International Journal of Progressive Education, Volume 17 Number 2, 2021
© 2021 INASED

Modelling of the Attitude-Achievement Paradox in TIMSS 2015 with respect to the Extreme Response Style Using Multidimensional Item Response Theory*

Munevver Ilgun Dibek [i]
TED University

Rahime Nukhet Cıkrıkcı [ii]
Istanbul Aydin University

Abstract

This study aims to first investigate the effect of the extreme response style (ERS) which could lead to an attitude-achievement paradox among the countries participating in the Trends in International Mathematics and Science Study (TIMSS 2015), and then to determine the individual- and country-level relationships between attitude and achievement by adjusting the effect of ERS. For the sample of this correlational study, 500 students were randomly selected from each of the 15 countries that participated in TIMSS 2015. The differences in the ERS tendency of the countries were determined by performing MANOVA. To determine the effect of ERS, two different multidimensional item response theory (MIRT) models were used: one did not include the ERS trait as a dimension while the other included this trait as a dimension. The results were analyzed with Latent GOLD 5.1 and WinBUGS software. To determine the relationship between attitudinal variables and achievement, the correlation values based on the observed scores and MIRT models were obtained. Whether there was any significant difference between these correlation values was determined by Fisher's rz transformation. The findings of this study were as follows: (a) the model in which the ERS trait was included as a dimension best fit the data and (b) the correlation values based on the observed scores were negative and those based on the MIRT models were positive, with the two statistically differing from each other. ERS is one of the factors causing the achievement-attitude paradox; however, it not sufficient to explain this paradox.

Keywords: Multidimensional Item Response Theory, TIMSS, Attitude-Achievement Paradox, Extreme Response Style

DOI: 10.29329/ijpe.2021.332.12

----

* This paper was produced from the Ph. D. thesis study of the first author.

[i] **Munevver Ilgun Dibek,** Assist. Prof. Dr., Faculty of Education, TED University, ORCID: 0000-0002-7098-0118

Correspondence: munevver.ilgun@tedu.edu.tr

[ii] **Rahime Nukhet Cıkrıkcı,** Prof. Dr., Faculty of Science and Literature, Istanbul Aydin University

INTRODUCTION

In addition to providing cognitive data related to science and mathematics achievement, the Trends in International Mathematics and Science Study (TIMSS) also provides data on non-cognitive variables (attitude, perception, interest) through questionnaires in different formats presented to students, teachers, parents, and school administrators. There are a large number of international survey studies and publications that used the results of the TIMSS applications, in which attitudes were addressed as one of the non-cognitive constructs and the differences in attitudes associated with the achievement of the students were explored in international comparisons. These studies also shed light on the differences in attitudes and values between different cultures.

Some of the international comparison studies (Kadijevich, 2008; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005) investigating the relationship between attitudes and achievement indicate that students with high success in a subject matter tend to exhibit positive attitudes toward the related course, while others (Buckley, 2009; Van de Gaer & Adams, 2010) emphasize that students' attitudes toward a course are negative despite their high achievement. A negative relationship between attitude and achievement has been observed in international comparison studies using the data obtained from TIMSS, International Student Assessment Program (PISA), and the Progress in International Reading Literacy Study (PIRLS). As opposed to motivation theories, such as expectancy-value theory (Atkinson, 1957), which emphasize that there is a positive relationship between attitude and success, the direction of this relationship varies according to the individual or group level. In other words, there may be a positive relationship between students’ attitudes and achievement within one country while a negative relationship may be observed between different countries (Van de gaer, Grisay, Schulz & Gebhardt, 2012). Therefore, the use of the correlations obtained at the individual and group levels interchangeably decreases the validity of the results obtained from research (Robinson, 1950).

In the literature, the fact that the relationship between attitude and achievement is positive within a country and negative between different countries is defined as an achievement-attitude paradox (Van de et al., 2012). One of the main reasons for this paradox is the possible response style differences between countries (Buckley, 2009). Since Cronbach developed the concept of response style 1941, this style has become the focus of much attention, in terms of the "tendency to systematically respond to the questionnaire items regardless of the content of the item" (Paulhus, 1991, p.17). The response style of individuals causes various psychometric problems in the data (Bolt & Newton, 2011). More specifically, it reduces the validity of test scores by causing systematic errors in the test scores of individuals with the same level of knowledge or similar attitudes or personality traits, leading to differences in scores (Cronbach, 1946). Conversely, individuals with the same raw score are considered to be at the same level in terms of the substantive trait to be measured. For example, assuming that two respondents (A and B) respond to eight items measuring extraversion, and these items have five response categories ranging from 0 to 4. If respondent A chooses the extreme responses of each item and displays a response pattern of 0-4-0-4-0-4-0-4, and respondent B chooses only the midpoint of response categories for each item and displays a response pattern of 2-2-2-2-2-2-2-2, these two respondents would obtain the same raw score of 16. If the raw score alone is taken into consideration, both A and B would receive the same level of extraversion. However, this raises the question of whether these two scores really mean the same. Do they both represent the level of extraversion, or rather do they indicate two different response styles (an extreme and a midpoint response style)?

Response styles threaten the validity of research results in two ways (Baumgartner and Steenkamp, 2001). The first is by affecting the univariate distribution of the variable discussed in the research. Specifically, it causes errors in the mean and variance of the scale scores. Therefore, when the effect of response style is not controlled, false results may be obtained from comparative tests, such as the t-test and F-test (Cheung & Rensvold, 2000). When this situation is considered especially in the context of international comparative studies, the differences in the substantive trait can be misinterpreted due to the differences in the countries' own response styles. Therefore, studies that do not take bias into account due to response styles in cross-cultural studies may make biased inferences about the mean difference in attitude scores between groups (Harumi, 2011). In other words, the differences observed in attitude scores can be interpreted as differences in the measured substantive trait although there are differences in response styles. Another situation in which response styles threaten validity is related to their effect on multivariate distribution. In particular, they lead to obtaining biased results from Cronbach's alpha, regression analysis, factor analysis, and structural equation modeling by distorting the correlation between variables (Reynolds & Smith, 2010).

In the literature, it has been stated that there are various response styles that negatively affect the validity of test scores. This study investigated the effects of the extreme response style (ERS), in which the tendency is to respond at one of the two extremes of the response scale (Baumgartner & Steenkamp, 2001). The reasons for examining ERS were: (i) lack of response categories due to the four-point Likert type of scales used to measure students' attitudes in TIMSS 2015 applications, (ii) the variance in individuals' responses being explained more than the variance in other response styles (De Jong, Steenkamp, Fox & Baumgartner, 2008), (iii) the effect of ERS being more reliably controlled by various methods (Engle, 2016), (iv) frequent expression of differences in exhibiting extreme response styles in intercultural comparisons, and (v) differences in educational achievement in relation to ERS (Lu & Bolt, 2015).

In the literature, the frequently recommended methods to determine the effect of ERS are based on Ad Hoc Extreme Response Style Measures and latent trait models. In the former, the ERS index is formed by using the frequency of responses of individuals in the extreme categories of the response scale or the standard error of item scores. The literature contains research that involves constructing an ERS index, determining whether there is a difference between the variables by performing variance analysis (Bachman, O'Malley & Freedman-Doan, 2010; İlgün Dibek, Yavuz & Çokluk Bökeoğlu, 2018), and examining the relationship between individual-level response styles and country-level characteristics using regression models (Harzing, 2006) and hierarchical linear models (Baumgartner & Steenkamp, 2001; Johnson, Kulesa, Llc, Cho, & Shavitt, 2005).

Research studies using classical methods (ad-hoc ERS measures, regression, or hierarchical linear models) have assumed a linear effect of response styles on bias in individuals' scale scores. In addition, these methods are inadequate to determine the extent to which the different levels of the substantive trait to be measured correlate with the created response style indices, and whether the actual measured property of the responses given at the ends is related to the end limits of the substantive trait itself or response styles. In addition to these classical methods, various families of stochastic models have been proposed to measure response styles in psychological assessments with Likert response formats. In research using latent trait models, it is not necessary to create an ERS index based on descriptive statistics. It has been seen that some of these studies (Cheung & Rensvold, 2000) perform confirmatory factor analysis (CFA) and some perform latent class analysis (LCA). However, since CFA models adopts a linear approach to the effects of response styles on responses to items, it is insufficient in determining non-linear effects. Therefore, while it can be used to study acquiescence response styles (Billiet and McClendon, 2000), it is not the preferred method of determining ERS. In the literature, when the studies examining the response styles with latent class analysis are examined, it is seen that they take the measurement level of ERS categorically. Some of these studies (Moors, 2004; van Rosmalen, van Herk, & Groenen, 2010) perform latent class multinominal logistic models while others (Eid & Rauber, 2000) analyze ERS with latent class mixed models. However, there are several limitations to the use of LCA in determining response styles. First, the response style is considered as a categorical variable; however, in the psychology literature, response styles are often seen as a continuous variable (Greenleaf, 1992a; Prediger, 1999). Another limitation is that the scores cannot be purified from the effect of the response style. Therefore, it is not a useful method for correcting the ERS bias on the substantive trait of measurement (Bolt & Johnson, 2009).

In determining response styles, the Multidimensional Item Response Theory (MIRT) model is used because of the limitation that the previously proposed methods cannot account for the simultaneous influence of the substantive trait and the response style on the selection of response categories. As one of the latent trait models, MIRT is used to model the relationship between two or more latent variables and the response to an item (Reckase, 2009). In this context, MIRT evaluates the response style statistically as another psychological dimension which has an effect on the individual's response and analyzes it together with the effect of the substantive trait. In MIRT models, the probability of selecting the k-response category of an item j at an ability level of θ1, θ2 „„ θn is determined as follows (Bolt & Johnson, 2009):

$$P(Y_j = k|\theta_1\theta_{ERS}) = \frac{exp(\alpha_{jk1}\theta_1+\alpha_{jk2}\theta_{ERS}+c_{jk})}{\sum_{h=1}^{k} exp(\alpha_{jk1}\theta_1+\alpha_{jk2}\theta_{ERS}+c_{jk})} \quad (1)$$

where

a: category slope of parameter of item j

c: intercept parameter

θ_1: substantive trait parameter;

$\theta_{ERS:}$ trait related to extreme response style

In order to model the response style, constant value constraints are applied to the category slope parameters along the items. For example, in a seven-point Likert type, in order for θ_1 to be interpreted as the substantive trait to be measured and θ_{ERS} as the trait of the extreme response style, for θ_1, the "a" parameters of the response categories of all items are fixed to the values -3, -2, -1, 0,1,2,3 while for θ_{ERS}, these values are fixed to the values 3, -1.2, -1.2, -1.2, -1.2, -1.2, 3. It should be noted that the response categories, which are symmetrical to each other, are fixed to the same value by absolute value and $\sum_k c_{jk} = 0$ is for each item j (Lu & Bolt, 2015).

In the literature, there are also studies involving the response styles with MIRT models. Of these studies, Bolt and Johnson (2009) used MIRT to determine and control the impact of ERS on individuals' motivation to smoke and the status of items' differential item functioning (DIF). The researchers stated that the effect of extreme response style also changed according to the level of motivation to smoke, and that the corrected scores of the individuals differed from the uncorrected scores according to the response style. In another study, the same researchers described how to use the multidimensional multinomial logit item response model to explain that this model was effective when the trait related to the response styles and the substantive trait to be measured were related to each other.

Extending a previous study (Bolt and Johnson 2009), Bolt and Newton (2011) analyzed the responses to multiple scales of American students participating in PISA 2006 in order to better predict the effect of ERS with the model developed by the researchers. At the end of the study, they found that the effect of ERS could be better predicted using multiple scales.

Lu and Bolt (2015) aimed to determine the role of ERS in the attitude towards science scores of students participated in PISA 2006. To this end, he included each variable related to attitude towards science and ERS into MIRT model separately in a multi-level context in order to determine and correct the impact of ERS on the responses of the countries participating in PISA 2006 to the attitude scales. The final outcome was that ERS had a negative effect on the attitudes towards science of the students having a high level of achievement at the country level. However, due to the complexity of the models the researchers could not include all attitudinal variables and ERS into the models at the same time, which may distort the results.

Although these studies using the MIRT model largely seem to compensate for the shortcomings of the other methods in terms of predicting the effect of ERS, these studies are insufficient to examine the effect of response styles in a multi-level context. Therefore, in order to contribute to filling the gap in the field of education and make more valid inferences, in this study, a multi-level MIRT model was used because it provides the opportunity to determine and correct the effect of ERS. In addition, intercultural comparison research and theories (Bandura, 1994; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005), which indicate that there is a positive relationship between the student's attitude toward a course and their success in that course, do not take into account the differentiation of each school system in a social, economic and cultural manner. Therefore, these studies and theories fail to explain the relationship between achievement and attitude and the change in the direction of this relationship from one country to another (Shen & Tam, 2008). In this context, the country-level comparison of the relationships between students' achievements and their perceptions and attitudes with multilevel analyses goes beyond the theories providing information at the individual level (e.g., motivation theories and self-efficacy theories). Therefore, the results of this study are valuable in terms of understanding the country-level relationships between achievement and attitudes, yielding a valid interpretation of these relationships. In this respect, it is considered that the results obtained from the research will assist in explaining the negative relationship between attitude and success at the country level. In addition, although various methods are used to control the effect of the response style PISA 2015 (OECD, 2017), which is one of the large-scale applications that take into account the results of the studies where various secondary analyzes are made by making use of its data, there is no information about this situation in TIMSS applications. In the light of this information, regarding the students from countries that participated in the TIMSS 2015 application and displayed an attitude-achievement paradox, the responses to the following research questions were sought in this study:

a) How do the two models, one including ERS and the other excluding ERS as a dimension, fit the data related to the responses of the students to the attitude scales?

b) How does the relationship between mathematics achievement and unadjusted and

ERS-adjusted attitude scores vary at both individual and country levels?

c) Is there any significant difference between the correlation coefficients indicating individual-level and country-level relationships between mathematics achievement and unadjusted and ERS-adjusted attitude scores?

METHOD

Research Model

This study aimed to determine the effect of ERS which might cause an attitude-success paradox among the countries participating in the TIMSS 2015 application and examine the relationship between attitude and mathematics achievement by correcting the effect of ERS on the attitude scores of eighth-grade students in different countries. In this context, it was a correlational study.

Population and Sample

The sample of the present study consisted of eighth-grade students from the countries participating in TIMSS 2015 and displaying the attitude-achievement paradox. The reasons for the selection of eighth-grade students participating in TIMSS 2015 were as follows: (i) students at this grade level are in a transition period from secondary school to high school, where the knowledge and skills required to be acquired in the mathematics curriculum differ greatly (Rodriguez, 2004) and (ii) fourth-grade students are not aware of their attitudes and cannot evaluate themselves effectively (Harter, 1999).

This study aimed to represent the pattern of the relationship between attitude and achievement in country selection among all countries participating in TIMSS 2015. To achieve this, students who do not like mathematics, are not confident in mathematics and do not value mathematics were selected. Based on the students' mathematics achievement scores, the countries were classified as those with high mathematics achievement but low attitude scores, those with moderate mathematics achievement and attitude scores, and those with low mathematics achievement but high attitude scores. A total of 15 countries, five countries representing each category, were included in the sample: Singapore, Republic of Korea, China-Taiwan, Hong-Kong, Japan, Sweden, Italy, Malta, Australia, Norway, Turkey, Chile, Kuwait, Egypt, and Saudi Arabia. Then, the missing values in the data set of each country were deleted and 500 sub-samples from each country were randomly selected.

Data Collection Tools

A student questionnaire and mathematics achievement test were used as data collection tools. In the student questionnaire, for the countries in which the students have higher mathematics achievement, the percentage of students who have negative attitudes in terms of students' liking of learning mathematics, self-confidence in mathematics, and valuing mathematics was higher than that of other countries where students have low mathematics achievement. Therefore, the scales of these variables were used in the present study. Each of the items in these scales has four response categories listed as “1” indicating "strongly agree” and "4" indicating “strongly disagree".

The Cronbach's alpha reliability coefficients obtained from the selected countries ranged between .70 and .96 (Martin, Mullis, Hooper, Yin, Foy & Palazzo, 2016). The reliability coefficients greater than .70 indicate that the scores obtained from the scales are reliable (Fraenkel & Wallen, 2006). In the study, the scores of eighth-grade students regarding mathematics achievement were obtained from the mathematics achievement tests applied in TIMSS 2015. In this study, as in similar studies (e.g., Buckley, 2009), the mean of five plausible values was used to represent students' mathematics achievement.

Data Analysis

The missing values for each country were deleted from the data set considering the number of individuals in the sample and the fact that multiple assignments would affect the individual's extreme response style scores. Since the response categories of the items in the scales were ordered in a way that a high score obtained from the item represented a negative attitude toward mathematics, reverse coding was undertaken for the items reporting positive attitude. Considering the time taken for the analysis in terms of the processor speed and memory of the computer, 500 sub-samples from each country's data set were selected randomly to avoid the complexity of analyses and to maintain an equal number of students in the samples. To determine whether the 500-sub-sample adequately represented the sample of the countries, the correlation values between the achievement scores and the scale mean scores of the sub-sample and the sample were examined. The correlation values between the related variables in the sub-sample and the correlation values between the variables in the whole sample were found to be similar, which is accepted as an indicator of the representation of the sub-sample.

For the first sub-goal of this study, two different models were developed with certain limitations. The first (basic) model consists of three dimensions: liking of learning mathematics (θ_L), self-confidence in mathematics (θ_C), and valuing mathematics (θ_V), which are dimensions related to the attitude to be measured. In the first model, the response category curve parameters (a_{jkm}) were fixed to equally spaced values (-3, -1, 1, 3) for each dimension. In the second model, as opposed to Lu and Bolt (2015), the fourth dimension ERS bias (θ_{ERS}) was also included. Otherwise,including only one attitudinal trait and one ERS trait at one time can cause a loss of information on the estimation of ERS ,which yield to biased estimation on the attitudinal trait. The category curve parameters for the fourth dimension were fixed to “1, -1, -1 and 1 for all items in the scales. Then, in order to determine the model that better fit the data, the model fit statistics for the models were obtained using the Latent Gold program (Vermunt & Magidson, 2008). The model-data fit is determined based on various model fit indices (the Bayesian information criterion- [BIC], Akaike information criterion [AIC], Akaike information criterion 3 [AIC3], and consistent Akaike information criterion [CAIC]). The lower values of these information criteria indicate a better fit (Vermunt & Magidson, 2005).

For the second sub-goal of the study, to determine the uncorrected and corrected attitude scores, firstly, the parameters of the items and individual related to the attitudinal dimensions and response style were estimated. For this purpose, a two-stage estimation process was undertaken using a multilevel MIRT model. In the first stage, the intercept parameters of the items in the scales were estimated with three-dimensional and four-dimensional models. Latent Gold was used to estimate the item parameters. In the second stage, using the estimated item intercept parameters for the cases where the effect of ERS was and was not taken into account, the attitude values of the students were estimated using the estimated parameters of the items. At this stage, theta value θ_{ERS} was not included in the model while estimating the uncorrected attitude scores of individuals with the three-dimensional model. In addition, the relationship between the attitudes and achievements of the countries was determined using mean achievement scores at the country level for the cases where the ERS was not controlled and controlled separately. At this stage, analysis was carried out using the WinBUGS (Spiegelhalter, Thomas & Best, 2004) software and the Markov Chain Monte Carlo (MCMC) method.

Similar studies were examined to determine the number of iterations required for the convergence of the parameter to not be affected by the initial values. In these studies, iterations were found to vary between 2000 and 10000. Therefore, in this study, as suggested by Gelman and Rubin (1996), a single chain was used and 8,000 iterations were performed for both models depending on the capacity of the computer where the analyses were performed. As suggested by Geyer (1992), 2% (160) of these iterations were used in the burn-in period to reduce the effect of initial values on posterior values.

It was examined whether the Markov chain converges in deciding the adequacy of the number of iterations and the accuracy of the parameter estimation. Time series graphs (Sinharay, 2004) and Monte Carlo (MC) errors (Ntzoufras, 2009), which are frequently employed to evaluate convergence in MCMC, were used. MCMC simulations performed for the countries participating in TIMSS 2015 did not show irregularity and convergence was achieved in the time series graphs.

In addition, the fact that MC errors were less than .05 (Ntzoufras, 2009) shows once again that convergence was achieved. After determining the corrected and incorrected attitude scores of eight-grade students in the countries participating in TIMSS 2015 according to the effect of ERS, the relationships between mathematics achievement and these latent trait scores and the observed scores were determined. In order to determine the within-country relations between the variables related to mathematics achievement and attitude toward mathematics based on the observed scores, firstly the mean of the scale scores of different attitude scales of the individuals in each country was calculated and four different values of each country (mathematics achievement score, mean scale scores for the liking of learning mathematics, self-confidence in mathematics, and valuing mathematics) were determined. It was then checked whether these values from 15 countries were normally distributed. Since the sample size was less than 35, the results of the Shapiro-Wilk normality test were considered (Shapiro & Wilk, 1965). Accordingly, it was determined that the country-level mean mathematics achievement scores of the selected countries were not normally distributed (p = .03). Therefore, the Spearman rank differential correlation coefficient (rs), which is one of the non-parametric correlation coefficients, was calculated to determine the relationship between the observed scores of attitude-related variables and mathematics achievement. At this point, the observed points were utilized. In addition, in order to determine the effect of controlling ERS on the relationship between attitude toward mathematics and mathematics achievement at the country level, three and four dimensional models were analyzed by WinBUGS and correlation values were obtained. Since the three-dimensional and four-dimensional models were in the common latent trait metric (Lu & Bolt, 2015), the correlation values obtained from the three-dimensional model at the country level were considered as a baseline to compare the correlation values obtained with the four-dimensional model.

Concerning the third research question, to determine whether there was a significant difference in the correlation values between the observed scores of attitude-related variables and mathematics achievement, and the correlation values obtained with the three-dimensional model and the four-dimensional model, MedCalc 18.2.1 (MedCalc Software bvba) (2018) was used. This program basically converts the correlation coefficients (r) to the z score. This transformation is called Fisher’s r to z transformation. In this analysis, the z observed score is determined at the next stage. In this analysis, the z observed score was determined to be higher or lower than the critical value at the .05 level, showing whether the correlation coefficients significantly differed.

RESULTS

Two different MIRT models were analyzed to determine the data fit of the models with and without ERS as a dimension. In this respect, the three-dimensional model including only attitudinal variables, such as liking of learning mathematics, self-confidence in mathematics, and valuing mathematics as dimensions and the four-dimensional model including ERS as a separate dimension were analyzed separately, and various model-data fit indices were obtained.

Table1. Model Comparison

Models	Fit Indices						
	LL	BIC	AIC	AIC3	CAIC	L^2	P
Three-dimensional Model	-209522.32	419820.91	419218.64	419305.64	419907.91	286174.19	7.7e-54656
Four-dimensional Model	-196641.21	394094.39	393464.43	393555.43	394185.39	260411.98	1.5e-49216

As shown in Table 1, the model fit index values estimated by the four-dimensional model were lower than those estimated by the three-dimensional model. Since lower values of model fit indexes indicate a better fit (Vermunt & Magidson, 2005), it was found that the four-dimensional model best fitted the data, showing that the effect of ERS on individuals' responses to items existed. In this case, when the effect of ERS was taken into consideration, the true values of the latent trait levels of the students' attitudes were more accurate and clear. At the same time, the limitations applied to the item response categories of the attitudinal traits and ERS in the four-dimensional model contributed to the increase of model-data fit. The within-country correlation values between the students' corrected and uncorrected attitude scores and mathematics achievement according to the effect of ERS are given in Table 2.

Table 2. Within Country Correlation Values

Countries	Liking Learning Mathematics		Self-confidence in Mathematics		Value on Mathematics	
	a	b	a	b	a	b
Singapore	.28	.33	.39	.50	.11	.22
Republic of Korea	.41	.47	.50	.63	.39	.42
China-Taiwan	.44	.49	.51	.63	.35	.41
Hong-Kong	.32	.32	.37	.44	.19	.25
Japan	.35	.39	.44	.54	.22	.28
Sweden	.42	.46	.63	.68	.19	.28
Italy	.36	.38	.52	.57	.19	.25
Malta	.29	.32	.41	.50	.11	.22

Australia	.29	.35	.50	.60	.19	.27
Norway	.37	.42	.61	.66	.20	.28
Turkey	.19	.23	.52	.59	.11	.15
Chile	.19	.26	.45	.53	.13	.21
Kuwait	.20	.23	.33	.44	.18	.27
Egypt	.24	.33	.37	.50	.14	.30
Saudi Arabia	.28	.25	.39	.54	.11	.26

a = Correlation value for the relationship between non-corrected attitude score and mathematics achievement, b = Correlation value for the relationship between corrected attitude score and mathematics achievement

In Table 2, the within-country correlation values for the relationship between the uncorrected and corrected attitude scores of the students with the mathematics achievement for each dimension are shown to be similar to each other with increment in addition to being positive. In other words, the students who had positive attitudes toward mathematics had a higher level of mathematics achievement in both cases (where the effect of ERS was not controlled and controlled).

To determine whether the increment in correlation is significant or not, the p values obtained from Fisher's z transformations were performed. The results about these transformations were given in Table 3.

Table 3.Differences between the Within-country Correlations of the Uncorrected and Corrected Attitude Scores of Countries with the Mathematics Achievement

Country	Dimension	p
Singapore	Liking	.38
	Confidence	.03*
	Value	.07
Republic of Korea	Liking	.24
	Confidence	.00*
	Value	.57
China-Taiwan	Liking	.31
	Confidence	.00*
	Value	.27
Hong-Kong	Liking	1.00
	Confidence	.19
	Value	.32
Japan	Liking	.46
	Confidence	.04*
	Value	.31
Sweden	Liking	.43
	Confidence	.17
	Value	.13
Italy	Liking	.71
	Confidence	.26
	Value	.32
Malta	Liking	.60
	Confidence	.07
	Value	.07
Australia	Liking	.29
	Confidence	.07
	Value	.07
Norway	Liking	.29
	Confidence	.02*
	Value	.18
Turkey	Liking	.35
	Confidence	.19
	Value	.18
Chile	Liking	.51
	Confidence	.11
	Value	.52
Kuwait	Liking	.61
	Confidence	.04*
	Value	.13
Egypt	Liking	.12
	Confidence	.01*
	Value	.00*
Saudi Arabia	Liking	.61
	Confidence	.00*
	Value	.01*

*significant at .05

As can be seen from Table 3, in general, there is a significant difference between the correlation scores of the uncorrected and corrected attitude scores regarding confidence in mathematics with mathematics achievement of students in the countries participating in TIMSS 2015 (p<.05).

The between-country correlation values between the students' corrected and uncorrected attitude scores and mathematics achievement according to the effect of ERS are given in Table 4.

Table 4. Between-Country Correlation Values

Variable Pair	Correlation Value based on Observed Scores (r_{s1})	Correlation value obtained from three-dimensional model	Correlation value obtained from four-dimensional model
Achievement-Liking	-.59	.08	.15
Achievement- Confidence	-.74	.04	.19
Achievement-Value	-.65	.01	.13
Achievement-ERS	-	-	-.09

Table 4 reveals that the correlation bases on the observed score were negative and those based on the MIRT models were positive. For example, there is a significant negative relationship between mathematics achievement and liking of learning mathematics at the country level (r_{s1} = -.59, p < .01); thus, the students with high mathematics achievement tended to dislike mathematics. On the other hand, according to the country-level correlation value of the estimated scores according to the three-dimensional MIRT model, a positive low-level relationship was found between these students' math achievement and their attitudes regarding their liking of learning mathematics. Similarly, according to the country-level correlation value of the four-dimensional MIRT model, which took into account the effect of ERS, a positive correlation was found between mathematics achievement and students' liking of learning mathematics. The change in the estimation of the correlation values at the country level according to different models was observed in other dimensions related to attitude as well.

The results of Fisher's rz transformation to determine whether there was a significant difference in the country-level relationships between mathematics achievement and the uncorrected and corrected attitude scores according to the effect of ERS are presented in Table 5.

Table 5. Differences between the country-level correlations

Correlation value pairs	Dimension	z	p
a-b	Self-confidence	-3.40	.01*
	Liking	-2.63	.01*
	Value	-2.75	.01*
a-c	Self-confidence	-3.43	.01*
	Liking	-2.64	.01*
	Value	-2.81	.00*
b-c	Self-confidence	-.61	.54
	Liking	-.06	.95
	Value	-1.16	.24

*significant at .05, a = Correlation value based on the observed scores, b = Correlation value based on the scores obtained from the three-dimensional MIRT model, c = Correlation value based on the scores obtained from the four-dimensional MIRT model

Table 5 shows the statistically significant difference between the correlation values based on observed scores, showing the relationship between mathematics achievements and the attitudinal variables at the country level and the correlation values predicted by the MIRT models (p < .05).

However, no statistically significant difference was found between the correlation values showing the relationship between mathematics achievement and each variable determined based on the scores predicted by the three-dimensional MIRT model and the correlation value obtained from the four-dimensional MIRT model.

When the findings obtained from this study using Likert-type scales based on self-statements of the students were evaluated as a whole, ERS was effective in the students' responses and the effect of ERS on the students' attitudes also affected the direction and level of their relationship with the achievement of the students.

DISCUSSION, CONCLUSION AND SUGGESTIONS

This study aimed to determine the effect of extreme response style on students' attitudes in international assessments and the role of this effect on their relationship between mathematics achievement and attitudes toward mathematics. Compared to Bolt and Johnson (2009), Bolt and Newton (2011) and Lu and Bolt (2015), the present study added a multilevel structure to the MIRT model by including all attitudinal variables and ERS at one time and implemented it with the data regarding attitude toward mathematics obtained from TIMSS 2015 assessment where students are nested in countries. In this respect, the three-dimensional MIRT model in which the effect of ERS was controlled and the four-dimensional MIRT model which included the effect of ERS were analyzed, and based on these models, the students' uncorrected and corrected attitude scores were determined according to the effect of ERS. The correlations of attitude scores with mathematics achievement were examined at the individual and country levels.

As a result of the analysis of the three- and four-dimensional MIRT models, it was concluded that ERS had an effect on the students' responses to the items related to attitude due to the better fit of the four-dimensional model with the data. A similar finding was also found in the study which examined the role of ERS in the responses of students from the United States participating in PISA 2006 (Bolt & Newton, 2011). In their study, it was stated that the model in which ERS was considered as a separate dimension better fit the data. This shows that ERS has an impact on students' responses in large-scale applications involving students from various cultures, such as TIMSS and PISA. There are several reasons for the effect of ERS in this study, which uses data from a large-scale application, such as TIMSS, where certain standardized test and scale development steps are followed and the validity and reliability of the scores of instruments are established. For example, students come from different community structures (Hosftede, 2001). More specifically, when the countries categorized according to their achievement and attitude scores are classified in terms of the dimensions of the society stated by Hofstede (2001), it can be stated that the countries in the 1st and 3rd groups are more collectivistic and have a high power distance, and the countries in the 2nd group are more individualistic and less in power. Group dependence, interpersonal relationships and group solidarity are important for individuals living in collectivist cultures, since these individuals determine their behavior and attitudes according to norms or demands of society (Hofstede, 2001). On the other hand, those living in an individualist society have more control over their actions, can take responsibility for their actions, and tend to compete more than cooperate. In this context, in an individualist society, people tend to have more extreme reactions (Johnson et al., 2005), while individuals in collectivist cultures tend to react at a medium level in order to achieve cohesion in society (Smith, 2004).

One of the factors affecting the different response styles of cultures is power distance. It is the degree to which powerless individuals in a society accept and expect that power is not evenly distributed (Hosftede, 2001). In societies with high power distances, the idea that inequality is necessary to maintain order in society is dominant, but it is important to have the same views as those of individuals with high power. In societies with low power distances, the views of each individual are respected. Therefore, individuals in societies with low power distances tend to respond in the middle, while societies with high power distances tend to react at extreme ends (Johnson et al., 2005). Thus, in the current study, it was seen that the respondents exhibited different response styles according to the dominant characteristics of the cultures in which they lived.

According to the results of the analysis of the scores of the eighth-grade students who participated in TIMSS 2015, the relationships between the corrected and uncorrected attitude scores according to the effect of ERS and the mathematics achievement at the individual and country levels and the observed scores and the scores obtained from the multidimensional item response theory models were analyzed. It was concluded that there was a high negative relationship between mathematics achievement and students' attitudes toward mathematics at the country level. In addition, it was found that these relationships were positively altered and decreased when predicted according to the three-dimensional model which produced latent trait scores. When the effect of ERS was taken into consideration, it was found that there were positive relations in similar amounts. In other words, when the relationship between mathematics achievement and attitude toward mathematics was determined on the basis of the scores estimated by MIRT models, it was in a positive direction similar to the relationships at the individual level. This finding is similar to the outcome of the study in which the relationship between science attitudes and science achievement of students participating in the PISA 2006 application, in which the effect of ERS was considered (Lu & Bolt, 2015). In these large-scale applications ERS is one of the factors causing this situation. On the other hand, it was concluded that the relationship between ERS and mathematics achievement was low and that the correlation values at the individual and country levels obtained from both three dimensional and four dimensional MIRT models were close to each other. This shows that unlike the three-dimensional model, controlling ERS in the four-dimensional model is not sufficient to explain the attitude-success paradox alone. This may be caused by other response styles (acquiescence response style-[ARS], disacquiescence response style-[DRS], etc.) that are effective in individuals' responses and affect the correlation between the variables (Reynolds & Smith, 2010). Another factor affecting the amount of the relationship between mathematics achievement and attitude toward mathematics may be the big fish and little pond effect (Marsh et al., 2008). More specifically, in this study, which was based on students' self-report, they may have evaluated themselves and expressed their attitudes differently than they would in a group of students. In this case, the direction and level of their relationship with mathematics achievement may have been affected.

It was concluded that the correlations between the observed scores regarding the attitude toward mathematics of students participating in TIMSS 2015 and mathematics achievements and the correlation values estimated by MIRT models differed at the country level. A similar finding was found in a study conducted by Lu (2012). Accordingly, in PISA 2006, the relationship between students' observed scores regarding attitudes toward science and their science achievement participating was found to be different with the case where the effect of ERS was taken into consideration in MIRT model. The reason why the correlation values determined based on the observed scores are different from the correlation values obtained by performing the MIRT models might be that the MIRT models are based on latent traits. Also, the correlation values computed by using observed scores are based on classical test theory and correlation values obtained with MIRT models are based on IRT. So, the correlation values calculated according to the two theories were different. In addition, the three-dimensional models providing uncorrected attitude scores and the four-dimensional models providing corrected attitude scores are based on the same theory, IRT, have similar algorithms in the background, and are on the same latent trait metrics, So, this fact may have had an effect on the finding that correlation values obtained from 3 and 4 dimensional MIRT model are similar to each other, which was the case in the study of Lu (2012).

The results of the present study should be evaluated within the following limits. Firstly, since the scales used in this study were of four-point Likert type, the mid-point response category did not exist. Therefore, the effect of selecting the mid-point response category on attitude scores was not examined. By increasing the number of response categories, the effect of the midpoint response style can also be examined. In addition, the effect of other response styles, such as acquiescence and disacquiescence can be included in the models established in the research. Another limitation of the study is that a Bayesian approach was adopted to ensure that all uncertainties in the parameter estimations were taken into consideration simultaneously and that standard errors were obtained more easily. However, Bayesian approaches require specific priors; for example, in this research, the identity matrix was used as a prior, but since the priors used may have affected the results, alternative methods that do not require the use of priors, such as the marginal maximum likelihood method can be used for the same purpose in future studies. The length of time required for the analyses in the research is one of the limitations of the current study. Depending on the capacity of the computer used, the analyses lasted approximately 120 hours each for each model due to the sample size being large to ensure the reliability of the results and the necessity for the variables related to attitude and ERS to be modeled simultaneously. To reduce the processing time, a two-dimensional model including a variable related to attitude and ERS as a separate dimension can be analyzed separately as it was done in the study of Lu and Bolt (2015). Similarly the mean of five different possible values was presented to represent students' mathematics achievement. In future research, to reduce the analysis time, five different plausible values can be included in the model in order to represent students' mathematics achievement. By overcoming these limitations, the number of countries can be increased to provide better representation.

REFERENCES

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.

Bandura, A. (1994). Self-efficacy. *Encyclopedia of Human Behavior*, *4*, 71–81.

Baumgartner, H. and Steenkamp, J. E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Billiet, J. B., and McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608-628.

Bolt, D. M., and Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*,335-352.

Bolt, D. M., and Newton, J. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71,* 814-833.

Buckley, J. (2009). *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. Retrieved from https://edsurvey.rti.org/PISA/

Cheung, G. W., and Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross cultural research using structural equation modeling. *Journal of Cross- Cultural Psychology*, *31,* 187-212.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.

De Jong M. G., Steenkamp J.-B. E. M., Fox J.-P., Baumgartner H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45*(1), 104-115.

Engle, P. J. (2016). *Response Style in the Political Survey*.(Unpublished doctoral dissertation).Unıversıty of Wisconsin Madison.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20-3.

Fraenkel, J.R., and Wallen, N.E. (2006). *How to design and evaluate research in education*. New York: McGraw-Hill.

Geyer, C. J. (1992). *On the convergence of Monte Carlo maximum likelihood calculations*. Technical Report 571, School of Statistics, Univ. Minnesota.

Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*, 176- 188.

Harter, S. (1999). *The construction of the self: A developmental perspective*. New York: Guildford Press.

Harumi, C. A. (2011). *Cross-cultural differences in response styles* (Unpublished doctoral dissertation). Washington State University.

Harzing, A. (2006). Response styles in cross-national survey research. *International Journal of Cross Cultural Management*, 6, 243-265.

Hofstede, G. H. (2001). *Cultures consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, California: Sage Publications, Inc.

İlgün Dibek, M., Yavuz, H. & Çokluk Bökeoğlu, Ö. (2018). Tutum - başarı paradoksunda tepki stillerinin rolü: dokuz ülkenin karşılaştırılması, *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 18*(2), 932-952.

Johnson, T.R., & Bolt, D. M. (2010). On the use of factor-analytic mutinomial logit item response models to account for individual differences in response style. *Journalof Educational and Behavioral Statistics*, *35*, 92-114.

Johnson, T., Kulesa, P., Cho, Y. I., and Shavitt, S. (2005). The relation between culture and response styles: evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264-277.

Lu, Y. (2012). *A multilevel multidimensional ıtem response theory model to address the role of response style on measurement of attitudes in PISA 2006,(*Dissertation).Universıty of Wisconsin-Madison.

Lu, Y. and Bolt, D.M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-scale Assessments in Education,3*(2), 1-18. doi:1.1186/s40536-015-0012-.

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

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

Marsh, H.W, Seaton, M., Trautwein, U., Ludtke, O., Hau, K.T., O'Mara, A.J., and Craven, R.G. (2008). The big fish little pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20, 319–35.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., and Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), Methods and Procedures in TIMSS 2015 (pp. 15.1-15.312). Retrieved from Boston College, TIMSS &PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html

MedCalc Software bvba (2018). MedCalc Statistical Software version 18.2.1, Ostend, Belgium. Retrieved from https://www.medcalc.org.

Moors, G. (2004). Facts and artifacts in the comparison of attitudes among ethnic minorities. A multilevel latent class structure model with adjustment for response style behavior. *European Sociological Review, 20*, 303-32.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Wiley Series in Computational Statistics, Hoboken, USA.

OECD (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised edition*, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264281820-en

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.

Prediger, D. J. (1999). Basic structure of work-relevant abilities. *Journal of Counseling Psychology, 46*, 172-184.

Reckase, M. (2009). *Multidimensional item response theory*. Dordrecht: Springer.

Reynolds, N., and Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: A simplified approach to eliminating the problem. *Journal of Service Research*, 13, 230–243. doi: 1.1177/1094670509360408

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, *17*(1), 1-24. doi: 1.1207/s15324818ame1701_1

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (Complete samples). Biometrika, *52*(3/4), 591-611.

Shen,C.and Tam, H.P. (2008) The paradoxical relationship between student achievement and self-perception: a cross-national analysis based on three waves of TIMSS data, *Educational Research and Evaluation, 14*(1), 87-100, DOI: 1.1080/13803610801896653

Sinharay, S. (2004). Experiences with markov chain monte carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*, 461– 488.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61.

Spiegelhalter, D., Thomas, A., and Best, N. (2004). *WinBUGS version 1.4* [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

van de Gaer, E. and 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. and 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.

van Rosmalen J., van Herk H., Groenen P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47, 157-172.

Vermunt, J. K., and Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent Gold 4.5 Syntax Module*. Belmont, MA: Statistical Innovations, Inc.