International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 326–341
https://doi.org/10.21449/ijate.718670
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examination of Wording Effect of the TIMSS 2015 Mathematical Self-Confidence Scale Through the Bifactor Models

Esra Oyar[1,*], **Hakan Yavuz Atar**[2]

[1]Gazi University, Department of Educational Science, Ankara, Turkey

ARTICLE HISTORY
Received: Apr. 12, 2020
Revised: Jan. 14, 2021
Accepted: Mar. 16, 2021

Keywords:
Wording effect,
Method factor,
Mathematical self-esteem,
TIMSS

Abstract: The aim of this study is to examine whether or not the positive and negative items in the Mathematical Self-Confidence Scale employed in TIMSS 2015 lead to wording effect. While examining whether the expression effect is present or not, analyzes were conducted both on the general sample and on a separate sample for female and male students. To this end, data of 5724 students from Turkey who participated in TIMSS 2015 were used. Six different measurement models were created in the analysis of data and tested with Confirmatory Factor Analysis. The study revealed that positive items have a higher mean than the negative ones. In addition, it was concluded that the bifactor models fit the data better compared to the traditional DFA model, in which the model where negative items were taken as a separate factor are those that best fit the data. This situation is verified both in the general sample and the subgroups of females and males. In conclusion, it is recommended that the scale items should be created carefully and whether the positive and negative items result in separate factors should be examined.

1. INTRODUCTION

Each measurement instrument is created for a specific purpose, under specific conditions and in a way to apply to specific individuals (Erkuş, 2003). Thus, one of the psychometric properties that are sought for in any measurement instrument is the degree to which it serves its purpose, in other words, its validity. Validity is the process of evidence collection with the aim of supporting the inferences to be drawn from the test scores obtained through measurement instruments (Cronbach, 1984). This process involves determining the degree to which the structure intended to be measured is being measured. However, some situations encountered during the measurement threaten validity and lead to errors in the measurement of the intended structure. One of the situations that threaten validity is the method factor (Ford & Scandura, 2018). Method factor occurs when participants systematically respond to the items differently due to the wording of the items in the scale (DiStefano & Motl, 2009). In case the measurement instrument includes method factors such as item characteristic (social desirability, etc.), item content (positive or negative items, etc.) and measurement content (time or place of

*CONTACT: Esra OYAR ✉ esra.tas18@gmail.com Gazi University, Department of Educational Science, Ankara, Turkey

ISSN-e: 2148-7456 /

measurement, etc.) (Podsakoff et al., 2003), the researcher cannot measure the intended trait due to difference from the real factor in the structure that is intended to be measured, which threatens the validity (Chen, 2017; Yang et al., 2012). If the test has negative and positive items, it causes a method factor due to the item content. This situtation is defined as the wording effect in the literature (Gu et al., 2015). It has been suggested in the literature that measuring various structures in social sciences including personality, attitude and anxiety requires the use of positive and negative items evenly (DeVellis, 2003; Weijters et al., 2013), which is argued to decrease the response bias (Weijters et al., 2013). When scale items include negative statements, participants read them more carefully, thereby eliminating responses that have the same response patterns (Podsakoff et al., 2003). The main assumption when using both types is that the negative items will represent the structure in the same way as their positive counterparts (Marsh, 1996). In other words, when the negative items are reverse coded, both item sets should be psychometrically indistinguishable. However, recent studies have revealed that the coexistence of positive and negative items in a scale results in systematic measurement error and thus leads to biased interpretation of results (Gu et al., 2015; Schriesheim et al., 1991). In addition, researchers state that a two-factor structure is produced when mixed items are used (Greenberger et al., 2003; Ibrahim, 2001), which jeopardizes the structure validity (Schmitt and Stuits, 1985; Woods, 2006), and that positive items have a higher mean compared to the negative items (Weems, Onwuegbuzie and Collins, 2006). In the measurement of the structure, wording effect not only poses a threat against the validity but also can decrease the reliability of both the scale items and the scores (Gu et al., 2015; Weems et al., 2003; Yang et al., 2012). Therefore, if the wording effect is modeled through a proper measurement model, researchers can assess the psychometric properties (validity, reliability, etc.) of the data more precisely based on this effect (Gu et al., 2015). When the literature review is examined, considering that the positive and negative items in scale development and adaptation studies may cause difficulties in construct validity, testing this situation has been deemed worthy of research.

Various methods are employed in modeling the wording effect. The most frequently used methods are Confirmatory Factor Analysis (CFA) models and bifactor models (DiStefano and Motl, 2006; Tomas & Oliver, 1999). Confirmatory Factor Analysis (CFA) is a type of Structural Equation Modeling (SEM) and helps to analyze measurement models that allow to establish relationships between the observed variables or indicators (items) that measure the same latent traits or factors (Brown, 2006). Another measurement model employed in the wording effect is the bifactor model (Wang et al., 2018). Bifactor models were developed by Holzinger and emerged as a type of confirmatory factor analysis (Jennrick & Bentler, 2011). In recent years, bifactor models have been increasingly used as an alternative but more advantageous approach in testing the multi-facet structures and in addressing the subject of dimensionality in psychological research (Chen & Zhang, 2018). This model includes one common factor that represents the shared variant in all scale items and an additional group factor that represents the shared variant in the items in a group (Reise, 2012). The common factor represents the individual differences in the target factor which is common to the items and the researcher deals with. Group factor, on the other hand, refers to the shared variant in item responses that cannot be explained by the common factor (Reise et al., 2010). Common factor and group factor are assumed to be orthogonal.

In studies examining the wording effects, (i) a model incorporating only the relevant factor, (ii) bifactor models incorporating positive and negative items as separate factors in addition to the common factor, and (iii) measurement models incorporating the correlation between the error terms of the positive and negative items are created (Chen et al., 2010; Gu et al., 2015; Horan et al., 2003; Marsh, 1996). Bifactor models in which positive and negative items are included as separate factors are also called correlated method (CM) (Lindwall et al., 2012). Similarly, measurement models including the correlation between the error terms of positive and negative

items are defined as correlated uniqueness (CU) (Lindwall et al., 2012). Both models are measurement models which attempt to identify the wording effect of positive and negative items; however, they have some differences. The CM model incorporates certain latent method factors underlying the scale items of the same method (in other words, item formats expressed as positive or negative) along with a latent factor. On the contrary, the CU models are based on establishing a correlation between the remains of positive and negative items (Lindwall et al., 2012; Wu, 2008). Thus, the CM model can be predicted by other factors or variables, but it is not the case in the CU model. Interpretation of method factors is easier and clearer in the CM model than the CU model (Wu, 2008).

In the light of the foregoing, in order to determine whether or not the test items referred to as negative measure a structure other than the intended one, Weems et al. (2006) conducted a study on 153 university students who studied education and psychology. The study revealed that the mean scores the students obtained from the positive items were higher than that from the negative items. In their study, Yang et al. (2012) examined whether or not the positive and negative items in the Attitude Toward Mathematics Learning Scale in TIMSS 2007 had a wording effect on the Taiwan and America sample. The sample of the study consists of the data of 4111 Taiwanese and 7831 American fourth-grade students. A series of CFA showed that there is a wording effect for both samples. Negative items are claimed to have lower reliability and approximately 25% of the score variance in the negative items are told to be caused by the measurement method, not the latent trait. In conclusion, the researchers stated that whether the items had wording effect should be examined and the negative statements should be worded as simple as possible. In another study, which examines the wording effect based on TIMMS scales, Michaelides (2019) performed some analyses by way of an 18-item motivation scale. The scale included three sub-scales. The measurement models that were created are, respectively, (1) one-dimensional model, (ii) three-dimensional model, (iii) second-degree factor model with three sub-scales, (iv) the model in which three dimensions are correlated and negative method factor is included, (v) a model in which the uniqueness variance of negative items are correlated, and (vi) the model in which negative and positive items are included as factors. When the fitting values of the measurement models are considered, the model in which the correlation between negative items was established yielded the best result.

Studies examining the method factor caused by wording are usually carried out on adults (DiStefano & Motl, 2009; Horan et al., 2003; Tomas & Oliver, 1999). When the verbal skills of younger participants are considered, however, this effect might be greater (Yang et al., 2012). Benson and Hocevar (1985) investigated the wording effect in the attitude scales on fourth- to sixth-grade children in the USA by way of the item sets consisting of 15 items. The first item set included only the positive items while the other one included only the negative items. At the end of the study, it was determined that students did not give the same response to the positive and negative items having the same content and were likely to demonstrate a less positive attitude in negative items. Researchers stated that little children cannot express agreement by giving a negative response to a negative statement or disagreement by giving a positive response to a negative statement. Thus, assessing whether using positive and negative items in combination results in the wording effect for younger participants is of importance (Yang et al., 2012). Based on these studies, it is important to investigate whether a separate latent structure is formed in the inclusion of negative statements in the analysis by reverse coding, especially in young age groups, to reveal the structure correctly. Another point examined in terms of wording effect is whether it differentiates depending on gender (DiStefano & Motl, 2009; Michaelides et al., 2016). Studies also tested measurement invariance by taking gender variable as a subgroup. However, this study did not attempt to determine whether the measurement model accepted based on general sample is similar for both female and male but rather to find out which measurement model fits the data better for both females and males.

1.2. Purpose

The aim of this study is to determine whether or not the responses of eighth-grade students to the scale items consisting of both positive and negative items in the Mathematical Self-Confidence Scale conducted in TIMSS 2015 have a wording effect by means of Confirmatory Factor Analysis based on the bifactor models that have been created. To this end, the presence of the wording effect will be investigated not only on the general sample but also on separate samples created both for male and female students by way of creating different measurement models (Models 1-6).

1.3. Research Questions

This study includes attempts to address the following research questions:

1. Is there a significant difference between the scores the students got from between the mean scores the students got from the positive and negative items in the Mathematical Self-Confidence Scale?
2. Do the positive and negative items in the Mathematical Self-Confidence Scale result in a wording effect?
    a. Is there a wording effect in the general sample?
    b. Is there a wording effect for female students?
    c. Is there a wording effect for male students?

2. METHOD

2.1. Research Design

The purpose of this study is to investigate whether there is a method/wording effect on the items in TIMSS 2015 Mathematical Self-Confidence Scale by way of CFA models and bifactor models. This is a descriptive study in that it aims to put forward the current situation (Büyüköztürk et al., 2017).

2.2. Study Group

In this study, students who participated in the 2015 TIMSS exam from Turkey constitute the working group. Among these students, data of 5724 8th grade students who responded to all items in the "Confidence in Mathematics" scale were used. 48.5% (2779 people) of these students are female students and 51.5% (2945 people) are male students.

2.3. Data Collection Tool

The measurement tool used in this study is the Scale of Self-Confidence in Mathematics, which was developed in a different language and adapted to Turkish (Table A1). Within the scope of the study, the effects of positive and negative items on the construct validity of the scale were examined. In the analyzes, it was tried to determine whether a separate structure was formed in the case of positive or negative matter. For this reason, it is thought that cultural effect from a scale obtained by adaptation study will not make a difference in the response pattern to positive and negative items.

There are a total of 9 items in the Mathematical Self-Confidence Scale administered in TIMSS 2015, which was designed to determine the self-confidence degree of students in the Mathematics class, these items consist of four positive and five negative items. Items and information related to them are available in ANNEX1.

Translation of the items in the scale has been obtained from the TIMSS 2011 final report. Students' responses to these items are evaluated on a 4-point Likert scale of 1) Completely agree, 2) Partially agree, 3) Partially disagree, 4) Completely disagree. In the analysis phase, positive items were reverse scored and the total score was calculated based on 9 items in the scale.

2.4. Data Analysis Procedures

In order to seek an answer to the first research question of this study, "Is there a significant difference between the scores the students got from the means of the positive and negative items in the Mathematical Self-Confidence Scale?" paired sample t-test was performed based on the mean scores of students for the positive and negative items (Kirk, 2007). Since the number of the positive and negative items in the scale is different, in order to ensure that both total scores will be in the same range, total scores of students for positive and negative items were divided by the total number of items in the relevant score. Cohen's d was used to calculate the effect size.

$$d = \frac{t}{\sqrt{N}}$$

Following the standard of Cohen (1988), effect size estimates of 0.2, 0.5 and 0.8 were considered as small, medium and large, respectively. In the study, six different measurement models were created in order to address the second research question and tested through confirmatory factor analysis. These models are as follows:

1. Model: Single-factor model for the Mathematical Self- Confidence variable.

2. Model: Bifactor model composed of both Mathematical Self- Confidence factor and positive and negative items.

3. Model: Bifactor model composed of both Mathematical Self- Confidence factor and positive items.

4. Model: Bifactor model composed of both Mathematical Self- Confidence factor and negative items.

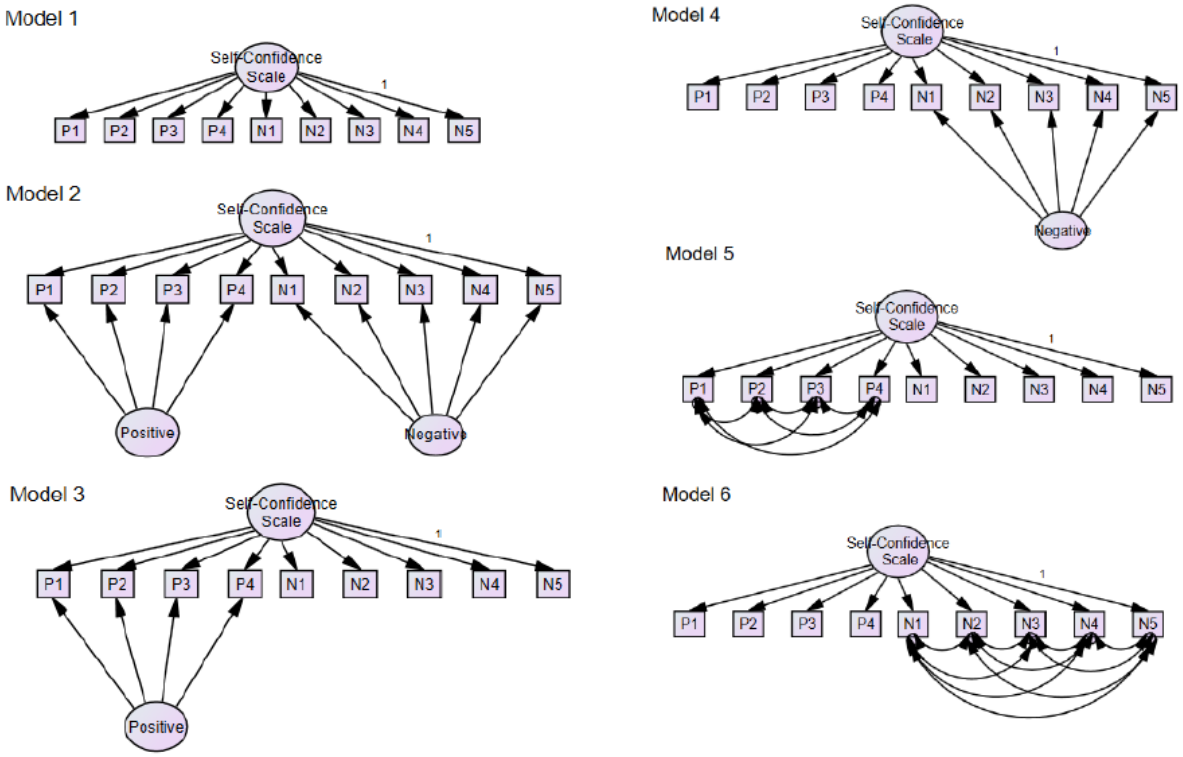
5. Model: A Mathematical Self- Confidence factor including correlated uniquenesses among positively worded items

6. Model: A Mathematical Self- Confidence factor including correlated uniquenesses among negatively worded items

The figural representations of the models are presented in Figure 1. The purpose of Model 1 is to create a measurement model for a single latent factor (Mathematical Self- Confidence). The measurement model was created assuming that all items in the scale fall under a single latent factor and their model fit indices were examined. In Model 2, positive and negative items are collected under a separate latent factor for each in addition to the Mathematical Self-Confidence latent factor and the bifactor model was created. Model 3 and Model 4 differ from Model 2 in that the bifactor model was created with the assumption that only the positive items and only the negative items fall under a latent factor, respectively, in addition to the Mathematical Self- Confidence latent factor. In Model 5 and Model 6, a correlation was established between latent variances of positive and negative items, respectively, and the model contained a single latent factor (Mathematical Self- Confidence). Goodness of fit indices obtained from all models was examined and the model that fits the data best was accepted. This process was carried out not only on the general sample but also on the sub-samples containing only females or only males, and efforts were exerted to find out which measurement model fits the data in the relevant sample. At this point, the primary aim is to determine which one of the measurement models created displays the best fit in each of the three data sets.

Model 2, Model 3 and Model 4 are CM models, whereas Model 5 and Model 6 are CU models. CM models incorporate positive and negative items as a distinct latent factor in addition to the common latent factor. CU models, on the other hand, create a measurement model by correlating residual variances (uniqueness variances) rather than gathering negative items under a latent factor for each.

Figure 1. *Model Representations*.



2.4.1. *Assessment criteria*

In the evaluation of measurement models, the values of $\chi 2$, RMSEA, SRMR, CFI and TLI in the MPlus package program output were examined.

• RMSEA value smaller than 0.08, CFI and TLI values greater than 0.95 and SRMR value smaller than 0.06 indicate that the data and the model represent a perfect fit, whereas RMSEA value smaller than 0.10, CFI and TLI values greater than 0.90 and SRMR value smaller than 0.08 indicate that the data and the model represent an adequate fit (Hu & Bentler, 1999).

• Low RMSEA and SRMR but high CFI and TLI values in a measurement model are interpreted as the model fits? the data better than other models.

2.4.2. *Testing of the assumptions*

In data analysis, the first missing data, extreme value and normality assumption checks were carried out. Since the missing data did not exceed 5%, students with missing data were excluded from the study. Information regarding the sample on which the analyses were performed is shown in Table 1.

Following the deletion of the missing data, the sample included data from 5724 students, 2779 (48.5%) females and 2945 (51.5%) males. Examination of z scores for the extreme value revealed that there is no student score out of ±3 range, so there is no extreme value in the data. Finally, skewness and kurtosis values were checked for normality assumption. The skewness and kurtosis values for total scores and scores obtained from positive and negative items are in the range of ±1. Thus, considering the sample size and skewness and kurtosis values, it can be said that the data has a normal distribution (Büyüköztürk, 2012).

Table 1. *Descriptive Statistics for Sampling.*

		f	%
Gender	Female	2779	48.5%
	Male	2945	51.5%
Total		5724	100%

3. RESULT / FINDINGS

Mean and standard deviation values were calculated not only for the entire sample but also for both subgroups of female students and male students for each item in the scale. Calculated values for items are as shown in Table 2.

Table 2. *Statistics for Items.*

Items	General		Female students		Male students	
	μ	SD	μ	SD	μ	SD
M1	2.96	.943	2.96	.954	2.97	.932
M2*	2.46	1.101	2.47	1.125	2.44	1.077
M3*	2.52	1.156	2.52	1.173	2.53	1.139
M4	2.79	.967	2.77	.953	2.81	.979
,M5*	2.60	1.130	2.59	1.150	2.61	1.111
M6	2.37	1.042	2.27	1.024	2.47	1.050
M7	2.62	1.055	2.61	1.056	2.64	1.055
M8*	2.17	1.149	2.16	1.163	2.18	1.136
M9*	2.32	1.143	2.33	1.153	2.31	1.132

*negative items

Examination of the values in the table reveals that item means obtained from the entire sample and the means of female and male students are close. In the scale, item 8, “Mathematics is harder for me than any other subject” has the lowest mean, while item 1 “I usually do well in mathematics” has the highest. Means obtained from the positive items are higher than the means obtained from the negative items both in the general sample and in the subgroups of females and males.

Paired sample t-test was employed to find out whether there is a significant difference between the mean scores students got from the positive and negative items in the scale, the results of which are shown in Table 3.

Table 3. *Paired Sample T-Test Results for Positive and Negative Items.*

		μ	SD	*t*	*p*
Items	Positive	2.69	.850	23.92	.000*
	Negative	2.41	.891		

*$p<0.05$

When table values are examined, it is seen that the scores students got from positive (μ =2.69) and negative items (μ = 2.41) differentiate significantly and this difference is in favor of the positive items (t= 23.92, $p<0.01$). In other words, students got higher scores from the positive items compared to the negative items. Cohen's d was calculated with the values obtained from the t-test result (d=0.32). It is seen that the value obtained from the analysis results has a

medium size effect. In the study, six different measurement model were created for the second research question. Goodness of fit indices obtained from the analyses are presented in Table 4.

Table 4. *Goodness of Fit Index Results for General Sample.*

	df	χ^2	RMSEA	CFI	TLI	SRMR
Model 1	27	4931.74	0.178	0.73	0.63	0.098
Model 2	21	2049.29	0.130	0.89	0.81	0.371
Model 3	25	2261.62	0.125	0.87	0.82	0.372
Model 4	24	1242.16	0.094	0.93	0.90	0.168
Model 5	21	631.97	0.071	0.96	0.94	0.036
Model 6	17	168.13	0.039	0.99	0.98	0.012

It seems that the data do not fit the single-factor structure (Model 1) for this model (χ^2 = 4931.74; RMSEA= 0.178; CFI= 0.73; TLI=0.63; SRMR=0.098). The results of the bifactor model, the model which was created second, fit the data better than the previous model (χ^2 = 2049.29; RMSEA= 0.130; CFI= 0.89; TLI=0.81; SRMR=0.371). However, the obtained values are not in the desired range for perfect fit. For Model 3, examination of the results revealed that the data fit the model better than the other models (χ^2 = 2261.62; RMSEA= 0.125; CFI= 0.87; TLI=0.82; SRMR=0.372). Model 4 has proven to fit the data best compared to previous models. (χ^2 = 1242.16; RMSEA= 0.094; CFI= 0.93; TLI=0.90; SRMR=0.168). Among Model 5 and Model 6, the model in which a correlation was established between the error terms of negative items (Model 6) showed the best fit (χ^2 = 168.13; RMSEA= 0.039; CFI= 0.99; TLI=0.98; SRMR=0.012).

Finally, considering the fit indices, the models that fit the values best were found to be Model 4 and Model 6. In both models, negative items were included in the measurement model. By adding negative items to the model, it can be said that negative items cause a wording effect as a result of obtaining the most suitable model for the data. Table 5 shows standardized factor loading values for each items obtained from the created models.

Table 5. *Standardized Factor Loading Values for General Sample.*

Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	.74	.66	.66		.66	.66	.89		.53	.82
y4	.71	.62	.63		.63	.63	.84		.48	.80
y6	.70	.63	.60		.63	.61	.82		.48	.78
y7	.69	.61	.61		.61	.61	.81		.46	.77
y2	.60	.70		.85	.74		.35	.75	.69	.40
y3	.73	.77		.13	.81		.58	.58	.77	.56
y5	.42	.59		.08	.61		.27	.60	.55	.22
y8	.67	.84		.02	.83		.48	.66	.79	.46
y9	.69	.87		-.03	.84		.52	.64	.80	.50

When the table values were examined, standardized factor loading values for Model 1 were predicted to be between .42 and .74. In Model 2, loading values for common factor were predicted to be between .59 and .87., and for negative and positive factors in Model 2, the factor loading values were predicted to be between .60 and .66 and between -.03 and .85, respectively. In Model 3 the factor loading values were predicted to be between .61 and .84 for general factor,

and between .61 and .66 for positive items. Loading values under common factor were predicted to be between .27 and .89 for Model 4, and factor loading values were predicted to be between .58 and .75 for negative items. In model 5 and 6, loading values for common factor were predicted to be between .46 and .80, .22 and .82, respectively.

As a result, according to the standardized factor load values, the results obtained from all models except Model 2 are at acceptable values. However, considering the fit indices, it can be said that the most suitable model for the data is Model 4 and Model 6. Among the measurement models, results of goodness of fit indices for the group of female and male students are presented in Table 6.

Table 6. *Goodness of Fit Index Results for Female and Male Students.*

Goodness of Fit Index Results for *Female* Students

	df	χ^2	RMSEA	CFI	TLI	SRMR
Model 1	27	1710.29	0.150	0.83	0.77	0.070
Model 2	21	1173.63	0.141	0.88	0.80	0.340
Model 3	25	1292.21	0.135	0.87	0.82	0.337
Model 4	24	775.94	0.106	0.92	0.89	0.166
Model 5	21	405.64	0.081	0.96	0.93	0.034
Model 6	17	113.94	0.045	0.99	0.98	0.012

Goodness of Fit Index Results for *Male* Students

	df	χ^2	RMSEA	CFI	TLI	SRMR
Model 1	27	3721.12	0.216	0.55	0.41	0.132
Model 2	21	1034.61	0.128	0.88	0.79	0.397
Model 3	25	1135.38	0.123	0.87	0.81	0.403
Model 4	24	558.10	0.087	0.94	0.90	0.173
Model 5	21	297.65	0.067	0.97	0.94	0.042
Model 6	17	64.45	0.031	0.99	0.99	0.012

Examining the table values for female students, the data does not seem to fit the single-factor structure model (χ^2 = 1710.29; RMSEA= 0.150; CFI= 0.83; TLI=0.77; SRMR=0.070). In the second model, the model fits the data better (χ^2 = 1173.63; RMSEA= 0.141; CFI= 0.88; TLI=0.80; SRMR=0.320). However, obtained values are not in the desired range for perfect fit. As the third model, examination of the results revealed that the data fit the model better than the other models (χ^2 = 1292.21; RMSEA= 0.135; CFI= 0.87; TLI=0.82; SRMR=0.316). In the next model, only negative items are included in the model as a factor and it is determined that it is the model that best fits the data compared to the previous models (χ^2 = 775.94; RMSEA= 0.106; CFI= 0.92; TLI=0.89; SRMR=0.161). Finally, between Model 5 and Model 6, the model in which a correlation was established among the error terms of negative items (Model 6) showed the best fit (χ^2 = 113.94; RMSEA= 0.045; CFI= 0.99; TLI=0.98; SRMR=0.012). Finally, considering the fit indices, the models that fit the values best were found to be Model 4 and Model 6. In both models, negative items were included in the measurement model. In this case, negative items in the scale items cause a wording effect in the subgroup consisting of female students.

Examining the table values for male students, the data does not seem to fit the single-factor structure model (χ^2 = 3721.11; RMSEA= 0.216; CFI= 0.55; TLI=0.41; SRMR=0.132). In the second model, the model fits the data better (χ^2 = 1034.61; RMSEA= 0.128; CFI= 0.88; TLI=0.79; SRMR=0.372). Fit indices are not in the acceptable range for both models. As the

results of the third model revealed ,the data fit the model better than the other models (χ^2 = 1135.39; RMSEA= 0.123; CFI= 0.87; TLI=0.81; SRMR=0.379). In the next model, it is determined that it is the model that best fits the data compared to the previous models (χ^2 = 558.10; RMSEA= 0.087; CFI= 0.94; TLI=0.90; SRMR=0.163). Finally, between Model 5 and Model 6, the model in which a correlation was established among the error terms of negative items (Model 6) showed the best fit (χ^2 = 64.45; RMSEA= 0.031; CFI= 0.99; TLI=0.99; SRMR=0.012). As a result, it is seen that Model 4 and Model 6 are the measurement models that show best fit, similar to the result obtained for the general sample and the subgroup of female students. In both models, negative items were included in the measurement model. In this case, negative items in the scale items cause a wording effect on the subgroup of male students. Table 7 shows standardized factor loading values for each items obtained from the measurement models created for female and male students.

Table 7. *Standardized Factor Loading Values for Female and Male Students.*

Standardized Factor Loading Values for Female Students

Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	0.78	0.66	0.66		0.67	0.67	0.89		0.63	0.84
y4	0.74	0.67	0.58		0.67	0.58	0.84		0.59	0.81
y6	0.73	0.66	0.56		0.66	0.55	0.81		0.58	0.78
y7	0.71	0.64	0.59		0.64	0.58	0.81		0.56	0.77
y2	0.65	0.75		0.83	0.75		0.44	0.71	0.71	0.51
y3	0.77	0.78		0.09	0.81		0.69	0.46	0.78	0.67
y5	0.48	0.60		0.03	0.61		0.35	0.53	0.56	0.34
y8	0.70	0.85		-0.02	0.84		0.55	0.63	0.80	0.55
y9	0.73	0.88		-0.07	0.85		0.60	0.59	0.81	0.59

Standardized Factor Loading Values for *Male* Students

Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	0.75	0.66	0.66		0.66	0.66	0.88		0.42	0.80
y4	0.73	0.60	0.66		0.60	0.66	0.84		0.37	0.80
y6	0.74	0.61	0.64		0.61	0.64	0.83		0.39	0.79
y7	0.71	0.59	0.64		0.59	0.63	0.80		0.37	0.76
y2	0.47	0.61		0.88	0.73		0.26	0.77	0.68	0.29
y3	0.62	0.77		0.18	0.82		0.47	0.66	0.77	0.45
y5	0.30	0.57		0.14	0.60		0.12	0.63	0.54	0.11
y8	0.57	0.82		0.08	0.82		0.40	0.69	0.77	0.38
y9	0.59	0.86		0.03	0.83		0.44	0.68	0.78	0.41

When the table values were examined for female students, standardized factor loading values for Model 1 were predicted to be between .48 and .78. In Model 2, loading values for common factor were predicted to be between .60 and .88. In Model 2, for positive and negative factors, the factor loading values were predicted to be between .56 and .66 and between -.07 and .83, respectively. In Model 3, loading values under common factor were predicted to be between .61 and .85, and between .55 and .67 for positive items. Loading values under common factor were predicted to be between .35 and .89 for Model 4, and factor loading values were predicted

to be between .46 and .71 for negative items. In Model 5 and Model 6, where correlations between errors were included in the model, factor loading values were predicted to be between .56 and .81 and between .34 and .84, respectively.

When the table values were examined for male students, standardized factor loading values for Model 1 were predicted to be between .30 and .75. In Model 2, loading values for common factor were predicted to be between .57 and .86. When positive and negative items were taken as factor, the factor loading values were predicted to be between .64 and .66 and between .03 and .88, respectively. Loading values under common factor were predicted to be between .59 and .83 for Model 3, and between .63 and .66 for positive items. Finally, loading values under common factor were predicted to be between .12 and .88 for Model 4, and factor loading values were predicted to be between .63 and .77 for negative items. In Model 5 and Model 6, where correlations between error terms were included in the model, standardized factor loading values were predicted to be between .37 and .78 and between .11 and .88, respectively.

As a result, for both samples, according to the standardized factor loading values, the results obtained from all models except Model 2 are within acceptable range. However, when evaluated together with the fit indices, it can be said that the most suitable model for the data is Model 4 and Model 6.

4. DISCUSSION and CONCLUSION

The aim of this study is to examine whether or not the positive and negative items in the Mathematical Self-Confidence Scale employed in TIMSS 2015 cause a wording effect. For this purpose, in addition to the general sample, subgroups of male and female students were examined separately. Based on the study, it was determined that there was a significant difference between the scores students got from the positive items and the scores they got from the negative items. The mean of the students from the negative items is lower than the mean they got from the positive items. Second, it was determined that the measurement models that best fit the data were the models incorporating the method factor for negative items (Model 4 and Model 6). Although negative items are considered as a separate factor in both models, Model 6 gives better results than Model 4. This may be due to the fact that CU models that allow residuals to be correlated consider not only the variance associated with the wording effect, but also unknown factors (Wu et al., 2017). However, Model 4 can be accepted as the measurement model for the relevant scale since it is easy to interpret (Wu, 2008). In conclusion, negative items for both the general sample and the groups of female and male students in this study cause a method factor in the respondents. The method factor generally represents the “nuance” variance that is not desired in the observed output related to the way the information is collected, rather than the variance intended to be measured (Maul, 2013).

In this study, it is seen that the mean of positive items is higher than the average of negative items because students do not agree more with negative items than positive items. In other words, while the students did not give negative responses such as "I partially disagree" or "I completely disagree" to the negative items; they give positive responses to positive items such as "I partially agree" or "I completely agree". One reason students prefer to respond less to negative items may be "social desire". Social desirability refers to the tendency of the participants to give socially desired answers instead of choosing answers that reflect their true emotions (Grimm, 2010). For example, "I usually do well in mathematics" is the item with the highest average ($\mu = 2.96$) and most of the students answered this item as "I partially agree" or "Strongly agree". However, the item with the lowest average in the scale is "Mathematics is harder for me than any other subject "$\mu = 2.17$). The students agreed with this item at a moderate level compared to the previous sample item. There are studies in the literature on the data obtained from TIMSS conducted in different years. Similarly, Marsh (1986) found that younger

students and students with poor reading skills could not respond appropriately to the negative items in the rating scales. As a result, it can be said that expressing negative items requires special attention, especially for students in the younger age group, and scale items should be formed with simpler expressions rather than a long and complex structure.

There are similar studies on TIMSS scales, in which positive and negative items cause a wording effect (Hooper et al., 2013; Wang et al., 2018). In their analysis on the Mathematical Self-Confidence Scale administered in TIMSS 2011, Hooper et al. (2013) put forth that there are differences in terms of psychometric properties between positive and negative items. Confirmatory Factor Analysis was adopted in this study for analysis. In their study, they stated that the model fit indices recorded a remarkable increase when correlations were established between the error terms of negative items in both fourth-grade data and eighth-grade data, which can be argued to cause a wording effect for the negative items in the scale. In another study carried out on the same scale, Wang et al., (2018) investigated the presence of wording effect through multi-level models in which students were divided into classes. As a result of this study, it was determined that there are both intra-level and inter-level wording effects in scale items. The results of both studies are similar to this study. In this study, bifactor models and the Mathematical Self-Confidence Scale administered in TIMSS 2015 were examined and it was determined that negative items caused a wording effect. Recent studies show that bifactor models are frequently used in determining the wording effect (Hyland et al., 2014; Wang et al., 2015).

Another finding obtained from the study is that the same measurement model was used both for the general sample and for the groups created only for female students or only for male students. In each of the three samples, the best result was obtained when the correlations between error terms of negative items were included in the model. Similar findings have been found in the literature (DiStefano & Motl, 2009; McLarty et al., 1989). In their study, DiStefano and Motl (2009) examined whether the wording effect differs by gender based on the Rosenberg Self-Esteem Scale (RSES) items. The study showed that there is a method factor for the items worded negatively in the RSES scale for both men and women, but this effect does not differ by gender.

As a result, this study examined whether the positive/negative items in the scale items cause the method factor, and whether the structure contains a method factor for female students and male students as well as for the general sample. This study can also be carried out with the data of English-speaking or non-English-speaking students or students in different countries speaking different languages. Similarly, it can be determined whether the scale items cause a wording effect based on different age groups. In addition, in the scale development process, negative items can be included by considering the group to which the scale will be administered. Similarly, if a scale is to be adapted, it can be examined whether the positive/negative items cause a wording effect and analyses can be made based on the appropriate measurement model.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Esra Oyar: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis and Writing, Supervision. **Hakan Yavuz Atar:** Methodology, Visualization, Supervision and Validation.

ORCID
Esra OYAR https://orcid.org/0000-0002-4337-7815
Hakan Yavuz ATAR https://orcid.org/0000-0001-5372-1926

5. REFERENCES

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231-240. https://doi.org/10.1111/j.1745-3984.1985.tb01061.x

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

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabi: İstatistik, araştirma deseni, SPSS uygulamaları ve yorum* (16. Baskı). Pegem Akademi. [Handbook of data analysis for social sciences: Statistics, research design, SPSS practice and interpretation (16. Edition). Pegem Academy].

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2017). *Bilimsel araştırma yöntemleri*. Pegem Yayıncılık.

Chen, Y. (2017). On the ımpact of negatıvely keyed ıtems on the assessment of the unıdımensıonalıty of psychologıcal tests and measures. [Doctoral dissertation, The University of British Columbia]. ProQuest Dissertations and Theses.

Chen, Y. H., Rendina-Gobioff, G., & Dedrick, R. F. (2010). Factorial invariance of a Chinese self-esteem scale for third and sixth grade students: evaluating method effects associated with positively and negatively worded items. *The International Journal of Educational and Psychological Assessment, 6* (1), 21-35.

Chen, F. F., & Zhang, Z. (2018). Bifactor models in psychometric test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 325–345). John Wiley, Sons Ltd.

Cronbach, L. J. (1984). *Essentials of psychological testing (4th edition).* Harper & Row.

Cohen, J. (1988). *Statistical power analysis for the behavioral* sciences *(2nd edition).* Lawrence Erlbaum.

DeVellis, R. F. (2003). *Scale development: Theory and applications (2nd edition).* Sage

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*, 440-464. https://doi.org/10.1207/s15328007sem1303_6

DiStefano, C. & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded ıtems: Investigating factorial ınvariance by sex. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 134-146. https://doi.org/10.1080/10705510802565403

Erkuş A. (2003). *Psikometri üzerine yazılar. (1. baskı).* Türk Psikologlar Derneği Yayınları.

Ford, L. R., & Scandura, T. A. (2018). A typology of threats to construct validity in item generation. *American Journal of Management, 18*(2). https://doi.org/10.33423/ajm.v18i2.298

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S.P. (2003). Item-wording and the dimensionality of the rosenberg self-esteem scale: Do they matter?. *Personality and Individual Differences, 35*(2003), 1241 1254. https://doi.org/10.1016/S0191-8869(02)00331-8

Grimm, P. (2010). *Social desirability bias.* Wiley International Encyclopedia of Marketing. Hoboken, Wiley.

Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the core self-evaluation scale (CSES): A bi-factor perspective. *Personality and Individual Differences, 83*, 142-147. https://doi.org/10.1016/j.paid.2015.04.006

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the job diagnostic survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468. https://doi.org/10.1037/0021-9010.70.3.461

Hooper, M., Arora, A., Martin, M. O., & Mullis, I. V. S., (2013, June*). Examining the behavior of "reverse directional" items in the TIMSS 2011 context questionnaire scales*. Paper Presented at the 5th IEA International Research Conference. National Institute of Education, Nanyang Technological University, Singapore.

Horan, P. M. , DiStefano, C. & Motl, R. W. (2003) Wording effects in self-esteem scales: methodological artifact or response style?. *Structural Equation Modeling, 10*(3), 435-455. https://doi.org/10.1207/S15328007SEM1003_6

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

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self Esteem Scale. *Personality and Individual Differences, 66*, 188-192. https://doi.org/10.1016/j.paid.2014.03.034

Ibrahim, A.M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports*, *88*, 497–500. https://doi.org/10.2466/pr0.2001.88.2.497

Kirk, R. (2007). *Statistics: an introduction.* Nelson Education.

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of personality assessment, 94*(2), 196-204. https://doi.org/10.1080/00223891.2011.645936

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37-49.

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors?. *Journal of Personality and Social Psychology, 70*, 810-819. https://doi.org/10.1037/0022-3514.70.4.810

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*, 169. https://doi.org/10.3389/fpsyg.2013.00169

McLarty, J. R., Noble, A. C., & Huntley, R. M. (1989). Effects of item wording on sex bias. *Journal of Educational Measurement, 26*(3), 285-293. https://doi.org/10.1111/j.1745-3984.1989.tb00334.x

Michaelides, M. P. (2019). Negative keying effects in the factor structure of TIMSS 2011 motivation scales and associations with reading achievement. *Applied Measurement in Education, 32*(4), 365-378. https://doi.org/10.1080/08957347.2019.1660349

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the rosenberg self-esteem scale. *Personality and Individual Differences, 97*, 13-18. https://doi.org/10.1016/j.paid.2016.03.011

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47* (5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment, 92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Schmitt, N., & Stuits, D.M. (1985). Factors defined by negatively keyed items: The result of careless respondents?. *Applied Psychological Measurement, 9*, 367-373. https://doi.org/10.1177/014662168500900405

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51*(1), 67-78. https://doi.org/10.1177/0013164491511005

Tomas, J. M. & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 84-98. https://doi.org/10.1080/10705519909540120

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement, 75*(1), 157-178. https://doi.org/10.1177/0013164414528209

Wang, Y., Kim, E. S., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and psychological measurement, 78*(2), 253-271. https://doi.org/10.1177/0013164417690858

Weems, G.H., Onwuegbuzie, A.J., & Collins, K.M.T. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation & Research in Education, 19*(1), 3-20. https://doi.org/10.1080/09500790608668322

Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. (2003). Profiles of respondents who respond inconsistently to positively-and negatively-worded items on rating scales. *Evaluation & Research in Education, 17*(1), 45-60. https://doi.org/10.1080/14664200308668290

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*, 320–334. https://doi.org/10.1037/a0032121

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factory analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189–194. https://doi.org/10.1007/s10862-005-9004-7

Wu, C. H. (2008). an examination of the wording effect in the rosenberg self-esteem scale among culturally chinese people. *The Journal of Social Psychology, 148* (5), 535-552. https://doi.org/10.3200/SOCP.148.5.535-552

Wu, Y., Zuo, B., Wen, F., & Yan, L. (2017). Rosenberg self-esteem scale: Method effects, factorial structure and scale invariance across migrant child and urban child populations in China. *Journal of personality assessment, 99*(1), 83-93. https://doi.org/10.1080/00223891.2016.1217420

Yang, Y., Chen, Y. H., Lo, W. J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment, 30*(5), 509-519. https://doi.org/10.1177/0734282911435461

6. APPENDIX

Table A1. *Items in the Mathematics Self- Confidence Scale.*

Codes	Items - English	Items - Turkish
BSBM19A	I usually do well in mathematics	Matematikte genellikle iyiyimdir.
BSBM19B	Mathematics is more difficult for me than for many of my classmates*	Matematik birçok sınıf arkadaşıma göre bana daha zor gelir.*
BSBM19C	Mathematics is not one of my strengths*	Matematik başarılı olduğum alanlardan biri değildir. *
BSBM19D	I learn things quickly in mathematics	Matematik konularını hızlı öğrenirim.
BSBM19E	Mathematics makes me nervous*	Matematik beni gerginleştirir/endişelendirir.*
BSBM19F	I am good at working out difficult mathematics problems	Zor matematik problemleri çözmekte iyiyimdir.
BSBM19G	My teacher tells me I am good at mathematics	Öğretmenim matematikte iyi olduğumu söyler.
BSBM19H	Mathematics is harder for me than any other subject*	Matematik benim için diğer alanlardan daha zordur.*
BSBM19I	Mathematics makes me confused*	Matematik benim kafamı karıştırır.*

*Reverse scored items.