International Journal of Assessment Tools in Education
2024, Vol. 11, No. 1, 20–38
https://doi.org/10.21449/ijate.1345549
Published at https://dergipark.org.tr/en/pub/ijate *Research Article*

The effects of reverse items on psychometric properties and respondents' scale scores according to different item reversal strategies

Mustafa İlhan[1*], **Neşe Güler**[2], **Gülşen Taşdelen Teker**[3], **Ömer Ergenekon**[4]

[1]Dicle University, Ziya Gokalp Faculty of Education, Department of Mathematics and Science Education, Diyarbakir, Türkiye.
[2]İzmir Democracy University, Faculty of Education, Department of Educational Sciences, Izmir, Türkiye.
[3]Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Türkiye.
[4]Nafiye Ömer Şevki Cizrelioğlu Kindergarten, Diyarbakir, Türkiye.

ARTICLE HISTORY

Received: Aug. 18, 2023
Accepted: Jan. 22, 2024

Keywords:
Antonyms,
Negations,
Item reversal strategy,
Item wording,
Reverse worded items.

Abstract: This study aimed to examine the effects of reverse items created with different strategies on psychometric properties and respondents' scale scores. To this end, three versions of a 10-item scale in the research were developed: 10 positive items were integrated in the first form (Form-P) and five positive and five reverse items in the other two forms. The reverse items in the second and third forms were crafted using antonyms (Form-RA) and negations (Form-RN), respectively. Based on the research results, Form-P was unidimensional, while other forms were two-dimensional. Moreover, although reliability coefficients of all forms were obtained as above .80, the lowest one was acquired for Form-RN. There were strong-positive relationships between students' scores in the three scale forms. However, the lowest one was estimated between Form-P and Form-RN. Finally, there was a significant difference between the students' mean scores obtained from Form–RN and other two versions, but the effect size of the said difference was small. In conclusion, all these results indicate that different types of reverse items influence psychometric properties and respondents' scale scores differently.

1. INTRODUCTION

Likert-type scales, which are introduced by Rensis Likert, have been frequently used to measure the complex psychological constructs by many researchers from diverse disciplines since 1932. In such scales, items related to the construct to be measured and ordered response options are presented to the individuals and they respond to the items by selecting the category that best reflects them. That is, Likert scales are of self-report type and therefore are open to the response bias. Biased responses can occur in varied forms such as central tendency, extreme responding, social desirability, acquiescence, and dissent with Likert scales.

Central tendency bias happens because of the respondent's reluctance to select extreme response positions (Brace & Bolton, 2022). Contrary to the central tendency, in extreme response style, individuals tend to use the end points of the scales than middle response options (Kline, 2005). Social desirability stems from the respondent's willingness to portray herself or

*CONTACT: Mustafa İlhan ✉ mustafailhan21@gmail.com Dicle University, Ziya Gokalp Faculty of Education, Department of Mathematics and Science Education, Diyarbakir, Türkiye.

e-ISSN: 2148-7456 /

himself favorably, regardless of his or her true characteristics; hence, social desirability bias is also called "faking good" (Furr, 2018). Acquiescence bias is the tendency to agree ("yea-saying") to all or majority of the items, regardless of their contents (Karandashev, 2021). On the other hand, dissent bias is the opposite tendency: it refers mainly to the respondent's propensity to disagree with statements (Stauesberg, 2002). These distortions in responses lead to construct irrelevant variance and threat the validity of measurements. Therefore, to obtain valid and reliable measures, precautions are needed to mitigate these biases.

Inclusion of reverse items (negative items) to the scale is a commonly proposed antidote to cope with response biases, especially acquiescence and dissent types (Ahlawat, 1985; Bandalos, 2018; Coolican, 2013; Bolt et al. 2020; Mayerl & Giehl, 2018). Reverse items function as cognitive "speed bumps" (Podsakoff et al., 2003) and can make respondents read the items more carefully, thereby reducing the probabilities of giving distracted and more generalized responses (Locker et al., 2007). Unfortunately, there is a price to pay for utilizing reverse items. For instance, reversals in item polarity may be confusing to respondents, especially when completing a long instrument (DeVellis, 2017). This situation can increase measurement error, reduce the validity and reliability of measures, and distort factor structure (Weijters & Baumgartner, 2012).

In general, two main strategies are available for item reversal. The first one involves adding negations to the item statements, namely, words like "not" or "no" and affixal morphemes like "un-," "non-," "dis," or "-less." In this case, the item's direction is changed without substantially modifying its wording (van Sonderen et al., 2013). The second strategy, on the other hand, is replacing the word or phrase in the original item with that of the antonym (Zhang et al., 2016). Reverse items created based on the first strategy are called *negatively worded negatively keyed*, and those formed based on the second one is named *positively worded negatively keyed* (Finney, 2001). Simply put, the items "*The conditions of my life are not good*" and "*The conditions of my life are bad*" are the two versions of the item "*The conditions of my life are good*" created based on the first and second strategies, respectively (Kam, 2023).

Notably, both strategies can bring some weak points. The reverse items formed with negations may cause erroneous data due to carelessness responses and more difficult judgmental process (Zhang et al., 2016). On the other side, in reverse items formed with antonyms, some respondents may not view an antonym intended as a reversal by the survey developer as being opposite in meaning to positive items (e.g., "relaxed" may not be perceived as an antonym for "stimulated") (Weijters & Baumgartner, 2012). Thereupon, respondents may give the same responses to a positive item and its reverse counterpart, and in this case, reversal ambiguity comes up. For instance, a respondent might agree with both the items *I like simple tasks* and *I like complex tasks* because *liking simple tasks* does not necessarily purport *disliking complex tasks* (Zhang et al., 2016). Swain et al. (2008) analyzed nearly 2000 Likert items in Bearden and Netemeyer's (1999) *Handbook of Marketing Scales* and ascertained that 81% of the reverse items were with negations, that is, items created based on the first strategy. The main point is that regardless of which strategy was created with, higher scores on the reverse items signify that the respondent has a low level of measured trait.

Although Likert scales have been utilized for about 90 years, researchers still experience some dilemmas about how they should act while developing these measurement tools. One of the foremost dilemmas encountered by researchers when constructing a scale concerns on either the integration of reverse items to the scale or its composition of items as being worded in the same direction or not. The dilemma about reverse items is not just about their incorporating into the scales or not. Another major dilemma is how they should be worded if reverse items are to be employed. Should the item reversal be achieved by using antonym expressions or negations (Weijters & Baumgartner, 2012).

A review of the literature disclosed that research about reverse items has a long history, and most studies focused on the first dilemma mentioned previously. The effects of reverse items on validity, internal consistency, factorial structure and factor loadings, item correlations, missing values, and respondents' ability estimation (mean scores on the scale) are principal subjects emphasized in these studies (e.g., Ahlawat, 1985; Bergstrom & Lunz, 1998; Boley et al., 2021; Bolt et al., 2020; Bulut & Bulut, 2022; Bulut, 2021; Chamberlain & Cummings, 1984; Conrad et al., 2004; Dooden, 2014; Dueber et al., 2021; Guyatt et al., 1999; Herche & England, 1996; Hooper et al., 2013, İlhan & Güler, 2017; Kula Kartal, 2021; Kula Kartal et al., 2022; Locker et al., 2013; Marsh, 1996; Schriesheim & Hill, 1981; Salazar, 2015; Schotte et al., 1996; Suárez-Alvarez et al., 2018; Vigil-Colet et al., 2020).

Conversely, the number of studies seeking answer to the second dilemma regarding reverse items is more limited. In simpler terms, there is a relatively small body of literature that deals with the comparison between the effect of different types of reverse items on the measurement qualities. With regard to this, Schriesheim et al. (1991) conducted a study on the effect of negation and polar opposite item reversals on reliability and validity and found out that the effect of reverse items on psychometric properties differed based on the strategy used for item reversal. Moreover, Salazar (2015) scrutinized the pros and the cons of combining reverse (negative) and regular (positive) items in scales in the Spanish context and concluded that individuals do not respond in the identical fashion to all types of reverse items. Similarly, Zhang et al. (2016) manipulated the types of reverse items (antonym vs. negation) while they investigated the effect of reverse items on the factor structure of the Need for Cognition Scale. As a result, they established that both the number and type of reverse items affect the factor structure of the scale.

In previous studies, variables that play a role in individuals' responses to reverse items have also been investigated. The results obtained demonstrate that the effects that arise from reverse items vary according to culture, age, linguistic features, and respondents' reading proficiency. For example, Marsh (1986) analyzed the bias of reverse items on a sample of preadolescent children and detected that younger children and children with low-reading proficiencies are clearly less able to respond to reverse items appropriately. Also, Williams and Swanson (2001) (cited in Weems et al., 2006) found a similar relationship for adults. Likewise, Bulut and Bulut (2022) revealed that the severity of item wording effect that emerges because of reverse items is related to the reading ability. Hooper et al. (2013) audited the behavior of reverse items on the Confidence in Mathematics Scale administered to students in TIMSS 2011. They proved that the effect of reverse items differs across grade levels and countries. Furthermore, Schmitt and Allik (2005) concluded that reverse items were interpreted differently across nations based on the data collected from 53 nations by means of Rosenberg Self-Esteem Scale. In the same vein, Wong et al. (2003) reported that the problems associated with reverse items do not occur in the same manner in all cultures and languages.

Considering the cultural backgrounds that affect the functioning of reverse items and the limited studies on the effect of types of reverse items (antonym vs. negation) on measurements, the authors believe that research conducted in different cultures on the subject would contribute to the literature. Indeed, researchers mentioned that it would be relevant to explore this phenomenon in diverse countries (e.g., Salazar, 2015). From this point of view, the effect of reverse items created with negations and with antonyms on the measurements in a Turkish-speaking sample was examined. Also, the current research was carried out on high school students unlike those studies in which the effect of different types of reverse items on psychometric qualities was tested on primary school students, undergraduates and older adults. In light of all these expositions, this research expectedly contributes to the literature and offers important insights into the debate about reverse items, which has a long history and is still a hot

topic today. As the scales are fundamental data collection tool in a wide range of scientific disciplines, it is thought that the present paper will appeal an extensive audience and expands the existed knowledge about the problems associated with reverse items.

Within the scope of this specific research, three scale forms were identified: the first one includes only 10 positive items (Form-P); the second one is the combination of five positive items and five reverse items achieved with antonyms (Form-RA); and the third one, on the other hand, comprises five positive items and five reverse items created with negations (Form-RN). By comparing the three scale forms, this study sought answers to the following research questions:

1) Do the three scale forms differ in terms of (a) factorial structure, (b) concurrent validity, and (c) internal consistencies?
2) Do the scores of the respondents vary from one scale form to another?

2. METHOD

2.1. Participants

This study is conducted in Diyarbakır, a province in the southeast of Türkiye. Because research in the relevant literature (e.g., Dagnew, 2017; Geddes et al., 2010; Verešová & Malá, 2016) reveals that students' attitude toward school is significantly related to their academic achievement, a study group with participants from different achievement levels was created. While deciding on the achievement levels of the schools, the placement scores of the national high-stake exams applied were taken in order to select students for high schools in Turkey as a reference. After randomly choosing a school from low, medium, and high achievement levels, a total of 1166 students, 666 girls and 500 boys, aged between 14 and 19 (M=15.06, SD=1.05) took part in the sample.

2.2. Instruments

The instrument to be used in the research must be convertible into reversed items created with negations and antonym expressions, without changing the meaning and wording severity of the regular items it contains. When the Turkish literature was reviewed, such a scale was not found. On that account, instead of drawing upon a scale with tested psychometric properties in Turkish culture, a new measurement tool was generated in the study. In line with the research purpose, the three versions of a School Attitude Scale (SAS) in the study were developed: Form-P, Form-RA, and Form-RN. While constructing the scale forms, 14 positive items were initially written to measure the attitude toward school. These items were composed in such a way that they can be converted into reverse items with antonyms and negations. Then, the draft form consisting of positive items was sent to three experts. While two of these experts were from the field of psychological counseling and guidance, the other one was from the field of measurement and evaluation in education. Relying on the experts' opinions, the content validity indices (CVI) of the items were computed and values ranging from –.33 to 1.00 were obtained. Two items with a CVI of –.33, that is, which two out of three experts deemed inessential, were removed from the scale. Thus, 12 items remained. The CVI for the entire scale based on these 12 items was detected as .83.

Next, seven of the 12 items in the scale were converted into reverse items to generate Form-RA and Form-RN. For instance, the item "*I try to attend school regularly*" in Form-P was transformed to "*I skip school whenever I get the chance*" in Form-RA and "*It is not important for me to attend school regularly*" in Form-RN. Then, the formed reverse counterparts of seven items were sent to six experts who evaluated their equivalence. Three of the experts were PhD candidates in the field of educational sciences, and their bachelor's degree was in Turkish language teaching. The fourth expert was an associate professor in the field of measurement

and evaluation in education, and her bachelor's degree was also in Turkish language teaching. The fifth and sixth experts were professors, one from the field of psychological counseling and guidance, and the other from the field of curriculum and instruction. Experts judged each positive item and its reverse counterparts as "equivalent" and "not equivalent", and also offered certain suggestions to enhance the equivalence of items in the three forms. The Fleiss' kappa statistic for the agreement between the experts was .506. Four experts reported that two of the items in the three scale forms were not equivalent in terms of the response category to be endorsed by any participants. Therefore, the two items were removed in question from the scale. Moreover, two experts pointed out that the items in the reversed form were relatively strict for the item "*I think the school supports my personal development*." So, this item was rearranged, which was initially stated as "*The school does not contribute to my personal development*" in Form-RN, as "*I do not think that the school contributes to my personal development*." Similarly, this item was converted, which was originally stated as "*School is just a waste of time for my personal development*" in Form-RA, into the sentence "*I see school as a waste of time for my personal development.*" Thus, three scale forms were drawn up, each consisting of 10 items. The items in these forms were presented in Table 1.

Table 1. *Items in three different forms of the SAS.*

Item Number	Form–P	Form–RA	Form–RN
1	I believe schools are important institutions for the progress of societies.	I believe schools are important institutions for the progress of societies.	I believe schools are important institutions for the progress of societies.
2	I look forward to the opening of schools while on holiday.	When I'm on holiday, I get depressed as the time for schools to open approaches.	I don't want schools to reopen while I'm on holiday.
3	I believe that school helps us to be responsible individuals.	I believe that school helps us to be responsible individuals.	I believe that school helps us to be responsible individuals.
4	I try to attend school regularly.	I skip school whenever I get the chance.	It is not important for me to attend school regularly.
5	Learning new things at school makes me happy.	Learning new things at school makes me happy.	Learning new things at school makes me happy.
6	I go to school willingly.	I go to school reluctantly.	I wouldn't go to school if I could.
7	I think I learned a lot of things in school that will benefit me.	I think I learned a lot of things in school that will benefit me.	I think I learned a lot of things in school that will benefit me.
8	I think school is an enjoyable place.	I think school is a boring place.	I don't see school as an enjoyable place.
9	I think that school supports my personal development.	I see school as a waste of time for my personal development.	I do not think that the school contributes to my personal development.
10	I believe that all children/young people should go to school.	I believe that all children/young people should go to school.	I believe that all children/young people should go to school.

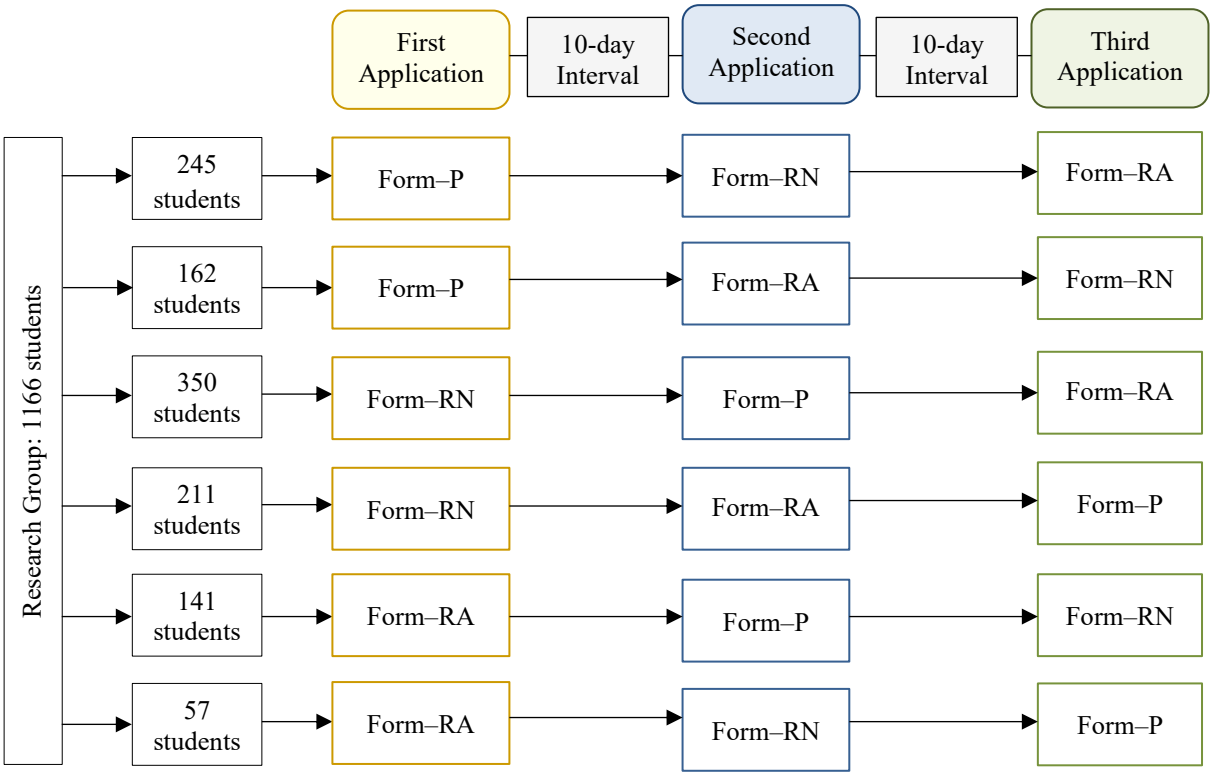
The table shows the commonality of items 1, 3, 5, 7, and 10 in all three forms. On the other side, Items 2, 4, 6, 8, and 9 (gray shaded in Table 1) were positive in Form-P, reversed with opposite meanings in Form-R, and reversed with negations in Form-RN. Also, Table 1 showed that negative items were given randomly order in the scale. The items were anchored with a five-point rating of *strongly disagree* (1), *disagree* (2), *somewhat agree* (3), *agree* (4) and *strongly agree* (5) in all three versions of the scale. Considering the possible effects of response option orders to participants' responses, the order of response options in ascending format (i.e., *strongly disagree* to *strongly agree*) was arranged in all three scale versions.

Before using the scale forms that were developed to collect the research data, a pilot study on Form-P was performed. To this end, Form-P was administered to 394 high school students other than the main sample. The pilot data were randomly divided into two parts to conduct exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA and CFA were performed after examining the relevant assumptions. Since the multivariate normality assumption was violated, principal axis analysis in EFA and Satorra-Bentler robust maximum likelihood (MLM) in CFA were utilized as the estimation methods. Based on the EFA results, a single-factor structure was acquired where factor loadings ranged from .55 to .72 and the extracted variance was 42.50%. As a result of CFA, on the other hand, the factor loadings belonging to single-factor model varied between .44 and .77, and the fit statistics were within acceptable limits (χ^2/df = 3.01, RMSEA = .10, [90% CI = .080, .124], SRMR = .058, TLI = .89, and CFI = .91). Furthermore, McDonald's ω was .878 (95% CI = .860, .896), and item-total correlations were within .47 and .66 for the dataset obtained by combining the data files from which EFA and CFA were conducted.

2.3. Procedure

The data collection process was fulfilled in three applications because of the three forms. A complete counterbalanced design was used to control for order effects that could arise from the sequence of administration of the scale forms, thus enhancing the internal validity of the research. The scale forms were applied to the students with 10-day intervals. Figure 1 summarizes the process followed for collecting the research data.

Figure 1. *The process followed in data collection.*



At the beginning of the scale forms, along with the demographics of gender and grade level, the students were asked the following: "*How many points would you give if you would rate your love of school between 1 and 10?*" The data obtained through this question was used to test and compare the concurrent validity of the three different versions of the SAS. The scales were administered to the students in pencil-and-paper format, in their actual classrooms, and in Turkish. Students' participation in the research was on a voluntary basis and they did not receive

any compensation. Furthermore, students were assured that their data would remain anonymous and would not be shared with any other person or institutions; however, they were required to use a nickname to match the three scale forms they responded. Fortunately, we did not encounter any students who refused to take part in the study in any of the classes where the administration was performed. In fact, data from 1329 participating students were collected. However, the 157 students who were absent in any of the three administrations and six students who had missing data in any scale forms although they participated in all the three applications were excluded. Consequently, a dataset of 1166 students was achieved. The compliance of the research with current ethical standards was approved by the Social and Humanities Ethics Committee of Dicle University, Diyarbakır, Türkiye.

2.4. Data Analysis

The analysis process started with the preparation of the data for analysis and the examination of the distribution characteristics. First, there was no missing value in the dataset. Because it is recommended in the literature to conduct EFA and CFA on different samples (Dawson, 2017; Fabrigar et al., 1999), the study sample was randomly split out into two halves before starting the analysis. Following this, the outliers was extracted from the datasets. As a result, 20 participants were removed from the sample used for EFA, leaving 563 students in this dataset. In addition, 26 participants were excluded from the dataset used for CFA, remaining 557 students in this dataset. In the next step, the skewness and kurtosis coefficients were examined to test the univariate normality and the Henze Zirkler tests for checking over multivariate normality. Notably, the skewness and kurtosis coefficients of the datasets were within ±1, and univariate normality was achieved. On the other hand, significant results of Henze Zirkler tests revealed that multivariate normality was violated. Hence, principal axis analysis in EFA and Satorra-Bentler robust maximum likelihood (MLM) in CFA were operated as the estimators.

In EFA, KMO values for Form-P, Form-RA, and Form-RN were .91, .88, and .86, respectively. Besides, Bartlett's test results were statistically significant for all three forms [χ^2_{Form-P} = 2040.49, $\chi^2_{Form-RA}$ = 2161.18, $\chi^2_{Form-RN}$ = 1579.70; df = 45; p < .001]. In conclusion, sample and data were satisfactory and continued with the factor analysis. Also, the parallel analysis method was employed to identify the number of factors to be extracted in EFA and the scree plots including parallel analysis results was presented in the Appendix.

Another evidence of validity that was investigated in the research was concurrent validity. Within this framework, the Pearson correlations between students' responses to the question of "*How many points would you give if you would rate your love of school between 1 and 10?*" and their mean scores from each scale version were calculated. To boot, the correlation coefficients obtained as a measure of concurrent validity were interpreted, and the variability of the three forms of the SAS with regard to concurrent validity was tested by analyzing the significance of the differences between the correlation coefficients.

To estimate the internal consistencies of the three scale versions and compare their reliability coefficients, the McDonald's ω was calculated for each form. Given that the presence of univariate normality, parametric tests were used to compare students' scores on three different versions of the SAS. Pearson correlation analysis was carried out to see the relationships between students' scores on three scale forms, and repeated measures ANOVA was executed to assess whether students' school attitude scores differ from one version of the scale to another. Mauchly's test of sphericity was checked before repeated measures ANOVA. Results revealed that the sphericity assumption had been met (Mauchly's W = .995, χ^2 = 5.15, df = 2, p > .005). Moreover, the eta squared value (η^2) was checked to appraise the effect size of the difference observed in ANOVA.

In the research, for detecting multivariate outliers and calculating the Henze Zirkler multivariate normality test, the web tool developed by Aybek (2021) with R software running in its background was employed. The significance of the differences between the correlation coefficients obtained for the concurrent validity of the three versions of the SAS was tested by means of the interface developed by Diedenhofen and Musch (2015). The functions in this web application are based on the tests implemented in the cocor package of the R programming language. All other analyses in the study were conducted in JASP 0.16.

3. RESULTS

The EFA outputs for the three versions of the SAS were primarily inspected. Table 2 summarizes the results of EFA.

Table 2. *EFA results for the three versions of the SAS.*

	Item Number	Form–P	Form–RA		Form–RN	
		Factor–1	Factor–1	Factor–2	Factor–1	Factor–2
Factor Loadings	1	.676	.740	.007	.693	.060
	2	.426	.106	.734	.098	.610
	3	.728	.807	.042	.704	.043
	4	.579	–.076	.413	–.196	.391
	5	.640	.628	–.063	.577	-.111
	6	.691	.068	.821	–.178	.610
	7	.792	.838	.023	.832	.048
	8	.518	.001	.801	.147	.662
	9	.665	–.302	.448	–.149	.468
	10	.604	.590	–.026	.608	.020
Variance Explained		40.90%	49.75%		41.20%	

Table 2 exhibits that Form-RA and Form-RN have a two-factor structure with positive items in one factor and reverse items in the other unlike the single-factor Form-P. Moreover, the variances explained are close to each other in Form-P and Form-RN. On the other hand, the extracted variance in Form-RA is clearly higher than that in the other two forms. Another remarkable result in Table 2 is as follows: Form-RA and Form-RN overlap in terms of the dispersion of the items to the factors, but the items' factor loadings are generally higher in Form-RA than those in Form-RN.

Under the CFA, the factor structures that emerged in the EFA for all three scale versions were tested. However, the unidimensional model in addition to the two-factor structure in Form-RA and Form-RN was also examined, because the SAS was prepared by forecasting a single-factor structure and Form-P was unidimensional. Table 3 comprises the fit indices reported in the CFA and shows the critical values for the fit indices in the second column.

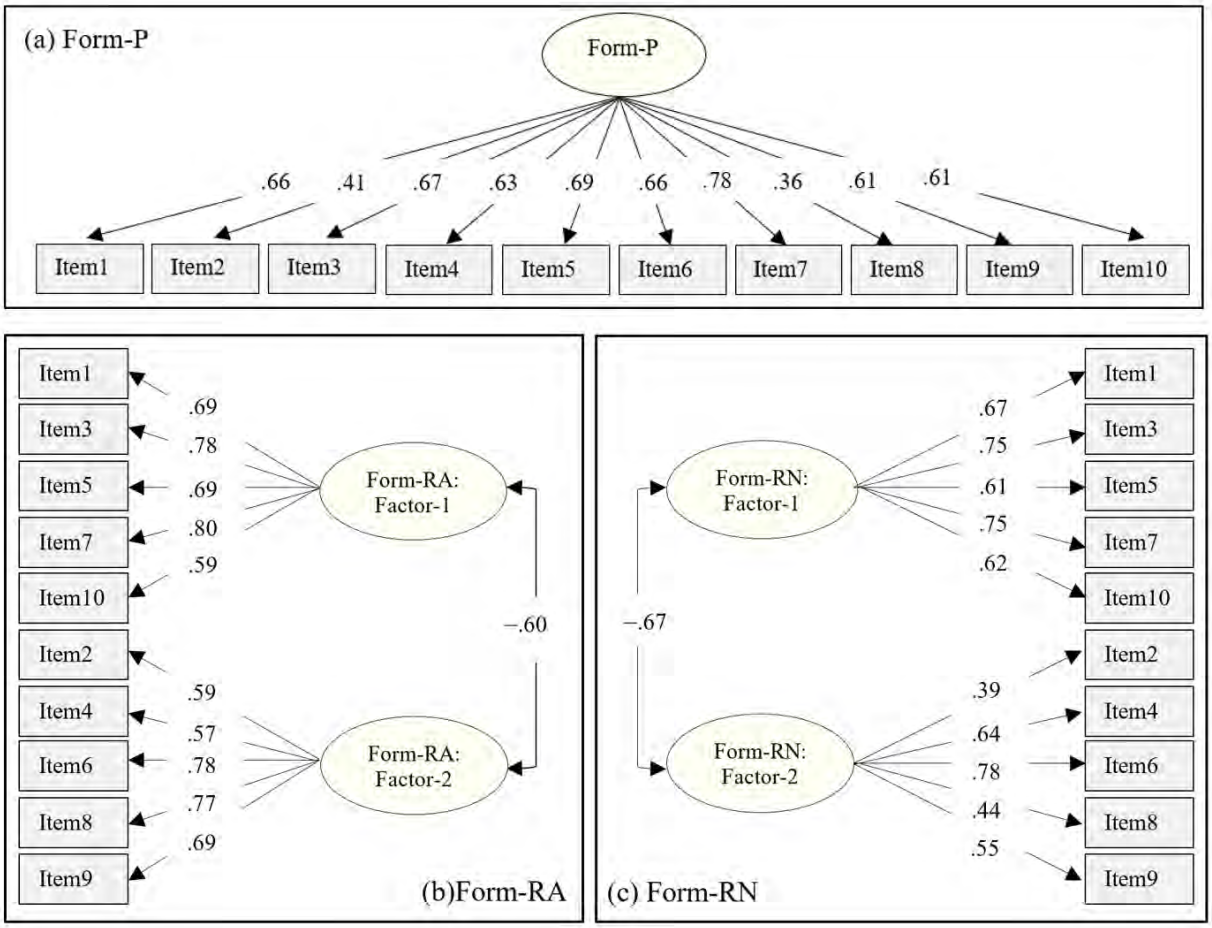
Table 3. *Fit statistics obtained in CFA for the three versions of the SAS.*

Fit Statistics	Critical values	Form–P	Form–RA		Form–RN	
			Single-factor model	Two-factor model	Single-factor model	Two-factor model
χ^2/df	≤ 5[a]	3.819	13.654	2.510	7.419	3.105
RMSEA	≤ .10[b]	.071 90% CI (.059, .084)	.151 90% CI (.139, .163)	.052 90% CI (.038, .066)	.107 90% CI (.095, .120)	.061 90% CI (.048, .075)
SRMR	≤ .10[c]	.042	.095	.037	.068	.042
MFI	≥ .90[c]	.915	.672	.955	.817	.938
CFI	≥ .90[c]	.944	.784	.975	.851	.953
NNFI (TLI)	≥ .90[c]	.928	.722	.967	.808	.937

[a] Marsh & Hocevar (1985), [b] Meyers at al. (2006), [c] Pituch & Stevens (2016)

Table 3 illustrates that the fit indices of the single-factor model are out of the critical values for Form-RA and Form-RN. By contrast, all fit statistics belonging to two-factor model are within acceptable limits in both Form-RA and Form-RN. Apparently, model-data fit was provided only in Form-P for single-factor analysis. Figure 2 represents the measurement models in which the model-data fit was achieved for the three versions of the SAS. It presents that, just like in EFA, the factor loadings of the items are generally higher in Form-RA than they are in Form-RN. Additionally, the correlation between the factor with positive items and the factor with reverse items is higher in Form-RN than it is in Form-RA.

Figure 2. *Measurement models for (a) Form-P, (b) Form-RA, and (c) Form-RN of the SAS.*



Having completed EFA and CFA, the concurrent validity results were analyzed. The correlation coefficients from concurrent validity of the three versions of the SAS were calculated as .592 (95% CI [.552, .628], $p < .001$), .644 (95% CI [.608, .677], $p < .001$), and .611 (95% CI [.573, .647], $p < .001$) for Form-P, Form-RA, and Form-RN, respectively. Salkind (2010) recommends a series of range for interpreting correlation coefficients as very weak (<.20), weak (.20–.40), moderate (.40–.60), strong (.60–.80), and very strong (>.80). Accordingly, the obtained correlation coefficients point to strong relationships for the concurrent validity of all three forms of the SAS. As a striking result, the correlation coefficient of Form-RA was calculated higher in proportion to the other two forms of the SAS. When the significance of the differences between acquired correlation coefficients was tested via Pearson and Filon's z statistic, there was a significant difference between Form-P and Form-RA ($z = -2.577$, 95% CI [−.092, −.013], $p < .01$). Conversely, there were no significant differences between Form-P and Form-RN ($z = -.893$, 95% CI [−.061, .023], $p > .05$) and between Form-RA and Form-RN ($z = 1.646$, 95% CI [−.006, .073], $p > .05$) in terms of their concurrent validity evidence.

Subsequent to comparing the three versions of the SAS in terms of validity proofs, the reliability estimations were performed. McDonald's ω coefficients of the measures were calculated as .861 (95% CI [.849, .874]), .855 (95% CI [.842, .868]), and .816 (95% CI [.800, .832]) for Form-P, Form-RA, and Form-RN, respectively. Evidently, the reliability values estimated for all three versions of the SAS are over .80. DeVellis (2017) proposed the following ranges when judging the reliability: below .60, unacceptable; between .60 and .65, undesirable; between .65 and .70, minimally acceptable; between .70 and .80, respectable; and between .80 and .90, very good. On the other hand, the coefficients above .90 signify redundant items and mean that the scale should be shortened. Relying on the intervals listed, all three scale forms yield quite reliable measurements. What stands out in reliability analysis results is that the internal consistency coefficient of Form-RN is lower than those of the other two forms.

After checking the three versions of the SAS against their psychometric properties, the respondents' scores on the different forms were compared. Based on the results of Pearson correlation analysis, there were significant relationships between students' scores in the three scale forms. To clarify, the correlation coefficients were estimated as .695 (95% CI [.633, .724]) between Form-P and Form-RA, .647 (95% CI [.611, .680]) between Form-P and Form-RN, and .702 (95% CI [.671, .731]) between Form-RA and Form-RN. Although there are strong positive relationships between students' scores in the three scale forms, the correlation between Form-P and Form-RN is relatively low. Eventually, repeated measures ANOVA was conducted to establish whether the students' mean scores differ across three scale forms, and the results are shown in Table 4.

Table 4 denotes that students' school attitude scores differ significantly from one scale form to the other ($F_{(2, 2238)} = 4{,}36$, $p < .05$, $\eta^2 = .004$). Although post-hoc test results revealed that the significant differences were observed between Form-RN and the other two versions, a closer inspection of the results suggests that the eta squared value, which is a measure of effect size, is quite low. In other words, Cohen (1988) proposed the following guidelines for interpreting the eta squared values: .01, small effect; .06, moderate effect; and .14, large effect (cited in Pallant, 2005). These benchmarks notify that the statistically significant differences detected are minor in practice.

Table 4. *Results of repeated measures ANOVA regarding the difference between the means across three versions of the SAS.*

Scale Version	Mean	*SD*	*F*(2, 2238)	Post-Hoc (LSD)	η^2
Form–P	3.47	.78	4.36*	Form–RN > Form–RA Form–RN > Form–P	.004
Form–RA	3.46	.82			
Form–RN	3.51	.76			

* $p<.05$

4. DISCUSSION and CONCLUSION

In the present research, the impacts of reverse items created with negations and with antonyms on the psychometric properties and respondents' scale scores were evaluated. The EFA results exposed that Form-P has a unidimensional structure, while Form-RN and Form-RA have a two-factor structure in which positive and reverse items are in separate factors. Furthermore, CFA outputs verified that the model where positive and reverse items were considered as distinct factors fit the data better than the single-factor solution for both Form-RA and Form-RN. More exactly, regardless of which reversal strategy is used, the reverse items caused an artificial factor that can also be called a method factor in addition to the trait factor. Consistent with this result, many studies in the literature revealed that reverse items in scales measuring a single unitary construct distort the factor structure of the scale by generating a separate dimension, which leads to an unintentionally multidimensional factor solution (Bulut & Bulut; 2022; DiStefano & Motl, 2006; Dunbar et al., 2000; Hazlett-Stevens et al., 2004; Herche & Engelland, 1996; Knight et al., 1998; Pilotte & Gable, 1990; Spector et al., 1997).

As a result of EFA, the factor loadings in most of the items of Form-RA and, in a parallel manner, its extracted variance were higher than those of the other scale forms. Besides, CFA results showed that Form-RA had higher factor loadings for many, if not all, items. The concurrent validity results were also in this direction and were higher for Form-RA. These results hint that the reverse items created with negations do not serve the purpose of increasing the validity of the measurements. The reverse items formed with antonyms, on the other hand, partially improve the validity even though they jeopardize the unidimensionality of the scale.

The research result regarding the variance ratio is contrary to that of Suárez-Alvarez et al. (2018). To sum up, the explained variance ratio in Form-RA was higher than that in Form-P. Conversely, Suárez-Alvarez et al. (2018) ascertained that the percentage of explained variance in the version containing both positive and reverse items with antonyms was lower than the form consisting of only positive items. This may be related to language properties, to the group in which the study was conducted, or to the differences between the scale forms used. As this research was conducted on a Turkish-speaking sample and that of Suárez-Alvarez et al. (2018) was carried out on a Spanish-speaking group, the aforementioned discrepancy could be attributed to language features. Indeed, Schmitt and Allik (2005) and Hooper et al. (2013) disclosed that reverse items were interpreted differently across countries. The mentioned contradiction may also be associated with the sample characteristics because the factors such as reading achievements (Bulut & Bulut, 2022; Michaelides, 2019), age levels (Bulut, 2021; Marsh 1986) and cognitive abilities (Gnambs & Schroeders, 2020) of the samples can affect their responses to the reverse items. Weems et al. (2003) analyzed profiles of individuals who respond inconsistently to positive and reverse items on Likert scales and remarked that certain characteristics of the participants made them more likely to generate differential patterns of responses to the positive and reverse items. In addition to the issues listed, the use of different scales in this paper and in Suárez-Alvarez et al.'s (2018) research might have caused discordant results concerning the explained variance ratio.

The results of the study indicated that internal consistency coefficient of Form-RN was lower than that of the other two versions. The facts that the reliability coefficient of Form-RN is lower than that of Form-P are coherent with those found by other researchers (Bourque & Shen, 2005; Carlson, 2011; Coleman, 2013; Ebesutani et al., 2012; Johnson et al., 2004; Roszkowski & Soven, 2010; Salazar, 2015; Suárez-Alvarez, 2018). Accordingly, we can infer that reverse items created with negations are more open to measurement errors. This is thought to reflect the fact that negatively worded phrases (i.e., reverse items) require additional cognitive efforts and usually cause confusion for respondents (Chyung et al., 2018). Justifiably, the foregoing situation is not true for reverse items built with antonyms as the reliability value of Form-RA

is similar to that of Form-P. The research results, in which Form-RA produces more valid and reliable measurements compared with Form-RN, accord with the results obtained in earlier studies. Weijters and Baumgartner (2012) executed a comprehensive literature review on reverse items and stated that negations are problematic at the judgment stage because they require additional mental steps during item processing. Further, they posited that negations should be employed sparingly and reverse items created with antonyms may be more beneficial than those with negations. Likewise, Zhang et al. (2016) criticized the reverse items derived from negations as follows: this type of reverse items might engender careless responding or judgmental complication for some individuals. Some respondents may fail to notice the presence of a negative particle in the item stem (e.g., misread I am not happy as I am happy), making errors due to carelessness. Moreover, a reverse item generated with negation makes it more difficult for the respondents to judge whether the item content is match with his or her own beliefs. Considering the results respecting "antonyms vs. negations" comparison in the light of the literature, arguably, using antonym is a better strategy for creating reverse items.

The analysis of comparing the scale scores of the participants in the three forms denoted that there were positive and strong correlations between the scores on the different forms. The correlation coefficients obtained reflect that the relative agreement between the scores of the participants in the three scale forms is high. Accordingly, the participants have, by and large, similar rankings in terms of their scores in the three forms, but there is no exact identical ranking. Moreover, the mean scores in Form-RN were significantly higher than those of the other two scale versions, and the difference between the scores of Form-P and Form-RA was not significant. The calculated effect size set out that the statistically significant difference was practically quite small. In simpler terms, even if there were not large differences between participants' mean scores on three scale forms, there was no full absolute agreement, either. On the basis of these results, it can be conceivably hypothesized that in cases where small score differences are important, the decision to be taken about the participants may change depending on the item wording effect. This inference matches with the ideas of Schotte et al. (1996) who stated that not only *what is asked* but also *how it is asked* influences the responses of participants in self-report instrument.

Comparing the study's results regarding the effect of reverse items on scale scores with the results of previous studies in the literature, a rather contradictory picture appears. First and foremost, similar to this research, Benson and Hocevar (1985), Weems et al. (2006), Hughes (2009), and Locker et al. (2013) established that reverse items change the scale scores of the respondents. Conversely, Greenberger et al. (2003) illustrated that item wording did not influence participants' mean scores, and Zhang et al. (2016) specified that the item means were similar across four scale versions with a different composition in terms of positive and reverse items. Second, a significantly higher mean score was detected in Form-RN than those of the other two scale versions. Parallel to this result, Taylor and Bowers (1972) (as cited in Schriesheim & Hill, 1981), Schotte et al. (1996), and Suárez-Alvarez et al. (2018) uncovered that reverse item generates a higher mean response than does the positive counterpart. On the other hand, Stewart and Frye (2004), Locker et al. (2013), Vigil-Colet et al. (2020), and Dueber et al. (2021) found that reverse items yield lower mean scores after coding them in the same direction of positive items. Hence, there is no consistency among the studies on whether the reverse items differentiate the scale scores, and if so, in which direction. These differences may originate because of the varieties between the samples of the studies and the measurement tools used. In particular, it seems possible that differences in mean scores due to reverse items are in opposite direction in scales where the measured attribute is negative in nature (e.g., depression) and in scales where it is positive (e.g., happiness).

The combination of the research results provides important suggestions for practice. Before anything else, an instrument developer should be cautious about the usage of a mix of positive and reverse items when constructing a scale. He or she must make up his or her mind whether reverse items are really necessary and avoid using them unless there is a clear justification. Considering that negated reverse items will attenuate the psychometric properties and elicit a difference, albeit small, in the scale scores of the participants, he or she should prefer items created with antonyms instead of items built with negations in cases where he or she utilizes reverse items. In addition, because item phrasing may differentiate scale scores, researchers should be careful when comparing scale scores that have diverse item word formats. Nonetheless, research limitations that restrict the generalizability of these implications must be inculcated. Corroborating studies are required to reach a more decisive conclusion on this issue.

4.1. Limitations and Future Avenues Research

The present study has certain limitations. First, this study was conducted on a Turkish-speaking sample, and the way reverse items are formed differs from one language to another. For example, in Turkish, the verb is at the end of the sentence, and suffixes in the form of "-me, -ma" are added to the verb to achieve reverse items with negations. In English, on the other hand, the verb is after the subject, and the words like "not or no" are used before the verb to form negated reverse items. In other languages the situation is likely to be different. In this sense, similar studies should be conducted in other languages regarding the impact of reverse items. Second, participants of this study were exclusively high school students whose age mean was approximately 15.06. Future research can focus on different age groups, as the effect of reverse items varies depending on the age and cognitive development of the respondents. Finally, the current study employed the SAS to investigate the influences of reverse items. Therefore, researchers must utilize other instruments when replicating this study.

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 authors. **Ethics Committee Number**: Dicle University, 27.12.2021-21445.

Authorship Contribution Statement

Mustafa İlhan: Investigation, creating of the instruments, data analysis, resources, visualization, and writing-original draft. **Neşe Güler**: Investigation, creating of the instruments, receiving experts' opinions and writing-original draft. **Gülşen Taşdelen Teker**: Creating of the instruments, resources and writing-original draft. **Ömer Ergenekon**: Data collection and writing-original draft.

Orcid

Mustafa İlhan https://orcid.org/0000-0003-1804-002X
Neşe Güler https://orcid.org/0000-0002-2836-3132
Gülşen Taşdelen Teker https://orcid.org/0000-0003-3434-4373
Ömer Ergenekon https://orcid.org/0000-0001-9964-5535

REFERENCES

Ahlawat, K.S. (1985). On the negative valence items in self-report measures. *The Journal of General Psychology*, *112*(1), 89–99. https://doi.org/10.1080/00221309.1985.9710992

Aybek, E.C. (2021). *Data preparation for factor analysis*. URL: https://shiny.eptlab.com/dp2fa/

Bandalos, D.L. (2018). *Measurement theory and applications for the social sciences*. The Guilford.

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

Bergstrom, B.A., & Lunz, M.E. (1998, 13–17, April). *Rating scale analysis: Gauging the impact of positively and negatively worded items*. Annual Meeting of the American Educational Research Association, San Diego.

Boley, B.B., Jordan, E., & Woosnam, K.M. (2021). Reversed polarity items in tourism scales: Best practice or dimensional pitfall? *Current Issues in Tourism*, *24*(4), 1-13. https://doi.org/10.1080/13683500.2020.1774517

Bolt, D., Wang, Y.C., Meyer, R.H., & Pier, L. (2020). An IRT mixture model for rating scale confusion associated with negatively worded items in measures of social-emotional learning. *Applied Measurement in Education*, *33*(4), 1-18. https://doi.org/10.1080/08957347.2020.1789140

Brace, I., & Bolton, I. (2022). *Questionnaire design: How to plan, structure and write survey material for effective market research* (5th ed.). Kogan Page.

Bulut, H.Ç. (2021). Item wording effects in psychological measures: Do early literacy skills matter? *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 239-253. https://doi.org/10.21031/epod.944067

Bulut, H.Ç., & Bulut, O. (2022). Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? *Studies in Educational Evaluation*, *72*(2). https://doi.org/10.1016/j.stueduc.2022.101126

Carlson, M., Wilcox, R., Chou, C.-P., Chang, M., Yang, F., Blanchard, J., Marterella, A., Kuo, A., & Clark, F. (2011). Psychometric properties of reverse-scored items on the CES-D in a sample of ethnically diverse older adults. *Psychological Assessment*, *23*(2), 558–562. https://doi.org/10.1037/a0022484

Chamberlain, V.M., & Cummings, M.N. (1984). Development of an instructor/course evaluation instrument. *College Student Journal*, *18*(3), 246–250.

Chyung, S.Y.Y., Barkin, J.R., & Shamsy, J.A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement*, *57*(3), 16–25. https://doi.org/10.1002/pfi.21749

Coleman, C.M. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study* [Unpublished doctoral dissertation, James Madison University]. https://commons.lib.jmu.edu/diss201019/73/

Conrad, K.J., Wright, B.D., McKnight P., McFall, M., Fontana, A., & Rosenheck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD Scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement*, *5*(1), 15–30.

Coolican, H. (2013). *Research methods and statistics in psychology* (5th ed.). Routledge.

Dagnew, A. (2017). The relationship between students' attitudes towards school, values of education, achievement motivation and academic achievement in Gondar secondary schools, Ethiopia. *Research in Pedagogy*, *7*(1), 30–42. https://doi.org/10.17810/2015.46

Dawson, J. (2017). *Analysing quantitative survey data for business and management students*. Sage.

DeVellis, R. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoSONE*, *10*(4), 1-12. https://doi.org/10.1371/journal.pone.0121945

Dooden, H. (2014). The effects of positively and negatively worded items on the factor structure of the UCLA loneliness scale. *Journal of Psychoeducational Assessment*, *33*(3), 259–267. https://doi.org/10.1177/0734282914548325

Dueber, D.M., Toland, M.D., Lingat, J.E., Love, A.M.A., Qiu, C., Wu, R., & Brown, A.V. (2021). To Reverse item orientation or not to reverse item orientation, that is the question. *Assessment*, *29*(7), 1422–1440. https://doi.org/10.1177/10731911211017635

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment*, *16*(1), 13–19. https://doi.org/10.1027//1015-5759.16.1.13

Ebesutani, C., Drescher, C.F., Reise, S.P., Heiden, L., High, T.L., Damon, J.D., & Young, J. (2012). The loneliness questionnaire-short version: An evaluation of reverse- worded and non-reverse-worded items via item response theory. *Journal of Personality Assessment*, *94*(4), 427–437. https://doi.org/10.1080/00223891.2012.662188

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Finney, S.J. (2001). *A comparison of the psychometric properties of negatively and positively worded questionnaire items* (Publication No. 3009716) [Doctoral dissertation, University of Nebraska]. ProQuest Dissertations & Theses Global.

Furr, R.M. (2018). *Psychometrics: An introduction* (3rd ed.). Sage.

Geddes, J.D., Murrell, A.R., & Bauguss, J. (2010). Childhood learning: An examination of ability and attitudes toward school. *Creative Education*, *1*(3), 170-183. https://doi.org/10.4236/ce.2010.13027

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, *27*(2), 404-418. https://doi.org/10.1177/1073191117746503

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*(6), 1241-1254. https://doi.org/10.1016/S0191-8869(02)00331-8

Guyatt, G.H., Cook, D.J., King, D., Norman, G.R., Kane, S.L., & Van Ineveld, C. (1999). Effect of the framing of questionnaire items regarding satisfaction with training on residents' responses. *Academic Medicine*, *74*(2), 192-194. https://doi.org/10.1097/00001888-199902000-00018

Hazlett-Stevens, H., Ullman, J.B., & Craske, M.G. (2004). Factor structure of the Penn State worry questionnaire: Examination of a method factor. *Assessment*, *11*(4), 361–370. https://doi.org/10.1177/1073191104269872

Herche, J., & England, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science*, *24*(4), 366-374. https://doi.org/10.1177/0092070396244007

Hooper, M., Arora, A., Martin, M.O., & Mullis, I.V.S, (2013, 24–25 June). *Examining the behavior of "reverse directional" items in the TIMSS 2011 context questionnaire scales* [Conference presentation]. 5th IEA International Research Conference. National Institute of Education, Nanyang Technological University, Singapore. https://www.iea.nl/sites/default/files/2019-04/IRC-2013_Hooper_etal.pdf

Hughes, G.D. (2009). The impact of incorrect responses to reverse-coded survey items. *Research in the Schools*, *16*(2), 76–88.

Ilhan, M., & Guler, N. (2017). The number of response categories and the reverse scored item problem in Likert-type scales: A study with the Rasch model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 321-343. http://dx.doi.org/10.21031/epod.321057

JASP Team (2022). *JASP (Version 0.16.3)* [Computer software].

Johnson, J.M., Bristow, D.N., & Schneider, K.C. (2004). Did you not understand the question or not? An investigation of negatively worded questions in survey research. *Journal of Applied Business Research*, *20*(1), 75–86. https://doi.org/10.19030/jabr.v20i1.2197

Kam, C.C.S. (2023). Why do regular and reversed items load on separate factors? Response difficulty vs. item extremity. *Educational and Psychological Measurement*, *83*(6), 1085–1112. https://doi.org/10.1177/00131644221143972

Karandashev, V. (2021). *Cultural models of emotions*. Springer.

Kline, T.J.B. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage.

Knight, R.G., Chisholm, B.J., Marsh, N.V., & Godfrey, H.P.D. (1998). Some normative, reliability, and factor analytic data for the revised UCLA Loneliness Scale. *Journal of Clinical Psychology*, *44*(2), 203-206. https://doi.org/10.1002/1097-4679(198803)44:2<203::AID-JCLP2270440218>3.0.CO;2-5

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. https://doi.org/10.24106/kefdergi.708968

Kula Kartal, S., Aybek, E.C., & Yaşar, M. (2022). Investigating the wording effect in scales based on the different dimension reduction techniques. *Journal of Uludag University Faculty of Education*, *35*(1), 44–67. https://doi.org/10.19171/uefad.1033284

Locker, D., Jokovic, A., & Allison, P. (2013). Direction of wording and responses to items in oral health related quality of life questionnaires for children and their parents. *Community Dentistry Oral Epidemiology*, *35*(4), 255–262. http://dx.doi.org/10.1111/j.1600-0528.2007.00320.x

Marsh, H.W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37-49. https://doi.org/10.1037/0012-1649.22.1.37

Marsh, H.W. (1996). Positive and negative global self- esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Marsh, H.W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin*, *97*(3), 562–582. https://doi.org/10.1037/0033-2909.97.3.562

Mayerl, J., & Giehl, C. (2018). A closer look at attitude scales with positive and negative items response latency perspectives on measurement quality. *Survey Research Methods*, *12*(3), 193–209. https://doi.org/10.18148/srm/2018.v12i3.7207

Meyers, L.S, Gamst, G., & Guarino, A.J. (2006). *Applied multivariate research: Design and interpretation*. Sage.

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

Pallant, J. (2005). *SPSS survival manual* (2nd ed.). Allen & Unwin.

Pilotte, W.J., & Gable R.K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603–610. http://dx.doi.org/10.1177/0013164490503016

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*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Roszkowski, M.J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113-130. http://dx.doi.org/10.1080/02602930802618344

Salazar, M.S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–200. https://doi.org/10.7334/psicothema2014.266

Pituch, K.A., & Stevens, J.P. (2016). *Applied multivariate statistics for the social sciences* (6th ed.). Routledge.

Schmitt, D.P., & Allik, J. (2005). Simultaneous administration of the Rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*(4), 623-642. https://doi.org/10.1037/0022-3514.89.4.623

Schotte, C.K.W., Maes, M., Cluydts, R., & Cosyns, P. (1996). Effects of affective-semantic mode of item presentation in balanced self-report scales: Biased construct validity of the Zung self-rating depression scale. *Psychological Medicine*, *26*(6), 1161–1168. http://dx.doi.org/10.1017/s0033291700035881

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

Schriesheim, C.A., & Hill, K.D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*(4), 1101–1114. http://dx.doi.org/10.1177/001316448104100420

Spector, P.E., Van Katwyk, P.T., Brannick, M.T., & Chen, P.Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, *23*(5), 659–677. https://doi.org/10.1016/S0149-2063(97)90020-9

Stauesberg, M. (2002). Survey and questionnaires. In S. Engler & Micheal Stausber (Eds,), *The Routledge handbook of research methods in the study of religion* (pp. 461–482). Routledge.

Stewart, T.J., & Frye, A.W. (2004). Investigating the use of negatively phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, *79*(10), 18–20. https://doi.org/10.1097/00001888-200410001-00006

Suárez-Alvarez, 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. https://doi.org/10.7334/psicothema2018.33

Swain, S.D., Weather, D., & Niedrich, R.W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1)*,* 116-131. https://doi.org/10.1509/jmkr.45.1.116

van Sonderen, E., Sanderman, R., & Coyne, J.C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE*, *8*(7), 1-7. https://doi.org/10.1371/journal.pone.0068967

Verešová, M., & Malá, D. (2016, 11–15October). *Attitude toward school and learning and academic achievement of adolescents* [Conference Presentation]. 7th International Conference on Education and Educational Psychology. https://www.europeanproceedings.com/article/10.15405/epsbs.2016.11.90

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema*, *32*(1), 108-114. https://doi.org/10.7334/psicothema2019.286

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. http://dx.doi.org/10.1080/14664200308668290

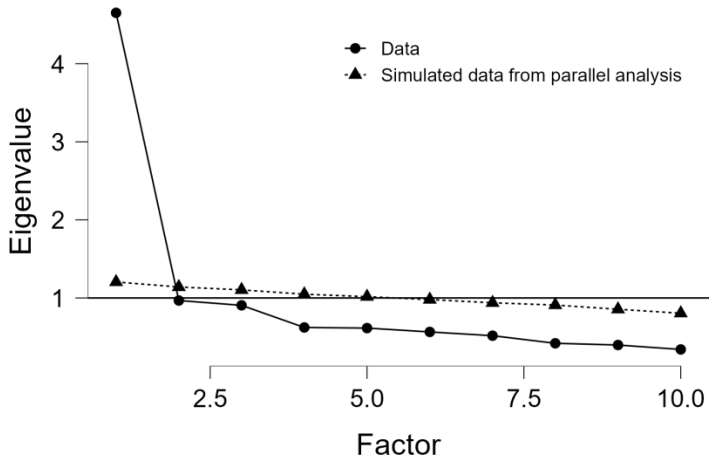
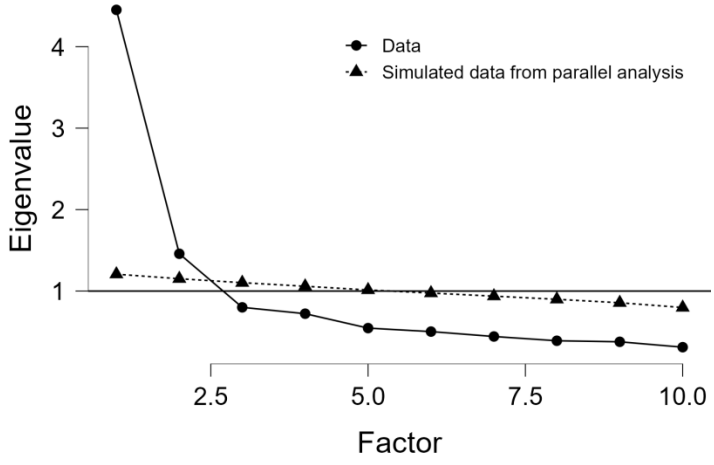
Weems, G.H., Onwuegbuzie, A.J., & Collins, K.M.T. (2006) The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation & Research in Education*, *19*(1), 3-20. https://doi.org/10.1080/09500790608668322

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Wong, N., Rindfleisch, A., & Burroghs, J.E. (2003). Do reversed-worded items confound measures in cross-cultural consumer research? The case of material values scale. *Journal of Consumer Research, 30*(1), 72–91. https://doi.org/10.1086/374697

Zhang, X., Noor R., & Savalei, V. (2016) Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PLoS ONE*, *11*(6), 1-15. https://doi.org/10.1371/journal.pone.0157795

APPENDIX

Scale version	Scree plot with parallel analysis results
Form–P	
Form–RA	
Form–RN