S3WantScold	0.96	0.13	0.54	0.05	0.82	0.14	1.02	0.27
S3WantShout	1.54	0.15	0.54	0.05	1.47	0.16	0.90	0.36
S4wantCurse	-0.22	0.12	0.54	0.05	-0.52	0.15	1.13	0.21
S4WantScold	0.65	0.12	0.54	0.05	0.66	0.15	0.49	0.23
S4WantShout	1.12	0.14	0.54	0.05	1.29	0.16	-0.10	0.27
S1DoCurse	-0.42	0.12	0.54	0.05	-0.51	0.16	0.69	0.21
S1DoScold	0.10	0.12	0.54	0.05	0.13	0.15	0.46	0.21
S1DoShout	1.01	0.13	0.54	0.05	1.17	0.16	-0.02	0.26
S2DoCurse	-0.27	0.12	0.54	0.05	-0.15	0.16	0.32	0.21
S2DoScold	0.43	0.12	0.54	0.05	0.45	0.15	0.47	0.22
S2DoShout	1.47	0.14	0.54	0.05	1.63	0.17	-0.09	0.30
S3DoCurse	0.65	0.12	0.54	0.05	0.44	0.14	1.20	0.24
S3DoScold	1.53	0.15	0.54	0.05	1.49	0.16	0.70	0.34
S3DoShout	2.69	0.21	0.54	0.05	2.71	0.22	0.21	0.63
S4DoCurse	0.00	0.12	0.54	0.05	-0.17	0.15	0.91	0.21
S4DoScold	0.69	0.12	0.54	0.05	0.64	0.15	0.66	0.24
S4DoShout	1.88	0.16	0.54	0.05	1.98	0.18	-0.02	0.37

Table 4 displays the estimated item parameters for the EPCM and the cross-classified EPCM. Each model decomposed the *no/perhaps* step thresholds based on behavior type (cursing, scolding, shouting), behavior mode (doing or wanting), and blame type (self or others). The main difference between the EPCM and the cross-classified EPCM was the additional random item residuals in the cross-classified EPCM. The top part of Table 4 indicates the location of the *no/perhaps* thresholds for the items based on behavior type, behavior mode, and blame type. For instance, the items associated with the cursing behavior type ( $\mu_{Curse} = -0.916$ ) were easier to endorse than the scolding ( $\mu_{Scold} = -0.073$ ) and shouting ( $\mu_{Shout} = 0.728$ ) behavior types for both EPCM and cross-classified EPCM. Also, for the items associated with blaming self over blaming others ( $\mu_{Self}$ ), endorsing the response category of *perhaps* over *no* was  $\exp(0.786) = 2.19$  times more difficult in the EPCM and  $\exp(0.82) = 2.27$  times more difficult in the cross-classified EPCM. Endorsing *perhaps* over *no* for the mode of doing over the mode of wanting ( $\mu_{Do}$ ) was estimated to be  $\exp(0.465) = 1.59$  times more difficult in the EPCM and  $\exp(0.51) = 1.67$  times more difficult in the cross-classified EPCM.

The bottom part of Table 4 shows the estimated step parameters for the distance from the first threshold (*no/perhaps*) to the second threshold (*perhaps/yes*), depending on the behavior type. The estimated step parameter for the cursing behavior indicated the largest value for both EPCM ( $\mu_{Curse} = 0.781$ ) and cross-classified EPCM ( $\mu_{Curse} = 0.8$ ). This finding suggests that selecting the response of

yes over perhaps and no was more difficult for the items related to cursing than the items related to scolding and shouting. The opposite of this statement is true for the shouting-related items. That is, selecting yes over perhaps and no was easier for the items related to shouting than those related to either cursing or scolding. When the top and bottom parts of Table 4 are compared, the same trend seems to be reversed. The distance from no/perhaps to perhaps/yes was the smallest for the shouting behavior ( $s_c = 0.007$ ) whereas the same distance for the cursing behavior was the largest ( $c_c = 0.781$ ). This finding suggests that endorsing yes over perhaps and no in the cursing items required high levels of verbal aggression, whereas endorsing yes over perhaps and no in the shouting items was much easier for the respondents.

**Table 4.** Summary of the EPCM and Cross-Classified EPCM

	EPCM			Cross-Classified EPCM		
	<i>b</i>	SE	exp( <i>b</i> )	<i>b</i>	SE	exp( <i>b</i> )
<i>Behavior – Curse, Scold or Shout</i>						
Curse	-0.916	0.082	0.400	-0.961	0.134	0.383
Scold	-0.073	0.079	0.930	-0.117	0.132	0.890
Shout	0.728	0.080	2.071	0.714	0.132	2.042
<i>Blame – Self or Others</i>						
Self	0.786	0.047	2.194	0.820	0.105	2.270
<i>Mode – Do or Want</i>						
Do	0.465	0.046	1.592	0.510	0.105	1.665
<i>No/Perhaps to Perhaps/Yes – Step x Behavior</i>						
Step x Curse	0.781	0.076	2.184	0.800	0.077	2.226
Step x Scold	0.395	0.110	1.484	0.440	0.111	1.553
Step x Shout	0.007	0.124	1.007	0.158	0.126	1.171

Table 5 displays the model fit results for the four IRT models. Comparing the RSM to the PCM, AIC favors the PCM, while BIC favors the more parsimonious RSM. Given the disagreement between AIC and BIC, there is not a robust agreement between the relative model fit statistics and thus we cannot make a decision regarding whether the distance between the no/perhaps and perhaps/yes should be equidistant across the items. The EPCM used three covariates to explain item difficulties and one covariate to explain step parameters. For the explanatory IRT models, both AIC and BIC favored the cross-classified EPCM, which is not a surprising outcome because the cross-classified EPCM includes more parameters. The model estimates fixed effects for the behavior type, behavior mode, and blaming as well as random effects for the individual items that represent the thresholds of no to perhaps.

**Table 5.** Summary of the Model-Fit Results from the Four IRT Models

Model	<i>df</i>	AIC	BIC
Rating Scale Model	26	11470	11656
Partial Credit Model	49	11450	11801
Explanatory Partial Credit Model	9	11521	11586
Cross-classified Explanatory Partial Credit Model	10	11469	11548

#### 4. DISCUSSION

Traditional polytomous IRT models provide information about the threshold locations of items and estimate the latent trait levels of respondents. Although traditional IRT models are capable of describing respondents and items, they often fail to explain why thresholds for certain items function in a different way. Polytomous explanatory IRT models presented in this study are an alternative that can provide a meaningful context to the response processes.

Previous studies that utilized explanatory IRT models with polytomous data estimated the location of all thresholds without reference to prior thresholds, making the coefficients difficult to interpret. This study took an alternative approach to parameterizing thresholds so that the distance between thresholds would have a simplified interpretation. By improving the interpretation of these parameters, it potentially allows for improved practices in developing measurement instruments – such as surveys, scales, and questionnaires. In addition to re-parameterizing the explanatory IRT models for polytomous data, this study also displayed the versatility of explanatory IRT models by comparing several models.

A total of four traditional and explanatory IRT models were fit to the verbal aggression data: RSM, PCM, EPCM, and cross-classified EPCM. AIC and BIC model fit indices were used to compare the models. Both model-fit indices favored explanatory IRT models over the more traditional RSM and PCM. The AIC statistic favored the cross-classified EPCM, which included both the first threshold parameters as random effects and additional covariates to explain the distance between the *no/perhaps* and *perhaps/yes* step threshold locations. When comparing the relative fit of the cross-classified EPCM to the other models, the findings suggested that cross-classified models could be useful, as the model showed better relative fit compared to its more restrictive EPCM counterpart. Despite the inconclusive relative fit of the cross-classified EPCM and EPCM, these models show great utility in explaining why an item may be difficult by using information from the collective assessment. For instance, the explanatory item response models indicated that respondents are less likely to select a *yes* response for an item associated with the shouting behavior than other behavior types, which is the type of information that would not otherwise be collected from either the RSM or the PCM.

While EIRM has not been typically used for explaining the distance between step thresholds in polytomous items, this study revealed a situation where the estimated step thresholds between response categories did not vary enough for some items. For instance, after fitting explanatory models to explain the difference between the location of the *no/perhaps* response threshold and the *perhaps/yes* threshold, the results showed that a group of items, specifically the items sharing the same shout behavior type did not show a statistically significant difference between the *no/perhaps* and the *perhaps/yes* step thresholds. This could be interpreted as the *perhaps* option not being as useful to understanding respondents underlying verbal aggression for the items where verbal aggression is expressed through shouting. This means respondents were likely to skip *perhaps* and go from selecting *no* to *yes* when reaching a certain level of aggression. This finding implies that when the more traditional PCM is fit to the data, there are instances where *perhaps* thresholds do not function properly. By using explanatory response models, the functionality of multiple response categories in polytomous items can be determined and the cases where of a response option not functioning could be explained by using an item-related covariate.

While this study only examined polytomous data structures with explanatory item response models, future studies can compare the conclusions drawn from polytomous explanatory IRT models against the findings from the models where polytomous item responses are dichotomized and dichotomous IRT models are applied. Additionally, the models in this study can be modified for different purposes, such detecting item parameter drift and construct shift in polytomously-scored items and improving test equating/linking results in both dichotomous and polytomous data settings.

## ORCID

Luke Stanke  <https://orcid.org/0000-0001-5853-1267>

Okan Bulut  <https://orcid.org/0000-0002-4340-6954>

## 5. REFERENCES

- Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*(4), 408–426. doi:10.1111/jedm.12026
- Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. doi:10.1177/0013164411432166
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4) 581–594. doi:10.1177/014662167800200413
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. doi:10.1109/TAC.1974.1100705
- Bates, D., Maechler, M., Bokler, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01
- Beretvas, S. N. (2008). Cross-classified random effects models. In A. A. O’Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161-197). Charlotte, SC: Information Age Publishing.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. doi:10.1007/BF02291411
- Bock, R. D., & Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. doi:10.1007/BF02293801
- Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89 - 118. <http://dx.doi.org/10.1080/08957340801926086>
- Bulut, O. (2019). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* [Computer software]. Available from <https://github.com/okanbulut/eirm>.
- Bulut, O., Palma, J., Rodriguez, M. C., & Stanke, L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *Sage Open, 5*(2), 1-18. doi:10.1177/2158244015586238
- Cawthon, S., Kaye, A., Lockhart, L., & Beretvas, S. N. (2012). Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. *Journal of School Psychology, 50*, 293–316. doi:10.1016/j.jsp.2012.01.002
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133–148. doi:10.1111/j.1745-3984.2005.00007
- De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*(3-4), 243–276. <http://dx.doi.org/10.1080/15305058.2002.9669495>

- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533–559. doi:10.1007/s11336-008-9092-x
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48(1), 1–28.
- De Boeck, P., & Wilson, M. (2004). Explanatory item response models: a generalized linear and nonlinear approach. *Statistics for Social Science and Public Policy*. New York, NY: Springer.
- Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton, FL: CRC Press.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197. <http://dx.doi.org/10.1037/0033-2909.93.1.179>
- Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds, *Cognitive Assessment* (pp. 107–135). Springer USA.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380–396. <http://dx.doi.org/10.1037/1082-989X.3.3.380>
- Embretson, S. E. (2006). *Cognitive models for the psychometric properties of GRE quantitative items*. Final Report. Princeton, NJ: Educational Testing Service.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E., & Yang, X. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 119–145). New York, NY: Cambridge University Press.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359–374.
- French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, 47(3), 299–317. doi:10.1111/j.1745-3984.2010.00115.x
- Ferster, A. E. (2013). *An evaluation of item level cognitive supports via a random-effects extension of the linear logistic test model*. Unpublished doctoral dissertation, University of Georgia.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: CRC Press.
- Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72(4), 665–686. doi:10.1177/0013164411430707
- Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, 35, 71–76. doi:10.1016/j.stueduc.2009.10.004
- Janssen, R. (2010). Modeling the effect of item designs within the Rasch model. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 227–245). Washington, DC, US: American Psychological Association.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York, NY: Springer-Verlag.
- Jiao, H., & Zhang, Y. (2014). Polytomous multilevel testlet models for testlet based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, 68(1), 65–83. doi:10.1111/bmsp.12035
- Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing*, 14(3), 245–264. <http://dx.doi.org/10.1080/15305058.2013.877911>
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions of performance. *Sociological Methods and Research*, 33, 188–229. doi:10.1177/0049124103262065



- Kubinger, K. (2008). On the revival of the Rasch model-based LLTM: from constructing tests using item generating rules to measuring item administration effects. *Psychological Science Quarterly*, *3*(3), 311–327.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *5*(1), 85–106.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. doi:10.1023/A:1008929526011
- Luppescu, S. (2012, April). *DIF detection in HLM item analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272
- Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educational and Psychological Measurement*, *70*(3) 420–439. doi:10.1177/0013164409355696
- Plieninger, H. & Meiser, T. (2014). Validity of multi-process IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. doi:10.1177/0013164413514998
- Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value added item difficulty effects. *Journal of Educational Measurement*, *44*(1), 69–87. doi:10.1111/j.1745-3984.2007.00027.x
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from <http://www.psychometrika.org/journal/online/MN17.pdf>
- Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136
- Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitability for practical test applications. *Psychology Science Quarterly*, *51*, 181–194.
- Thissen, D., Chen, W., & Bock, D. (2003). *MULTILOG 7* [Computer software]. Chicago, IL: Scientific Software International.
- Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75–109). New York: Springer-Verlag.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55.
- Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics and Data Analysis*, *11*(3), 275–295. doi:10.1111/j.2044-8317.1990.tb00925.x
- Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach*. Electronic Theses, Treatises and Dissertations. Paper 4588.

- Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28(4), 369–386. doi:10.3102/10769986028004369
- Van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 167–187). New York, NY: Springer-Verlag.
- van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York: Springer
- Vansteelandt, K. (2000). *Formal models for contextualized personality psychology*. Unpublished doctoral dissertation, K.U. Leuven, Belgium.
- Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the Partial Credit Model. *Psicologica: International Journal of Methodology and Experimental Psychology*, 29(2), 229–254.
- Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, 67(4), 583 - 605. doi:10.1177/0013164406296974
- Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29(4), 296 - 318. doi:10.1177/0146621605276281
- 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
- Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, 48(4), 441-456. doi:10.1111/j.1745-3984.2011.00154.x
- Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, 30, 22–42. doi:10.1177/0146621605279867
- Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In Hartig, J., Klieme, E., Leutner, D. (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 91-120). Göttingen, Germany: Hogrefe & Huber.
- Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, 13(1), 1–22.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, 56(4), 589–600.

---

**Appendix**

R codes for estimating the explanatory IRT models with the verbal aggression dataset

**# Install and load the required packages**

```
install.packages("devtools")
devtools::install_github(repo = "okanbulut/eirm")
library("eirm")
```

**# Reformat the VerbAgg dataset for polytomous EIRM**

```
data("VerbAgg")
VerbAgg2 <- polyreformat(data=VerbAgg, id.var = "id", long.format = FALSE,
  var.name = "item", val.name = "resp")
```

**# Rating Scale Model**

```
mod1 <- eirm(formula = "polyresponse ~ -1 + item + polycategory + (1|id)",
  data = VerbAgg2)
```

```
print(mod1, Easiness = FALSE)
```

**# Partial Credit Model**

```
mod2 <- eirm(formula = "polyresponse ~ -1 + item + item:polycategory +
  (1|id)", data = VerbAgg2)
```

```
print(mod2, Easiness = FALSE)
```

**# Explanatory Partial Credit Model**

```
mod3 <- eirm(formula = "polyresponse ~ -1 + btype + situ + mode +
  polycategory + polycategory:btype + (1|id)",
  data = VerbAgg2)
```

```
print(mod3, Easiness = FALSE)
```

**# Cross-Classified Explanatory Partial Credit Model**

```
mod4 <- eirm(formula = "polyresponse ~ -1 + btype + situ + mode +
  polycategory + polycategory:btype + (1|item) + (1|id)",
  data = VerbAgg2)
```

```
print(mod4, Easiness = FALSE)
```

## Thematic Content Analysis of Studies Using Generalizability Theory

Gül en Ta delen Teker<sup>1\*</sup>, Ne e Güler<sup>2</sup>

<sup>1</sup> Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

<sup>2</sup> zmir Demokrasi University, Faculty of Education, Measurement and Evaluation Department, zmir, Turkey

### ARTICLE HISTORY

Received: 11 March 2019

Revised: 14 May 2019

Accepted: 22 May 2019

### KEYWORDS

Educational sciences,  
Generalizability theory,  
Thematic content analysis

**Abstract:** One of the important theories in education and psychology is Generalizability (G) Theory and various properties distinguish it from the other measurement theories. To better understand methodological trends of G theory, a thematic content analysis was conducted. This study analyzes the studies using generalizability theory in the field of education in Turkey by using the method of thematic content analysis. It reviews 60 studies, including 31 articles and 29 theses published from 2004 to 2017. The selected studies underwent thematic content analysis using parameters including tagged information, aim, G Theory type, number of facets used in the study, Turkish word for “facet,” object of measurement, sample size, design type, mixed-design availability, shared results of G and D studies, computer programs, method of calculating negative variance, availability of fixed facets, and design balance. The data were interpreted on the basis of frequencies; both table and figures are included in the study. According to the results, there is an increase in the number of studies conducted by using G theory by years. Of these, many compare theories; most of them applying univariate G Theory and consider two-faceted measurement situations. While a small subset of studies features mixed design, a large group features crossed design, with individuals as the object of measurement. The computer program most commonly used in analyses is EduG. The majority of studies use balanced design. Recommendations are provided accordingly with the results.

## 1. INTRODUCTION

One of the most important steps taken in any scientific study is the measurement process used to obtain information needed to analyze a particular object or property. However, the data obtained in this process may contain various types of “error.” These errors, which differ in accordance with the measurement conditions, have a different meaning from the one that is traditionally assumed. Errors occur naturally in measurement; it is therefore essential to determine how and under what conditions to carry out “ideal” acts of measurement, given this reality. In education and psychology, this issue is discussed as an aspect of “reliability,” which

**CONTACT:** Gülşen TAŞDELEN TEKER ✉ [gulsentasdelen@gmail.com](mailto:gulsentasdelen@gmail.com) 📧 Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

may be defined as the extent to which the observed scores are consistent (or inconsistent) (Brennan, 2011).

One of the theories concerning reliability in education and psychology is the Generalizability (G) Theory. This theory enables a researcher to determine the source and number of inconsistencies in the observed scores. Another theory, Classical Test Theory (CTT), consists of observed scores (X), real scores (T), and error scores (E) ( $X = T + E$ ). Although only one error term appears in this model for CTT, the term contains all probable errors. In this context, one of the most important advantages of G Theory is that it enables the investigation of different sources of error within the model it is based on. For instance, the relation between G Theory and the process of measurement where there are K number of sources of error can be described as follows:

$$X = \mu_s + E_1 + E_2 + \dots + E_K \tag{1}$$

Here,  $\mu_s$  in Equation 1 is the universe score, interpreted in a similar way to the real score in CTT. The universe score is defined as the expected value of the observed scores obtained through repetitive measurements (Brennan, 2001). One of the properties that makes G Theory important and different is its conceptual framework. The concepts in this framework are the *universe of admissible observations*, *Generalizability (G) study*, the *universe of generalization*, and *decision (D) study*. The present study uses a sample situation to ensure that these G Theory concepts are understood better. For example, consider a measurement process in which the mathematical problem-solving skills of students are measured in different tasks (t) and scored by more than one rater (r). This process contains two *facets*, labelled “tasks” and “raters.” Facets represent similar situations in measurement. Let us assume that tasks (one facet in this measurement process) contain an infinite number of tasks, while raters (another facet in this measurement process) contain an infinite number of raters. Both facets have been selected from an infinite universe of admissible observations. If each rater scores each task carried out by every student in the sample process, the measurement design is called a crossed design, and the process is represented as *sxtxr*. If, however, each task carried out by all students in the process is scored by different raters, the raters are said to be “nested” in the tasks and the study design is known as a “nested design,” represented as *sx(r:t)*. A crossed design is usually preferred in studies conducted using G Theory. The reason for this is that all sources of error, associated with all probable facets and the interactions between those facets, can be estimated in crossed-design studies. This situation gives D studies great flexibility.

A careful analysis of the example above makes clear that students also participate in the process, alongside tasks and raters; they too are considered variance sources of the measurement process. Any individuals, students, objects, or situations constituting the subject matter being measured are called the *object of measurement* in G Theory. While the term *universe* is used to denote the facets of measurement in G Theory, *population* is preferred for the object of measurement (Brennan, 1992). Observable scores, obtained by evaluating a task in the population or universe of admissible observations by a rater, are represented in Equation 2:

$$X_{str} = \mu + s + t + r + st + sr + tr + str \tag{2}$$

In Equation 2,  $\mu$  represents the average within the universe and population, while  $s$ ,  $t$ ,  $r$ ,  $st$ ,  $sr$ ,  $tr$ , and  $str$  represents each of the seven unrelated components. This is a linear model of *sxtxr* (Brennan, 2011; Güler, Uyanık, & Teker, 2012). A model of this design contains seven sources of variance, known as, “G study variance components.” Once these variance components have been estimated, the values can be used in estimates of universe score variance, error variance, various generalizability universe coefficients with similar interpretations, and various D-study designs. Variance components in a G study can be estimated using the expected values of squares average in the variance analysis. As is clear from here, a variance analysis (ANOVA) appears

in the statistical structure of Equation 2. However, the F test is not used in G Theory. This case reflects one of the operational differences that distinguish G Theory from traditional variance analysis (Brennan, 1992; Güler et al., 2012).

The variance components obtained through the G study are used to design various D studies. The above-mentioned example can help to explain this situation. Let us assume that there is a process of measurement in which students' mathematical problem-solving skills are evaluated by three raters (r) using five different tasks (t). Each level of the two facets (tasks and raters) is called a *condition*. In this study, there are five conditions for the facet of tasks and three conditions for the facet of raters. Firstly, the variance components are calculated using the data obtained through the G study. After that, various D studies can be set up to decide on designs containing the same or different numbers of conditions of the facets made available by the G study. For instance, in D studies organized on the basis of a G study with five available tasks, designs can be created in which the same number (5), a smaller number (1, 2, 3 or 4), or a larger number (6, 7 etc.) of tasks is available; such designs can also include the same or a smaller or larger number of raters. One point to take careful note of here is that the variance components obtained through the G study are values estimated using a single task and rater (one condition). Thus, estimates made for various numbers of facet and task conditions also constitute D studies.

The universe score variance for a randomly crossed D study with the same structure as the G study in the example above is as follows:

$$\sigma^2(\tau) = \sigma^2(s) \tag{3}$$

The relative error variance is:

$$\sigma^2(\delta) = \frac{\sigma^2(s)}{n_t} + \frac{\sigma^2(s)}{n_r} + \frac{\sigma^2(s)}{n_t n_r} \tag{4}$$

The absolute error variance is:

$$\sigma^2(\Delta) = \frac{\sigma^2(t)}{n_t} + \frac{\sigma^2(r)}{n_r} + \frac{\sigma^2(t)}{n_t n_r} + \frac{\sigma^2(s)}{n_t} + \frac{\sigma^2(s)}{n_r} + \frac{\sigma^2(s)}{n_t n_r} \tag{5}$$

The generalizability coefficient is:

$$E(\rho^2) = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \tag{6}$$

The dependability coefficient is:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \tag{7}$$

As equations 3, 4, 5, 6 and 7 make it clear, G Theory is based on average score metrics (Brennan, 2011), unlike CTT, which is based on total score metrics. In addition, the relative error variance and generalizability coefficient are interpreted in a similar way to the error variance and dependability coefficient, which are based on a relative comparison of individuals in CTT. Another point worth noting is that the generalizability coefficient and the dependability coefficient are not equal in G Theory. The error variance and dependability coefficient, based on absolute decisions that can be calculated in G Theory, cannot be calculated in CTT. However, both the absolute error variance and the error variance in CTT are derived from the random sampling assumption; when more than one facet is taken from the generalizability universe, the CTT error variance can be estimated at very low levels (for further details, see Brennan, 1997).

The above equations can also be applied to different generalizability universes and D studies. For instance, let us suppose that the raters assessing student mathematical problem-solving skills are constant. In other words, let us suppose that the purpose is not to generalize the raters

in this study into a larger universe (alternatively, assume that all of the raters in the universe are the raters in this study). In this case, the universe score variance is:

$$\sigma^2(\delta) = \sigma^2(s) + \frac{\sigma^2(s_r)}{n_r^2} \tag{8}$$

The relative error variance is:

$$\sigma^2(\delta) = \frac{\sigma^2(s)}{n_t^2} + \frac{\sigma^2(s_r)}{n_t^2 n_r^2} \tag{9}$$

And the absolute error variance is:

$$\sigma^2(\Delta) = \frac{\sigma^2(t)}{n_t^2} + \frac{\sigma^2(t_r)}{n_t^2 n_r^2} + \frac{\sigma^2(s)}{n_t^2} + \frac{\sigma^2(s_r)}{n_t^2 n_r^2} \tag{10}$$

It is clear that, when the rater facet is constant, estimated error variances decrease and universe score variance increases. This means that dependability coefficients will be estimated at high levels. However, the gain obtained by the increase in dependability values restricts the interpretations that can be made in relation to generalizability. In a similar way, as the sample size increases in D studies (that is to say, as the number of conditions increases) and/or when the study has a nested design, error variance decreases; the increase in the number of nested facets in the design restricts the interpretations that can be made in relation to the generalizability of measurements.

G Theory is basically a theory of measurement based on random facets. Therefore, at least one facet should be taken at random in the measurement process. The measurement models in which constant (as well as random) facets are available are known as mixed models in G Theory (Brennan, 1992).

All the above-mentioned examples relate to univariate G Theory. However, some studies are considered in the context of multivariate G Theory. In the first example above, let us suppose that the students are expected to deal with algebraic and analytic problems. Universes of admissible observations correspond to each context and each universe corresponds to one single constant case. In other words, a univariate mixed model can be formulated with a multivariate model, which requires a constant facet; a more flexible representation of the constant facet is thus assured. At a statistical level, multivariate G Theory analyses involve not only variance components, but also co-variance components (Brennan, 2011). In the case of a simpler explanation, the univariate G Theory may be used to analyze the scores obtained from a single test; multivariate G Theory is used to determine the generalizability of scores obtained from a test composed of different sub-tests (Atılgan, 2004; Brennan, 2001; Deliceoglu, 2009).

Three fundamental theories can be used to determine reliability: CTT, Item Response Theory (IRT), and G Theory. Of these, CTT is generally preferred because its underlying mathematical model is easier to understand and its assumptions are flexible (Hambleton & Jones, 1993). Although IRT contains a more complicated mathematical model and its assumptions are difficult to meet, it takes precedence over individual measurement applications because it can generate independent estimates of item and ability parameters. CTT and G Theory focus on test results, while IRT focuses on responses to items (Brennan, 2011).

Various properties distinguish G Theory from the other two theories. Although the mathematical structures of the basic equations in CTT and G Theory are similar, with unobservable values on the right ( $X = T + E$ , and equation, respectively), CTT has only one error term, while G Theory permits the division of error terms to reflect different sources of error. G Theory also has a richer conceptual framework than CTT. Two points are particularly relevant: (1) in G Theory, it is possible to distinguish between constant and random facets; (2) G Theory makes it possible to carry out different types of decision studies (Brennan, 2011).

Cronbach et al., (1972) and Brennan (2001) argue that G Theory removes the difference between reliability and validity. One of the most important differences between G Theory and IRT is that, while G Theory focuses on test scores, IRT focuses on item scores. Although items are a constant facet in IRT, they are almost always considered random in G Theory (Brennan, 2011).

In recent years, Turkey has seen an increase in studies conducted using G Theory. Researchers in education and other fields have shown more interest in G Theory because it differs from CTT and IRT and can be advantageous in a number of situations. Given this context, the present study uses various criteria to analyze G Theory-based research carried out in Turkey. Its main purpose is to determine the general tendency of G Theory-based studies and to provide new resources and information to researchers who may have doubts about using G Theory in their own research. To enhance the quality of future academic work, examination themes are also explained in detail. For researchers hoping to contribute to literature on any topic, general trends in current studies in the relevant field, gaps in the literature, and research characteristics presented are very important.

The literature included a review of studies on the use of G Theory: Rios, Li, and Faulkner-Bond (2012), examined 58 studies published in the field of psychology and education between 1997 and 2012, focusing on sample size, the handling of missing data, the question of balance (or unbalance), multiple group comparisons, analysis trends (e.g., computer programs used, methods of estimating variance components), and reporting results. Other than this study, no published research had explored studies conducted using G Theory, indicating a gap in the literature. The present study sets out to provide detailed information on the theoretical and conceptual bases of G Theory to guide researchers aiming to conduct research on the deficiencies of or mistakes made in published studies. The present study makes an important contribution to the literature by promoting the widespread, correct use of G Theory, which is now widely and increasingly studied, by introducing this theory to researchers in the main branches of science, beyond the educational sciences.

The present study therefore examines research carried out using G Theory in the field of education in Turkey using the thematic content analysis method. The following questions guided the current study:

1. What were the aims of the research studies analyzed?
2. Which types of G Theory are used more often in measurement situations?
3. How many faceted designs are used in the study?
4. How is the term “facet” translated into Turkish?
5. Did the object of measurement specify?
6. What are the sample sizes used in the studies?
7. What types of designs are covered?
8. What types of mixed design exist and how can they be used correctly?
9. What proportion of studies fall into the G and D studies categories?
10. Which computer programs are used most frequently?
11. What is the preferred way of explaining negative variance when analyzing research results?
12. What are the various types of fixed facet and how are they discussed?
13. At what rate do studies use balanced or unbalanced patterns?



## 2. METHOD

### 2.1. Research Model

The present study has carried out a thematic content analysis of theses and articles based on G Theory in the field of education in Turkey in 2004–2017; the various themes covered here were selected to reveal their similarities and differences. A thematic content analysis involves the synthesis and interpretation of different research findings on the same subject (Au, 2007, Çalık & Sözbilir, 2014, Finfgeld, 2003, Walsh & Downe, 2005). Studies that conduct a thematic content analysis provide a very rich resource to researchers working in related fields, who cannot access all the work in the field or systematically examine those studies (Çalık, Ayas, & Ebenezer, 2005; Ültay & Çalık, 2012). Compared to meta-analyses and descriptive content analysis studies, relatively few studies offer thematic content analyses (Çalık & Sözbilir, 2014).

### 2.2. Data Collection

All of the education articles incorporating G Theory published in Turkey between 2004 and 2017 were obtained using the Google Academic search engine and/or which were reached in journals indexed by ULAKBIM (national index) and Social Science Citation Index (SSCI). All of the theses in the Council of Higher Education's National Thesis Centre Database of Turkey were also included in the scope of this study. There are no studies carried out in Turkey used G Theory before 2004. For this reason, the starting point for this study was set as 2004. There were 41 articles from 23 different journals and 29 theses published in six different universities. Ten of the articles analyzed were derived from Master's or Ph.D. theses; they were compared with the original M.A. or Ph.D. theses and found to be no different. The reason for excluding articles derived from Master's or doctoral theses was to avoid duplicating studies. Excluding them made it possible to present a more accurate picture of G Theory studies. The elimination of such studies left a total of 60 studies, 31 articles, 20 Master's theses, and 9 Ph.D. theses for content analysis. The investigated studies are listed in Appendix.

### 2.3. Data Analysis

Before carrying out the content analysis, the researchers developed a checklist to help them analyze studies incorporating G Theory. The purpose of the checklist was to set standard criteria for analyzing the articles. The checklist had two main parts: "study tag" and "theoretical information." Expert opinions were obtained from three measurement and evaluation specialists, who had carried out studies on G Theory and were able to evaluate the checklist. The specialists recommended including key words and author names in the tags used to describe studies under analysis. The checklist was updated to reflect these views; the version shown in Figure 1 was ultimately used by two researchers in this study.

To ensure consistency across different researchers, five randomly selected studies were examined independently by two researchers. Using the data obtained, Equation 11 (suggested by Miles and Huberman (1994)) was used to calculate consistency between researchers, as follows:

$$\text{Reliability} = \frac{N_{\text{or}}}{N_{\text{or}} + N_{\text{od}}} = .86 \quad (11)$$

The interrater consistency obtained using Equation 11 was calculated as .86. This value should be .80 or above (Miles & Huberman, 1994, Patton, 2002). This result compared to the criterion drawn from the literature, sufficient coherence is obtained. Within the scope of this study, 60 studies were reviewed by researchers, in accordance with the themes in Figure 1, to identify any inconsistencies in the data. Articles or theses with inconsistencies were reviewed by the researchers again independently, to see whether there was any disagreement. For just one study researchers had a disagreement. The researchers came together to discuss the issue and tried to

reach agreement, as well as obtaining the opinion of a third independent researcher. As a result of this process, once consent of researchers has been obtained, all the data were combined. Frequency and percentage analyses of codes were carried out for each theme.

	<b>Criteria</b>	<b>Coding</b>
<b>Study tags</b>	Study Number	.....
	Title of the study	.....
	Type of study	(1) article (2) Thesis (M.A) (3) Thesis (PhD)
	Author(s)	.....
	Journal / University	.....
	Year of publication	.....
	Key words	.....
	Aim of the study	.....
	Type of G Theory for measurement	(1) Univariate (2) Multivariate
	Number of facets	(1) 1 facet (2) 2 facets (3) 3 facets (4) 4 facets
<b>Theoretical information</b>	Naming the term "facet"	(1) Yüzey (in Turkish) (2) Değişkenlik/Varyans kaynağı (in Turkish) (3) Facet (4) other
	Stating the object of measurement	(0) No (1) Yes
	Describing the object of measurement as a facet	(0) No (1) Yes
	Object of measurement	(1) Individuals/ students (2) Items / Tasks / Raters (3) Other
	Type of design	(1) Crossed (2) Nested
	Availability of Mixed design	(0) No (1) Yes
	Whether it is used correctly when mixed design is available	(0) No (1) Yes
	Whether the results for G study are presented	(0) No (1) Yes
	Whether the results for D study are presented	(0) No (1)Yes
	Computer programs used	(1) GENOVA/mGENOVA/urGENOVA (2) SPSS (3) EduG (4) G-String (5) Other (6) Not stated
Availability of negative variances	(0) No (1) Yes	
If available, whether negative variances are described	(0) Described (1) Not described	
Availability of constant facets	(0) No (1) Yes	
If available, whether constant facets are described	(0) Not described (1) Described	
Whether the design is balanced	(1) Balanced (2) Unbalanced	

Figure 1. Checklist used in the study

### 3. RESULT / FINDINGS

The research findings are presented in two parts. The first section headings refer to the tags used to categorize the articles and theses; the second section focuses on theoretical information.

#### 3.1. Findings Related to Tagged Information in the Studies Analyzed

##### 3.1.1. Year of publication

Figure 2 shows the distribution of articles and theses by publication year. Although the increase has not been steady, there has clearly been an increase in the number of articles and theses written using G Theory since 2004. While the increase in the number of articles has reached a peak in recent years, the number of Master’s theses reached its highest value in 2015; although G Theory continues to be used regularly, frequency has decreased in the last few years. Among doctoral theses, there was an increase between 2012 and 2014; after that date, there is no doctoral thesis conducted on G Theory.

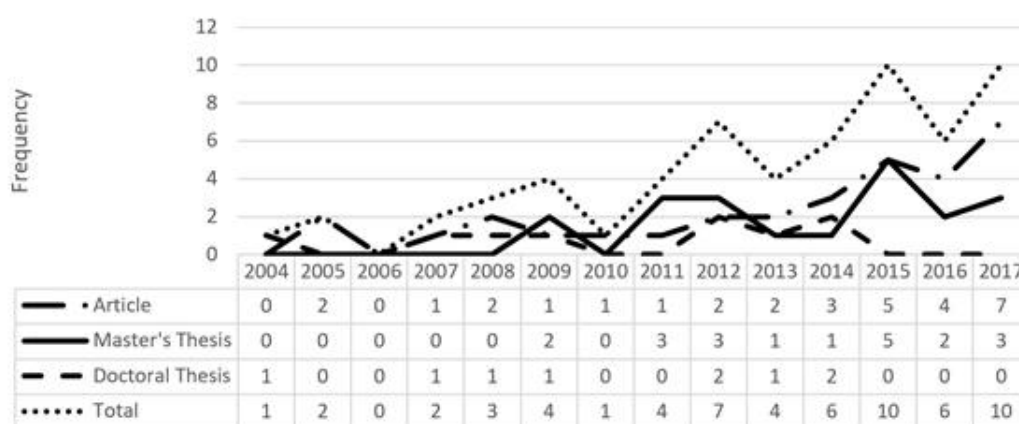


Figure 2. Distribution of articles and theses by year

As is clear from Figure 2, the first Turkish doctoral thesis to use G Theory was completed in 2004. It was followed by one doctoral thesis per year in 2007, 2008, 2009, and 2013, and two doctoral theses in 2012 and in 2014. The first Master’s thesis was written in 2009 (f=2). While no Master’s theses used G Theory in 2010, it was used in 1–5 theses every year after 2011.

##### 3.1.2. Keywords

The keywords in the 31 articles and 29 theses were reviewed to determine their frequencies. 109 different key words were used in the studies. The most frequently used word was *Generalizability Theory* (f=56) – as expected. *Reliability* (f=23), *Classical Test Theory* (f=13), *interrater reliability* (f=8) and *decision study* (f=8) were the most frequently used key words. These were followed by *generalizability coefficient* (f=6), *generalizability (G) study* (f=5) and *Phi coefficient* (f=5), as the basic concepts in G Theory. In addition, 75 key words were used just one time each. “Rating/rater/scoring” concepts were included among the 39 key words; of these concepts, “scoring key/rubric” appears in 12 of them. Computer programs used in G Theory analysis, including EduG, GENOVA, mGENOVA, SAS, and SPSS were used 13 times.

#### 3.2. Findings Involving Theoretical Information

**Aim of the study:** G Theory can be used in a range of different academic fields. According to the topics, theory comparison was the most commonly studied subject throughout the studies. 14 studies compared G Theory to CTT and Many-Facet Rasch Model (MFRM). In addition, six studies compared performance assessment tools (checklists and analytical and holistic rubrics). In nine studies set out to establish the most appropriate number of raters, the quality of raters involved in evaluation, and interrater reliability. They examined the reliability of scores

obtained from measurement tools including the Vee diagram, concept map, multiple-choice tests, structured grids, and performance tasks (f=10). Other studies investigated instructor, peer, and self-evaluation results; standard setting methods; computer software used in analyses (SPSS, GENOVA, EduG, and SAS), and the results obtained from the crossed and nested designs. Various reliability methods were also compared.

**Type of G Theory used for measurement:** Generalizability Theory can be univariate or multivariate, depending on the measurement situation involved. Most of the studies examined (f=54) included univariate G Theory analyses. Because of its complexity, multivariate G Theory was used in just three Ph.D. theses and three articles written by the same authors.

**The number of facets used in the study:** Measurement situations with two facets appeared in 45 studies; nine studies included measurement situations in which one and two facets were considered together. It has been observed that one facet in three studies, three facets in two studies, and four facets in one study.

**Translating the term “Facet”:** Since the term was translated into Turkish in different ways in G Theory studies in Turkey, the naming of this concept was also considered. It was most frequently used as “the source of variability (de i kenlik kayna ı)” in the studies analyzed in this study (f= 18). It was used as “surface (yüzey)” in seven studies, as “the source of variance (varyans kayna ı)” and “variable (de i ken)” in four studies, and “component (bile en)” and “variance component (varyans bile eni)” in one study each. It was called “facet” in one study written in Turkish without finding Turkish concept for it. Only one of the 13 works written in English used "variance source" instead of "facet". In 11 studies, however, it was observed that the term was not considered although the G Theory was used.

**Presenting and describing the object of measurement as a facet:** Only two studies considered the object of measurement as a facet. While 44 studies clearly defined the object of measurement, 14 gave no explanation. When observations related to the state of the measurement object, only 25 studies clearly defined the object of measurement. Of the remaining 35 studies, 23 provided no information and 10 provided the correct usage. Two studies failed to mention the measurement object and used the concept in the wrong way.

**Sample size:** Table 1 shows the sample sizes used in these studies.

Of the studies investigated, 16 had sample sizes below 30, while 25 had sample sizes between 30 and 100. Only 19 studies had samples larger than 100.

**Table 1.** Object of Measurement Sample Size

Object of measurement	Sample size	Frequency	Total
Person	<30	11	55
	30–100	25	
	102–187	10	
	203–249	5	
	309	1	
	689	1	
	1000	1	
	1500	1	
Item	16	1	3
	18	1	
	20	1	
Occasion	7	1	1
Task	6	1	1

**Type of design:** The most frequently used design was a crossed design (f=48). Nested designs were used in seven studies, while five studies used both crossed and nested designs.

**The availability and correct use of mixed designs:** Only six of the 60 studies used mixed designs. One study that claimed to have a mixed design actually had a random design.

**The presentation of G and D study results:** G study results were presented as expected in most studies except one. In addition, D study results were not given in only three studies. It was observed that the D study was not performed because the purpose of these studies was not to estimate the reliability for different measurement situations.

**Computer programs used:** An evaluation of the computer programs used in the analyses, revealed that the most frequently used program was EduG (f=32). The second most frequently used program was SPSS (f=16). GENOVA was used in nine studies, mGENOVA in five studies, and G-String, R, and SAS were used in two studies each. In five studies, the computer program was not specified. Since some studies used more than one program (e.g., SPSS-EduG and SPSS-GENOVA), the sum of the frequency values above exceeds the number of studies examined.

**The availability and description of negative variance:** In 21 studies, a negative variance was observed. Adopting the approach of Cronbach et al., the negative variance was treated as zero in 11 studies; zero was also used to estimate other variance components. In 10 studies, adopting Brennan's approach, the negative variance was regarded as zero and used as it was in estimates of other variance components.

**The availability and description of fixed facets:** Six of the studies analyzed had constant facets. Only two explained these constant designs to readers.

**Design balance:** Only six of the 60 Turkish studies had an unbalanced design. Of these, two were Ph.D. theses and one was an article written by one of the Ph.D. authors.

#### 4. DISCUSSION, CONCLUSION and SUGGESTIONS

The present study analyzed 60 Turkish studies in two stages, using tag information and their theoretical foundations. An examination of the years in which G Theory studies were published revealed an overall increase in publications, despite occasional decreases. The theoretical structure of G Theory is complex and difficult to analyze at elementary levels; for this reason, it is used primarily in doctoral theses. However, beginning with the year 2009, it has also appeared in Master's theses. One of the reason for this may be that G Theory Master's level analyses are now being conducted by means of user-friendly computer programs, such as EduG, rather than the more advanced GENOVA. Another reason may be the increasing number of workshops are held at congresses and courses are taught in Master's and doctoral programs. User friendly computer programs will make it possible to carry out more analyses based on G theory for various studies. Another explanation for this result is the increase in the number of researchers working in the area of measurement and evaluation. In particular, applications for research-assistant posts in the field of measurement and evaluation have been increased since 2002, resulting in a larger number of researchers working in the field and therefore more studies based on G Theory.

The keywords presented in the study tags were also analyzed, revealing that G Theory was used most often in studies involving interrater dependability and standard settings. The fact that 88 words appeared only once or twice appears to show that a range of studies on diverse topics have been carried out using G Theory. Among studies that feature rating and rater keywords (f = 39), G Theory is frequently discussed in relation to rater reliability, consistency, the rater effect, the number of raters, the reliability of ratings, and rating methods. Although CTT is used

more frequently than G Theory in the literature, it cannot be used to determine the number of raters needed to obtain more reliable results or the number of criteria to include in scoring keys. The information available to G Theory through D studies may make it a better choice than CTT, especially in such studies.

An investigation of the aims of the studies in question found that they made a great number of theoretical comparisons. By comparison, the most frequently used theory is CTT, which has been used for many years in the literature and is better known than G Theory in the theoretical literature. Another theory often used in theory comparisons is MFRM. It is thought that this model, which is covered in IRT, tends to be preferred because it allows analyses to be carried out using fewer parameters than other IRT models. Like G theory, MFRM is frequently used to determine the reliability of a rater; it can consider more than one error source at the same time. These can also be cited as reasons for comparing G Theory to MFRM. As well as being used in theory comparisons, G Theory was also preferred when researchers wished to determine the reliability of scores obtained using various measurement tools. G Theory has the advantage of being able to simultaneously handle many sources of error in a measurement process at the same time. While this topic is not new in the literature, many studies have investigated the reliability of self and peer assessments, which have been discussed more frequently in recent years. G Theory makes it possible to evaluate the rater as a facet, while also evaluating the points of self, peers, and instructors as conditions of this facet. Since G Theory makes it possible to estimate the magnitude of the variance between evaluations of different raters and the reliability coefficient, it may be preferred in such studies. A final category of studies compared the results obtained using different types of computer software able to carry out G Theory analyses. Because of free and user-friendly software and their manuals, the use of G theory potentially increases.

Most of the studies analyzed in the course of this research were produced using univariate G Theory. Only five studies used multivariate G Theory, potentially reflecting the following two factors: (1) the situations considered by researchers were better suited to the use of univariate G Theory, and (2) researchers preferred not to use multivariate G Theory because it was relatively difficult and complicated to analyze.

It was found that the most of the studies investigated used two-faceted measurement designs. Frequency measurement situations with two surfaces may reflect standard educational practice (items and raters as facets). Very few of these studies had three- or four-faceted designs. As the number of facets increases in G Theory, the number of estimated variance values for each facet and the interactions between facets increase. Interactions are therefore difficult to interpret. For example, where a two-faceted crossed design consists of seven components of variance, in a three-faceted design, this number increases to 14. Researchers tend to avoid highly faceted designs because it is difficult to interpret the large number of variance components that result from the increased number of facets.

The concept of “source of variability” was used in almost half of the studies analyzed instead of the English term “facet” available in the G Theory. The use of agreed on words can be supported by reaching an agreement on Turkish equivalents to the English terms and compiling them in a glossary, and thus comprehensibility of the G Theory studies can be increased. Indeed, there is such a glossary study conducted in Turkey by the Association of Measurement and Evaluation in Education and Psychology, and accordingly, it is recommended that the words “yüzey (facet)” or “değişkenlik kaynağı (source of variability)” be used as corresponding to English word “facet”. Yet, it was observed that using “source of variability” as Turkish equivalence to the English word “facet” could cause confusions in studies conducted in Turkish. The sentence “A one-facet design has four sources of variability” from an important resource book, “Generalizability Theory: A Primer” by Shavelson and Webb (1991, p.4) would

exemplify our claim because the translation of the sentence into Turkish would also cause confusion. Considering this situation, it is thought by researchers that using “yüzey” rather than “de i kenlik kayna ı” in Turkish as equivalence to English “facet” would be more appropriate.

In G Theory, the objects to be measured such as students, individuals, methods etc. and decisions will be made on it known as the object of measurement. Differences in the object of measurement are defined as “the sources of error” in CTT, since variance that depends on the object of measurement is a desired situation. These differences are not considered to be facets in G Theory either. Two of the studies examined presented this idea inaccurately. First, the difference between the concepts of facet and object of measurement is difficult to understand. Second, the fact that “facet” is used to express the source of variability in the metric target state, a distinction that does not exist in MFRM (a theory that G Theory is often compared to) may add to this confusion. Researchers should therefore be encouraged to clarify which sources of variability discussed in their studies are facets or objects of measurement. In 54 of the 60 studies examined, the object of measurement was the individual. Items, tasks, situations, and raters can all be objects of measurement, depending on the type of measurement. It is very important for researchers to clearly define the object of measurement and to accurately define it as a measured object to ensure an accurate interpretation of the findings.

Sample size is quite important in most statistical methods, as it influences the accuracy of estimates and can increase or reduce errors. Of the studies examined here, 44 had a sample size of 30 or more. This ratio was considerably higher than the value obtained when Rios, Li and Faulkner-Bond (2012) conducted 58 studies using G Theory between 1997 and 2012. They found that the mod of the sample size was 20. Atılğan (2013) examined the effect of sample size on the G and Phi coefficients and found that it was impossible to make stable predictions if the sample size was 30 or below. Results could be considered sufficiently unbiased if the sample size was 50, 100, 200, or 300. With a sample size of 400, the result can be considered definite. In the context of educational sciences studies undertaken using G Theory in Turkey, the majority of results are based on an adequate sample size. However, the samples are small in some of these studies. Due to logistical, economic, and time constraints related to data collection, some studies based on G Theory have been carried out using smaller samples (Rios, Li, & Faulkner-Bond, 2012). At this point, a balance must be established between the need to increase sample sizes to achieve a correct estimate of variance components and the pressures of staff, time, and cost limitations. While a G Theory sample consisting of 20 individuals is considered the lower limit (Webb, Rowley, & Shavelson, 1988), more accurate estimates can be obtained from larger samples. In future studies, it may be advisable to use the largest sample size possible. In particular, researchers struggling with unbalanced, complex designs may end up deleting data and changing to a balanced format. Such situations reduce the size of study samples. In recent years, it has become possible to overcome these difficulties via user-friendly software, capable of carrying out unbalanced pattern analyses.

Many studies in the literature have been conducted using G Theory; the majority of studies examined here ( $f = 47$ ) have adopted a crossed design, possibly because all possible sources of variance can be estimated in fully crossed designs. Only in fully crossed designs can researchers access the variance values of each source of variability, as well as their interactions. In nested designs, it is not possible to estimate the variances of nested facets alone. For this reason, crossed designs may be preferred to nested designs. In the studies investigated here, the nested facet in nested designs was generally raters. Particularly in a performance assessment, it may not be possible for each individual to be evaluated by all of the raters. Such measurement situations should be monitored in relation to variables including cost and the effective use of

the resources. If it is not possible to use a crossed design, the study should be carried out using nested designs.

G Theory is basically a theory of measurement with random facets; for this reason, at least one facet should be random (Güler et al., 2012). If at least one facet in a design is constant and the other facets are random, the design is said to be mixed. The present investigation of studies conducted using G Theory in Turkey found that only six of the 60 studies examined used the term "mixed design". In only one of these six studies, the mixed design was defined as "the combined use of crossed and nested facets". On the other hand, in basic sources of G theory in the literature mixed models are defined as the the measurement models in which constant (as well as random) facets are available are known as mixed models in G Theory (Brennan, 1992; Shavelson & Webb, 1991), and the other five studies used the mixed model in accordance with this definition. Actually, those five studies used multivariate G Theory. While multivariate G Theory is used to generalize scores obtained from a test containing different sub-tests, the sub-tests are regarded as constant facets (Brennan, 2001). For this reason, multivariate G Theory studies are essentially mixed design studies. Although one study analyzed here had a random design, it was presented as mixed, possibly because the crossing and nesting of facets in the same design was wrongly perceived as indicating a mixed design. In describing a design as mixed, it is important not to use crossed and nested designs together; at least one of the facets must be fixed. Taking this into consideration will enable the terminology of G Theory to be used more accurately.

In G Theory, it is possible to investigate dependability in two stages, via a Generalizability (G) study and a Decision (D) study (Brennan, 2001; Goodwin, 2001; Shavelson & Webb, 1991). Except in one of the studies examined, G study results were shared. It is appropriate to share the results of analyses obtained to serve the purpose of a study. If the variance components estimated as a result of the G study are not evaluated, it may be sufficient to share only the predicted G and Phi coefficients. The present study also considered the D studies carried out within G Theory by 56 of the studies examined. In education, researchers frequently investigate ways to reduce error and improve the reliability of measurements designed for particular purposes. The fact that most of the examined studies carried out D studies, which serve this purpose within the framework of G Theory, may reflect a general effort to increase credibility.

An evaluation of computer programs revealed that a majority of studies used EduG. This may reflect the fact that EduG is free and relatively user-friendly. The common use of G Theory depends not only on the need for available reference materials that clearly and comprehensibly explain its theoretical foundations, but also on the availability of user-friendly computer programs to perform analyses. The first computer program developed to carry out G Theory analyses was GENOVA, developed by Brennan in 1983. Brennan wrote a detailed explanation of both univariate and multivariate G Theory in "Generalizability Theory," released in 2001; the author developed mGENOVA for multivariate analyses, and urGENOVA to estimate balanced and unbalanced designs and random effect variance components. The syntax in which G Theory analyses could be performed on the SPSS, SAS, and MATLAB programs was released in 2006. The fact that this syntax has been organized for use with relatively common programs is a positive step toward expanding the use of G Theory. In 2006, Jean Cardinet released EduG Program. Finally, in 2011, the G-String Program—which can also be used for unbalanced designs—was produced by Bloch and Norman. In addition to these software tools, it is possible to carry out a G Theory analysis via the "gtheory" package using R, which is a free software. It is possible to see the impact of software by examining the yearly distribution of G Theory studies (see [Figure 2](#)). Improved software compatible with G Theory analyses and the publication of user guides will increase its use among researchers from various disciplines. In this context, computer programs are clearly important.



As the literature indicates, variance estimates can sometimes be negative. Negative estimates are caused by erroneous measurement models or sampling errors (Güler et al., 2012). Since a negative variance indicates the wrong choice of models or samples, precautions should be taken in cases where the variance is negative. Cronbach et al., (1972) initially said that the negative variance should be replaced with zero and that zero should be used to calculate other variance components. Brennan (1983, 2001), however, argued that this suggestion could cause biased calculations of variance components. Cronbach responded by saying that, although the negative variance should be replaced by zero, the negative value itself should be used to calculate other variance components (Atılgan, 2004). When the variance is negative, the value is either replaced by zero or used as it is. The decision-makers are not researchers, but computer programs. None of the analyzed studies explained this situation. Depending on the software that are used, researchers can access the analysis results or carry out estimates using both approaches.

Another point of importance in G Theory is whether or not the design is balanced. In a balanced design, the number of observations is equal at every level of the source of variability. However, observations per variable are not equal in an unbalanced design, due to lost data or differences between the number of observations and the levels of variables (Brennan, 2001). Let us consider, for instance, a measurement situation in which individuals respond to two different testlets, and in which the items are nested within the testlets and the individuals are crossed with them. If the testlets have an equal number of items, the design is balanced. If each testlet has a different number of items, then the design is unbalanced. In other words, if there are three items in each testlet, the design is balanced—but if there are two items in one testlet and four in the other, the design is unbalanced. The fact that unbalanced designs have been used in studies based on G Theory in recent years indicates that researchers are considering using the theory in different measurement situations. The very small number of studies using unbalanced designs ( $f=3$ ) may reflect the fact that designs involving unbalanced data are relatively complex. Another explanation may be that the researchers have removed some data or filled in the missing data to change their unbalanced designs into balanced ones. One final explanation may be that, previously, G Theory analyses of unbalanced data could only be carried out using the urGENOVA program—which was very complex and avoided by researchers. The G String program produced in 2011 by Bloch and Norman (used in three unbalanced-design studies) is an easy-to-use and useful program. This program may become widespread and commonly used with the researches conducted with unbalanced datasets.

Since the study has offered both detailed conceptual and theoretical explanations, as well as information on general biases, it can serve as a resource for prospective researchers. One limitation of the present research is the fact that it focused on studies in the field of educational sciences. Despite this, it provides information that could be of substantial value to the researchers in other fields, helping to promote the widespread and accurate use of G Theory in many other scientific fields.

Initiatives designed to increase the use of G Theory, such as organizing seminars and workshops, supervising post-graduate theses, and writing books and articles to inform researchers, will raise awareness among scientists, encourage them to use G Theory and increase its use. G Theory can also be introduced to all departments in educational faculties and medical schools. Many departments within faculties of education, health sciences (a field in which G Theory studies are relatively common), and educational sciences fields that carry out measurement and evaluation research will also benefit from using G Theory. Since there can be multi faceted measurement designs in the above mentioned fields, carrying out the researches by using G Theory can improve the qualifications of those studies. For instance, there are studies related to special education (Pekin, Çetin, & Güler, 2018), science education

(Shavelson, Baxter, & Pine, 2009; Yin & Shavelson, 2008), mathematics education (Kersting, 2008; Lane, Liu, Ankenmann, & Stone, 1996), medical education (Lafave & Butterwick, 2014; Turner, Lozano-Nieto, & Bouffard, 2006) and dentistry education (Ta delen Teker & Odaba 1, 2019; Gadbury-Amyot et al., 2014).

## ORCID

Gül en Ta delen Teker  <https://orcid.org/0000-0003-3434-4373>

Ne e Güler  <https://orcid.org/0000-0002-2836-3132>

## 5. REFERENCES

- Arık, R. S. & Türkmen, M. (2009). *Examination of the articles in the scientific journals published in the field of educational sciences*. Paper presented at I. International Congress of Educational Research, Çanakkale, Turkey.
- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok de i kenlik kaynaklı Rasch modelinin kar ıla tırılmasına ili kin bir ara tırma* [A research on the comparison of the generalizability theory and many facet Rasch model] (Doctoral Dissertation). Hacettepe University, Ankara.
- Atılğan, H. (2013). Sample size estimation of G and Phi coefficients in generalizability theory. *Eurasian Journal of Educational Research*, 51, 215–228.
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36, 258–267.
- Bekta , M., DüNDAR, H. & Ceylan, A. (2013). Investigation of several variables papers national classroom teacher education symposium. *U ak University Journal of Social Sciences*, 6(2), 201–226. DOI: <http://dx.doi.org/10.12780/UUSB167>
- Bloch, R. & Norman, G. (2011). *G String 4 user manual* (Version 6.1.1). Hamilton, Ontario, Canada. Retrieved from [http://fhsperd.mcmaster.ca/g\\_string/download/g\\_string\\_4\\_manual\\_611.pdf](http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf)
- Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, 24, 1–21. doi:10.1080/08957347.532417.
- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.
- Brennan, R. L. (1997). A perspective on the history of Generalizability Theory. *Educational Measurement: Issues and Practice*, 16(4), 14–20. <https://doi.org/10.1111/j.1745-3992.1997.tb00604.x>
- Brennan, R. L. (1992). *Elements of Generalizability Theory*. NY: Springer-Verlag.
- Çalık, M. & Sözbilir, M. (2014). Parameters of content analysis. *Education and Science*, 39(174), 33–38. doi:10.15390/EB.2014.3412
- Çilta , A. (2012). Content analysis of the graduate thesis and dissertations in mathematics education in Turkey between 2005-2010. *The Journal of Academic Social Science Studies*, 5(7), 211–228.
- Çilta , A., Güler, G. & Sözbilir, M. (2012). Mathematics education research in Turkey: A content analysis study. *Educational Sciences: Theory & Practice*, 12(1), 565–580.
- Deliceo lu, G. (2009). *The Comparison of the reliabilities of the soccer abilities'rating scale based on the Classical Test Theory and Generalizability*. (Doctoral Dissertation). Ankara University, Ankara.
- Do ru, M., Gençosman, T., Ataalkın, A. N. & eker, F. (2012). Fen bilimleri e itiminde çalı ılan yüksek lisans ve doktora tezlerinin analizi [Analysis of master's and doctoral theses in science education]. *Journal of Turkish Science Education*, 9(1), 49–64.
- Finfgeld, D. L. (2003). Metasynthesis: The state of the art-so far. *Qualitative Health Research*, 13(7), 893–904. DOI: 10.1177/1049732303253462

- Gadbury-Amyot, C. C., Kim, J., Palm, R. L., Mills, G. E., Noble, E. & Overman, P. R. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program. *Journal of Dental Education*, 67(9), 991-1002.
- Gökta , Y., Küçük, S., Aydemir, M., Telli, E., Arpacık, Ö., Yıldırım, G. & Reiso lu, . (2012). Educational technology research trends in Turkey: A content analysis of the 2000–2009 decade. *Educational Sciences: Theory & Practice*, 12(1), 191–196.
- Gülbahar, Y. & Alper, A. (2009). Trends and issues in educational technologies: A review of recent research in TOJET. *The Turkish Online Journal of Educational Technology – TOJET*, 8(2), 124-135.
- Güler, N., Kaya Uyanık, G. & Ta delen Teker, G. (2012). *Genellenebilirlik Kuramı* [Generalizability Theory]. Ankara: PegemA Yayıncılık.
- Güler, N. (2008). *A research on classical test theory, generalizability theory and Rasch model* (Doctoral Dissertation). Hacettepe University, Ankara.
- Günay, R. & Aydın, H. (2015). Inclinations in studies into multicultural education in Turkey: A content analysis study. *Education and Science*, 40(178), 1–22. DOI: <http://dx.doi.org/10.15390/EB.2015.3294>
- Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 3847. <http://dx.doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Kaleli Yılmaz, G. (2015). Analysis of technological pedagogical content knowledge studies in Turkey: A meta-synthesis study. *Education and Science*, 40(178), 103–122. DOI: <http://dx.doi.org/10.15390/EB.2015.4087>
- Karada , E. (2009). E itim bilimleri alanında yapılmı doktora tezlerinin incelenmesi [A Thematic Analysis on Doctoral Dissertations Made In the Area of Education Sciences],. *Ahi Evran Üniversitesi E itim Fakültesi Dergisi* 10(3), 75–87.
- Kersting, N. (2008). Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers’ Knowledge of Teaching Mathematics. *Educational and Psychological Measurement*, 68(5), 845-861. DOI:10.1177/0013164407313369
- Kılıç Çakmak, E., Çebi, A., Mihçi, P., Günbatar, M. S. & Akçayır, M. (2013). *A content analysis of educational technology research in 2011*. 4th International Conference on New Horizons in Education. INTE 2013 Proceedings Book, 397–409.
- Lafave, M. R. & Butterwick, D. J. (2014). A generalizability theory study of athletic taping using the Technical Skill Assessment Instrument. *Journal of Athletic Training*, 49(3), 368-372. doi: 10.4085/1062-6050-49.2.22
- Lane, S., Liu, M., Ankenmann, R. D. & Stone, C. A. (1996). Generalizability and Validity of a Mathematics Performance Assessment. *Journal of Educational Measurement*, 33(1), 71-92. <https://doi.org/10.1111/j.1745-3984.1996.tb00480.x>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded Sourcebook*. (2nd ed). Thousand Oaks, CA: Sage.
- Patton, M.Q. (2002). *Qualitative research and evaluation methods* (3rd Ed.). London: Sage Publications, Inc.
- Pekin Z., Çetin S. & Güler N. (2018). Comparison of Interrater Reliability Based on Different Theories for Autism Social Skills Profile. *Journal of Measurement and Evaluation in Education and Psychology*, 9(2), 202-215. <https://doi.org/10.21031/epod.388590>
- Rios, J.A., Li, X., & Faulkner-Bond, M. (2012, October). *A review of methodological trends in generalizability theory*. Paper presented at the annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.
- Saban, A. (2009). Content analysis of Turkish studies about the multiple intelligences theory. *Educational Sciences: Theory & Practice*, 9(2), 833–876.

- Shavelson, R. J. & Webb, N. M. (1991). *Generalizability Theory: A primer*. Newbury Park, CA: Sage.
- Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education*, 4(4), 347-362. DOI: [10.1207/s15324818ame0404\\_7](https://doi.org/10.1207/s15324818ame0404_7)
- Ta delen Teker, G. & Odaba 1, O. (2018). Reliability of scores obtained from standardized patient and instructor assessments. *European Journal of Dental Education*, 23, 88-94. DOI: [10.1111/eje.12406](https://doi.org/10.1111/eje.12406)
- Turner, A. A., Lozano-Nieto, A. & Bouffard, M. (2006). Generalizability of extracellular-to-intracellular fluid ratio using bio-impedance spectroscopy. *Physiological Measurement*, 27(4), 385-397. DOI: [10.1088/0967-3334/27/4/005](https://doi.org/10.1088/0967-3334/27/4/005)
- Yalçın, S., Yavuz, H. Ç. & Iğün Dibek, M. (2015). Content analysis of papers published in educational journals with high impact factors. *Education and Science*, 40 (182), 1–28. DOI: [10.15390/EB.2015.4868](https://doi.org/10.15390/EB.2015.4868)
- 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](https://doi.org/10.1080/08957340802161840)
- Walsh, D. & Downe, S. (2005). Meta-synthesis method for qualitative research: A literature review. *Journal of Advanced Nursing*, 50(2), 204–211.
- Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using Generalizability Theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, 21, 81–90.

## 6. APPENDIX: List of studies included in research

- Atılğan, H. (2004). *A Research on comparisons of Generalizability Theory and many facets Rasch measurement* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Akta, M. (2013). *An investigation of the reliability of the scores obtained through rating the same performance task with three different techniques by different numbers of raters according to Generalizability Theory* (Unpublished master's thesis). Mersin University, Mersin, Turkey.
- Alkan, M. (2013). *Comparison of different designs in scoring of PISA 2009 reading open ended items according to Generalizability Theory* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Anadol, H. Ö. (2017). *The examination of reliability of scoring rubrics regarding raters with different experience years* (Unpublished master's thesis). Ankara University, Ankara, Turkey.
- Arsan, N. (2012). *Investigation of the raters' assessment in ice skating with Generalizability Theory and Rasch measurement* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Bacı, V. (2015). *The comparison of different designs in generalizability theory with Classical Test Theory in the measurement of mathematical reasoning ability* (Unpublished master's thesis). Gazi University, Ankara, Turkey.
- Bağcı, B. (2015). *A generalizability analysis of the reliability of measurements, applied in the sixth grade science course "Let's circuit electric" unit* (Unpublished master's thesis). Gaziosmanpaşa University, Tokat, Turkey.
- Büyükkıdık, S. (2012). *Comparison of interrater reliability based on the Classical Test Theory and Generalizability Theory in problem solving skills assessment* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Çakıcı Eser, D. (2011). *Comparison of interrater agreement calculated with Generalizability Theory and logistic regression* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Deliceoğlu, G. (2009). *The comparison of the reliabilities of the soccer abilities' rating scale based on the Classical Test Theory and Generalizability* (Unpublished doctoral dissertation). Ankara University, Ankara, Turkey.
- Güler, N. (2008). *A research on Classical Test Theory, Generalizability Theory and Rasch Model* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish* (Unpublished master's thesis). Boaziçi University, İstanbul, Turkey.
- Gündoğdu, C. (2012). *A comparison of Angoff, Yes/No and Ebel standard setting methods* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Kaya Uyanık, G. (2014). *Investigation of two facets design with generalizability in item response modeling* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Kaya, G. (2011). *Application of Generalizability Theory to fill-in concept map assessment* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Kızıltoprak, F. (2016). *The comparison of reliability of PISA mathematical literacy items' competency needs scores obtained with competency scheme based on Generalizability Theory and Classical Test Theory* (Unpublished master's thesis). Gazi University, Ankara, Turkey.
- Küçük, F. (2017). *Assessing academic writing skills in Turkish as a foreign language* (Unpublished master's thesis). Boaziçi University, İstanbul, Turkey.

- Nalbanto lu Yılmaz, F. (2012). *Comparison of balanced and unbalanced designs in Generalizability Theory* (Unpublished doctoral dissertation). Ankara University, Ankara, Turkey.
- Nalbanto lu, F. (2009). *Comparison of different designs in accordance with the Generalizability Theory in performance measurements* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Özberk, E. H. (2012). *Comparing different coefficients in Generalizability Theory decision studies* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Öztürk, M. E. (2011). *The comparison of reliability of the "volleyball abilities observation form" (vaof) points to the Generalizability and the Classical Test Theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Pekin, Z. (2015). *Comparison of interrater reliability based on Classical Test Theory and Generalizability Theory for autism social skills profile* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- algam, A. (2016). *The comparison of reliability of the Generalizability Theory and the test-retest technique for the short answered maths exam* (Unpublished master's thesis). Gazi University, Ankara, Turkey.
- Ta delen Teker, G. (2014). *The effect of testlets on reliability and differential item functioning* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.
- Ta delen, G. (2009). *A comparison of angoff and nedelsky cutting score procedures using generalizability theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Ta tan, Z. (2017). *Investigation of multi-surface patterns in generalizability* (Unpublished master's thesis). Mersin University, Mersin, Turkey.
- Yelbo a, A. (2007). *The examination of reliability according to classical test and generalizability theory on a job performance scale* (Doctoral Dissertation). Hacettepe University, Ankara, Turkey.
- Yıldıztekin, B. (2014). *The comparison of interrater reliability by using estimating techniques in classical test theory and generalizability theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.
- Yüksel, M. (2015). *Comparison of scores obtained from different measurement tools used in the determination of student achievement* (Unpublished master's thesis). Gaziosmanpa a University, Tokat, Turkey.
- Acar Güvendir, M. & Güvendir, E. (2017). The determination of an english speaking exam's data reliability using Generalizability Theory. *Trakya University Journal of Education*, 7(1), 1-9.
- Anıl, D. & Büyükkıdık, S. (2012). An example application for the use of four facet mixed design in Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, 3(2), 291-296.
- Atılğan, H. (2005). Generalizability Theory and a sample application for inter-rater reliability. *Educational Sciences and Practice*, 4(7), 95-108.
- Atılğan, H. (2008) Using Generalizability Theory to assess the score reliability of the Special Ability Selection Examinations for music education programmes in higher education. *International Journal of Research & Method in Education*, 31(1), 63-76, DOI:10.1080/17437270801919925
- Atılğan, H. (2013). Sample size for estimation of G and Phi coefficients in Generalizability Theory. *Eurasian Journal of Educational Research*, 51, 215-228
- Atılğan, H. & Tezba aran, A. A. (2005). An investigation on consistency of G and Phi coefficients obtained by Generalizability Theory alternative decisions. *Eurasian Journal of Educational Research*, 5(18), 28-40.

- Can Aran, Ö., Güler, N. & Senemo lu, N. (2014). An evaluation of the rubric used in determining students' levels of disciplined mind in terms of Generalizability Theory. *Dumlupınar University Journal of Social Sciences*, 42, 165-172.
- Çetin, B., Güler, N. & Sarıca, R. (2016). Using Generalizability Theory to examine different concept map scoring methods. *Eurasian Journal of Educational Research*, 66, 212-228. <http://dx.doi.org/10.14689/ejer.2016.66.12>
- Do an, C. D. & Anadol, H. Ö. (2017). Comparing fully crossed and nested designs where items nested in raters in Generalizability Theory. *Kastamonu Education Journal*, 25(1), 361-372.
- Do an, C. D. & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, 17, 631-651. <http://dx.doi.org/10.12738/estp.2017.2.0321>
- Gözen, G. & Deniz, K. Z. (2016). Comparison of instructor and self-assessments on prospective teachers' concept mapping performances through Generalizability Theory. *International Journal on New Trends in Education and Their Implications*, 7(1), 28-40.
- Güler, N. (2009). Generalizability Theory and comparison of the results of G and D studies computed by SPSS and GENOVA packet programs. *Education and Science*, 34(154), 93-103.
- Güler, N. (2011). The comparison of reliability according to Generalizability Theory and Classical Test Theory on random data. *Education and Science*, 36(162), 225-234.
- Güler, N. & Gelbal, S. (2010). Studying reliability of open ended mathematics items according to the Classical Test Theory and Generalizability Theory. *Educational Sciences: Theory & Practice*, 10(2), 989-1019.
- Güler, N. & Ta delen Teker, G. (2015). The evaluation of rater reliability of open ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 12-24.
- Güler, N., Ero lu, Y. & Akbaba, S. (2014). Reliability of criterion-dependent measurement tools according to Generalizability Theory: Application in the case of eating skills. *Abant İzzet Baysal University Journal of Education*, 14(2), 217-232.
- Han, T. (2017). scores assigned by inexpert efl raters to different quality EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education*, 13(1), 136-152.
- Han, T. & Ege, . (2013). Using Generalizability Theory to examine classroom instructors' analytic evaluation of EFL writing. *International Journal of Education*, 5(3), 20-35.
- İhan, M. & Gezer, M. (2017). A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. *Pegem Eğitim ve Öğretim Dergisi*, 7(4), 637-662, <http://dx.doi.org/10.14527/pegegog.2017.023>
- Kamı , Ö. & Do an, C. D. (2017). How consistent are decision studies in G Theory? *Gazi University Journal of Gazi Educational Faculty*, 37(2), 591-610.
- Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalizability Theory. *Perceptual and Motor skills*, 105, 891-905.
- Kara, Y. & Kelecio lu, H. (2015). Investigation the effects of the raters' qualifications on determining cutoff scores with Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 58-71.
- Nalbanto lu Yılmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice*, 17, 395-409. <http://dx.doi.org/10.12738/estp.2017.2.0098>

- Nalbanto lu Yılmaz, F. & Ba usta, B. (2015). Using Generalizability Theory to assess reliability of suturing and remove stitches skills station. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 107-116.
- Ö retmen, T. & Acar, T. (2014). Estimation of generalizability coefficients: An application of structural equation modeling. *Journal of Education and Practice*, 5(14), 113-118.
- Polat Demir, B. (2016). The examination of reliability of vee diagrams according to Classical Test Theory and Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 419-431.
- Ta delen Teker, G., Güler, N. & Kaya Uyanık, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for Generalizability Theory. *Educational Sciences: Theory & Practice*, 15(3), 635-645. DOI: [10.12738/estp.2015.3.2278](https://doi.org/10.12738/estp.2015.3.2278)
- Ta delen Teker, G., ahin, M. G. & Baytemir, K. (2016). Using Generalizability Theory to investigate the reliability of peer assessment. *Journal of Human Sciences*, 13(3), 5574-5586. doi:[10.14687/jhs.v13i3.4155](https://doi.org/10.14687/jhs.v13i3.4155)
- Yelbo a, A. (2008). The assessment of reliability with Generalizability Theory: An application in industrial and organizational psychology. *Studies in Psychology*, 28, 35-54.
- Yelbo a, A. (2012). Dependability of job performance ratings according to Generalizability Theory. *Education and Science*, 37(163), 157-164.
- Yelbo a, A. (2015). Estimation of generalizability coefficient: An application with different programs. *Archives of Current Research International*, 2(1), 46-53.



## Examination of the Extreme Response Style of Students using IRTree: The Case of TIMSS 2015

Munevver Ilgun Dibek <sup>1</sup>\*

<sup>1</sup> Department of Educational Sciences, TED University, Ankara, Turkey

### ARTICLE HISTORY

Received: 01 March 2019

Revised: 13 May 2019

Accepted: 18 May 2019

### KEYWORDS

Attitude,  
Extreme Response Style,  
Item Response Tree,  
TIMSS,

**Abstract:** In the literature, response style is one of the factors causing an achievement-attitude paradox and threatens the validity of the results obtained from studies. In this regard, the aim of this study is two-fold. Firstly, it attempts to determine which item response tree (IRTtree) models based on the generalized linear mixed model (GLMM) approach (random intercept, random intercept with fixed effect of extreme response and random intercept-slope model) best fit the Trends in International Mathematics and Science Study (TIMSS) 2015 data. Secondly, it purports to explore how the extreme response style affects students' attitudes toward mathematics of students. This study is both basic research and descriptive research in terms of seeking for answers for two different research questions. For the sample of this research, 15 countries were randomly selected among countries participated in TIMSS 2015. The students' responses to items measuring attitude in the student questionnaire were analyzed with the packages "lme4" and "irtrees" in R software. When the model fit indices were evaluated, the random intercept-slope model was found to be the best fit to the data. According to this model, the extreme response style explains a significant amount of variances in the students' attitude toward mathematics. Additionally, students with a negative attitude toward mathematics were found to have an extreme response style. It was concluded that an extreme response style had an effect on students' attitude.

## 1. INTRODUCTION

International comparative studies investigating the relationship between attitude and achievement have reported conflicting results. Some researchers (Kadijevich, 2008; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005) indicated that students with a high level of achievement in a domain tended to hold positive attitudes toward mathematics while others (Buckley, 2009; Van de Gaer & Adams, 2010) found that these students had negative attitudes toward the course despite their high achievement. The negative relationship between attitude and achievement is also observed in international comparison studies concerning student performance, such as TIMSS and The Programme for International Student Assessment (PISA).

In contrast to motivational theories, such as the expectancy value theory (Atkinson, 1957), which emphasizes the positive relationship between attitude and achievement, the direction of

---

CONTACT: Munevver Ilgun Dibek ✉ [munevver.ilgun@tedu.edu.tr](mailto:munevver.ilgun@tedu.edu.tr) 📧 Department of Educational Sciences, TED University, Ankara, Turkey

the relationship between attitude and achievement varies according to the investigation being conducted at an individual or group level. In other words, there may be a positive relationship between the attitudes of students toward a domain within a country, but a negative correlation may be found between student attitude and achievement between countries (Bofah & Hannula, 2015; Van de gaer, Grisay, Schulz, & Gebhardt, 2012). Therefore, the interchangeable use of correlations identified at the individual and group levels reduces the validity of the results obtained from the studies (Robinson, 1950).

In the literature, the attitude-achievement paradox is defined as the relationship between attitude and achievement being positive at the individual level but negative at the group level (Van de et al., 2012). Another reason is the response style differences between countries (Buckley, 2009). Response style is “the tendency to respond systematically to the items of a questionnaire regardless of their content” (Paulhus, 1991, p.17). The response style of individuals creates various psychometric problems in the data (Bolt & Newton, 2011). More specifically, it reduces the validity of test scores by producing a systematic error in the test scores of individuals with the same level of knowledge, attitude or similar personality characteristics (Cronbach, 1946). When focus is narrowed from the response style to extreme response style (ERS), ERS pulls the response away from the center (midpoint) and therefore increases the estimated variance. Additionally, when one of the end point (extreme response categories) is more chosen, bias can occur. More precisely, when people are more prone to choose positive extreme category than negative extreme category, a positive bias may occur. On the other hand, if people are more prone to choose negative extreme response category when compared to positive extreme response category, bias will be in the negative way (Liu, 2015). Since correlation and variance of the scores are partially related to each other, correlation between the variables is also affected because of the extreme response style. Specifically, since ERS causes the increased variance, the correlation between the variables of interest decreases as the tendency of choosing extreme end points the individuals increases (Heide & Gronhaug, 1992). Additionally, due to the fact that several statistical techniques such as regression analysis, canonical correlation analysis, factor analysis are based on correlation, ERS will affect the results obtained from them (Peterson, Rhi-Perez, & Albaum, 2012). Also, within-country correlations are affected by the ERS since the amount of the degree of ERS changes from one country to another from one culture to another

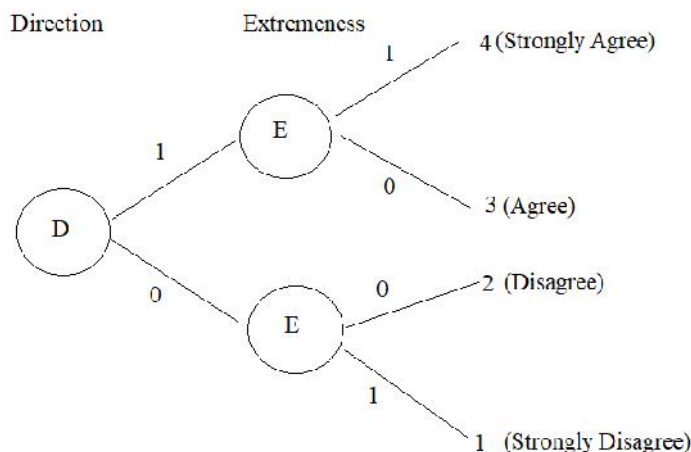
The fact that response styles lead to erroneous inferences and misapplications on educational decisions and policies at the national level makes it important to correct the effects of these response styles on the scores of psychological structures, such as attitudes. In this regard, there are several methods proposed in the literature with a number of model-free and model-based approaches being suggested as a way to address response style in rating data. The first methods are based on getting frequencies of certain response categories which are selected (Bachman & O'Malley, 1984). When there are finite number of response categories, dependencies among them may be observed. In this case, the separate effects of them will be difficult to interpret. Due to these dependencies among these measures, it is valuable to examine whether model-based approaches give rise to similar results. In this regard, the item response tree (IRTtree) model was used in this study because it focuses on a response process that address how response style may affect the selection of a response category (Böckenholt, 2017). The general rationale for selecting this model is twofold: (i) IRTtree models are more flexible and informative, which helps them to solve problems that are not fixed by using other approaches, and (ii) IRTtree models can be seen as the generalized linear mixed models (GLMM), which allows the use of the available user friendly software R and package, namely lme4 (Bates, Maechler, Bolker, & Walker, 2015)

### 1.1. IRTree Model

Response tree models are used for categorical data. In these models, the categorical response categories can be converted to binary responses presented in a binary response tree. In this situation, the response process can be accepted as a sequential process of passing through the tree to its end nodes (Jeon & de Boeck, 2016). The model is referred to as an item response tree model because it utilizes a tree structure (Boeckenholt, 2012; de Boeck & Partchev, 2012). It contains sub-trees, internal nodes, and branches split off from these nodes and leaves. The leaves can be seen as terminal nodes representing the observed categorical item responses. In a tree structure, nodes and branches are represented by circles and arrows, respectively.

The IRTree model is can be used to handle extreme response tendencies in the multidimensional item response theory framework. With this model, when individuals respond to ordinally scaled items, it is assumed that s/he engages in a two stage decision-making process (Böckenholt, 2012). For instance, from an item with response options “1 (Strongly Disagree)”, “2 (Disagree)”, “3 (Agree)”, and “4 (Strongly Agree)”, a person may choose response categories depending on two processes: s/he may first decide on in which the direction s/he should give a response (positive or negative), and then decide on the extremeness of the response (Thissen-Roe & Thissen, 2013). Each of these processes is referred to as a pseudo-item that is modelled with a one- or two-parameter IRT model (Böckenholt & Meiser, 2017). In other words, for the estimation of the multiple response models, pseudo-items are used to represent the outcomes of each response process (Böckenholt, 2012)

An IRTree model is used to measure the sequential decision-making response process. In figure 1, IRTree model developed for a four-category Likert-scaled item is presented. The probability of the direction of response (either agree or disagree) can be represented as a function of a latent trait,  $\theta_1$ , which indicates the substantive trait of interest. The probability of extremeness of the response can be represented as a function of a latent variable  $\theta_{ERS}$ , which refers to person’s tendency to choose extreme responses (Thissen-Roe & Thissen, 2013). In this situation, the probability of response extremeness is assumed to be independent from the first decision.



**Figure 1.** IRTree Model for a four-category item

This tree is called a nested tree since every node is connected to another node by branches (Jeon & de Boeck, 2016). A two-parameter logistic (2PL) is used to model the first decision:

$$P(D_1=1|\theta_1)=\frac{1}{1+e^{-(b_1+a_1\theta_1)}} \text{ and} \tag{1}$$

$$P(D_1=0|\theta_1) = 1 - P(D_1=1|\theta_1) \tag{2}$$

where  $b_1$  refers to the intercept parameter and  $a_1$  is the discrimination parameter (Thissen-Roe & Thissen, 2013). A modified 2PL model is used to model the second decision:

$$P(D_1=1 | \theta_{ERS}, \theta_1) = \frac{1}{1 + e^{-(b_2 + u_2 \theta_E - \nu \text{var}(b_1 + a_1 \theta_1))}} \text{ and} \tag{3}$$

$$P(D_1=0 | \theta_{ERS}, \theta_1) = 1 - \frac{1}{1 + e^{-(b_2 + u_2 \theta_E - \nu \text{var}(b_1 + a_1 \theta_1))}} \tag{4}$$

where  $b_2$  refers to the intercept parameter and  $a_2$  refers to the slope parameter indicating the item-specific probability of extreme responding (Thissen-Roe & Thissen, 2013). The  $\nu$  parameter is used to represent compensatory characteristics of the two traits. This shift term,  $\nu \text{var}(b_1 + a_1 \theta_1)$ , is used as an additive term when the response categories “3” and “4” (i.e.,  $k=3$  and  $k=4$ ) and subtractive term when the response categories “1” and “2” (i.e  $k=3$  and  $k=4$ ). When  $\nu$  is positive, respondents with moderate tendencies will only give an extreme response when their position has a strong intensity (Leventhal & Stone, 2018).

The model formulation of IRTree is based on two main assumptions: (i) the outcomes of the internal nodes are independent of each other, and (2) each observed outcome is associated with only one path. More precisely, according to these assumptions, each particular sequence of conditionally independent internal decisions are resulted in a different observed outcome (‘1’, ‘2’, ‘3’, ‘4’). For example, the probability of a response given to an item is computed as the product of the probability of decision 1 and the probability of decision 2. To explain it with a formula, as stated by Leventhal and Stone (2018), the probability of selecting response option  $k$  given  $\theta_1$  and  $\theta_{ERS}$  is

$$P(U=k)=P(D_1)*P(D_2) \text{ for } k=1,2,3, a \quad 4 \tag{5}$$

In general, for each internal node of the tree, a different latent variable for each split between the categories are allowed in IRTree models (linear or nested response trees). In addition, a different set of item parameters can be used depending on the split. All these facilities make it possible for these models to measure latent variables with a different manner when compared to other methods using a simple correct-incorrect scoring and other classical ordered-category models (such as partial credit model-PCM and graded response model) (Boeck & Parthchev, 2012)

### 1.2. Other Related models

The main characteristics of IRTree models are that (i) they can be represented as a tree structure and (2) they take into consideration of multiple sources of personal differences. In IRT, to model categorical item responses a tree structure is exploited implicitly. For instance, in sequential models proposed by Tutz (1990), all options for an item are reviewed sequentially. These models include attempt-specific parameters to account for different probabilities of success over repeated attempts. In a study conducted by Culpepper (2014), item responses some of which were partially ordered and others were repeatedly attempted were modelled using a sequential decision rule. Yavuz, Bulut, Ilgun Dibek and Kursad (2018) used sequential models for repeatedly attempted item responses to determine the effect of this modelling on the students’ performance. In addition, in different models were used, such as the rating scale model

(Andrich, 1978), partial credit model (Masters, 1982), generalized partial credit model (Muraki, 1992), and a divide-by-total scoring rule indicating possible options are reviewed immediately prior to final response. For example, in a study conducted by Ilgun Dibek, Bulut, Kursad and Yavuz (2018), students' responses were modelled utilizing this rule in PCM. However, these models mentioned above address a single source of individual differences in responding to scale. Apart from these models, there are other IRT models that take into consideration multiple sources of individual differences in students' responses to items. For example, Huang (2016) used the mixture random effect model to investigate the effect of ERS on rating scales by identifying several latent classes from different ERS levels and detecting the possible items which function differentially due to ERS. Johnson (2007) merged multiple latent traits to address personal differences in response styles. Bolt, Wollack and Suh (2012) extended the nested logit model to multidimensional model which can be used for multiple latent traits to be applied to the choice of distractors for multiple-choice items. To narrow down these studies, De Boeck and Partchev (2012) and Boeckenholt (2012) proposed item response models which are represented as a tree structure and allow for the handling of multiple causes of personal differences.

To sum up, when the literature and related methods for response style were examined it is clear that the negative effects of the extreme response style on results obtained from several techniques and international comparison studies occur. In addition, there are several the handicaps of different methods to determine the effect of ERS. Therefore, using relatively new method which is more efficient to determine effect of it is necessary.

The purpose of this study is to determine the best IRTree model based on the GLMM approach. It is also aimed to predict how extreme response style (ERS) affect students' attitude scores under the best model using TIMSS 2015 data. In this context, the questions that are sought to be answered in the study are:

1. Which of the IRTree models (random intercept model, random intercept model with ERS effect, random intercept-slope model) is best fitted to the TIMSS 2015 subdata?
2. What is the effect of ERS on the students' scores regarding attitude-related constructs (liking mathematics, self-confidence in mathematics, and value on mathematics) based on the model that best fits the data?

## **2. METHOD**

This is a basic research study in terms of determining the model that best fits the data by analyzing different IRTree models based on the GLMM approach, and thus contributing to the information necessary for test development theories (Kidd, 1959) as well as a descriptive research study in terms of determining the effect of ERS among students and items and thus providing accurate description of the phenomenon (Johnson & Christensen, 2008).

### **2.1. Population and Sample**

The sample of the present study consisted of eighth-grade students of the countries in which the attitude-achievement paradox was observed in TIMSS 2015. A two-stage stratified sampling procedure was used to select the students. In the first stage, schools were chosen randomly in accordance with their proportion in the population. In the second stage, at least one class was randomly chosen from each of these schools. All the students in these classes were included in the study (LaRoche, Joncas, & Foy, 2016). The reason why eighth grade students were chosen is that fourth grade students, who also participated in TIMSS 2015, are not considered to be aware of their own competences and attitudes, and thus cannot evaluate themselves effectively (Harter, 1999).

To determine which countries would be included in this study, all the countries were ranked according to their mathematics achievement. Also, the percentage of the students whose attitudes were negative were taken into consideration. Accordingly, in the five of these countries, the students' scores were above the average mathematics achievement of all countries that participated in TIMSS 2015, but the percentage of the students with a negative attitude toward mathematics regarding three attitudinal constructs were higher compared to the other countries. In another five countries, the students had low mathematics achievement, but the percentage of the students who had a negative attitude toward mathematics regarding three attitudinal constructs was lower than the other countries. Therefore, these 10 countries were selected since they better displayed the paradoxical relationship between attitude and achievement. Then, to better represent the pattern of the relationship between attitude and achievement of all countries participated in TIMSS 2015 and to equate the number of countries in each segment, five countries in which the students had moderate mathematics achievement and attitude toward mathematics were also selected. As a result, 15 countries were chosen.

For the selected countries, the mathematics achievement scores and percentages of the students who had a negative attitude toward mathematics are given in [Table 1](#) (Mullis, Martin, Foy, & Hooper, 2016).

**Table 1.** Mathematics Achievement and Percentages of the Students

Countries	Mathematics Achievement	Do Not Like Learning Mathematics (%)	Not Confident in Mathematics (%)	Do Not Value Mathematics (%)
Singapore	621	33	46	8
Korea	606	58	55	24
Taipei	599	56	60	41
Hong-Kong	594	46	54	29
Japan	586	59	63	29
Norway	512	48	29	8
Australia	505	50	43	12
Sweden	501	52	41	14
<b>International Average</b>	<b>500</b>	<b>38</b>	<b>43</b>	<b>13</b>
Italy	494	51	43	24
Malta	494	49	49	11
Turkey	458	30	54	12
Chile	427	50	52	12
Kuwait	392	36	38	12
Egypt	368	20	34	7
Saudi Arabia	392	42	33	15

The population and sample of these countries are presented in [Table 2](#) (LaRoche & Foy, 2016). As it can be seen from [Table 2](#), the number of the schools included in sample and sample size of the students changes from 48 to 285 and from 3759 to 10338, respectively. Moreover, some of the countries (Singapore and Malta) included all schools in their sample.

**Table 2.** Population and Sample

	Population		Sample	
	School	Student	School	Student
Singapore	167	47626	167	6116
Korea	3007	587190	150	5309
Taipei	931	285714	190	5711
Hong-Kong	477	463863	133	4155
Japan	10406	1162528	147	4745
Norway	1000	61174	142	4795
Australia	2436	272115	285	10338
Sweden	1616	95438	150	4090
Italy	5718	554401	161	4481
Malta	48	4004	48	3817
Turkey	15583	1298955	218	6079
Chile	5390	240740	171	4849
Kuwait	327	39997	168	4503
Egypt	9900	1300305	211	7822
Saudi Arabia	7343	402639	143	3759

## 2.2 Data Collection Tools

In the current study, the data collection tool was a student questionnaire including the items concerning the demographic information of the students, their home environment, learning, school environments, their perceptions and attitudes (Hooper, Mullis & Martin, 2013). In this study, the variables related to attitude, such as students' liking learning mathematics, self-confidence in mathematics, and value on mathematics were addressed in order to examine the attitude achievement paradox mentioned in the literature and in the TIMSS report (Mullis et al., 2016). The items related to these variables have four response categories, ranging from 1 (strongly agree) to 4 (strongly disagree). Therefore, a high score obtained from these scales in TIMSS 2015 shows a negative attitude toward mathematics, while low scores indicate a positive attitude. The Cronbach alpha reliability coefficients of the scores of the scales obtained from the selected countries varied between .70 and .96 (Martin, Mullis, Hooper, Yin, Foy, & Palazzo, 2016). The fact that the reliability coefficients were greater than .70 indicates that the scores obtained from the scales are reliable (Nunnally, 1978).

## 2.3 Data Analysis Procedures

The missing values in the data set of each country were deleted considering the high number of individuals in the samples and the possibility of multiple imputation affecting response categories (Mooi, Sarstedt, & Mooi-Rec, 2018) selected by students, which is crucial and main focus for this study. As the categories of response to the items in the scales are ranked as higher values representing negative attitude, a reverse coding was undertaken in order that the higher values obtained from the scales would indicate positive attitude toward mathematics. The students' responses for each item were modeled by the IRTree given in Figure 1, and the responses in this figure were converted to pseudo items presented in Table 3.

In Table 3, the pseudo-items and the category probabilities for this IRTree model are given. For each item and student, two responses were assigned. For example, if the student's responses to attitudinal item was "1", namely "strongly disagree", s/he received a score of "0" for node D and "1" for node "E". The same procedure was implemented for all responses to the items of the three attitudinal constructs.

**Table 3.** Pseudo-items for four-category model

Response Categories	D	E	Category Probability
1	0	1	$\left(1 - \frac{1}{1 + e^{-(b_1 + a_1\theta_1)}}\right) \left(\frac{1}{1 + e^{-(b_2 + a_2\theta_E + \bar{\tau}(b_1 + a_1\theta_1))}}\right)$
2	0	0	$\left(1 - \frac{1}{1 + e^{-(b_1 + a_1\theta_1)}}\right) \left(1 - \frac{1}{1 + e^{-(b_2 + a_2\theta_E + \bar{\tau}(b_1 + a_1\theta_1))}}\right)$
3	1	0	$\left(\frac{1}{1 + e^{-(b_1 + a_1\theta_1)}}\right) \left(1 - \frac{1}{1 + e^{-(b_2 + a_2\theta_E + \bar{\tau}(b_1 + a_1\theta_1))}}\right)$
4	1	1	$\left(\frac{1}{1 + e^{-(b_1 + a_1\theta_1)}}\right) \left(\frac{1}{1 + e^{-(b_2 + a_2\theta_E + \bar{\tau}(b_1 + a_1\theta_1))}}\right)$

Once the scores were assigned to nodes, three different IRTree models based on GLMM were applied and analyzed separately for three attitudinal constructs. Model 1 was created by including the fixed effects of students. In this model, each subject is assigned a different intercept value. In other words, this model accounts for baseline-differences in attitude toward mathematics, and it is referred to as the random intercept model. Model 2 was conducted by including fixed effects of students and the fixed effect of nodes; thus, it takes into consideration of the effect of students’ extreme response style on their attitudes toward mathematics. In Model 3, the subjects are allowed to have both differing intercepts and different slopes for the effect of extreme response style, and this shows how the effects of extreme response style varies within the student population. This is called the random intercept-slopes model. All models were estimated using the R packages of lme4 (Bates et al., 2015) and irtrees (Boeck & Partchev, 2012) (see for related codes in Appendix).

After running all the three selected models, ML estimation using likelihood-based fit statistics, such as the likelihood-ratio (LR) statistics, Akaike’s information criterion (AIC), and the Bayesian information criterion (BIC) were performed. The LR statistics to compare the nested tree models was utilized since LR tests can be used to determine the significance of node main effects (Jeon & Boeck, 2016) as follows: suppose  $L_0$  and  $L_1$  are the likelihood of the data for Model 1 with  $p_0$  (number of parameters) and for Model 2 with  $p_1$  (number of parameters), respectively. When Model 1 is nested within Model 2, to compare these models, the following procedure was employed:  $\chi^2 = -2 \times (\log L_0 - \log L_1)$  follows a Chi-squared distribution with  $p_1 - p_0$  degrees of freedom. This test rejects that the null hypothesis if  $\chi^2$  is greater than a Chi-square percentile with  $p_1 - p_0$  degrees of freedom.

To determine how much of the variability in the dependent variable (attitude) was attributable to other variables, such as personal differences and extreme response style, intra-class correlation (ICC) was computed. ICC is calculated by dividing the between-group-variance (random intercept variance) by the total variance. It can be considered as “the proportion of the variance explained by the grouping structure in the population” (Hox, 2002, p.15).

### 3. RESULT / FINDINGS

Analyses conducted to determine the most appropriate IRT model for TIMSS 2015 data resulted in some model fit indices being discussed. Some indices, such as likelihood- (LL), the degree of freedom (df), BIC and AIC are presented in Table 4.



**Table 4.** Model Fit Indices

Variables	Models	AIC	BIC	LL	Deviance	df
Like	Model 1	163473.90	163572.00	-81726.90	163453.90	10
	Model 2	160608.60	160716.50	-80293.30	160586.60	11
	<b>Model3</b>	<b>138874.30</b>	<b>139001.90</b>	<b>-69424.20</b>	<b>138848.30</b>	<b>13</b>
Self-confidence	Model 1	170379.80	170477.90	-85179.90	170360	10
	Model 2	167464.80	167572.70	-83721.40	167443	11
	<b>Model3</b>	<b>153184.30</b>	<b>153311.90</b>	<b>-76579.20</b>	<b>153158</b>	<b>13</b>
Value	Model 1	151333.40	151431.60	-75656.70	151313	10
	Model 2	136347.30	136455.20	-68162.60	136325	11
	<b>Model3</b>	<b>130778.20</b>	<b>130905.80</b>	<b>-65376.10</b>	<b>130752</b>	<b>13</b>

As shown in Table 2, the three IRT models examined with the LL, BIC and AIC values, the model that best fits is the third model for three attitude-related constructs since lower values of these indices indicate a better fit to the data. In addition to these indices,  $-2 \log^2$  values can be compared to determine which model better fits the data. For example, for the variable “students’ liking of mathematics”, Chi-Square statistics, the degree of freedom and the difference between the values of  $-2 \log^2$  belonging to the Model 1 and Model 2 were evaluated first. Since the calculated value ( $\chi^2 = 81726.90 - 80293.3 = 1433.60$ ) is greater than the table value ( $\chi^2(1; .001) = 10.83$ ), the difference between  $-2 \log^2$  values is significant. In this case, it can be said that the Model 2 is more suitable for the data. Then, the same comparison for Model 2 and Model 3 was undertaken. Since the calculated value ( $\chi^2 = 80293.3 - 69424.2 = 10869.10$ ) is greater than the table value ( $\chi^2(2; .001) = 13.82$ ), the difference between  $-2 \log^2$  values is significant. In this case, it can be stated that Model 3 was more suitable for the data. The similar logic is also valid for the other attitude related-constructs.

The estimates of the predictors (items and node 2) for students’ liking of mathematics and the random effects obtained from analyzing model 2 are given in Table 5:

**Table 5.** Model Results

Liking Learning Mathematics			Self-Confidence in Mathematics			Value on Mathematics		
Predictor	Est.	CI	Predictor	Est.	CI	Predic	Est.	CI
item1	.90	.87 - .93	item1	.73	.70 - .76	item1	2.14	2.11 - 2.17
item2	.75	.72 - .78	item2	.31	.28 - .34	item2	1.51	1.48 - 1.54
item3	.26	.23 - .29	item3	.26	.23 - .29	item3	2.32	2.29 - 2.35
item4	.88	.85 - .91	item4	.35	.32 - .38	item4	2.00	1.97 - 2.03
item5	.80	.77 - .83	item5	.42	.39 - .45	item5	.57	.54 - .60
item6	.12	-.05 - .01	item6	.11	.08 - .14	item6	1.85	1.82 - 1.88
item7	.46	.43 - .49	item7	.20	.17 - .23	item7	2.25	2.22 - 2.28
item8	.20	.17 - .23	item8	.54	.51 - .57	item8	2.65	2.62 - 2.68
item9	.51	.48 - .54	item9	.34	.31 - .37	item9	2.91	2.88 - 2.94
node 2	-.95	-.98 - -.92	node 2	-.85	-.88 - -.82	node 2	-1.97	-2.00 - -

Random Effects		Random Effects		Random Effects	
00 person	6.30	00 person	3.49	00 person	3.38
11 person.node2	9.12	11 person.node2	5.52	11 person.node2	3.15
01 person	-.71	01 person	-.66	01 person	-.38
ICC	.41	ICC	.39	ICC	.51

Est.= estimation,  $p < .001$

According to Table 5, for example, for item 1 of the scale concerning students' liking learning mathematics, a one unit increase in the score of item 1 is associated with a .90 unit increase in the expected log odds of students' liking mathematics. Similarly, students who chose extreme response categories are expected to have .95 lower log odds of liking mathematics than students who do not choose extreme response categories. More specifically, tendency of displaying extreme response style decreases their attitude scores regarding liking mathematics by almost 3-fold ( $e^{.95} = 2.56$ ). Additionally, the same logic was found to be valid for the other attitude-related constructs.

For the random effects, the variance at the second node was higher than the variance for an individual. The same was also valid for the "students' self-confidence in mathematics". That is, the variability in the score of students' liking learning mathematics and self-confidence at mathematics was mostly caused by students' extremeness tendency. According to the results concerning the students' self-confidence in mathematics construct, ICC was found to be .41. That is, 41% of the variance of students' attitude scores regarding liking learning mathematics was explained by students' extreme response style and their individual differences. In addition, it was found that there was a negative correlation between students' scores of attitude-related constructs (liking learning mathematics, self-confidence in mathematics and value of mathematics) and node 2 specific traits ( $\beta_{11} = -.71$ ,  $\beta_{12} = -.66$ ,  $\beta_{13} = -.38$ , respectively). This means that students who display a more extreme response style tend to have a lower score regarding attitude toward mathematics. In other words, a student whose attitude is negative tended to more choose categories "1" or "4" since node 2 represents the propensity for selecting an extreme response.

#### **4. DISCUSSION and CONCLUSION**

The first aim of this study was to determine which IRTree models based GLMM approach is best fitted to analyze the TIMSS 2015 subdata. The second aim was to investigate the effect of ERS on students' attitude toward mathematics depending on the analysis of the model that best fitted the data. To achieve these aims, predictions were made by utilizing three different models for each attitudinal constructs.

The third model, which was more complex including both random effect and random slopes for students, as well as the fixed effect of nodes, was concluded to be the best fit to the TIMSS 2015 subdata for three constructs regarding attitude. Similar findings were also found in the study by De Boeck and Wilson (2004), who investigated the role of admission and affirmation in the individuals' responses to items measuring verbal aggression. To achieve this, they tested different models by excluding and including the fixed effect of two nodes and random effect of the individuals. In their tree structure, the first node represents admitting the aggressive reactions and the second node concerned affirmation. They concluded that the most complex model including the fixed effect of the nodes and random effect of the individuals was best fitted to the data.

It was concluded that students' extremeness tendency explained a significant amount variability in students' attitude toward mathematics; thus, an extreme response style had an effect on students' attitude. This result was also supported by a study by Bökhenholt and Meiser (2017), in which different IRT models (mixed polytomous Rasch models and item response tree models) were used to control response styles in rating scales. They indicated that response styles affect students' response to personal need for structure construct and the models used in their study differed in presenting response styles as multidimensional sources of individuals' variances.

In addition, students whose attitude was positive tended to choose mid-points. This result can be related to cultural dimensions of the selected countries. In other words, structure of their

societies may shape their responses to Likert items. For example, according to Hofstede (2001), except for Australia and European countries (Norway, Australia, Sweden, Italy and Malta), the majority of the selected countries are considered to be collectivistic. As emphasized by Hofstede, in collectivist societies, people generally act as members of group or organization. In such cultures, the interconnectedness between individuals plays an important role in their life with loyalty in these societies being at the forefront. Those from collectivistic cultures are more likely to choose responses at midpoints as a result of their desire to maintain harmony in society.

The presence of the effect of the response style in large scale assessments, which was demonstrated in this study, requires all educational stakeholders be more conscious and careful for practice in educational field. Especially, policy makers who cares the results of international assessments must be aware that differences in attitudes of the students coming from different countries may be caused from response style and take several steps by keeping this issue in their mind. Although this study has caught up some valuable points, it has several limitations. Firstly, considering the role of response style on attitude-achievement paradox, only the Likert scales measuring attitudinal constructs have been addressed in this study. Since response style can affect the responses of the students to the items related to other constructs, future researchers can test the model-data fit for the data of different scales used in TIMSS 2015. Also, the approach used in this study could be easily expanded to analyze the effect of other response styles, such as midpoint response style, acquiescence response style. The items used in this study has four response categories. To put it in different words, none of the items have midpoint response categories. This issue may lead the students to choose extreme end-points of the response categories. Therefore, the same approach can be used for items having mid-point response categories to determine whether the presence of this categories change the result. In addition, in this study IRTree models based on the GLMM approach were used due to their flexibility; however, further studies can be conducted to compare other models used for polytomous items and to determine which model is best fitted to the data.

## ORCID

Munevver Ilgun Dibek  <https://orcid.org/0000-0002-7098-0118>

## 5. REFERENCES

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64, 359-373.
- Bachman, J. G., O'Malley, P. M., & Freedman-Doan, P. (2010). *Response styles revisited: Racial/ethnic and gender differences in extreme responding* (Monitoring the Future Occasional Paper No. 72). Ann Arbor, MI: Institute for Social Research.
- Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01
- Bofah, E. A. and Hannula, M. S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-Scale Assessments in Education*, 3(1), doi:1.1186/s40536-015-0014-y
- Bolt, D. M., & Newton, J. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, 71, 814-833.
- Bolt, D., Wollack, J., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, 77, 339–357.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17(4), 665-678.

- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22(1), 69–83. doi:10.1037/met0000106
- Böckenholt, U. & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, 70, 159–181. doi:10.1111/bmsp.12086
- Buckley, J. (2009, June). *Cross-national response styles in international educational assessment: Evidence from PISA 2006*. NCES Conference on the Program for International Student Assessment: What we can learn from PISA, Washington, DC.
- Büyüköztürk, . (2005). *Sosyal Bilimler için veri analizi el kitabı [Data analysis handbook for social sciences]*. 5. baskı. Pagem A Yayıncılık.
- Bybee, R., & McCrae, B. (2007). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education*, 33, 7-26.
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- Culpepper, S. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, 38, 632–644.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48, 1–28.
- De Boeck P & Wilson M (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer-Verlag, New York.
- Harter, S. (1999). *The construction of the self: A developmental perspective*. New York: Guilford Press.
- Heide, M. & Gronhaug, K. (1992) The impact of response styles in surveys: a simulation study. *Journal of the Market Research Society*, 34, 215-231.
- Hofstede, G. H. (2001). *Cultures consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, California: Sage Publications, Inc.
- Hooper, M, Mullis. I. V. S., & Martin, M.O. (2013). TIMSS 2015 Context Questionnaire Framework. Mullis, I.V.S. and Martin, M.O. (Eds.) *TIMSS 2015 Assessment Frameworks*. Retrieved January 15, 2019, from Boston College, TIMSS and PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2015/frameworks.html>
- Hox J. 2002. *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum
- Huang H-Y. (2016) Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers Psychology*, 7(1706), 1-15. doi: 10.3389/fpsyg.2016.01706
- Ilgun Dibek, M., Bulut, O., Sahin Kursad, M., & Yavuz, H. C. (2018, July). *Should students with disabilities have multiple opportunities in answering items?* Paper presented at the International Testing Commission Conference, Montreal, QC, Canada
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior research methods*, 48(3), 1070–1085. doi: 10.3758/s13428-015-0631-y
- Johnson, T.R. (2007). Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika*, 72, 489–504.
- Johnson, R.B. and Christensen, L.B. (2008) *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. 3rd Edition, Sage Publications, Inc., Los Angeles.
- Kadijevich, D. (2008). TIMSS 2003: Relating dimensions of mathematics attitude to mathematics achievement. *Zbornik instituta za Pedagogical Research*, 40(2), 327–346. doi: 1.2298/ZIPI0802327K
- Kidd, C. V. (1959). Basic research: Description versus definition. *Science*, 129, 368-371.

- 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. 5.1-5.175). Retrieved January 8, 2019, from Boston College, TIMSS & PIRLS International Study Center website: <http://timss.bc.edu/publications/timss/2015-methods/chapter-5.html>
- LaRoche, S., Joncas, M., and Foy, P. (2016). Sample Design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 3.1-3.37). Retrieved January 10, 2019, from Boston College, TIMSS & PIRLS International Study Center website: <http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html>
- Leventhal, B.C & Stone, C.A (2018). Bayesian analysis of multidimensional item response theory models: A discussion and illustration of three response style models, *Measurement: Interdisciplinary Research and Perspective*, 16(2), 114-128, doi: [10.1080/15366367.2018.1437306](https://doi.org/10.1080/15366367.2018.1437306)
- Liu, M. (2015). *Response Style and Rating Scales: The Effects of Data Collection Mode, Scale Format, and Acculturation* (Unpublished doctoral dissertation). The University of Michigan.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397-416.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). *Market research: The process, data, and methods using Stata*. Singapore: Springer.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Retrieved January 10, 2019, from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2015/international-results/>
- Nakagawa, S., and H. Schielzeth. 2013. A general and simple method for obtaining  $R^2$  from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2): 133-142. doi: [10.1111/j.2041-210x.2012.00261.x](https://doi.org/10.1111/j.2041-210x.2012.00261.x)
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightman (Eds.), *Measures of Personality and Social Psychological Attitudes* (Vol. 1). San Diego, CA: Academic Press.
- Peterson, R.A, Rhi-Perez, P. & Albaum, G. (2012). A cross-national comparison of extreme response style measures. *International Journal of Market Research*, 56(1), 89-110.
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type Items. *Journal of Educational and Behavioral Statistics*, 38(5), 522-547.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39–55.
- Van de Gaer, E. & Adams, R. (2010, May). *The Modeling of Response Style Bias: An Answer to the Attitude-Achievement Paradox?*, paper presented at the annual conference of the American Educational Research Association, Denver, Colorado, USA.
- Van de gaer, E., Grisay, A., Schulz, W. & Gebhardt, E. (2012). The reference group effect an explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, 43(8), 1205-1228
- Yavuz, H. C., Bulut, O., Ilgun Dibek, M., & Sahin Kursad, M. (2018, July). *Providing revision opportunities in alternate assessments: An application of sequential IRT*. Paper presented at the International Testing Commission Conference, Montreal, QC, Canada.

## **Appendix**

```
library(irtrees)
library(glmertree)
library(reshape)
library(haven)
data <- read_sav("C:/Users/computer/Desktop/data.sav")
View(data)
data<-data.matrix(data)
datamap <- cbind(c(0, 0, 1, 1), c(1, 0, 0, 1))
dataT <- dendrify(data, datamap)
model1 <- glmer(value ~ 0 + item + (1|person) , family = binomial, data = nesrespT, control =
  glmerControl(optimizer = "bobyqa"))
model2 <- glmer(value ~ 0 + item + node + (1 | person) , family = binomial, data = nesrespT,
  control = glmerControl(optimizer = "bobyqa"))
model3 <- glmer(value ~ 0 + item + node + (1+node| person) , family = binomial, data =
  nesrespT, control = glmerControl(optimizer = "bobyqa"))
> anova(model1, model2, model3)
```

## Analyzing the Views of Teachers and Prospective Teachers on Information and Communication Technology via Descriptive Data Mining

Ozge Can Aran<sup>1</sup>, Ahmet Selman Bozkir<sup>2,\*</sup>, Bilge Gok<sup>3</sup>, Esed Yagci<sup>1</sup>

<sup>1</sup> Hacettepe University, Department of Educational Science, Ankara, Turkey

<sup>2</sup> Hacettepe University, Department of Computer Engineering, Ankara, Turkey

<sup>3</sup> Hacettepe University, Department of Primary Education, Ankara, Turkey

### ARTICLE HISTORY

Received: 07 March 2019

Revised: 23 May 2019

Accepted: 20 June 2019

### KEYWORDS

Communication technology,  
Data mining,  
Teachers,  
Prospective teachers,  
Clustering

**Abstract:** This study aims to determine the overt and covert patterns that teachers' and prospective teachers' views on the use of information and communication technology (ICT) instruments contain by using the method of data mining. The study group was composed of 192 prospective teachers attending a state university in Ankara, Turkey and 101 teachers working in Ankara-all of whom took part in the study on the basis of volunteering. Teachers' and prospective teachers' views were obtained by means of a scale. Clustering and association rules - algorithms for data mining - were applied to the data collected, and thus the frequently held patterns for teachers' and prospective teachers' views on ICT instruments were found. Consequently, cluster analysis suggested that prospective teachers considered themselves more competent than teachers in terms of computer skills but that teachers were the group having the most positive views. In addition to this, the results of association rules analysis indicated that the prospective teachers and teachers held the opinion that ICT instruments added variety to the teaching-learning process and ensured students' focusing their attention on lessons, also stated that using ICT instruments would increase students' participation in classes.

## 1. INTRODUCTION

The skill of using technology, one the most important skills of twenty-first century (P21, 2019) is becoming more and more important day by day (Baytekin, 2004). Technology promotes the improvement of students' self-regulation and higher order thinking skills and thus it is considered as an effective instrument forming the basis of student-centered learning (Ha laman, Ku kaya Mumcu, & Usluel, 2007; Kottler & Brookhart Costa, 2009). Also technology is as important as the disciplines of science and mathematics in raising researchers, educators and leaders who can solve problems encountered in daily life and who can think in depth (STEM, 2019). Curricula are of great importance in this respect in raising technology literate individuals who can adapt to the requirements of the age (Yanpar, 2005). Employing technological support in implementing curricula will ensure that many students understand a subject in depth (Can Aran & Senemo lu,

**CONTACT:** Ahmet Selman Bozkir ✉ [selman@cs.hacettepe.edu.tr](mailto:selman@cs.hacettepe.edu.tr) 📍 Hacettepe University, Department of Computer Engineering, Ankara, Turkey

ISSN-e: 2148-7456 / © IJATE 2019

2014) because technology-assisted education makes it possible to enrich learning environments with different activities. A teacher teaching the unit of living things in Life Studies course, for example, can present the geographical and biological properties of living organisms to their students by using sounds, pictures and graphs by means of technology (Yanpar Yelken, 2012). In history classes, on the other hand, students can research the causes of First World War on the internet. Students can also travel in space in technological environment and investigate planets while learning about the solar system. Students studying architecture at university, for instance, can furnish rooms according to the size of the rooms by using technology while decorating the house (Anthony, 2012). Enriching learning environments with technology assistance and with different activities will ensure retention in learning (Akkoyunlu & Yılmaz, 2005; Kürüm, 2016) and thus it will affect students' achievement in positive ways (Yanpar Yelken, 2012).

Using the technology effectively is highly important for technology-assisted education to attain success (Karaman & Kurfallı, 2008; Borich, 2014). Being knowledgeable about technology can be regarded as a part of teaching it. Yet, knowledge of technology on its own is not sufficient to teach technology (Bybee & Loucks-Horsley, 2000; Usluel & A kar, 2003). The reason for this is that teachers need to know first how to use technology and then how to integrate technology into their classes so that they can use technology effectively in their classes (Sert, Kurto lu, Akıncı, & Sefero lu, 2012). It is commonly thought that technology cannot improve badly planned teaching and that teachers' skill in including technology in their teaching is important in order for technology to be influential in teaching (Borich, 2014). In this respect, technological literacy is an important quality for teachers to possess (Ça ıltay, Çakıro lu, Ça ıltay, & Çakıro lu, 2001; Bayazıt & Sefero lu, 2009).

Research demonstrates that teachers who have positive attitudes towards Information and Communication Technology (ICT) integrate it more into their teaching practice (Moseley et al., 1999; Mümtaz, 2000). Additionally, some research studies point out that teachers who have computers and internet connection at home and in the classroom –that is to say, their knowledge of technology use - use technology more (Varner, 2003; Karaman & Kurfallı, 2008). However, other research studies find out that elementary or secondary school teachers do not use technology sufficiently (Ça ıltay et al., 2001; Kurtdede Fidan, 2008). Research has found that teachers' reasons for not using computers included problems in access to computers and low level computer skills (Mümtaz, 2000; Jenson, Lewis, & Smith, 2002; Buabeng-Andoh, 2012).

Efforts to improve physical conditions to support the use of technology are continuing today. Besides, more and more emphasis is laid to technology in curricula. Even if curricula facilitate the use of technology, improving those skills of teachers during pre-service training is very important (Wicklein, 1993; Yanpar, 2005). Therefore, including activities improving those skills in teacher training programmes assures that teachers become well-equipped with knowledge of technology in their training (Bayazıt & Sefero lu, 2009). Prospective teachers' views in relation to ICT are considered important in including ICT in the process of teaching (Can Aran, Derman, & Ya cı, 2016). Review of literature indicates that prospective teachers use information technologies more than teachers do (Sefero lu, Akbıyık, & Bulut, 2008). In comparing computer using skills of new teachers with those of experienced teachers, however, it was found that new teachers felt more comfortable in using computers than experienced teachers and that they used computers more often in lesson preparation (Russell, Bebell, O'Dwyer, & O'Connor, 2003). On examining prospective teachers' attitudes towards computers, studies found that there were positive correlations between their skills in using computers and having a computer (Deniz & Köse, 2003). In addition to that, positive correlations were also found between having a computer and being computer literate (Kıyıcı, 2008). Despite this, Çavu and Gökda (2006) concluded that prospective teachers' levels of using computers were low. Besides, technology is considered a gender-specific issue related to men (Lewis, 1999; Ku kaya Mumcu & Koçak Usluel, 2004;



Çoklar, 2008). However there are also studies in the literature demonstrating that whether or not teachers and prospective teachers use computers in their classes does not cause a difference in terms of gender (Hill Less, 2003; Deniz & Köse, 2003; Gerçek, Köseo lu, Yılmaz, & Soran, 2006; Çoklar, 2008). Considering the differences in the conclusions of research studies, it is given priority to investigating technology in terms of gender as Lewis (1999) emphasized. Therefore, the literature review shows that the use of technology can be explained with lots of variables and that teachers' positive views on the use of technology in teaching-learning processes will make teaching environments more effective. This study aims to describe various characteristics of teachers and prospective teachers and to reveal their views concerning the use of technology in classroom settings. This study analyzes various characteristics and views of teachers and prospective teachers via data mining and reveals the internal patterns along with contributing to the literature.

## 2. METHOD

This study aims to determine the overt and covert patterns that teachers' and prospective teachers' views on the use of information and communication technology (ICT) instruments. Thus, the study used a descriptive method so as to reveal the existing situation.

### 2.1. Population and Sample Selection

The study group was composed of 192 university students attending the educational faculty of a state university in Ankara and 101 teachers working in Ankara selected on a voluntary basis. Teachers' and prospective teachers' views were obtained by means of a questionnaire. Clustering and association rules- algorithms for data mining- were applied to the data collected, and thus the frequently held patterns for teachers' and prospective teachers' views on ICT instruments were found.

### 2.2. Data Collection

In this study, an ICT scale developed by the researchers, was used. The pilot form of the ICT scale had the 24-item and it was conducted to 172 undergraduate students attending various departments of the educational faculty of Hacettepe University. Then, factor analysis was conducted for the data obtained from pilot form so as to test the construct validity of the scale. The data obtained for this purpose were exposed to factor analysis. The Kaiser-Meyer-Olkin (KMO) test value for the fit of the data to the factor analysis was found as 0.80 and Barlett's sphericity test was found as  $\chi^2(105) = 435.70, p < 0.01$ . These results showed that the data fit the factor analysis. 9 items having similar values (Floyd & Widaman, 1995; Tabachnick & Fidell, 2007) and having no loads above 0.30 in two factors were removed from the scale. Having removed the above mentioned 9 items, analyses were repeated so as to decide on the number of factors. Following exploratory factor analysis (EFA), it was found that the scale had two factors with eigenvalue bigger than 1 and that 45.08% of the total variance was explained. In one-factor scales, having 30% or more explained variance can be considered sufficient but in multi-factor scales the explained variance is expected to be higher (Büyüköztürk, 2006). On examining the factor loads in exploratory factor analysis, it was found that the factor loads were between 0.53 and 0.76 in the first factor and that they were between 0.59 and 0.78 in the second factor. An examination of the items showed that the 10 items in the first factor were related to the direct effects of ICT on students' learning while 5 items in the second factor related to the indirect effects of ICT on students' learning. Accordingly, the results of the analysis indicated that the scale was valid and reliable. The developed scale had 15 items in total. Cronbach's Alpha for the ultimate scale was found to be as .76. The items in the scale aiming to discover teachers' and prospective teachers' views on the use of ICT instruments were in 5-point Likert type. Additionally, demographic information was also included in the first part of the scale. This

includes type of participant, gender, having taken a course related with ICT before, having a computer of one's own, computer using skills and length of computer use (daily).

### 2.3. Data Analysis

The methodology of data mining was employed in analyzing the research data. Data mining can be defined as the process of exploring meaningful knowledge within data sets by making use of such methodology as artificial intelligence, statistics and machine learning (Tan, Steinbach, & Kumar, 2006). Data mining processes are divided into two categories as predictive and descriptive (Bozkir, Gok, & Sezer, 2008). In predictive methods, the data labelled as numeric or discrete are divided into training, test and validation groups. The models to be created is trained by utilizing the training data in order to predict unseen test data. Various approaches (such as decision trees, support vector machines, statistics based naïve bayes and hidden Markov models) are suggested in this area in the literature.

Descriptive data mining, on the other hand, aims to uncover the hidden, meaningful and useful patterns in the data. Such methods as clustering and association rules mining come into prominence in the processes of descriptive data mining. The two most important elements distinguishing descriptive data mining from predictive data mining are as in the following: (1) aiming to uncover the hidden patterns in the data instead of predicting, and (2) conducting the processes based on unlabeled data and in unsupervised manner.

This study investigates the characteristics of teachers and prospective teachers and their views on the use of information and communication technology instruments by using clustering and association rules-which are among the methods of descriptive data mining. The analyses were carried out by using SPSS Clementine 12 data mining software. While K-means algorithm was used in clustering, the generalized rule induction (GRI) algorithm was used in finding the association patterns available in views concerning the items in the scale. The individuals answering the questionnaire were divided into 3 groups by means of clustering and the differences between the groups were analyzed; and frequently observed association patterns in teachers' and prospective teachers' views were uncovered through association rules. The analysis methods used in this study are described below.

#### 2.3.1. K-means Clustering

Clustering, in simplest terms, is the process of grouping the elements in a data set according to their qualities. Even though humans manage to do the clustering according to a few qualities, it becomes more and more difficult for them to do this as the number of qualities increase (Bozkir, Gok, & Sezer, 2009). Clustering methods have been used in innumerable areas (computer vision, geoscience, education, etc.) so far. Clustering is a useful method for the discovery of some knowledge from a dataset and an exploratory method for helping to solve classification problems (Kıray, Gok, & Bozkir, 2015).

K-means clustering approach, in particular, is indeed an unsupervised machine learning method and it was suggested by MacQueen (1967). Accordingly,  $k$  amount of examples are selected randomly or with another approach from a sample. The selected examples are regarded as centroids and the other examples are assigned to the relevant sets according to their distance from the first selected centroids. Distance is calculated by considering the qualities that the examples have. Having done the first assigning process, centroids are updated by considering the newly assigned examples. Thus, centroids continuously change at smaller intervals. The process continues until the total change is set to zero.

In K-means clustering method, objective function  $J$  minimises error squares as is seen in Equation (1). Accordingly,  $\|x_i^{(j)} - c_j\|^2$  is described as a distance function,  $x_i^{(j)}$  is described as an

example in the data, and,  $c_j$  is described as  $j$  centroid; the objective function  $J$  can be represented as in Equation (1).

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

The objective in cluster analysis is that sets have high intra-class similarities and low inter-class similarities- independently of the method (Tan, Steinbach and Kumar, 2006). The SPSS Clementine software enables users to construct their models in a visual environment by dragging and dropping appropriate nodes. This has been sometimes referred as visual data mining. In order to apply K-means clustering schema, we first loaded the data as from a Microsoft Excel file into SPSS Clementine 12 software. Following to pre-process stage (i.e. setting data types and filtering out the unnecessary variables) we have picked the K-means module from the toolbox and connected it to “Type” node as shown in Figure 1. For the next stage, we have selected algorithm parameters (i.e. cluster count) and made it run over the source data. Following to the completion of the process, we have analyzed and gathered the results sourced from the clustering study.

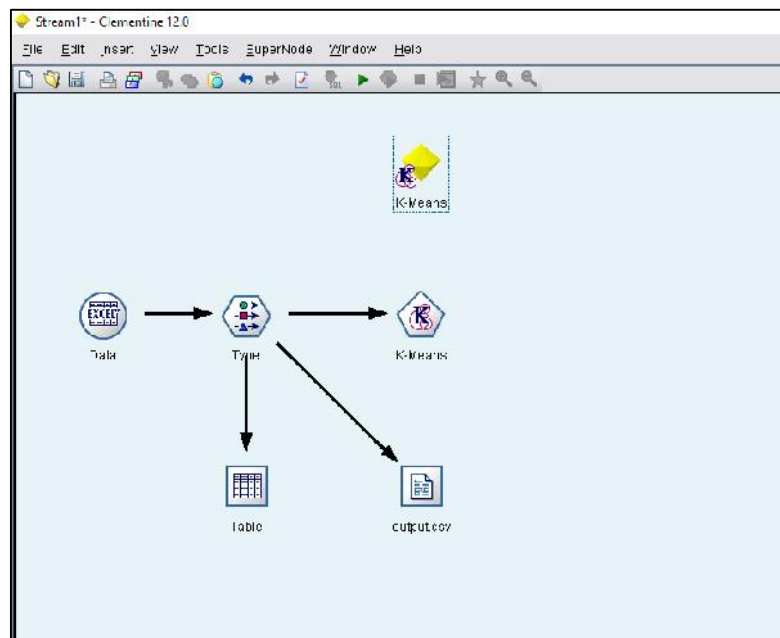


Figure 1. K-means clustering application in SPSS Clementine 12

### 2.3.2. Association rules mining and GRI

Association rules analysis - as a method of descriptive and unsupervised learning-based data mining - is a process of discovering hidden associations seen in activities, situations and observations in a data set according to relation and pattern associations. The method became popular with market basket application revealing what products customers buy in association with other products. Association rules analysis is also used today in such areas as medicine, education and engineering (Bozkir, Gok, & Sezer, 2008). Agrawal and Srikant (1994) recommended Apriori algorithm which is the most commonly known algorithm in this field. Each transaction contains at least one component according to association rules analysis and Apriori algorithm (for instance,  $t_i = \{\text{apples, pears}\}$ ). In this way, each transaction will have one or more than one element. Apriori algorithm works at two stages. First, it determines frequent item sets and prunes the irrelevant item sets in a bottom-up approach. At this stage user should determine the minimum support (*min-sup*) value as a parameter. Second, it calculates the strong rules as the predicate and the consequent on the basis of confidence value which is presented as a parameter.

The process can be represented in mathematical terms as:  $I = \{i_1, i_2, \dots, i_n\}$  and it shows all the elements. On the other hand, all the transactions in the data set are represented as  $T = \{t_1, t_2, \dots, t_n\}$ . In this case, an item set having zero or more than zero element will be a sub-set of  $X$ . Then the number of an item set is stated as  $\Omega(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$ . An association rule with  $X \rightarrow Y$  is described as  $X$  (predicate)  $\rightarrow Y$  (consequent) (Lai & Cerpa, 2001). According to these descriptions, the support value of an association rule in a data containing  $N$  number of transactions is shown in Equation 2 and confidence value is shown in Equation 3. It would be more appropriate to take into account the confidence value rather than the support value in interpreting the usefulness of the rules determined. This is because consequent is calculated according to the frequency of predicate in calculating the confidence value. In other words, posterior probability is taken into consideration here. Yet, there is no such computation throughout the calculation of support value. As a result of this fact, important and useful rules having relatively lower support values can be identified. This situation is illustrated in details in Tables 1 and 2 below.

$$\text{Support, } s(X \rightarrow Y) = \frac{\Omega(X \cup Y)}{N} \tag{2}$$

$$\text{Confidence, } g(X \rightarrow Y) = \frac{\Omega(X \cup Y)}{\Omega(X)} \tag{3}$$

Apriori algorithm functions with discrete data. Since the data set used in the study was numeric, GRI algorithm given in SPSS Clementine 12 data mining software was used. The GRI algorithm was preferred because it adopts Apriori as a principle and because it can also accept numerical data. Table 1 shows the calculations about the support value and the confidence value used.

In order to apply GRI algorithm over the dataset, we first loaded the data from a Microsoft Excel file into SPSS Clementine 12 software. Next, we have picked the GRI algorithm and make it run over the dataset by defining minimum support and confidence scores. Upon the model construction, generated rules were listed according to the confidence value in descending order. At this stage, we have selected the robust rules.

**Table 1.** Support and confidence value computation for a sample dataset

TID	Items	Support = Observation / Total Records	When X → Y, Confidence = Observation (Y) / Observation (X)
1	AB		
2	ABCD	Total Records = 6	Confidence {B → C} = 3 / 5 = 60 %
3	BCD	Support {BC} = 3 / 6 = 50%	Confidence {A → C} = 2 / 4 = 50%
4	BC	Support {AD} = 3 / 6 = 50%	Confidence {AB → C} = 1 / 3 = 33.3%
5	ACD	Support {ABD} = 2 / 6 = 33.3 %	Confidence {AB → D} = 2 / 3 = 66.7%
6	ABD	Support {BCD} = 2 / 6 = 33.3%	

As is seen from Table 1,  $T_i = \{A, B, C, D\}$  and it includes different elements. Six item sets in total are shown to exemplify what has been stated. Different associations of the elements are shown in second column of the table. Accordingly, example association rules are shown in the column on the right hand side. By considering the example item sets in the second column, various support and confidence values were calculated in third and fourth columns respectively. Since the item set {B, C} appeared three times, the support value has been computed as 3/6=50%. However, if one examines the rule “B → C”, it can be easily seen that the item {B}, which was the predicate of the rule, appears only 5 times along with the item {B, C}, which appears only 3

times. Consequently, the confidence values for the rule “B → C” has been calculated as  $3/5=60\%$ . The study considers cluster analysis and association rules analyses together. While the individuals taking part in the questionnaire were divided into 3 groups with the help of clustering and thus the differences were analysed, patterns frequently observed in teachers’ and prospective teachers’ views were revealed via association rules. This study analysed the variables shown in [Table 2](#) and the properties related to those variables in clustering method.

**Table 2.** Properties of variables constituting the data set

Name of the Attribute	Type	Values and Number of Distributions
Type of participant	Discrete	Prospective teachers (191 – 65.6%), teachers(100 - 34.4%)
Gender	Discrete	Male (97 – 33.3%), Female (194 – 66.7%)
Having taken a course related with ICT before	Discrete	Yes (269 – 92.4%), No(22 – 7.6%)
Having a computer of one’s own	Discrete	Yes(268 – 92%), No (23 – 8%)
Computer using skills	Discrete	Very good (52 – 17.8%), Good (132 – 45.3%), Medium (93 – 31.9%), Low (10 – 3.4%), Very low(4 – 1.3%)
Length of computer use (daily)	Discrete	More than 7 hours (13 – 4.4%), 5-6 hours (38 – 13%), 3-4 hours (88 – 30.2%),less than 2 hours (145 – 49.8%), Not any (6 – 2%)
Total positive values in views	Continuous	Min: 26, Max: 72, Mean: 55.1, SD: 7.037

While qualities such as the type of participants, gender, having a computer of one’s own and having taken a course related with ICT before are binary data; length of computer use and computer using skills exist as nominal data. Whereas the mean was 55.1 for total positive values in views, standard deviation was 7.037. In association rules analysis - which was done in addition to cluster analysis - the data collected through a scale containing various views were analysed. Minimum support was found to be 32% and minimum confidence was found to be 85% in GRI analysis-which was done for association rules in this study.

### 3. RESULT and DISCUSSION

The findings obtained in this study are presented under two headings as findings obtained from cluster analysis and findings obtained from association rules analysis. While differences between natural groups formed by considering prospective teachers and teachers together were included in findings obtained from cluster analysis, views held by teachers and prospective teachers were considered separately in findings obtained from association rules analysis.

#### 3.1. Results and Discussion for Clustering Analysis

Cluster analysis was carried out on data collected from teachers and prospective teachers. Values 2, 3 and 4 were tried in determining the clusters regarding k as the number of clusters, and 3- which was observed to be the most discriminative- was selected. The properties of clusters arising as a result of cluster analysis are shown in [Table 3](#).

An examination of [Table 3](#) makes it clear that the biggest cluster in terms of computer using skills and length of computer usage is Cluster 1 with 144 individuals in it- which was followed by Cluster 2 and Cluster 3. If we examine the details of Cluster 1, it was found that the cluster is composed of prospective teachers and that the distribution of the number of female and male participants is in balance. It is clear that Cluster 1 is the most active cluster with percentages of total number of individuals in categories of “good” and “very good” in computer using skills and with percentages of using computers more than 3 hours daily. Accordingly, the individuals in Cluster 1- 94.4% of whom had taken a course in computer- spent longer hours on computer than

the ones in the other clusters. As a result of this, the individuals in Cluster 1 had higher computer using skills than the ones in the other two clusters. This was a finding supportive of the conclusion Sefero lu, Akbıyık and Bulut (2008) reached indicating prospective teachers used information technologies at higher levels than teachers did. An examination of the findings for the clusters demonstrated that the length of computer use had positive correlations with computer using skills. Similar conclusions were reached in studies investigating the use of technology in education (Yanık, 2010).

**Table 3.** Clustering results for teachers and prospective teachers

Clusters	Cluster 1	Cluster 2	Cluster 3
Population	144	67	80
Type of user	Prospective teachers: 144 (100.0%) Teachers:0 (0%)	Prospective teachers: 0 (0%) Teachers 67 (100%)	Prospective teachers: 47 (58.75%) Teachers: 33 (41.25%)
Gender	Female: 75 (52.08%) Male: 69 (47.92%)	Female: 56 (83.58%) Male: 11 (16.4%)	Female: 63 (78.75%) Male: 17 (11.25%)
Length of computer use	None: 4 (2.7%) Less than 2 hours: 42 (29.16 %) 3-4 hours: 61 (42.36%) 5-6 hours:27 (18.75%) More than 7 hours:10 (6.9%)	None: 0 (0%) Less than 2 hours 29 (43.2%) 3-4 hours 26 (38.8%) 5-6 hours:9 (13.43%) More than 7 hours:3 (4.47%)	None: 3 (3.75%) Less than 2 hours: 74 (92.5%) 3-4 hours: 1 (1.25%) 5-6 hours:2 (2.5%) More than 7 hours:0 (0%)
Having a computer	Yes 138 (95.8%) No: 6 (4.2%)	Yes: 65 (97.01%) No: 2 (2.99%)	Yes: 65 (81.25%) No: 12 (18.75%)
Having taken a course related with ICT before	Yes: 136 (94.4%) No: 8 (5.6%)	Yes: 60 (89.55%) No: 7 (10.45%)	Yes: 73 (91.25%) No: 6 (8.75%)
Computer using skills	Very low: 2 (1.38%) Low: 2 (1.38%) Medium: 9 (6.25%) Good: 92 (63.8%) Very good:39 (27.08%)	Very low: 2 (2.9%) Low: 6 (8.9%) Medium: 6 (8.9%) Good: 40 (59.7%) Very good: 13 (19.4%)	Very low: 0 (0%) Low: 2 (2.5%) Medium: 78 (97.5%) Good: 0 (0%) Very good: 0 (%)
View scores in total	Average: 54.97 SD: 7.25	Average: 57.27 SD: 7.36	Average: 53.51 SD: 5.90

As is evident from Table 3, Cluster 2 was composed of teachers only. The participants in this group were all practicing teachers and they were mostly (83.8%) female. 57% of the individuals in this group- who had similarities with Cluster 1 in the length of computer use and in computer using skills- used computers more than 3 hours a day. In addition to that, the majority of them (78.2%) said that they considered themselves to be at medium level or above in terms of computer using skills. Besides, Cluster 2 was found to be the group holding the most positive views about using ICT instruments in the classroom. This was a finding consistent with the finding that teachers developed positive attitudes towards technological developments and educational technologies obtained in Halderman (1992), Ça ıltay et al. (2001) and Kurtdede Fidan (2008). Computer using skills of the individuals in Cluster 2 who were practicing teachers might have positively influenced the views of those individuals about using ICT instruments. The evidence for this may be the differences observed between Cluster 2 and Cluster 3. Although the numerical difference was not big between positive views, positive contributions were observed generally. It was also remarkable that the percentage of having a computer was higher in this cluster- of which the members were practicing teachers. Thus, it was thought that owning a computer might

affect teachers in having positive views about using ICT instruments. This was a result similar to the one found in a study conducted by Deniz and Köse (2003) with the participation of prospective teachers.

Interpreting the findings for Cluster 1 along with the ones for Cluster 2, it was concluded that owning a computer had direct influence on computer using skills and length of computer use. This mirrors the conclusion reached by Kıyıcı (2008) that for prospective teachers possessing a computer is correlated with computer literacy. This current study is also supportive of the conclusions reached by Çavu and Gökda (2006) that the level of computer usage skills of prospective teachers who do not have computer is inadequate, by Ku kaya Mumcu and Koçak Usluel (2004) that the 81,8% of teachers who have own computer stated that they use it and also by Winnans and Brown (1992) that elementary teachers' having own computer or using them for personal reasons at home or school is seen as a factor that affect teachers' use of the computer. First, very few teachers themselves. Cluster 3 containing 80 individuals was the group with the lowest values in terms of computer using skills and the length of computer use. The group was composed of prospective teachers as well as teachers and had shorter length of computer use. The great majority (97.5%) of the individuals in Cluster 3 said they had medium level of computer using skills. They (more than 96%) were also found to spend less than 2 hours using computer daily. Approximately one fifth of the group- of which the majority (81.25%) was female- did not have a computer of their own. Cluster 3- which was well behind the other clusters in terms of information and communication technologies- ranked the last in positive views on using ICT instruments in the classroom.

The clustering study reveals some other interesting findings regarding the relation between gender and computer using skills. If the properties of participant distributions of each cluster are carefully investigated it can be seen that the male participants in especially cluster 1 constitute much larger proportion compared to other clusters. Moreover, regarding the cluster 1 results, it has been found that this cluster involves the highest computer skill abilities when *good* and *very good* scores are summed up. If the proportions of the male and female participants are reviewed, it can be also seen that, the cluster 1 has much larger proportion of men compared to other clusters. Combination of these two findings supports that technology is considered as a gender-specific issue related to men (Lewis, 1999; Ku kaya Mumcu & Koçak Usluel, 2004). One another finding regarding to the length of computer usage duration, the cluster 1 also involves the longest duration if the values are investigated. Furthermore, the cluster 1 has been solely composed of prospective teachers. However, it is noteworthy that, the findings of some of the studies in the literature pointed out that prospective teachers' attitudes about computer usage does not differ in terms of gender (Deniz & Köse, 2003; Gerçek et al., 2006; Çoklar, 2008).

### 3.2. Results and Discussion about Association Rules Mining

The data collected from teachers and prospective teachers was then subjected to the association rules mining. The information written in "Consequent" column is in cause and effect relation with the information written in "Predicate" column along with the "support" and "confidence" scores. Following the analysis, the views held by prospective teachers are shown in Table 4. According to Table 4, 95.52% of the prospective teachers stating that ICT instruments added variety into teaching-learning process and that those instruments ensure that students focused their attention on lessons said that making use of ICT instruments would increase students' participation in classes. This finding was similar to the one obtained by Güngör and A kar (2004). The prospective teachers participating in the above mentioned study said that ICT instruments added variety into classes, as a result, it increased interest and efforts in classes and that classes gained continuity instead of being restricted into class hours. Besides, 90.77% of the participants who stated that ICT instruments ensures that students focused their attention on lessons, that learning with ICT instruments was more effective and that ICT instruments enable to concretize

what students learn in the classroom also, stated that students’ effective use of ICT instruments would influence their achievement in positive ways. This finding was similar to that of Sadi, ekerci, Kurban, Topu, Demirel, Tosun, Demirci, & Gökta (2008).

**Table 4.** Some of the results for association rules for prospective teachers

Predicate	Consequent	Support	Confidence
ICT instruments add variety into teaching-learning processes= I agree, Using ICT instrument in classes ensures that students focus their attention on lessons” = I agree	I think that making use of ICT instruments increases students’ participation in classes = I agree	35.08%	95.52%
Using ICT instrument in classes ensures that students focus their attention on lessons = I agree, ICT instruments ensure that students concretize what they have learnt in the classroom = I agree	Students’ effective use of ICT instruments influences their achievement in positive ways ICT = I agree	34.03%	90.77%
Learning is more effective in schools equipped with ICT= I agree, ICT instruments ensure that students concretize what they have learnt in the classroom = I agree	Students’ effective use of ICT instruments influences their achievement in positive ways ICT = I agree	41.36%	86.08%
ICT instruments ensure that students concretize what they have learnt in the classroom = I agree, View scores in total >52.5	Students’ effective use of ICT instruments influences their achievement in positive way ICT= I agree	35.6%	86.76%
Learning is more effective in schools equipped with ICT= I agree, Using ICT instrument in classes ensures that students focus their attention on lessons = I agree	Students’ effective use of ICT instruments influences their achievement in positive ways ICT= I agree	32.98%	85.71%

The majority of the prospective teachers taking part in the above mentioned study said that use of technology made learning permanent and ensured better comprehension and that it increased the quality of education and motivation in classes. In addition to that, nel, Evrekli and Balım (2011) also reached similar conclusions. Accordingly, the prospective teachers pointed out that using technology would be beneficial in science and technology teaching and that it could also have such effects on students as ensuring audio and visual learning, increasing interest and attention, facilitating learning, concretizing abstract concepts and increasing retention in



learning. Also, the finding that prospective teachers' views on the use of technology are positive in general is also available in the literature. Following the interviews with prospective teachers Yavuz and Co kun (2008) and Yılmaz, Ulucan and Pehlivan (2010) found that prospective teachers held positive views on the use of technology; and Usta and Korkmaz (2010) found that they had positive perception in this respect and Özgen and Obay (2008) found that they had positive attitudes. In addition to the finding that prospective teachers had positive views on the use of ICT, the data collected from teachers were put to association rules analysis. Following the analysis, the teachers' views are shown in Table 5.

**Table 5.** Some of the results for association rules for teachers

Predicate	Consequent	Support	Confidence
ICT instruments add variety into teaching-learning processes = I agree, View scores in total > 50.5	I think that making use of ICT instruments increases students' participation in classes = I totally agree	32%	100%
Using ICT instrument in classes ensures that students focus their attention on lessons = I totally agree	I think that making use of ICT instruments increases students' participation in classes = I totally agree	32%	96.88%
Using ICT instrument in classes ensures that students focus their attention on lessons = I agree	Students' effective use of ICT instruments influences their achievement in positive ways = I agree	33%	93.94%
ICT instruments add variety into teaching-learning processes = I agree, Using ICT instrument in classes ensures that students focus their attention on lessons = I agree, I think that making use of ICT instruments increases students' participation in classes = I agree	Learning is more effective in schools equipped with ICT = I agree	33%	90.91%
ICT instruments add variety into teaching-learning processes = I agree	Using ICT instrument in classes ensures that students focus their attention on lessons = I agree	41%	85.37%

According to Table 5, teachers stating that ICT instruments added variety into the teaching-learning processes also said that making use of ICT instruments would increase students' participation in classes. This was a finding parallel to the one obtained in Ertmer et al (2012). Accordingly, the teachers participating in the study said that technology enriched curriculum and thus more active student participation was ensured. Tekinarslan (2007) also pointed out that using technology in teaching environments would result in richness and thus would ensure students to participate more actively in classes. 96.88% of teachers stating that using ICT instruments ensured that students focused their attention on lessons also said that making use of ICT instruments would increase students' participation in classes. This finding was supportive of the one obtained in Kurtdele Fidan (2008) because the study concluded that education performed by using equipment and aids increased students' interest in classes and that it ensured active participation. In addition to this finding of research, 93.44% of the teachers stating that ICT instruments enabled students to focus their attention on lessons and that they concretized what students have learnt in classes also said that students' effective use of ICT instruments would influence their achievement in positive ways. The views that ICT enables students to focus their attention on lessons and thus increase achievement (Gu, 2017), that it increases achievement by increasing participation in classes (Yanpar Yelken, 2012), that it affects achievement in positive

ways by giving the impression that one uses real equipment (Çelik, 2007) which are reported in the literature are all in parallel to this finding of this study. Apart from that, teachers who had taken part in the study conducted by Üredi, Akba lı and Ulum (2016) also pointed out that ICT concretized what students had learned, that it attracted students' attention and that it brought about effective and permanent learning. At the same time, 90.91% of the teachers thinking that ICT instruments added variety into teaching-learning processes, that they enabled students to focus their attention on lessons and that they would increase students' participation in classes also stated that learning was more effective in schools equipped with ICT. Studies available in the literature suggest that ICT makes learning effective by enriching the teaching-learning process (Baytekin, 2004; Yanpar, Yelken 2012; Üredi, Akba lı & Ulum, 2016). Additionally, the study of Kurtdede Fidan (2008) made out that teachers believed that using equipment and aids in teaching learning process sparked students' interest towards the lesson and ensure both students' learning while having fun and participating more actively. Also teachers participated to the the same study stated that lessons were more effective and efficient in this way. Furthermore, having stated that ICT instruments added variety into the teaching-learning processes, teachers also stated that ICT instruments ensured that students focused their attention on lessons. Russell, Bebell, O'Dwyer and O'Connor (2003) pointed out that experienced teachers used technology to attract their students' attention.

#### **4. CONCLUSION and RECOMMENDATIONS**

Teaching environments in which information and communication technologies are used, aim to put students into the center and to actualize learning in the best way. Concerning the variables for teachers and prospective teachers and concerning their views on the use of ICT instruments, cluster analysis and association rules mining have significant advantages for exploring the patterns and relations hidden in raw data. Thus, 3 clusters were distinguished with cluster analysis. Views on the use of ICT instruments and many other properties (such as having a computer of one's own, gender, etc.) that those views could influence were analyzed in those clusters. The cluster having the highest level of computer using skills and the length of daily computer use was Cluster 1, which was composed of only prospective teachers. Considering the same properties, Cluster 2 followed Cluster 1, and Cluster 3- which was composed of teachers and prospective teachers- ranked the last. In this context, prospective teachers considered themselves as being more proficient than practicing teachers in computer using skills. In addition to that, teachers were found to be the group having the most positive views on the use of ICT instruments in the classroom setting. On interpreting the findings concerning teachers along with the findings concerning prospective teachers, it was concluded that having a computer of one's own had linear correlations with computer using skills and with the length of computer use. It was also found in cluster analysis that the number of male prospective teachers having computer using skills is much more than the female prospective teachers' according to the properties of cluster 1 when compared to other clusters.

The data collected in relation to the views of teachers and prospective teachers were given to association rules analysis in addition to cluster analysis. The results of association rules analysis indicated that ICT instruments added variety into the teaching-learning process, that they helped students to focus their attention on lessons and that making use of ICT instruments would also increase students' participation in classes. It was remarkable that the rules having high confidence values (95% and above) obtained from both teachers and prospective teachers touched on the same points. Motivation to use ICT instruments in classrooms can be raised by emphasizing those properties-which teachers and prospective teachers stressed- in teacher training and in in-service training.

## Acknowledgement

This manuscript is the extended version of the papers presented in 4th International Eurasian Educational Research Congress and 13 th International Conference on ICT in The Education of the Balkan Countries.

## ORCID

Ozge Can Aran  <https://orcid.org/0000-0003-3229-4325>

Ahmet Selman Bozkir  <http://orcid.org/0000-0003-4305-7800>

Bilge Gök  <https://orcid.org/0000-0002-1548-164X>

Esed Yagci  <https://orcid.org/0000-0002-5418-1172>

## 5. REFERENCES

- Agrawal, R., & Srikant. R. (1994). *Fast algorithms for mining association rules in large databases*. Paper presented at 20th International Conference on Very Large Data Bases, San Fransisco.
- Akkoyunlu, B., & Yılmaz, M. (2005). Türetimci çoklu ortam ö renme kuramı [Generative theory of multimedia learning]. *Hacettepe University Journal of Education*, 28, pp.9-18.
- Bayazıt, A., & Sefero lu. S. S. (2009). *Türkiye'deki teknoloji politikalarında e itimin yeri ve ö retmen yeti tirme politikaları*. 12. Bili im Teknolojileri I 1 nda E itim Kongresi (BTIE'2009) Bildiriler Kitabı. Ankara: Türkiye Bili im Derne i.
- Baytekin, Ç. (2004). *Ö renme-ö retme teknikleri ve materyal geli tirme*. Ankara: Anı Yayıncılık.
- Borich, G. D. (2014). *Etkili ö retim yöntemleri*. (B. Acat, Trans.). Ankara: Nobel Yayınları.
- Bozkir, A. S., Gök, B., & Sezer, E. (2008). *Üniversite ö rencilerinin interneti e itimsel amaçlar için kullanmalarını etkileyen faktörlerin veri madencili i yöntemleriyle tespiti*. Paper presented at Bilimde Modern Yöntemler Sempozyumu, Osmangazi Üniversitesi, Eski ehir.
- Bozkir, A. S., Gök, B. & Sezer, E. (2009). *Determination of the factors influencing student's success in student selection examination (OSS) via data mining techniques*. Paper presented at 5<sup>th</sup> Uluslararası leri Teknolojiler Sempozyumu, Karabük.
- Buabeng-Andoh, G. (2012). Factors influencing teachers' adoption and integration of information and communication technology into teaching: A review of the literature. *International Journal of Education and Development Using Information and Communication Technology (IJEDICT)*, 8(1), pp.136-155.
- Büyüköztürk, . (2006). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem A Yayıncılık.
- Bybee, R. W., & Loucks-Horsley, S. (2000). Advancing technology education: The role of professional development, *The Technology Teacher*, 60(2), pp.31-34.
- Can Aran, Ö. & Senemo lu, N. (2014). Disiplinli zihin özellikleri açısından fen e itiminin incelenmesi [An investigation of science education in terms of disciplined mind characteristics]. *Hacettepe University Journal of Education*, 29(4), pp.46-59.
- Can Aran, Ö., Derman, ., & Ya cı, E. (2016). Pre-service science and mathematics teachers' thoughts about technology. *Universal Journal of Educational Research*, 4(3), pp.501-510.
- Ça ıltay, K., Çakıro lu, J., Ça ıltay, N. & Çakıro lu, E. (2001). Ö retimde bilgisayar kullanımına ili kin ö retmen görü leri [Teachers' perspectives about the use of computers in education]. *Hacettepe University Journal of Education*, 21, pp.19-18.
- Çavuş, H. & Gökda , . (2006). E itim fakültesi'nde ö renim gören ö rencilerin internetten yararlanma nedenleri ve kazanımları [Education faculty students' benefiting reasons from internet and their gains]. *Yuzuncu Yil University Journal of Education*, 3(2), pp. 56-78.
- Çelik, L. (2007). *Ö retim materyallerinin hazırlanması ve seçimi*. In Ö retim teknolojileri ve materyal tasarımı edited by Özcan Demirel ve Eralp Altun. Ankara: Pegem Yayınları.

- Çoklar, A. N. (2008). *Ö retmen adaylarının e itim teknolojisi standartları ile ilgili öz yeterliklerinin belirlenmesi* [Assessing the self-efficacy of teacher candidates concerning the educational technology standards]. Doctoral dissertation, Anadolu University, Eski ehir.
- Deniz, L & Köse, H. (2003). Ö retmen adaylarının bilgisayar ya antıları ve bilgisayar tutumları arasındaki ili kiler [The relationship between computer attitudes and computer experiences of student teachers]. *Marmara Üniversitesi Atatürk E itim Fakültesi E itim Bilimleri Dergisi / Journal of Educational Sciences*, 18, pp. 39-64.
- Ertmer, P. A, Ottenbreit-Leftwich, A., Olgun, S., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers & Education*, 59(2), pp. 423-435.
- Floyd, F. J., & Widaman. K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), pp. 286-199.
- Gerçek, C., Köseo lu, P., Yılmaz, M. & Soran, H. (2006). Ö retmen adaylarının bilgisayar kullanımına yönelik tutumlarının çe itli de i kenler açısından incelenmesi [An analyses of the attitudes of teacher candidates towards computer use]. *Hacettepe University Journal of Education*, 30, pp.130-139.
- Güngör, C., & A kar, P. (2004). E-ö renmenin ve bili sel stilin ba arı ve internet öz-yeterlik algısı üzerindeki etkisi [The effects e-learning and cognitive style on achievement and perceived internet self-efficacy]. *Hacettepe University Journal of Education*, 27, pp. 116-125.
- Gu, P. (2017). *Promoting students' motivation and use of srl strategies in online mathematics learning*. Doctoral dissertation, University of Kansas.
- Halderman, C. F. (1992). *Design and evaluation of staff development program for technology in small schools*. Doctoral dissertation, University of North Texas.
- Haslamani, T., Mumcu, F. K. & Usluel, Y. K. (2007). The integration of information and communication technologies in learning and teaching process: A lesson plan example. *E itim ve Bilim*, 32(146), pp.54.
- Hill Less, K. (2003). *Faculty adoption of computer technology for instruction in North Carolina community college system*. Doctoral dissertation, East Tennessee State University.
- nel, D., Evrekli, E., & Balım, A. G. (2011). Ö retmen adaylarının fen ve teknoloji dersinde e itim teknolojilerinin kullanılmasına ili kin görü leri [Views of science student teachers about the use of educational technologies in science and technology course]. *Journal of Theoretical Educational Sciences*, 4(2), pp.128-150.
- Jenson, J., Lewis, B., & Smith, R. (2002). No one way: Working models for teachers' professional development, *Journal of Technology and Teacher Education*, 10(4), pp.481-496.
- Karaman, M. K., & Kurfalı. H. (2008). Sınıf ö retmenlerinin bilgi ve ileti im teknolojilerini ö retim amaçlı kullanım düzeyleri [Elementary school teacher's ICT usage level for instructional purposes]. *Journal of Theoretical Educational Sciences*, 1(2), pp.43-56.
- Kıray, S. A., Gök, B., & Bozkır, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health (JESEH)*, 1(1), pp.28-48.
- Kıyıcı, M. (2008). *Ö retmen adaylarının sayısal okuryazarlık düzeylerinin belirlenmesi* [Identifying digital literacy level of teachers candidates]. Doctoral dissertation, Anadolu University, Eski ehir.
- Kottler, E., & Brookhart Costa, V. (2009). Integrate technology to enrich learning. In *Secrets to success for science teachers* (181-202). Corwin: Thousand Oaks.

- Kurtdede Fidan, N. (2008). İlkö retimde araç gereç kullanımına ili kin ö retmen görü leri [Teachers' views with regard to the use of tools and materials in the primary level]. *Journal of Theoretical Educational Sciences*, 1(1), pp.48-61.
- Ku kaya Mumcu, F., & Koçak Usluel, Y. (2004). Mesleki ve teknik okul ö retmenlerinin bilgisayar kullanımları ve engeller [Use of computers by vocational and technical schools' teachers and obstacles]. *Hacettepe University Journal of Education*, 26, pp.91-99.
- Kürüm, D. (2016). Ö retim Materyallerinin De erlendirilmesi. In *Ö retim teknolojileri ve materyal tasarımı*, edited by Kıymet Selvi. Ankara: Anı Yayıncılık.
- Lai, K., & Cerpa, N. (2001). *Support vs confidence in association rule algorithms*. Paper presented at Conference on the Chilean Operations Research Society, Chile.
- Lewis, T. (1999). Research in technology education-some areas of need. *Journal of Technology Education*, 10(2), pp.41-56.
- MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations*. Paper presented at 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability, Berkeley.
- Moseley, D., Higgins, S., Bramald, R., Hardman, F., Miller, J., Mroz, M., Stout, J. (1999) *Ways Forward with ICT: effective pedagogy using information and communications technology for literacy and numeracy in primary schools* (Report No: ED458652). United Kingdom: Durham Univ. (England). Curriculum, Evaluation, and Management Centre. Retrieved from <https://files.eric.ed.gov/fulltext/ED458652.pdf>
- Mumtaz, S. (2000). Factors affecting teachers' use of information and communications technology: A review of the literature. *Journal of Information Technology for Teacher Education*, 9(3), pp.319-342.
- Özgen, K. & Obay, M. (2008). *Ortaö retim matematik ö retmen adaylarının e itim teknolojisine ili kin tutumları*. Paper presented at international educational technology conference (IETC), Anadolu Üniversitesi, Eski ehir.
- P21 Partnership for 21st Century Learning. Available online: <http://www.battelleforkids.org/networks/p21/frameworks-resources> (accessed on 15 June 2019).
- Russell, M., Bebell, D., O'Dwyer, L., & O'Connor, K. (2003). Examining teacher technology use: Implications for preservice and inservice teacher preparation. *Journal of Teacher Education*, 54(4), pp. 97-310.
- Sadi, S., ekerçi, A. R., Kurban, B., Topu, F. B., Demirel, T., Tosun, C., Demirci, T. & Gökta , Y. (2008). Ö retmen e itiminde teknolojinin etkin kullanımı: Ö retim elemanları ve ö retmen adaylarının görü leri [Effective technology use in teacher education: The views of faculty members and preservice teachers]. *International Journal of Informatics Technologies*, 1(3), pp. 43-49.
- Sefero lu, S. S., Akbıyık, C. & Bulut, M. (2008). Elementary school teachers' and teacher candidates' opinions about computer use in learning/teaching process. *Hacettepe University Journal of Education*, 35, pp.273-283.
- Sert, G., Kurtoglu, M., Akıncı, A., & Seferoglu, S. S. (2012). Overview of research on teachers' technology usage: a content analysis study. *Computers & Education*, 14(46), pp.1-8.
- STEM (Science, Technology, Engineering and Math) (2019, May). Science, Technology, Engineering and Math: Education for Global Leadership. Retrieved from <http://www.ed.gov/stem>.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5<sup>th</sup> ed.). New York: Allyn and Bacon.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Addison Wesley.

- Tekinarslan, E. (2007). E itimde internet kullanımı. In *Ö retim teknolojileri ve materyal tasarımı*, edited by Özcan Demirel and Eralp Altun. Ankara: Pegem Yayınları.
- Tosun, N. (2006). *Bilgisayar destekli ve bilgisayar temelli ö retim yöntemlerinin, ö rencilerin bilgisayar dersi ba arısı ve bilgisayar kullanım tutumlarına etkisi: Trakya Üniversitesi E itim Fakültesi ö rne i [The effects of computer assisted and computer based methods on the success of the students and the attitude of using computer in computer classes: A sample study at Education Faculty of Trakya"]*. Master thesis, Trakya University, Edirne.
- Usluel, Y. K., & A kar, P. (2003). Teachers' stages at the innovation-decision process related to the use of computers: Changes in two years. *Hacettepe University Journal of Education*, 24, pp.119-128.
- Usta, E., & Korkmaz, Ö. (2010). Ö retmen adaylarının bilgisayar yeterlikleri ve teknoloji kullanımına ili kin algıları ile ö retmenlik mesle ine yönelik tutumları [Pre-service teachers' computer competencies, perception of technology use and attitudes toward teaching career]. *Journal of Human Sciences*, 7(1), pp.1335-1349.
- Üredi, L., Akba lı, S. & Ulum, H. (2016). Investigating the primary school teachers' perspectives on the use of education platforms in teaching. *Educational Research and Reviews*, 11(15), pp.1432-1439.
- Varner, S. V. (2003). Attitudes and perceptions of secondary language arts teachers towards computer technology and its use in curriculum and instruction. Doctoral dissertation, University of Alabama, Alabama.
- Wicklein, R. C. (1993). Identifying critical issues and problems in technology education using a modified-delphi technique. *Journal of Technology Education*, 5(1), pp.54-71.
- Winnans, C., & Brown, D. S. (1992). Some factors affecting elementary teachers' use of the computer. *Computers & Education*, 18(4), pp.301-309.
- Yanık, C. (2010). Azeri ö retmen adaylarının bilgisayar okuryazarlık algıları ve internet kullanımına yönelik tutumları. In *Uluslararası Ö retmen Yeti tirme Politikaları ve Sorunları Sempozyumu II Bildiri Kitabı* (pp.191-201), Azerbaycan Devlet Pedagoji Üniversitesi, Bakü.
- Yanpar Yelken, T. (2012). *Ö retim teknolojileri ve materyal tasarımı*. Ankara: Anı Yayıncılık.
- Yanpar, T. (2005). *Ö retim teknolojileri ve materyal geli tirme*. Ankara: Anı Yayıncılık.
- Yavuz, S., & Co kun, E. A. (2008). Sınıf ö retmenli i ö rencilerinin e itimde teknoloji kullanımına ili kin tutum ve dü ünceleri [Attitudes and perceptions of elementary teaching through the use of technology in education]. *Hacettepe University Journal of Education*, 34, pp.276-286.
- Yılmaz, ., Ulucan, H. & Pehlivan, S. (2010). The attitudes and thoughts of the students attending physical education teaching program about using technology in education. *KEFAD*, 11(1), pp.105-118.