SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 2944-2948
Open Access. Online & Print

DOI: 10.5281/zenodo.122.126226

EFFECT OF NEGATIVELY WORDED ITEMS ON THE FACTORIAL STRUCTURE AND INTERNAL RELIABILITY OF PSYCHOLOGICAL SCALES

Hamzeh Dodeen[1*]

[1]Cognitive Sciences Department, Psychology Program, UAE University, United Arab Emirates (UAE)
Email: hdodeen@uaeu.ac.ae, Mobile: 00971507931665, ORCID ID: https://orcid.org/0000-0002-9761-198

Received: 20/12/2025
Accepted: 30/01/2026

Corresponding Author: Hamzeh Dodeen (hdodeen@uaeu.ac.ae)

ABSTRACT

This study investigated the effects of negatively worded items on the factor structure and internal reliability of psychological scales. Five scales (Helping Attitude, Life Orientation, ENRICH Marital Satisfaction, Self-Esteem, and State Self-Esteem) with both negatively and positively worded items were used with five different samples (588, 436, 355, 480, and 418) from the (masked for review) in the UAE. The statistical analysis included exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and internal reliability. The results showed that the Helping Attitude scale has two factors: the first can explain 32.57% of the total variance and represents the positively worded items, while the second can explain more than 14% of the total variance and was created by the negatively worded items. The results showed that the reliability of the entire scale was affected by the inclusion of negatively worded items. Similar results were observed for the other scales examined in the study. In conclusion, negatively worded items are problematic, negatively impact the scales factor structure, and decrease internal reliability.

KEYWORDS: Negatively Worded Items, Factor Structure, Factor Analysis, Reliability, Psychological Scales.

1. INTRODUCTION

Negatively worded items are expressed in a reverse semantic way as most other items in a measure or a scale (Barnette, 2000). The including of the negatively worded items in scales and surveys in social science research and applications in general, and in psychology in particular, is a common practice for identifying invalid responses or response effects. Response effects indicate that the answers to survey questions are based on irrelevant information rather than on the context or meaning of the question. Negatively worded items are more likely to make participants read the questions precisely and provide accurate responses (Schriesheim & Eisenbach, 1995). Although their use can help identify the presence of response effects, this practice has been shown to create other issues in the scale's construction and performance, especially its factorial structure and reliability. Research has indicated that negatively worded items create new factors that differ from the originally proposed ones (Chen, et. al, 2010; Dodeen, 2014; and Woods, 2006).

The purpose of the current study was to empirically investigate the effects of including negatively worded items on the factor structure and the internal reliability of psychological scale. It has been observed that negatively worded items usually gather separately from the other positively worded items on the scale because of their direction. This mean that the direction of the items not the items themselves (content, meaning, …) is affecting the structure of the scale. Ratability is an important psychometric characteristic of scales, and it indicates accuracy. Because reliability reflects the relationships among the items and the performance of the items in the scales, it is expected that negatively worded items affect the scale reliability too. So, this study tried to answer the following research: What are the effects of including negatively worded items on the factor structure and the reliability of psychological scales?

2. LITERATURE REVIEW

How negatively worded items affect a scale's structure and the effect of the item direction on the psychometric properties of scales and test in psychology, and mainly on the factor structure, has been investigated in many studies in related literature. Solís-Salazar (2015) empirically administered a questionnaire to a sample of 699 people from Costa Rica using different formats of positively and negatively worded items. The study reported that positive questionnaires demonstrated a better fit to the theoretical factor structure and higher reliability values, whereas the mixture of positively and negatively worded items undesirably influenced the internal reliability of the scales. Dodeen (2014) investigated how negatively worded items impact the factorial structure of the UCLA Loneliness Scale. Using responses from 1,429 college students, the results indicated that some of the received factors displayed the wording direction of the scale items. Henn et al. (2016) investigated the factorial construction of the Ryff Scales of Psychological Well-being using samples (202 adults and 226 students) from South Africa. Using both two types of factor analysis (exploratory confirmatory) it was found that the data had two factors: all positively and negatively worded items were gathered in the first and second one, respectively. The study concluded that the efficiency of negatively worded items must be reassessed.

Similarly, Roszkowski and Soven (2010) examined the psychometric behavior of including two negatively worded items (14 and18), with all the remaining items formulated in a positive way. Using a total of 3,605 undergraduate students who evaluated the university courses, the two negatively worded items had lower correlations with the total score than the other items. Moreover, the inclusion of these negatively worded items decreased the internal reliability of the scales and created a new factor that disappeared after the negatively worded items were rewritten in a positive direction. Chen et al. (2010) reported a biased response associated with the inclusion of negatively worded items in the Chinese Self-Esteem Scale. The data was collected from elementary school students in Taiwan. Using CFA, the study found that method effects introduced regular measurement errors that changed the explanations of the results.

Although several studies have reported its undesirable effects on data collection using scales, as mentioned above, negatively worded items are still included in scales and surveys. Most newly developed scales have items in two directions: positive and negative. According to Chen et al. (2010, p. 31), "several measurement textbooks have continued to note the advantages of a balance of negatively and positively worded items". Cultural differences are another factor associated with the use of negatively worded items. Various cultures deal with negative sentences differently (Dodeen, 2014; Zeng, 2020). This suggests that research on this topic should be expanded to allow more evidence and indications to be presented to researchers and practitioners, especially when examining the possible wording effects on psychological scales (DiStefano & Motl, 2006) and in different populations

and cultures.

2.1. Research Rational

This study differs from other studies in the area in several ways. First, the samples were from a non-Western population in which English was the second language. Second, it used five scales and five different samples, whereas most studies used only one scale. Additionally, these scales were carefully selected because they are frequently used in psychology, are relatively short, and include both negatively and positively worded items. Third, the study was administered to university students who could read and understand the questions or items on the scale. It has been observed that participants generally need higher levels of verbal reasoning when answering negatively worded items than directly to positively worded ones (Marsh, 1986). Thus, negatively worded items are problematic, especially for young and uneducated participants (Roszkowski & Soven, 2010).

3. METHODOLOGY

3.1. Procedure

The goal of the statistical analysis was to study how using negatively worded items affected both the factorial structure and the reliability of the psychological scales. Five scales containing both negatively and positively worded items and are commonly used in psychological research and real-life applications were selected. These scales are: Helping Attitude Scale, Life Orientation Scale, ENRICH Marital Satisfaction Scale, Self-Esteem Scale, and State Self-Esteem Scale. Real data were collected by applying these five psychological scales to different samples of college students. These scales represent a small sample of many similar scales used in general in social sciences and particularly in psychology.

Prior to statistical analysis, the data were checked for extreme values, outliers, or missing values. There were no outliers or missing data. Finally, coding of negatively worded items was reversed, and the responses were analyzed by conducting an EFA on each scale. The selected rotation method was oblique direct oblimin rotation, whereas the extraction method was principal axis factoring (PAF). PAF is a more appropriate method when the goal of conducting factor analysis is to detect the structure of the data or the factors of scales (Grieder & Steiner, 2022). The suitability of the data for factor analysis in each analysis was assessed applying two common tests: the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. KMO values that range from 0.8 to 1.0 indicate that the sample is adequate, while a significant value (< 0.05) of Bartlett's test of sphericity indicates that factor analysis is suitable for the data set (Hair et al. 2006; Sharma, 1996; Tabachnick & Fidell, 2007). An eigenvalue criterion greater than one was applied to extract and retain the primary factors of each scale.

To collect information about the participants, a few demographic variables (e.g., age, sex, and GPA) were added to each scale. Google Forms was used to administer each scale to a separate random sample of students at different times. Data were collected and analyzed using SPSS and EQS6.1 Windows. Several statistical analyses were applied to the data of each scale, including descriptive statistics and assessment of the reliability (internal reliability as estimated by Cronbach's alpha) of negatively worded items, positively worded items, and both. The factorial structure of each scale was examined by both EFA and CFA.

3.2. Instruments

The Helping Attitude Scale (Nickell, 1998) has a total of 20 items and uses Likert-type scale that ranges between 1 (strongly disagree) to 5 (strongly agree) so the total score of the scale ranges between 20 and 100. The scale has five negatively worded items; for example, "Helping others is usually a waste of time" (item 1). The Life Orientation Scale (Scheier et al., 1994) has only six items that assess optimism versus pessimism using a 5-point Likert-type scale ranges from 0 (strongly disagree) to 4 (strongly agree). The scale comprised three negatively worded items; for example, "I rarely count on good things happening to me" (item 9). ENRICH Marital Satisfaction Scale (Fowers & Olson, 1993) consists of 10 items and uses a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree), with the total score ranging from 10 to 50. The scale comprised five negatively worded items; for example, "I am not satisfied with the way we each handle our responsibilities as parents" (item 12). Self-Esteem Scale (Rosenberg, 1965) has a total of 10 items that assess personal self-esteem on a 4-point Likert-type scale that ranges between 1 (strongly disagree) to 4 (strongly agree), with a total score that ranges between 10 to 40. The scale has five negatively worded items, for example, "I wish I could have more respect for myself" (item 8). Finally, the State Self-Esteem Scale (Heatherton & Policy, 1991) includes a total of 20 items and uses a 5-point Likert-type scale that ranges between 1 (not at all) to 5 (extremely). It includes 13 negatively worded items, for example, "I am worried about what other people

think of me" (item 13).

3.3. Participants

The study conducted on the students at the (Masked for review) which is a public university in the United Arab Emirates (UAE) with an enrollment of approximately 15,000 students. Students voluntarily participated in this study by responding to one of the five scales which were conducted separately and in different times. Students were clearly informed that they can withdraw from the project anytime without any consequences. Moreover, the participants were guaranteed that the collected data would be confidential and used only for research purposes. The Research Ethics Committee at the (masked for review) approved the study and the used scales.

Even though English is a second language for the majority of students at the University, they have a strong grasp of it. In fact, English serves as the medium for teaching and communication within this university community. To enroll in their chosen academic majors or programs, students are required to achieve high scores on international equivalency exams like The International English Language Testing System (IELTS). Therefore, employing psychological scales written in English posed no problem for this study.

Five different samples (588, 436, 355, 480, and 418 students for a total of 2,277 students) were randomly selected and participated in the study. Data on sex, age, and GPA were collected from the participating students, and the results are shown in Table 1.

Table 1: Demographic Variables of the Participating Students.

c		Scale				
		Helping Attitude	Life Orientation	Marital Satisfaction	Self-Esteem	State Self-Esteem
Sex	Males Number	127	98	43	95	39
	Percentage	(21.6%)	(22.5%)	(12.1%)	(19.8%)	(9.3%)
	Females Number	481	338	312	385	379
	Percentage	(78.4%)	(77.5%)	(87.9%)	(80.2%)	(90.7%)
Age	Mean (SD)	20.85 (2.32)	20.85 (2.42)	21.38 (2.11)	21.65 (2.48)	20.29 (2.46)
GPA	Mean (SD)	3.06 (.52)	3.16 (1.18)	3.00 (.64)	2.86 (1.05)	2.77 (1.10)

Table 1 shows that both sexes were represented in each sample; however, there were more females than males in each one. This approximately reflects the actual ratio (70% females vs. 30% males) between males and females in this university. The average age is between 20 and 21 years and the average GPA was around 3.0 (on a 4.0-points scale).

4. RESULTS

EFA was applied to the data of each of the five scales using PAF as the extraction procedure and oblique rotation. The appropriateness of the current data to factor analysis was checked and confirmed using two common tests: KMO and Bartlett's Test of Sphericity. For the Helping Attitude Scale, the EFA data produced three factors. However, the third factor was minor, with an eigenvalue of 1.05, which is close to one. Thus, this factor was ignored and only two factors were considered (see the scree plot in Figure 1).

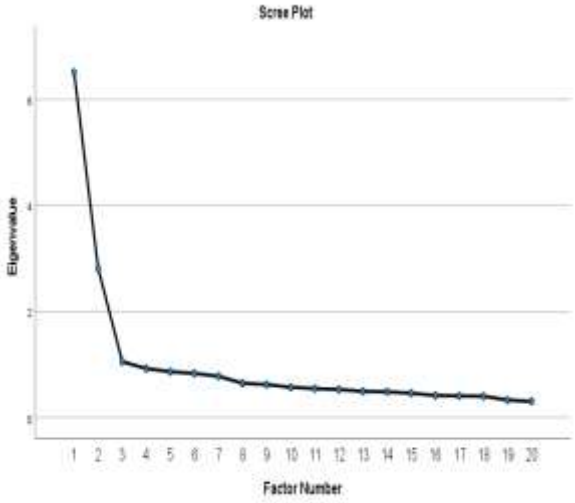

Figure 1: Scree Plot Showing the Number of Extracted Factors for the Helping Attitude Scale.

The first factor had an eigenvalue of 6.51 and explained 32.57% of the total variance, whereas the second factor had an eigenvalue of 2.18 and explained 14.07% of the total variance. The first factor represents the positive items. This can be observed from the loading values. All positively worded items had high loadings on this factor. In contrast, the negatively worded items created another factor and were highly loaded on it. The six negatively worded items had loadings of .44, .64, .66, .74, .65, and .63, respectively, on the second factor. The direction of the negatively worded items changed the factor structure of the scale and created a second factor. The internal reliability of the entire scale (measured by Cronbach's alpha) was .87. When this analysis was separated by the direction of the items, a higher reliability value (alpha = .89) was obtained for positively worded items, although the number of items decreased by dividing the scale (generally,

reliability is related to the number of items in the scale). This clearly shows that the negatively worded items decreased the reliability of the scale.

Table 2: EFA and Reliability Results of the Helping Attitude Scale.

	Factor 1	Factor 2		Internal Reliability Analysis		
Eigenvalue	6.51	2.18		Scale	No of Items	Cronbach's Alpha
Percentage of Variance	32.57 %	14.07 %		All Items	20	.87
				Positively Worded Items	14	.89
				Negatively Worded Items	6	.81
Loadings						
Item	Factor 1	Factor 2	Item	Factor 1	Factor 2	
Item1-negative	.14	**.44**	**Item11-negative**	.06	**.74**	
Item2	.57	.15	Item12	.69	.10	
Item3	.51	.12	Item13	.41	-.07	
Item4	.62	.08	Item14	.57	-.12	
Item5-negative	.01	**.64**	Item15	.77	-.01	
Item6	.68	.04	Item16	.60	.04	
Item7	.70	.07	Item17	.73	-.01	
Item8-negative	.08	**.66**	**Item18-negative**	-.07	**.65**	
Item9	.69	.08	**Item19-negative**	-.05	**.63**	
Item10	.44	-.16	Item20	.66	-.04	

The second scale is the Life Orientation Scale. The results of the EFA and reliability examination of this scale are presented in Table 3. The EFA created two clear factors; the first had an eigenvalue of 1.90 and could explain 31.59% of the variance, while the second extracted factor had an eigenvalue of 1.30 and explained 21.61%. Figure 2 presents a scree plot of the extracted factors of this scale.

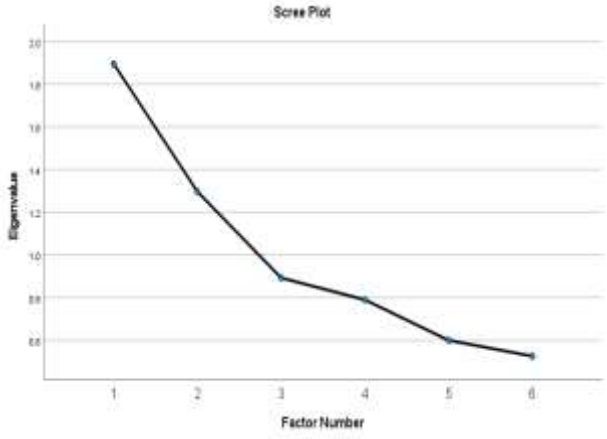

Figure 2: Scree Plot Showing the Number of Extracted Factors for the Life Orientation Scale.

Examining the items and their loadings on each extracted factor showed that the first factor included only positive items (items 1, 3, and 6), while the second factor had only negatively worded items (items 2, 4, and 5). Items were loaded into factors based on their direction, not on their content. This means that the item direction changed the factorial structure of the scale. The reliability analysis supported the EFA results. The internal reliability of the whole scale (.51) was lower than that of the positively worded items (.70), although the number of items decreased.

Table 3: EFA and Internal Reliability Results of the Life Orientation Scale.

	Factor 1	Factor 2	Internal Reliability Analysis		
Eigenvalue	1.90	1.30	Scale	No of Items	Cronbach's Alpha
Percentage of Variance	31.59 %	21.61 %	All Items	6	.40
			Positively Worded Items	3	.70
			Negatively Worded Items	3	.36
Loadings					
Item	Factor 1	Factor 2	Item	Factor 1	Factor 2
Item1	.66	-.01	**Item4_negative**	-.01	**.30**
Item2_negative	-.05	**.38**	**Item5_negative**	-.07	**.54**
Item3	.71	.01	Item6	.60	.00

The EFA results and the reliability analysis of the third scale, ENRICH Marital Satisfaction Scale, are presented in Table 4. The EFA created two factors; the first had an eigenvalue of 3.97 and explains 39.66% of the total variance, while the second factor had an eigenvalue of 1.48, which explains 14.79%. Figure 3 presents a scree plot of the extracted factors.

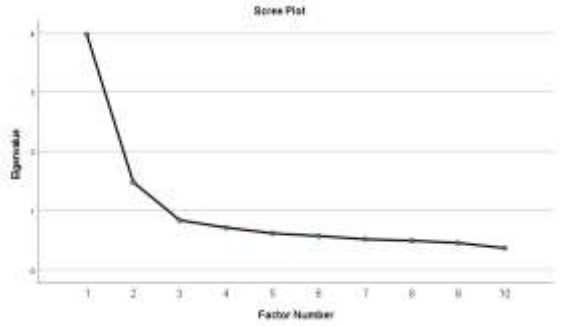

Figure 3: Scree Plot Showing the Number of Extracted Factors for the ENRICH Marital Satisfaction Scale.

Examining the items and their loadings on each

factor revealed that the first factor consisted of positively worded items (except item 9), while the second factor consisted of negatively worded items (items 1, 3, 5, and 8). The items were loaded into factors based on their direction, not on their content. This indicates that the direction of the items changed the factor structure of the scale. The reliability of the full scale (10 items) (.73) was higher than that of the positively (six items) and negatively (four items) worded items, which were .68 and .69, respectively. The reliability of this scale was not affected by the direction of the items.

Table 4: EFA and Internal Reliability Results of the Marital Satisfaction Scale.

	Factor 1	Factor 2	Internal Reliability Analysis		
Eigenvalue	3.97	81.4	Scale	No of Items	Cronbach's Alpha
Percentage of Variance	39.66	14.97	All Items	10	.73
			Positively Worded Items	6	.68
			Negatively Worded Items	4	.69
		Loadings			
Item	Factor 1	Factor 2	Item	Factor 1	Factor 2
Item1_negative	.16	**-.46**	Item6	.64	-.02
Item2	.68	-.01	Item7	.71	-.07
Item3_negative	.19	**-58**	**Item8_negative**	-.02	**-.66**
Item4	.72	.08	Item9	.05	.64
Item5_negative	-.06	**-.57**	Item10	.67	-.05

The results of the EFA and reliability analysis of the fourth scale, Self-Esteem Scale, are summarized in Table 5. The EFA created two factors; the first had an eigenvalue of 2.56 and explains 25.58% of the total variance, and the second factor had an eigenvalue of 1.60 and explains 16.03% of the total variance. Figure 3 presents a scree plot of the extracted factors.

Figure 4: Scree Plot Showing the Number of Extracted Factors for the Self-Esteem Scale.

Examining the items and their loadings on each factor showed that the first factor consisted of positive items, whereas the second factor consisted of negatively worded items (items 2, 5, 6, 8, and 9). The items on the scale were loaded into factors based on their direction, and not on their content. This indicated that the direction of the items changed the factor structure of the scale. The reliability analysis supported the results of the EFA. The reliability of the entire scale was .65, lower than that of the negatively worded items (.71), although the number of items decreased.

Table 5: EFA and Internal Reliability Results of the Self-Esteem Scale.

	Factor 1	Factor 2	Internal Reliability Analysis		
Eigenvalue	2.56	1.68	Scale	No of Items	Cronbach's Alpha
Percentage of Variance	26.03	16.80	All Items	10	.65
			Positively Worded Items	5	.56
			Negatively Worded Items	5	.71
		Loadings			
Item	Factor 1	Factor 2	Item	Factor 1	Factor 2
Item1	.36	.30	**Item6_negative**	**.58**	-.06
Item2_negative	**.42**	.03	Item7	.07	.38
Item3	.04	.43	**Item8_negative**	**.61**	.13
Item4	-.08	.48	**Item9_negative**	**.59**	.08
Item5_negative	**.63**	.01	Item10	.14	.58

Finally, Table 6 shows the results of the EFA and reliability examination for the State Self-Esteem Scale. There were five factors extracted with eigenvalues larger than one; however, the last two factors were ignored because they have eigenvalues close to one (1.04, and 1.02). This results in only three essential factors; the first had an eigenvalue of 6.40 and can explain 31.98% of the variance, the second factor had an eigenvalue of 2.33 and explains 11.63%, and the third factor has an eigenvalue of 1.33 and explains 6.63%. Figure 5 shows a scree plot of the extracted factors.

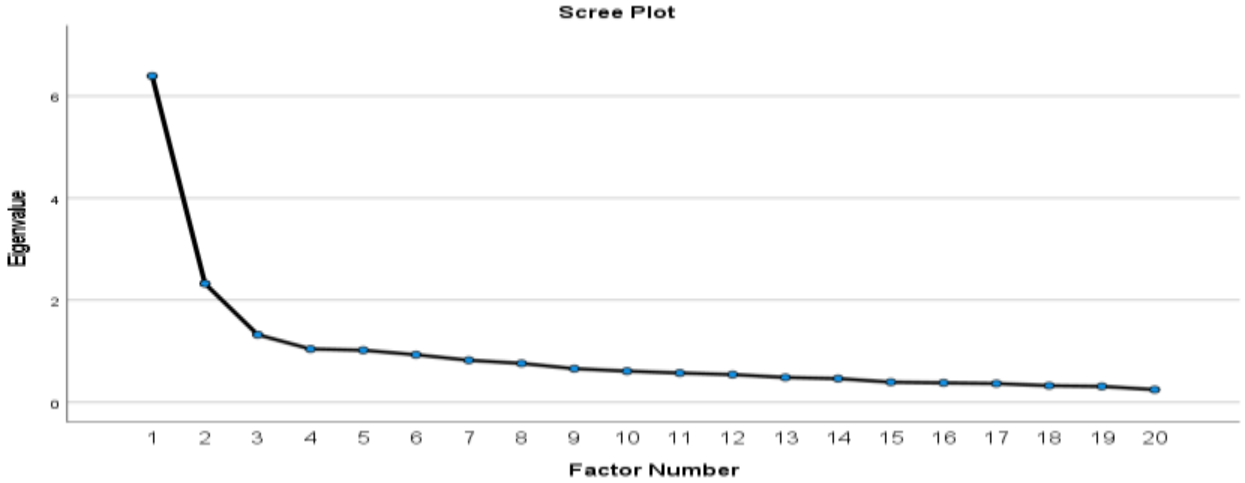

Figure 5: Scree Plot Showing the Number of Extracted Factors for the State Self-Esteem Scale.

Examining the items and their loadings on the first two factors showed that the first factor consisted mainly of the positive items, whereas the second factor contained mainly the negatively worded items. The direction of the items changed the factorial structure of the scale because factor loadings were based on the item direction and not the item content. However, few items (such as 3, 7, and 17) have exceedingly small loadings and are identical in both factors. The reliability of the whole scale (.88) was higher than that of the positively worded items (7) and the negatively worded items (13) items, which was .85. The reliability of this scale was not influenced by the direction of item direction as the case of the other scales.

Table 6: EFA and Internal Reliability Analysis of the State Self-Esteem Scale

	Factor 1	Factor 2	Internal Reliability Analysis		
Eigenvalue	6.40	2.33	Scale	No of Items	Cronbach's Alpha
Percentage of Variance	31.98	11.63	All Items	20	.88
			Positively Worded Items	7	.85
			Negatively Worded Items	13	.85
Loadings					
Item	Factor 1	Factor 2	Item	Factor 1	Factor 2
Item1	.18	-.59	Item11	.08	-.50
Item2_negative	**.47**	-.06	Item12	.06	-.43
Item3	-.13	-.20	**Item13_ negative**	**.21**	-.12
Item4_ negative	**.46**	-.01	Item14	.02	-.79
Item5_ negative	**.64**	-.18	**Item15_ negative**	**.56**	.05
Item6	.11	-.43	**Item16_negative**	**.39**	-.08
Item7_ negative	**.12**	.15	**Item17_ negative**	**-.05**	-.07
Item8_ negative	**.10**	.05	**Item18_ negative**	**.40**	-.09
Item9	-.08	-.76	**Item19_ negative**	**.74**	-.14
Item10_ negative	**.61**	.04	**Item20_ negative**	**.45**	-.04

4.1. Confirmatory Factor Analysis (CFA)

To support the results of the EFA described in Tables 2-6 above, a CFA was conducted to test the 1-factor model for each scale. Five commonly used indices were applied to test the goodness-of-fit of the data of each scale: nonnormed fit index (NNFI), comparative fit index (CFI), goodness-of-fit index

(GFI), standardized root-mean-square residual (SRMR), and root mean square error of approximation (RMSEA). By comparing the fit index values summarized in Table 7 with each corresponding norm, it can be concluded that the data for each scale did not fit the hypothesized model (1-factor model). This supports the EFA results and emphasizes that the inclusion of negatively worded items changed the factorial constitution of each of the five scales.

Table 7: Goodness-of-fit Indexes of the Factor Structure Model of the Five Scales.

	Fit Index				
	NNFI	*CFI*	*GFI*	*SRMR*	*RMSEA*
Helping Attitudes Scale	.69	.72	.75	.11	.11
Life Orientation Scale	.84	.91	.98	.06	.08
Marital Satisfaction Scale	.76	.81	.86	.09	.12
Self-Esteem Scale	.63	.71	.92	.09	.10
Self-State Scale	.62	.66	.71	.10	.12

Note: NNFI = nonnormed fit index; CFI = comparative fit index; GFI = goodness-of-fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation. NNFI, CFI, and GFI values (.90), SRMR (.06), and RMSEA (.05) indicated a good fit.

4.2. Discussion

The results were similar to those of the five psychological scales used in this study. The negatively worded items in each scale generated a new factor besides the first main factor that has the positively worded items. Negatively worded items change the factorial structure of each scale. For example, the Helping Attitude Scale comprises two factors, each representing the direction of the items. The first factor represents the positively worded items, whereas the negatively worded items are gathered to create the second factor. The direction of the items changed the factor structure of the scale and created a second factor. The same was observed for the other scales used in this study.

Regarding reliability analysis, the study showed that, in general, negatively worded items decreased the reliability of the psychological scales. For example, the results showed that the reliability value of the full Helping Attitudes Scale, as measured by Cronbach's alpha, was .87. At the same time, the reliability value for the positively worded items only was .89, which is higher than the reliability value of the entire scale, although the number of items decreased when the scale was broken down into positive and negatively worded items. However, item wording did not affect the reliability of other psychological scales, as was observed with the Marital Satisfaction and State Self-Esteem scales.

When developing a scale, all items are assumed to measure the same targeted construct and in the same direction. If the wording affects the performance of the item, then the item is not functioning in measuring the construct as it is supposed to be. Scales are tools to collect data from participants. These tools are supposed to be accurate, reliable, and valid. The results of this study revealed that both validity and reliability of scales are undesirably affected by the including of the negatively worded items. Negatively worded items change the factor structure of scales which affects their construct validity. At the same time, the overall reliability of scales decreased. Accordingly, the use of these items should be avoided or at least limited.

The fifth scale (the State Self-Esteem Scale) clearly showed a slightly different picture than the other four scales because five factors have been extracted. Additionally, the reliability of this scale was not affected by the direction of item wording. The clear difference between this scale and the other four scales is the relatively high number (percentage) of negatively worded items. The State Self-Esteem Scale has 13 (65%) negatively worded items, which is higher than the percentage of the negatively worded items in any of the other four scales used in the study. However, this conclusion cannot be confirmed from the current study results.

The findings of this study raise an important related question: Why do negatively worded items affect the factor structure and the reliability of psychological scales? Searching the related literature provided some explanations. One of these explanations is the difficulty in understanding negatively worded items compared to positively direct ones. As they are negative, they are indirect, and thus, more difficult to comprehend. In this regard, it has been observed that some participants had difficulty interpreting questions or items on scales (e.g., Bors et al., 2006) or had limited reading ability (Cordery & Sevastos, 1993). Participants require a greater level of verbal reasoning when responding to negatively worded items than when responding directly to positively worded items (Marsh, 1986). Cultural differences also affect the meaning of items when they are stated negatively. This is because different cultures deal with negative

sentences differently, yielding different responses depending on the culture (Dodeen, 2014). This issue is important especially when dealing with translated scales from culturally different countries and in multi-cultural research and studies. Another explanation is carelessness which is sometimes the cause of the problem when using negatively worded items (Roszkowski & Soven, 2010).

5. CONCLUSION

In conclusion, the study examined the effects of the use of mixed wording in scales, which is extremely common and crucial for improving the quality of our measures. Results confirmed that negatively worded items have an adverse impact on the psychometric evidence associated with a scale. Both factor structure and reliability of scales are seriously affected by negatively worded items. Researchers and practitioners are highly recommended not to include negatively worded items when developing or using scales. Several alternative procedures are available in the related literature.

5.1. Practical Implications

The study recommends avoiding the use of negatively worded items in psychological scales as they undesirably affect their psychometric properties. It has been clearly described above that negatively worded items change the factors' structure of the scales and decrease their internal reliability. As stated by Roszkowski and Soven (2010, p. 128), "negatively phrased items have been found to function aberrantly, producing artificial factors and lowering the internal consistency of scales." Therefore, the use of negatively worded items should be questioned. Unless there is a good reason for including negatively worded items, it is more recommended to use positively or directly worded items (Barnette, 2000).

It is advised that researchers, psychologists, and practitioners exclude negatively worded (reverse-coded) items when developing new scales. Furthermore, the mixed wording technique can result in misleading factor solutions, where the factors represent item wording rather than the actual construct.

Consequently, researchers might mistakenly determine that a scale is multidimensional when it is not. Factor analyses that have been published using mixed wording may require reinterpretation, particularly if negatively worded items load onto separate factors or if cross-loadings occur. Based on the findings of this study, it is advisable to reassess widely used measures and scales that include negatively worded items by employing all positively worded items or by separately modeling wording effects. Lastly, for practitioners, educators, and psychologists, decisions based on such scales might be less precise or biased. Individual scores could indicate comprehension difficulties rather than genuine traits.

Alternative procedures have been recommended to manage response bias in survey measurements. These procedures included the use of phrase completion (Hodge & Gillepsie, 2003) and mixed response options (response options were mixed in positive and negative directions, but all items were worded in the same direction). That is, half of the items could be assessed by scores ranging from "strongly disagree" to strongly agree" and the other half could use "strongly agree" to "strongly disagree." Another suggestion is to use fewer negatively worded items (filler items) to control for response bias and not use these items in the calculation of the scores (Forsterlee & Ho, 1999).

5.2. Limitation and Future Research

This study has a limitation in that it focuses on a narrow group, specifically college students from one university. Another constraint involves the data collection process, as data was gathered at a single time point for each scale, which did not allow for the evaluation of test-retest reliability and the longitudinal stability of wording effects. Lastly, the study only considered five psychological scales. Although these are commonly used, they represent just a small portion of psychological assessments. Another limitation is the selection of the scales used in the analysis. The five scales were not randomly selected to represent available psychological scales. This might limit the generalizability of the results of this study. Future research can include important variables that might affect the results and give better understanding of the effects of including negatively worded items.

For example, the number of negatively worded items (percentage) varies from one scale to another. In this study we have a scale with only three negatively worded items and another scale with 13. This can be investigated as a variable that affects the factor structure and reliability of the scales.

Future studies can also examine more psychological scales to confirm the undesirably effect of using mixed wording on their psychometric properties. Additionally, the influence of some variables such as gender, age, and culture of respondents can be investigated for better

understanding and knowledge.

Data Availability Statement: The data that support the findings of this study are openly available in [figshare] at https://figshare.com/account/items/24448417/edit

Institutional Review Board Statement: The Ethical Committee of the (masked for review) has granted approval for this study on 20/5/20, Ref. No. ERS-2020-6132.

REFERENCES

Barnette, 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), 361–70.

Bors, D. A., Vigneau, F., & Lalande, F. (2006). Measuring the need for cognition: Item polarity, dimensionality, and the relation with ability. Personality and Individual Differences, 40, 819-828.

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

Cordery, J., & Sevastos, P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively worded items? Journal of Applied Psychology 78(1), 141–3.

DiStefano, C., & Motl, R. (2006). Further investigating method effects associated with negatively worded items on Self-report surveys. Structural Equation Modeling: A Multidisciplinary Journal, 13(3), 440–464.

Dodeen, 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.

Forsterlee, R., & Ho, R. (1999). An examination of the short form of the need for cognition scale applied in Australian sample. Educational and Psychological Measurement, 59(3), 471-480.

Grieder, S., & Steiner, M. D. (2022). Algorithmic jungle: A comparison of implementations of principal axis factoring and Promax rotation in R and SPSS. Behavior Research Method, 54, 54–74.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). Multivariate data analysis. New Jersey: Pearson University Press.

Heatherton, T. F., & Policy, J. (1991). Development and validation of a scale for measuring state self-esteem. Journal of Personality and Social Psychology, 60, 895-910.

Henn, C. M., Hill, C., & Jorgensen, L. I. (2016). An investigation into the factor structure of the Ryff Scales of Psychological Well-Being. South African Journal of Industrial Psychology, 42(1), a1275.

Hodge, D. R., & Gillepsie, D. (2003). Phrase completion: An alternative to the Likert scale. Social Work Research, 27(1), 45–55.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon. Developmental Psychology, 22, 37–49.

Nickell, G. (1998). The helping attitudes scale. Paper presented at the 106th Annual Convention of the American Psychological Association.

Rosenberg, M. (1965). Society and adolescent self-image. Princeton, New Jersey: Princeton University Press.

Roszkowski, M., & Soven. J. (2010) Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. Assessment and Evaluation in Higher Education, (35), 113–130.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the Life Orientation Test. Journal of Personality and Social Psychology, 67, 1063-1078.

Schriesheim, C., & Eisenbach, R. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. Journal of Management, 21(6), 1177-1193.

Sharma, S., (1996). Applied multivariate techniques. John Wiley and Sons, Inc., New York.

Solís-Salazar, M. (2015). The dilemma of combining positive and negative items in scales. Psicothema, 27(2), 192-199. doi:10.7334/ psicothema2014.266

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed). Pearson Education Inc. Boston, MA.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. Journal of Psychopathology and Behavioral Assessment, 28(3), 186-191.
Zeng B, Wen H, Zhang J. (2020). How does the valence of wording affect features of a scale? The method effects of the undergraduate learning Burnout Scale. Front Psychol, 28; 11:585179.