Kastamonu Education Journal, 2021, Vol. 29, No:1, 201-209
doi: 10.24106/kefdergi.708968

| Research Article / Araştırma Makalesi |

Examining Scale Items in Terms of Method Effects Based on the Bifactor Item Response Theory Model

Ölçek İfadelerinin Ölçme Yöntemi Etkisi Açısından İki Faktör Modeline Dayalı Olarak İncelenmesi

Seval Kula Kartal[1]

Keywords
1.Method effects
2.Model comparison
3.The bifactor model

Anahtar Kelimeler
1.Ölçme yöntemi etkisi
2.Model karşılaştırma
3.İki faktör modeli

Received/Başvuru Tarihi
25.03.2020

Accepted / Kabul Tarihi
02.09.2020

Abstract

Purpose: The current study aims to apply a one-dimensional (the graded response model) and a multidimensional (the bifactor model) item response theory model to evaluate the presence of method effects on the data obtaining from the administration of the scale measuring emotional school engagement of students as a part of 2015 Programme for International Student Assessment (PISA).

Design/Methodology/Approach: The model data fits of the one-dimensional and bifactor model conceptualizations were compared on the data obtained from a large sample of Turkey school children.

Findings: The item parameters and model data fit statistics provided evidences for that the school engagement scale of the PISA 2015 measures a primary construct (the emotional school engagement) with two nuisance factors (the negative and positive wording effects).

Highlights: The results of the present study supported the use of bifactor model in evaluating the presence of method effects. Therefore, researchers using the Emotional School Engagement Scale of PISA are recommended to utilize more sophisticated statistical techniques such as the bifactor item response theory model.

Öz

Çalışmanın amacı: Bu araştırmanın amacı, PISA 2015 Okula Yönelik Duygusal Bağlılık Ölçeği'nin uygulanmasından elde edilen veri üzerinde ölçme yöntemi etkisinin tek boyutlu (aşamalı tepki modeli) ve çok boyutlu (iki faktör modeli) madde tepki kuramı modellerine dayalı olarak incelenmesidir

Materyal ve Yöntem: Tek boyutlu ve iki faktör modellerinin model veri uyumları Türkiye öğrenci örnekleminden elde edilen veri üzerinde karşılaştırılmıştır.

Bulgular: Madde parametreleri ve model veri uyumu istatistikleri, PISA 2015 Okula Yönelik Duygusal Bağlılık Ölçeği'nin bir baskın boyutun (duygusal bağlılık) yanında biri olumlu diğeri olumsuz maddelerden oluşan iki boyutu daha ölçtüğüne ilişkin bulgular sağlamıştır.

Önemli Vurgular: Bu araştırmanın bulguları ölçme yöntemi etkisinin analizinde iki faktör modelinin kullanılmasını desteklemiştir. Bu nedenle, PISA Okula Yönelik Duygusal Bağlılık Ölçeği'ni kullanacak araştırmacılara iki faktör modeli gibi daha ileri istatistiksel teknikler kullanmaları önerilmiştir.

[1] Pamukkale University, Faculty of Education, Measurement and Evaluation in Education, Denizli, TURKEY; https://orcid.org/0000-0002-3018-6972

Citation/Alıntı: Kula Kartal, S. (2021). Examining Scale Items in Terms of Method Effects Based on the Bifactor Item Response Theory Model. *Kastamonu Education Journal, 29*(1), 201-209. doi: 10.24106/kefdergi.708968

INTRODUCTION

Behavioral sciences mostly deal with complex psychological constructs that cannot be observed by researchers directly. Therefore, researchers frequently utilize self-report surveys to observe individuals' behaviors and get information regarding the traits of interest, such as personality, attitude, or motivation. Both negatively and positively worded items are commonly used in surveys because of concerns with preventing response bias. Paulhus (1991) defines response bias as systematic tendency to respond to scale items on some basis other than the item content. Inclusion of negatively and positively worded items on a survey might cause response bias, since respondents' answers to survey items might be affected by wording direction. If wording direction of the item systematically affects responses, then the data might be contaminated by a method effect that is used to collect the information (DiStefano & Motl, 2009).

Using negatively and positively worded items on the same survey is based on the assumption that respondents answer negatively and positively worded items in the same way. For example, Tomas and Oliver (1999) utilized the Rosenberg Self-Esteem Scale and found that all of the models positing method effect provided better fit than the uni-dimensional model. The study revealed the presence of method effects associated with item wording, especially for negatively worded items. Wang, Siegal, Falck, and Carlson (2001) investigated generalizability of Tomas and Oliver's (1999) conclusions on an untraditional population. Similarly, they found that wording direction of items affect respondents' answers to the scale items. Horan, Distefeno and Motl's (2003) study reported better fit when wording effects on the scale were taken into consideration. DiStefeno and Motl (2006) revealed that wording effect associated with negatively worded items on the scale were present. Supportively, Supple and Plunkett (2011) concluded that wording direction of items affect respondents' answers to the items of the Rosenberg Self-Esteem Scale

In addition, researches examined whether method effects could be substantiated within other scales measuring different contents. For example, Ye (2009) analyzed the factor structure of the General Health Questionnaire, and the findings of the study pointed out the presence of a method factor associated with negatively worded items. Gu, Wen, and Fan (2015) examined method effects on the Core Self-Evaluation Scale, and the findings evidenced the existence of a method factor contributing to the variance among negatively worded items. Wouters, Booysen, Ponnet, and Loon (2012) investigated the method effect on the items of the Hospital Anxiety and Depression Scale. Researchers revealed that the addition of a method factor improved model data fit in all tested models. Wang, Kim, Dedrick, Ferron, and Tan (2018) utilized the Students Confident in Mathematics Scale of 2011 The Trends in International Mathematics and Science Study (TIMSS), and found that the models taking into consideration both positive and negative wording effects provided the best fit statistics.

Method effects has been mostly analyzed by utilizing confirmatory factor analysis framework, based on either correlated traits-correlated methods (CTCM) or correlated traits-correlated uniqueness (CTCU) methods (Horan et al., 2003; Supple & Plunkett, 2011; Tomas & Oliver, 1999; Wang et al., 2001). In the CTCM framework, the method effect is accepted as a distinct latent trait and incorporated into the analysis as a separate factor. In the CTCU framework, it is accepted as irrelevant variance and represented by correlating error terms among similarly worded items. Both methods have their own strengths or weaknesses. The main problems with the CTCM are the ill-defined solutions and the confounding of method variance with trait variance. Although the CTCU overcomes these problems, it does not allow the method effects to be examined as a distinct factor (DiStefeno & Motl, 2006; Kumar & Dillon, 1992; Tomas & Oliver, 1999). Both methods provide practical ways to model method effects; however, they are both limited in examining the strength of the method effects.

A more recent method, the bifactor model, provides an alternative way to examine method effects. In a bifactor model, covariance among a set of item responses can be accounted for by a single general factor reflecting the common variance among all scale items and group factors reflecting additional common variance among items (Reise, 2012). Therefore, the bifactor model allows modeling one or two group factors representing method effects caused by negatively and positively worded items, in addition to the general factor underlying all scale items. Since group factors are orthogonal to the general factor and each others (DeMars, 2006), contributions of each variance source to the explained variance could be easily examined. This feature of bifactor modeling can be valuable in the development of scales including negatively and positively worded items and analyzing contributions of the substantive construct and wording effect (Wang et al., 2018).

Few studies utilizing the bifactor modeling reveal its superiority on simultaneously modeling the primary construct of interest and method effects. For example, Wang et al. (2018) analyzed the fit of various bifactor models that take into consideration method effects associated with only negatively worded items, only positively worded items, or both. The findings of the study revealed that the bifactor model with the positive and negative wording effects factors both at the within and between levels provided the best fit statistics. Researchers recommended evaluating independent contributions of the primary construct and method effects to tease out wording effects from the measurement of the primary construct. The findings of another study, conducted by Gu et al. (2015), evidenced the presence of a method factor contributing to the variance among the negatively worded items over and above the general self-esteem factor. In addition, researchers revealed that the method effect caused inaccurate estimation of reliability and criterion-validity coefficients. McKay, Boduszek, and Harvey (2014) tested a one-factor model, a two-factor model, a hierarchical model and a bifactor model on the data obtained from the administration of the Rosenberg Self-Esteem Scale. The study showed that the bifactor structure, reflecting a primary construct with two nuisance factors, provided a better fit to the data than the alternative solutions. In line with the related studies, Hyland, Boduszek, Dhingra,

Shevlin, and Egan (2014) studied on the data obtained from the administration of the Self-Esteem Scale, and observed improvements across all fit indices for the bifactor solution.

The current study builds on aforementioned researches, but it extends them in different ways. The related studies mostly focus on examining the wording effect on scales measuring self-esteem (DiStefeno & Motl, 2006; Horan et al., 2003; Supple & Plunkett, 2011; Tomas & Oliver, 1999; Wang et al., 2001). However, it is necessary to examine the presence of method effects within other scales measuring different contents. Along this line of research, the major purpose of the current study is to examine the presence of method effects on the data obtaining from the administration of the scale measuring school engagement of students as a part of 2015 Programme for International Student Assessment (PISA). The scale includes three negatively worded items, and three positively worded items. School engagement is widely accepted as a multidimensional psychological construct comprising of multiple components, such as behavioral, emotional, and cognitive engagement (Appleton, Christenson, Kim & Reschly, 2006; Fredericks, Blumenfield & Paris, 2004). However, the scale that is utilized in PISA focuses on measuring only the emotional component of school engagement. Emotional engagement is associated with the positive or negative reactions towards school, and feeling a sense of belongingness to the school and classroom (Archambault, Janosz, Fallu & Bagani, 2009; Craft and Capraro, 2017). One of the reasons of selecting the emotional engagement scale of 2015 PISA is that the presence of method effects might be observed more easily, since the scale measures only one aspect of a more complex construct. Since the construct targeted with the scale does not have a complex-structure, method effects is more likely to be the reason of any multidimensionality that might be observed in the data. Another reason of selecting the emotional engagement scale is that it is the only scale containing both negatively and positively worded items. In 2015 PISA, only negatively or positively worded items were used to measure other affective variables, such as attitude towards science, relationships with classmates and teacher.

In addition, findings of the studies pointed out the superiority of the bifactor structure in modeling the primary construct targeted with the scale and potential method effects confounding with the construct of interest. The examination of the related literature (Gu et al., 2015; Hyland et al., 2014; McKay et al., 2014) revealed that researchers conducted bifactor analyses based on the confirmatory factor analysis framework to investigate method effects. As a second contribution to the literature, the current study investigates method effects by conducting bifactor analysis based on the item response theory (IRT) framework. In the current study, the IRT framework was preferred over the confirmatory factor analysis framework for several reasons. Firstly, factor analytic methods use total scores and correlation matrices to provide information regarding the data structure. However, the IRT uses more information provided by the data, since it allows utilizing the whole information obtained from the response patterns of respondents (Li, Jiao, & Lissitz, 2012; Thissen & Wainer, 2001).

Secondly, the IRT can provide valuable information for analysis of method effects. Researches showed that the direction of wording might affect item response patterns of the respondents (Ray, Frick, Thornton, Steinberg & Cauffman, 2016; Weems, Onwuegbuzie, Schreiber, & Eggers, 2003). Therefore, examining whether distributions of item endorsement for negatively and positively worded items differ might provide an evidence for the presence of method effects. Confirmatory factor analysis framework is limited in giving information on how response categories of negatively and positively worded items function. The second reason of applying the IRT to examine method effects is that the IRT item parameters can describe the relationships among response categories of items and the latent trait. These strengths of the IRT framework might enable making more informed decisions regarding the functioning of scale items and the presence of method effects.

In sum, the inclusion of positively and negatively worded items on the scale might cause method effects variance that might lead to the inaccurate decisions regarding the factorial structure of scales. It is essential to take into account potential method effects, and utilize appropriate methods that can provide enriched information when the data is analyzed in terms of the presence of method effects. Therefore, the aim of this study is twofold. Firstly, the current study aims to provide an insight as to how the item response theory framework adds to analysis of method effects. Second aim of the present study is to apply a one-dimensional (graded response model) and a multidimensional (the bifactor graded response model) item response theory model to evaluate the presence of method effects.

METHOD/MATERIALS

Design

The current study is a fundamental research aiming to investigate the presence of method effects on the research data by comparing parameter estimations and model data fit statistics of a unidimensional and a multidimensional item response theory model (Karasar, 2005).

Participants

The study group of the current research was obtained from the PISA 2015 Turkey sample. The Turkish sample of PISA 2015 included 5,895 students from 231 different schools. Some of the students were excluded from the study group since they had missing information on the variables examined in this study. After excluding some of the students because of missing data, the study group consisted of 5,609 students. The distribution of students according to their gender and grade levels is presented in Table 1.

Table 1. The distribution of study group according to the gender and grade variables

	Gender		
Items	Female	Male	Total
7th grade	6	9	15
8th grade	36	50	86
9th grade	466	732	1,198
10th grade	2,189	1,937	4,126
11th grade	115	62	177
12th grade	4	3	7
Total	2,816	2,793	5,609

Instruments

The research data was obtained from answers provided by students to the Students Questionnaire of PISA 2015. The Students Questionnaire contained questions measuring non-cognitive outcomes like attitudes, beliefs, motivation, and learning-related behaviour, such as invested learning time. Students' emotional engagement to their schools is another non-cognitive variable addressed within the questionnaire. The questionnaire includes 3 negatively and 3 positively worded items measuring students' sense of belonging at school based on their reports about their feelings of social connectedness at school. Students give answers to the items on a four-point Likert scale. PISA 2015 asked students to report whether they "strongly agree", "agree", "disagree" or "strongly disagree" that they feel like an outsider or left out of things, that they make friends easily, that they feel like they belong, that they feel awkward and out of place, that other students seem to like them, or that they feel lonely (Organization for Economic Co-operation and Development [OECD], 2013).

Data Analysis

The analyses were all conducted after negatively worded items were reverse coded so that high scores indicate high level of school engagement. The research data was analyzed based on the graded response model (GRM) and the bifactor graded response model (B-GRM). The GRM models a nonlinear relationship between the probability of selecting a particular response category and latent trait of interest. The model computes category response probabilities based on a two-step process. Firstly, operating characteristic curves that represent the probability of an examinee responding in or above a particular response category are computed. Secondly, adjacent operating characteristic curves are subtracted to obtain probability of responding in a particular response category. In the GRM, one slope parameter is estimated for each item. The number of the threshold parameters for each item is equal to the number of the response categories minus 1. Slope parameters provide information for the discriminating power of the item (de Ayala, 2009; Embretson & Reise, 2000). Items with the greater slope parameters and narrower range of threshold parameters provide better discrimination among respondents (Fletcher & Hattie, 2004). Therefore, in the current study, scale items were compared based on the slope and threshold parameters provided by the GRM and B-GRM.

The bifactor model is a parametric model enabling modeling multilevel and multidimensional data structure. The bifactor model parameters can be estimated based on two different frameworks: structural equation modeling or item response theory modeling (Berkeljon, 2012; Brown, 2006; Qinn, 2014; Thissen & Wainer, 2001). As explained before, in the current study, item response theory framework was preferred for the analysis of method effects. The bifactor model (confirmatory) constraints that each item loads on the general dimension and on only one of the sub-dimensions. The general dimension and the sub-dimensions are uncorrelated with each other. The general dimension usually reflects the targeted trait with the measurement tool and accounts for the commonality among all of the items. The sub-dimensions explain the commonality among items belonging to the same subset of items. In the B-GRM, two slope parameters are estimated for each item; one is on the general dimension and the other is on the one of the sub-dimensions. The slopes reflect the degree to which items are good indicators of the general dimension, or sub-dimensions. Furthermore, intercept parameters are estimated for the item response categories. Similar with the unidimensional graded response model, item intercept parameters reflect the location on the latent trait where the probability of selecting a particular response category or higher is 0.50. However, in the bifactor model context, the threshold parameters are estimated on a scale defined by composite of the latent traits in the model (Berkeljon, 2012; Chen, West, & Sousa 2006; Immekus & Imbrie, 2008; Qinn, 2014).

Item response frequencies, threshold parameters estimated by the GRM, and the category characteristic curves were examined to evaluate functioning of negatively and positively worded items. Another method of evaluation of method effects is to compare model data fits of the models. Comparisons of item response theory models in terms of the general model data fit statistics were done based on the likelihood ratio test, Akaike (AIC), the Bayesian (BIC) and the adjusted Bayesian (A-BIC) information criteria. It was accepted that the model providing the lowest fit statistics is the model that best fits to the data among compared models (De Ayala, 2009; Li et al., 2012). The third method of evaluation is to examine item slope parameters estimated on the sub-dimensions by the B-GRM. High slope parameters both on the general dimension and sub-dimensions were accepted as further evidence for the presence of multidimensionality caused by method effects. The "mirt" package (Chalmers, 2012) on the R program was utilized to estimate fit statistics and item parameters for both models.

FINDINGS

The item response frequencies and threshold parameters were estimated in order to examine whether negatively and positively worded items function equivalently. The response frequencies and threshold parameters are presented in Table 2.

Table 2. Item Response Frequencies and the Threshold Parameters

	Item Response Frequencies				The GRM threshold parameters		
Items	1	2	3	4	b_1	b_2	b_3
1	1799	1806	0860	1144	-1.02	-0.44	0.57
2	1169	2318	1413	0709	-1.57	-0.40	1.08
3	1144	2300	1413	0752	-1.48	-0.37	1.08
4	1604	1914	1048	1043	-1.09	-0.39	0.67
5	0847	2741	1449	0572	-1.82	-0.48	1.44
6	1799	1806	860	1144	-1.02	-0.44	0.57

According to the frequencies presented in Table 2, approximately 30 percent of the respondents endorsed with the lowest rating on the negatively worded items (1, 4, and 6). However, approximately 20 percent of the respondents obtained the lowest score on the positively worded items (2, 3, and 5). Item response distributions reveal that the negatively worded items had a higher frequency of endorsement at the lowest response category. Supportively, item threshold parameters estimated by the GRM indicate that the positively worded items required lower levels of school engagement to endorse the lowest response category compared with the negatively worded items. As stated by Reise et al. (2011), threshold parameters represent the location on the latent trait continuum where the probability of responding in a response category is 0.5. For example, using the scale item 1, the threshold parameter shows that the trait level necessary to be rated above response category 1 is -1.02. Accordingly, the positively worded items require a respondent to be lower than approximately 1.5 or 2 standard deviations below the mean to have 50% probability of responding in the lowest category. However, being lower than 1 standard deviation below the mean is enough to have 50% probability of responding in the lowest category for the negatively worded items. Therefore, the negatively worded items seem more likely to be endorsed with the lowest rating.

Item frequencies reveal that approximately 12 percent of respondents endorsed with the highest rating on the positively worded items, while 18 percent of respondents obtained the highest score on the negatively worded items. Threshold parameters estimated for the highest response category vary between 1.08 and 1.44 for the positively worded items, between 0.55 and 0.67 for the negatively worded items. The GRM predicted that the positively worded items required higher levels of school engagement to endorse the highest response category compared with the negatively worded items. The negatively worded items are more likely to be endorsed with the highest rating. Interestingly, respondents tend to choose middle categories on the positively worded items more than they do on the negatively worded items. The item characteristic curves of the scale items supported this finding. There is not much utility in the middle categories of the negatively worded items (category 2 and 3), because their curves overlap quite a bit with the curves of category 1 and 4. There is fairly limited ability range in which “disagree” or “agree” responses are more likely to be preferred. However, middle categories of the positively worded items function better compared with the negatively worded items.

The threshold parameters estimated by the GRM indicate that the negatively and positively items function differently on the basis of the response patterns of respondents. The differences observed in the responses given to the negatively and positively worded items provide evidence to the potential presence of method effects. However, it is essential to reveal whether method effects causing different item functioning can be treated as nuisance or not. To examine the strength of method effects, the general model data fit statistics, and item slope parameters were estimated. Item slope parameters are presented in Table 3.

Table 3. Item Slope Parameters

Items	a_{GRM}	a_{GEN}	a_{NW}	a_{PW}
1	2.27	2.25	1.29	
2	2.38	2.33	1.14	
3	2.82	3.19	1.88	
4	1.87	2.35		1.84
5	1.97	2.27		1.41
6	1.76	1.90		1.30

a_{GRM}= item slope parameter estimated by the graded response model, a_{GEN}= item slope parameters estimated on the general dimension, a_{NW} and a_{PW}= item slope parameters estimated on the dimensions of negative and positive wording direction, respectively.

According to parameters given in Table 3, item slope parameters estimated by the GRM are quite high. In addition, the slope parameters provided by the B-GRM indicated that scale items have high slope parameters on the sub-dimensions defined based on the wording direction of items. The high slope parameters estimated on the general dimension and sub-dimensions (positive or negative wording direction) by the B-GRM revealed the presence of multidimensionality caused by method effects. Findings

based on the item response frequencies, threshold and slope parameters indicated that the negatively and positively items function differently on the basis of the response patterns of respondents. However, it is essential to reveal whether method effects causing different item functioning can be treated as nuisance or not. The general model data fit statistics of the models were estimated to examine the strength of method effects. The fit statistics are presented in Table 4.

Table 4. General Model Data Fit Statistics

Models	-2 lnL	Number of Parameters	AIC	BIC	A-BIC
GRM	75296.3	24	75344.3	75503.5	75427.2
B-GRM	73606.3	30	73666.3	73865.2	73769.9

GRM= the Graded Response Model, B-GRM= the Bifactor Graded Response Model, lnL= Log-likelihood values off the models, AIC= Akaike Information Criterion, BIC= Bayesian Information Criterion, A-BIC= Adjusted Bayesian Information Criterion

The GRM is nested within the B-GRM because imposing the constraint that all items have a common slope parameter on the sub-dimensions on the B-GRM produces the GRM. Therefore, likelihood ratio test statistic (G^2) was calculated to evaluate relative fit of hierarchically related models based on the -2 lnL value given in Table 4. The ratio test revealed that at the instrument level, the B-GRM represented a significant (df=6, χ^2=1690, $p<0.05$) improvement in fit over the GRM. Using the more complex B-GRM model achieved 2% increase in the general model data fit. It is suggested to use various fit statistics when examining model data fit, since models with more parameters tend to fit data better than models with fewer parameters (De Ayala, 2009). Therefore, AIC, BIC, A-BIC statistics were also examined to evaluate fit of the models. Supportively, according to statistics given in Table 4, the B-GRM displayed improvement in fit over the GRM. The better fit of the B-GRM over the GRM at the instrument level, indicated that the B-GRM improved the data fit by taking into consideration the nuisance factors explained by the wording direction of items in addition to the common factor underlying the item responses.

DISCUSSION

An important goal of the current study was to test whether any multidimensionality caused by the wording direction of the scale items observed in the data. Several findings of the current study suggested that item response patterns of respondents differ according to the item orientations. Supportively, Weems et al. (2003) found that the scores from the positively worded scale and the negatively worded scale differed significantly. Similarly, Ray et al. (2016) found that the negatively and positively worded items of the Inventory of Callous-Unemotional Traits showed different distributions. Researchers evidenced that average score of the negatively worded items was significantly higher than the means for positive items. The negatively worded items had a much higher frequency of endorsement at the two highest levels indicating the targeted trait. However, researchers found that positively worded items were much more likely to be endorsed with the lowest response category. In accordance with the researchers' findings, the current study evidenced that differences in the wording directions of the school engagement items resulted in differences in response patterns of the respondents. However, in the current study, negatively worded items had more endorsement on the lowest or the highest response categories, while the middle response categories were preferred more frequently on the positive items.

In the current study, item parameter estimations revealed that the negatively worded items had a narrower range of threshold parameters. In addition, more respondents endorsed with the lowest or the highest rating on the negatively worded items, while they were more likely to choose middle categories on the positively worded items. The response frequencies and threshold parameters pointed out that the negatively worded items provided better discrimination among respondents. In accordance with the findings based on the response frequencies and threshold parameters, it was found that both item response theory models estimated higher slope parameters for the negatively worded items. This finding is in contrast to the findings of previous researches (Barnette, 2000; Cole, Turner, & Gitchel, 2018), which evidenced that positive items discriminate better among respondents.

Cole et al. (2018) found that positive items tended to have a higher estimated discrimination value than the negative items. The negatively worded items were less informative than the positively worded items, and positive items provided a more reliable estimate of the level of the targeted trait. Barnette (2000) found that negative items reduced the internal consistency reliability. As stated by Wells and Wallock (2003) items which discriminate well among respondents are desirable and will improve reliability. Therefore, similar to the findings presented by Cole et al. (2018), the researcher concluded that negative items did not provide a reliable estimate of the trait. Conversely, the findings of the current study indicated that negatively worded items had higher slope parameters than the positive items.

One possible explanation for contradictory results regarding the psychometric qualities of the negative items might be related with how item reversal has been achieved. In practice, item reversal is mostly achieved through item negation (Weijters & Baumgartner, 2012). That is, researchers generally insert "not" into the stem of the positively worded items to negate the items. However, previous research has suggested that using words with opposite meanings rather than negation is a better strategy to

create negatively worded items (Weijters & Baumgartner, 2012). Both studies (Barnette, 2000; Cole et al., 2018), in which negative items had lower slope parameters, used negation to create negative items. However, the three negative items of the scale used in the current study *("I feel like an outsider at school." "I feel awkward and out of place in my school." "I feel lonely at school."*) do not include any negations, and they are associated with greater item discrimination parameter estimates compared to the positive items. Supportively, Suárez-Álvarez, Pedrosa, Lozano, García-Cueto, Cuesta, and Muñiz (2018) avoided negations while creating the regular and the reversed forms of the items. Researchers found that slope parameters of the negative items are very close to the slope parameters of the positive items. Another possible explanation for the inconsistent findings regarding the discriminating power of the negative items might be related with the latent trait of interest. For example, Min, Zickar and Yenkov (2018) found that discriminating power of negative items compared to the positive items varied across different personality dimensions.

Model comparisons revealed that the bifactor model fitted to the data better than the unidimensional model, and improvements were observed for all of the fit statistics for the bifactor model. In addition, item parameter estimations indicated the superiority of the bifactor model over the unidimensional item response model. Overall, item slopes for each sub-dimension were comparatively weaker than those on the general dimension however it should be noted that slopes of items were not low on the sub-dimensions. The slope parameters on the two the wording effects dimensions varied between 1.14 and 1.84 (a_{mean}= 1.5). It has been suggested that items with slope parameters lower than 0.9 can be considered as poorly functioning, while items with slopes higher than 1.7 can be accepted as highly discriminatory (Baker, 2001; De Ayala, 2009). Therefore, further examination of the slope parameters for the sub-dimensions provided important evidence regarding the appropriateness of taking into account the two wording effects dimensions while modeling the data.

The results of the current study supported the supremacy of the school engagement general dimension, and the presence of two wording effects dimensions. In line with these findings, results of the related studies in which the bifactor model has been utilized suggested to include method factors in addition to the general factor to have better model data fit. For example, Hyland et al. (2014) revealed the necessity to consider positive and negative wording directions as important method factors when applying the self-esteem scale in research contexts. Supportively, Gu et al. (2015) found that the factor loadings of the negatively worded items on the group factor were almost the same as those on the general factor. Researchers concluded that there is a method factor contributing to the method variance among the items over and above the general self-evaluation factor.

CONCLUSION AND RECOMMENDATIONS

The inclusion of positively and negatively worded items on the scale might cause method effects variance that might lead to the inaccurate decisions regarding the factorial structure of scales. It is essential to take into account potential method effects, and utilize appropriate methods that can provide enriched information when the data is analyzed in terms of the presence of method effects. Therefore, the currents study aimed to analyze the presence of the method effect based on the item response theory framework. The current study was carried out in order to examine functioning of scale items having different wording directions, and to reveal the presence of method effects based on the bifactor item response theory model. The unidimensional and bifactor model conceptualizations were investigated within a large sample of Turkey school children. The item parameters and model data fit statistics provided evidences for that the school engagement scale of the PISA 2015 measures a primary construct (the emotional school engagement) with two nuisance factors (the negative and positive wording effects).

The results of the present study supported the use of bifactor model in evaluating the presence of method effects. Therefore, for researchers using the Emotional School Engagement Scale of PISA, more sophisticated statistical techniques such as bifactor item response theory model could be considered. The present study provided evidences that negatively and positively worded items of the scale function differently. Researchers are recommended to carefully examine whether negative and positive scale items function equivalently while using self-report scales. The current study was carried out on the data obtained from Turkish sample of the PISA 2015. Students from different cultures might follow different cognitive or psychological processes while answering items of The School Engagement Scale. Therefore, the impact of method effects on the scale items might be examined within a more culturally representative sample of students. In addition, in the current study, the models were compared based on the item parameters and model data fit statistics. The criterion validity and the reliability of the person parameter estimations of the models were not evaluated in the present study. Therefore, future studies might compare the GRM with the B-GRM in terms of the validity and reliability of estimations.

REFERENCES

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44*, 427-445.

Archambault, I., Janosz, M., Fallu, J. S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence, 32,* 651-670.

Baker, F. (2001). *The basics of item response theory.* University of Maryland: College Park: ERIC Clearinghouse on Assessment and Evaluation.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361-370.

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

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

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Cole, K. L. M., Turner, R. C., Gitchel, W. D. (2018). A study of reverse-worded matched item pairs using the generalized partial credit and nominal response models. *Educational and Psychological Measurement, 78*(1), 103-127.

Craft, A. M., & Capraro, R. M. (2017). Science, technology, engineering, and mathematics project-based learning: Merging rigor and relevance to increase student engagement. *Electronic International Journal of Education, Arts, and Science, 3*(6), 140-158.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

DeMars, C. (2010). *Item response theory.* New York: Oxford University Press, Inc.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

DiStefano, C., & Motl, R. W. (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem scale. *Personality and Individual Differences, 46*, 309-313.

Embretson, S. E., & Reise, S.P.(2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associate, Inc.

Fletcher, D., & Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise*, 5, 423-446.

Fredericks, J. A., Blumenfield, P.C., & Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

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.

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.

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.

Immekus, J., & Imbrie, P. K. (2008). Dimensionality assessment using the full information item bifactor analysis for graded response data an illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement, 68*(4), 695-709.

Karasar, N. (2005). *Bilimsel araştırma yöntemi: Kavramlar, ilkeler, teknikler*. Ankara: Nobel Yayın Dağıtım.

Kumar, A., & Dillon, W. R. (1992). An integrative look at the use of additive and multiplicative covariance structure models in the analysis of MTMM data. *Journal of Marketing Research, 29*(1), 51-64.

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large scale science assessment. *Journal of Applied Testing Technology, 13*(2), 2-27.

McKay, M. T., Boduszek, D., & Harvey, S. A. (2014). The Rosenberg Self-Esteem Scale: A bifactor answer to a two-factor question? *Journal of Personality Assessment, 96*(6), 654-660.

Min, H., Zickar, M., & Yankov, G. (2018). Understanding item parameters in personality scales: An explanatory item response modeling approach. *Personality and Individual Differences,* 128, 1-6.

Organisation for Economic Co-operation and Development (2013). PISA 2012 results: Ready to learn students' engagement, drive and self-beliefs (Volume III). Retrieved from https://www.oecd.org/pisa/keyfindings/PISA-2012-results-volume-III.pdf

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of Callous Unemotional Traits. Psychological Assessment, 28(4), 394-404.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696.

Reise, S. P., Ventura, J., Keefe, R. S. E., Baade, L. E., Gold, J. M., Green, M. F., … Bilder, R. (2011). Bifactor and Item Response Theory Analyses of Interviewer Report Scales of Cognitive Impairment in Schizophrenia. *Psychol Assess., 23*(1), 245-261.

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. Psicothema, 30(2), 149-158.

Supple, A. J, & Plunkett, S. W. (2011). Dimensionality and validity of the Rosenberg Self-Esteem Scale for use with Latino adolescents. *Hispanic Journal of Behavioral Sciences, 33*(1), 39-53.

Thissen, D., & Wainer, H. (2001). *Test scoring.* NJ: Lawrence Erlbaum Associates, Inc.

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.

Wang , J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001) Factorial structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling, 8*(2), 275-286.

Wang, Y., Kim, E. U., 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.

Weems , G. H., Onwuegbuzie, A. J., Schreiber, J. B., Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*(6), 588-607.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 59*(5), 737-747.

Wells, C., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability.* Madison: University of Wisconsin, Testing & Evaluation Services.

Wouters E, Booysen F. L. R., Ponnet K, Baron Van Loon F. (2012). Wording effects and the factor structure of the Hospital Anxiety & Depression Scale in HIV/AIDS patients on antiretroviral treatment in South Africa. *PLoS ONE, 7*(4), 1-10.

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences, 46,* 197-201.