International Journal of Assessment Tools in Education
2019, Vol. 6, No. 2, 300–313
https://dx.doi.org/10.21449/ijate.534118

Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

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

Munevver Ilgun Dibek [1*]

[1] 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 (IRTree) 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 Department of Educational Sciences, TED University, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

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 ocur. One the other hand, if people are more prone to choose negative extreme response categoy 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 ontained from them (Peterson, Rhi-Perez, & Albaum, 2012). Also, within-country correlations are affect 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, dependincies 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 (IRTree) 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) IRTree models are more flexible and informative, which helps them to solve problems that are not fixed by using other approaches, and (ii) IRTree 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, θ_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 θ_{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.

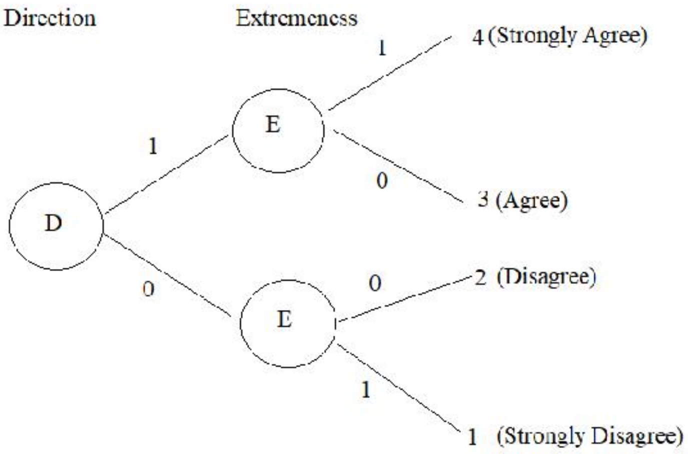

Figure 1. IRTree Model for a four-category item

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:

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

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

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:

$$P(D_1=1|\ \theta_{\mathrm{ERS}}, \theta_1) = \frac{1}{1+e^{-(b_2+a_2\theta_E \quad \mp \mathrm{va2}(b_1+a_1\theta_1))}} \text{ and} \tag{3}$$

$$P(D_1=0|\ \theta_{\mathrm{ERS}}, \theta_1) = 1- \frac{1}{1+e^{-(b_2+a_2\theta_E \quad \mp \mathrm{va2}(b_1+a_1\theta_1))}} \tag{4}$$

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 ν parameter is used to represent compensatory characteristics of the two traits. This shift term, $\nu\ _2(b_1+a_1\theta_1)$, 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 ν 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 θ_1 and θ_{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 comprasion studies occur. In addition, there are several the handicaps of different methods to determine the effect of ERS. Therfore, 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 phenemonen (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 paradoxial 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
International Average	**500**	**38**	**43**	**13**
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	$\left(1-\frac{1}{1+e^{-(b_1+a_1\theta_1)}}\right)\left(\frac{1}{1+e^{-(b_2+a_2\theta_E\quad\mp(b_1+a_1\theta_1))}}\right)$
2	0	0	$\left(1-\frac{1}{1+e^{-(b_1+a_1\theta_1)}}\right)\left(1-\frac{1}{1+e^{-(b_2+a_2\theta_E\quad\mp(b_1+a_1\theta_1))}}\right)$
3	1	0	$\left(\frac{1}{1+e^{-(b_1+a_1\theta_1)}}\right)\left(1-\frac{1}{1+e^{-(b_2+a_2\theta_E\quad\mp(b_1+a_1\theta_1))}}\right)$
4	1	1	$\left(\frac{1}{1+e^{-(b_1+a_1\theta_1)}}\right)\left(\frac{1}{1+e^{-(b_2+a_2\theta_E\quad\mp(b_1+a_1\theta_1))}}\right)$

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 χ^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
	Model3	**138874.30**	**139001.90**	**-69424.20**	**138848.30**	**13**
Self-confidence	Model 1	170379.80	170477.90	-85179.90	170360	10
	Model 2	167464.80	167572.70	-83721.40	167443	11
	Model3	**153184.30**	**153311.90**	**-76579.20**	**153158**	**13**
Value	Model 1	151333.40	151431.60	-75656.70	151313	10
	Model 2	136347.30	136455.20	-68162.60	136325	11
	Model3	**130778.20**	**130905.80**	**-65376.10**	**130752**	**13**

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 χ^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 χ^2 belonging to the Model 1 and Model 2 were evaluated first. Since the calculated value (χ^2 = 81726.90-80293.3= 1433.60) is greater than the table value (χ^2(1; .001) = 10.83), the difference between -2 log χ^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 (χ^2 = 80293.3- 69424.2= 10869.10) is greater than the table value (χ^2(2; .001) = 13.82), the difference between -2 log χ^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**		
τ_{00} person		6.30	τ_{00} person		3.49	τ_{00} person		3.38
τ_{11} person.node2		9.12	τ_{11} person.node2		5.52	τ_{11} person.node2		3.15
ρ_{01} person		-.71	ρ_{01} person		-.66	ρ_{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 ($\rho_{01} = -.71$, $\rho_{01} = -.66$, $\rho_{01} = -.38$, 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. Expecially, 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 catched 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 respose 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 mid-point 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, Ş. (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: Guildford 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 application*s. 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

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. *App lied 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

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)
```