Psicothema

Psicothema 2018, Vol. 30, No. 2, 149-158
doi: 10.7334/psicothema2018.33

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Using reversed items in Likert scales: A questionable practice

Javier Suárez-Alvarez[1], Ignacio Pedrosa[2], Luis M. Lozano[3], Eduardo García-Cueto[4], Marcelino Cuesta[4] and José Muñiz[4]

[1] Organization for Economic Cooperation and Development, [2] CTIC Technologic Center, [3] Universidad de Granada and [4] Universidad de Oviedo

Abstract

Background: The use of positively worded items and reversed forms aims to reduce response bias and is a commonly used practice nowadays. The main goal of this research is to analyze the psychometric implications of the use of positive and reversed items in measurement instruments. **Method:** A sample of 374 participants was tested aged between 18 and 73 (M=33.98; SD=14.12), 62.60% were women. A repeated measures design was used, evaluating the participants with positive, reversed, and combined forms of a self-efficacy test. **Results:** When combinations of positive and reversed items are used in the same test the reliability of the test is flawed and the unidimensionality of the test is jeopardized by secondary sources of variance. In addition, the variance of the scores is reduced, and the means differ significantly from those in tests in which all items are either positive or reversed, but not combined. **Conclusions:** The results of this study present a trade-off between a potential acquiescence bias when items are positively worded and a potential different understanding when combining regular and reversed items in the same test. The specialized literature recommends combining regular and reversed items for controlling for response style bias, but these results caution researchers in using them as well after accounting for the potential effect of linguistic skills and the findings presented in this study.

Keywords: Reversed items, recoded, validity, responses bias, item response theory.

Resumen

El uso de ítems inversos en las escalas tipo Likert: una práctica cuestionable. **Antecedentes:** el uso de ítems formulados positivamente junto con otros inversos es una práctica habitual para tratar de evitar sesgos de respuesta. El objetivo del presente trabajo es analizar las implicaciones psicométricas de utilizar ítems directos e inversos en la misma prueba. **Método:** se utilizó una muestra de 374 participantes con edades comprendidas entre 18 y 73 años (M=33.98; DT=14.12), con un 62,60% de mujeres. Mediante un diseño de medidas repetidas se evaluó a los participantes en una prueba de autoeficacia con tres condiciones: todos los ítems positivos, todos negativos y un combinado de ambos. **Resultados:** cuando se utilizan en la misma prueba tanto ítems positivos como negativos su fiabilidad se deteriora, y la unidimensionalidad de la prueba se ve comprometida por fuentes secundarias de varianza. La varianza de las puntuaciones disminuye, y las medias difieren significativamente respecto de las pruebas en las que todos los ítems están formulados positiva o negativamente. **Conclusiones:** los resultados de este estudio presentan una disyuntiva entre un posible sesgo de aquiescencia cuando los ítems tienen una redacción positiva y una comprensión potencialmente diferente cuando se combinan ítems regulares e invertidos en la misma prueba. La literatura especializada recomienda combinar ítems regulares e invertidos para poder controlar el sesgo del estilo de respuesta, pero estos resultados advierten a los investigadores que los usen también después de tener en cuenta el potencial efecto de las habilidades lingüísticas y de los hallazgos presentados en este estudio.

Palabras clave: ítems invertidos, recodificación, validez, sesgo de respuestas, teoría de respuesta al ítem.

The answers to the items of a test can be influenced by personal factors that may affect both the scores and the validity of the interpretations (Cronbach, 1946, 1950; Ferrando & Lorenzo-Seva, 2010; Fonseca-Pedrero & Debanné, 2017; Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016). This is known as response bias, and refers to any individual tendency to respond independently of the content that the item is evaluating, distorting the score in the trait being measured. Two types of response bias can be distinguished, called response set, and response style (Chiorri, Anselmi, & Robusto, 2009; van Sonderen, Sanderman, & Coyne, 2013). *Response set bias* refers to the item content, for instance choosing a socially desirable answer rather than the truth (social desirability). Various alternatives have been proposed in an attempt to avoid this kind of bias, most notably forced-choice questionnaires in which examinees choose between two items with similar social desirability (Brown, 2015). In addition, in recent years solutions have been proposed within the framework of Item Response Theory (IRT; Brown & Maydeu-Olivares, 2012). *Response style bias* is the tendency to respond to items without paying enough attention to their content. Acquiescence or the tendency to agree with statements is an example of response style bias. This type of bias can produce answer patterns which do not reflect the real profile of the examinees, and constitute a significant

Received: February 2, 2018 • Accepted: March 22, 2018
Corresponding author: Javier Suárez-Alvarez
Organization for Economic Cooperation and Development
2, rue André Pascal - Paris Cedex 16
75775 Paris (Francia)
e-mail: javier.suarez-alvarez@oecd.org

threat to the validity of the interpretations based on the self-report scores (van Sonderen et al., 2013). The use of items in both regular (positively worded) and reversed forms was introduced decades ago with the aim of reducing response style bias (Nunnally, 1978; Paulhus, 1991). There are two main strategies for reversing items in order to reduce acquiescence bias. The first consists of adding negation in such a way that the meaning of the item is changed without substantially changing the text (e.g. "*I consider myself a good person*" vs "*I do not consider myself a good person*"). The second can be achieved by using an antonymic expression (e.g. "*I consider myself a bad person*"). To reduce response style bias, test developers recommend that some of the items making up a test are reverse-keyed (Abad, Olea, & Ponsoda, 2011; Nunnally, 1978; Paulhus, 1991; Prieto & Delgado, 1996). More specifically, the most common practice is to include items with negations (Swain, Weathers, & Niedrich, 2008). However, there are several reasons to criticize this strategy, and claim there are more disadvantages than advantages (Weijters & Baumgartner, 2012; Weijters, Cabooter, & Schillewaert, 2010; Weijters, Geuens, & Schillewaert, 2009). The first, and most obvious disadvantage, is that the strategy of including reverse-keyed items contradicts one of the principal guidelines for item development: try to avoid negative formulations (Haladyna, Downing, & Rodríguez, 2002; Haladyna & Rodríguez, 2013; Lane, Raymond, & Haladyna, 2016; Moreno, Martínez, & Muñiz, 2004, 2006, 2015). In addition, inverting items by using an antonymic expression can produce problems of interpretation because the meaning of the item can change substantially (for example, does "*I am not a good person*" mean the same as "*I am a bad person*"?). Previous research suggests that the cognitive processing of these two types of items is not necessarily the same, even more so when reading skills are poor (Marsh, 1986, 1996). Furthermore, the use of reversed items together with direct items implies that responses to reversed items have to be recoded in order to obtain the total score of the scale. This process assumes that the two extremes of a Likert-type item (e.g. "*Completely disagree*" and "*Completely agree*") give exactly the same score and have the same semantic meaning in the construct being measured, and these assumptions are questionable and affect the psychometric properties of the test. Essau et al. (2012) carried out a cross-cultural research in five European countries to analyze the factorial structure of the *Strength and Difficulties Questionnaire* (SDQ), concluding that when reversed items are removed the model fit to the data improves significantly, both for the whole sample and by country. van Sonderen et al. (2013) compared the psychometric properties of a set of regular items with a set of items containing both regular and reverse-keyed items. Their hypothesis is that if reversing items reduces response bias, it would be expected that two identical items with respect to content but different in direction would be more strongly related than two items formulated in the same direction but with slightly different content. The results showed that the reversed items did not reduce response bias. Furthermore, some answer patterns suggest that the scores were affected by participants' inattention and confusion when items were combined. In addition, different studies show that including these types of items especially affects unidimensional instruments, making the model fit worse, and increasing the rejection of unifactorial models in favor of multidimensional ones (Dunbar, Ford, Hunt, & Der, 2000; Horan, DiStefano, & Motl, 2003; Woods, 2006). Additionally, when these types of items are included, the internal consistency of the test is flawed, and atypical response patterns appear (Bourque & Shen, 2005; Carlson et al., 2011; Hughes, 2009). These results converge with those found within the IRT framework (Ebesutani et al., 2012), in which items that are not reverse-keyed demonstrate better precision (Information Function) and discriminatory power (parameter *a*).

Despite the fact that in recent years the inclusion of regular and reversed items in the same test has begun to be questioned, little work has addressed the topic systematically and, in fact, it is still recommended (Weijters, Baumgartner, & Schillewaert, 2013). Most measurement instruments used both in research as well as in the different areas of applied psychology still include both types of items in the same test. From a methodological point of view, one of the main limitations found in previous research is the use of different samples to assess the different type of items, which does not guarantee the comparability of the results, confounding items and participants effects. The most rigorous way to assess the effect of combining regular and reversed items is evaluating the same examinees at different times, using a repeated measures design. To date, this design has not been used, and this will be our purpose. The main objective of this research is to analyze the effect of using reversed items on the psychometric properties of the test. All participants were evaluated three times in different ways with a self-efficacy test.

Previous research suggests that cognitive processing is not the same for positive and negative formulated items (Marsh, 1986, 1996; Mestre, 1988). To analyze if the formulation of the items (positively, negatively, combined) influences the results, a general intelligence test (abstract reasoning) and two verbal comprehension scales were administered, which allowed a more detailed analysis of the participants' responses. Another aspect to highlight is that, in addition to the Classical Test Theory perspective, IRT models have been used, which allow for more precise analysis of the measuring instruments' psychometric properties (De Ayala, 2009; van der Linden & Hambleton, 1997; Wilson, 2005). In short, this study was conducted to compare psychometric properties of the self-efficacy test across three forms (Form A: positively worded items; Form B: reversely worded items; and Form C: both type of wording combined) via reliability coefficients, item discrimination indices, goodness of fit of the one-factor model, measurement invariance tests, and mean comparisons and correlations.

Within this research context, and according to the previous reasoning, six main hypotheses guide our research. The combination of regular and reversed items in the same test was originally introduced, among other reasons, for the purpose of improving the psychometric properties of the test. However, contrary to this original motive, it seems that combining both regular and reversed items in the same test introduces noise to the assessment. As a consequence, our first hypothesis is that the discrimination indices of the items and the reliability of the test scores will be flawed when regular and reversed items are combined in the same test. As a logical consequence of the decrease in the internal consistency of the test (test reliability and discrimination indices of the items) when the items are combined, the second hypothesis proposed is that the goodness of fit of the one-factor model will be worse for tests composed of regular and reversed items. If the second hypothesis is confirmed, a third hypothesis related to the inexistence of a strong factorial invariance for the three forms of the test used is proposed: regular, reversed, and combined items. The combination of regular and reversed items

in the test requires the responses to reversed items to be recoded, assuming that the two extremes of a Likert-type scale can give exactly the same score. However, examinees' responses tend to disagree with reversed items more than they agree with regular items (Solís-Salazar, 2015). Therefore, the fourth hypothesis is that there are statistically significant differences in the average scores between the regular, reversed, and combined forms of the same test. Specifically, we expect Form B to show the highest mean and Form A to show the lowest mean. The cognitive process used by respondents for regular and reversed items is not necessarily the same since the comprehension of a reversed item requires better linguistic skills. The difficulty in comprehension is aggravated when people have to alternate between processing regular and reversed items. Therefore, our fifth hypothesis is that score differences between regular, reversed, and combined forms disappear when controlling for verbal comprehension. Finally, the logical consequence derived from the combination of regular and combined items is that if the combination of items does, in fact, reduce the acquiescence bias, the variability of the responses should be greater in the combined form than in the forms in which all of the items are either regular or reversed. Therefore, the sixth hypothesis is that combining regular and reversed items in the same test would increase the variability of the responses.

Method

Participants

The sample used is incidental and was composed of 374 participants from the general Spanish population, evaluated at three different times. The ages ranged from 18 to 73 years old (M=33.98; SD=14.12), 62.60% were women. In terms of educational level, 12.10% had completed compulsory secondary education, 31.50% had finished further education, 10.60% had vocational training and 45.80% had been in higher education.

Instruments

Self-efficacy questionnaire

This test was originally developed in Spanish (Suárez-Álvarez, Pedrosa, García-Cueto, & Muñiz, 2014). The test comprises 20 Likert-type items (positively worded) with a scale 1 to 5, in which *1* means *completely disagree* with the statement and *5* means *completely agree*. The questionnaire shows adequate psychometric properties in a sample of Spanish adolescents (α=.98; 30% of the total variance explained by the first factor; Suárez-Álvarez et al., 2014). The one-dimensional structure of the self-efficacy questionnaire has been confirmed in different samples (Muñiz, Suárez-Álvarez, Pedrosa, Fonseca-Pedrero, & García-Cueto, 2014; Suárez-Alvarez et al., 2014). The psychometric properties of the participants assessed in this study are presented in the Results section.

In order to test the hypotheses proposed, three different forms of the same test were developed. The first, Form A, was made up of 20 items, all positively formulated (e.g. "*I am able to overcome obstacles*" or "*I make use of resources around me*"). The second, Form B, was made up of 20 items, all negatively formulated. Part of the reversed items used words with opposite meanings, and others use direct negations of the regular items (4 negation and 16 antonymic expressions). As all the items are reversed, a high score on each item means a low score in self-efficacy (e.g. "*I feel unable to overcome obstacles*" or "*I do not make use of resources around me*"). For the construction of Form C, only the reverse-keyed items, which used words with opposite meanings rather than negation, were selected. Once this criterion was applied for the selection of reverse-keyed items, the regular items were randomly selected. This indicates that the selection of the regular items was not completely random but conditioned by the criteria of avoiding negations, as previous research has suggested that this is a better strategy (Weijters & Baumgartner, 2012). Finally, 10 regular items positively formulated and 10 reversed items were included.

Abstract Reasoning test

The *Primary Mental Abilities* abstract reasoning scale (PMA; Thurstone, 1996) was used. It is composed of 30 items of logical letter series with 6 answer options. The reliability coefficient of the scale in the current sample was .95. The first factor explains 36.23% of the variance, and the data shows a modest unidimensional scale (GFI [Goodness of Fit Index] = .93; Standardized Root Mean Square of Residuals [SRMSR] = 0.14).

Verbal Comprehension tests

For the evaluation of verbal comprehension, two classic tests (Antonyms and Sayings) were used (García-Cueto, Muñiz, & Yela, 1984; Muñiz, Sánchez, & Yela, 1986; Yela, 1987). The *antonyms* test was made up of 37 multiple choice items with four answer options in which the participant was asked to select the word which meant the opposite to the word underlined in the statement (e.g. "*The vase ended up being very fragile*"; a) *durable*, b) heavy, c) cheap, d) old-fashioned. Various qualitative and quantitative pilot studies were done, via expert judgment to ensure the content validity and through preliminary estimation of the psychometric properties of the test in different samples. Following these pilot studies, one item was eliminated, and the test was finally composed of 36 items. In the sample used in this research, all of the items had adequate indexes of discrimination and factorial weights (above .20). The mean of the indexes of the difficulty of the items was .76 and the reliability coefficient was high (α = .90). The first factor explained 22% of the variance and the indexes of fit show that the data modestly fit an essentially unidimensional structure (GFI = .90; SRMSR = 0.10). In relation to the evidence of relationships to other variables, the antonym test had a correlation of .37 (p < .001) with the PMA abstract reasoning scale and .49 (p < .001) with the sayings test used in this study.

The *sayings* test was composed of 22 multiple choice items with four answer options (e.g. "*All that glitters is not gold*": a) Gold glitters a lot, b) There are metals which glitter that are not gold, c) *Don't be fooled by appearances*, d) It's always good to have gold). Various qualitative and quantitative pilot studies were performed, both based on expert judgment to ensure the validity of content, and through preliminary estimation of the psychometric properties of the test in different samples. Five items were removed following the pilot studies due to the deficient psychometric functioning and the test ended up being composed of 17 items. In the sample used in this study, all of the items have adequate indexes of discrimination and factorial loadings (above .20). The mean of the indexes of the difficulty of the items was .82 and the reliability coefficient was acceptable (α = .77). The first factor

explains 28.14% of the variance and the indexes of fit confirm that the data fits an essentially unidimensional structure (GFI = .96; SRMSR = 0.10). Regarding validity evidence in relation to other variables, the sayings test has a correlation of .49 ($p < .001$) with the PMA abstract reasoning scale and .49 ($p < .001$) with the antonyms test used in this study.

Design

A repeated measures design was used in which all participants were evaluated at three different times by three forms of a self-efficacy test (Form A: regular items; Form B: reversed items; Form C: combined items). Each form was administered with a gap of at least one week to avoid memory effects. At the same time, the previously mentioned abstract reasoning and verbal comprehension scales were administered. To control the effect of administration order, 6 test booklets with different combinations of the test forms were randomly assigned (ABC=59; ACB=48; BAC=59; BCA=53; CAB=78; CBA=77).

Procedure

In order to standardize the test administration, a protocol was created giving instructions for the application of the test. This was given to the test administrators along with the test booklet. The test was done in paper and pencil format (75.40%) and online (24.60%) but in the latter case the tests of abstract reasoning and verbal comprehension were omitted due to the difficulty of controlling the test conditions (i.e. time, external help in answering, etc). The participants did not receive any compensation for taking part, their participation was voluntary and the confidentiality of their data was assured. The evaluation was carried out in compliance with current ethical standards, and the research was approved by the ethics committee of the University of Oviedo. All the assessment materials were administered in Spanish.

Data analyses

Firstly, all participants who missed one of the three administrations were deleted listwise. Missing values were imputed using the EM algorithm following the procedure described in Fernández-Alonso, Suárez-Álvarez, & Muñiz (2012). Reverse-worded items were reverse coded before analysis. Examination of the discrimination indices of the items in the three forms of the self-efficacy test was carried out using corrected item-test correlation. Estimation of the reliability coefficient was done using the Cronbach's Alpha coefficient and the test-retest reliability using the Spearman-Brown formula to obtain the reliability of the whole test (i.e. using the 10 common items between forms). The differences between the alpha coefficients of the three forms of the test were examined using the *w* statistic (Feldt, 1969). Various Confirmatory Factor Analyses (CFA) were carried out to confirm the fit of each of the forms to a unidimensional structure. The mean and variance adjusted maximum likelihood (MLMV) estimation method was used as the data is treated as continuous (items have five categories). The evaluation of fit of data to the model was done using multiple criteria: CFI > .90; TLI > .90; RMSEA < 0.08; SRMR < 0.08 (Kline, 2010). The Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) were also used for study the loss of information, so the lower the better (Kline, 2010). Measurement Invariance across the forms was analyzed using the one trait–three form single CFA model. The base model consists of 60 indicators and three correlated factors (Form A, Form B, and Form C), with 20 indicators loading on each factor. For the factor loading invariance test, equality constraints across the three factor loadings for all 20 items were settled simultaneously, i.e., item 1 = item 21 = item 41 to item 20 = item 40 = item 60. In addition, the item parameters were estimated along with the Test Information Function in the framework of IRT using the Graded Response Model (Samejima, 1969). The standardized residual errors of the items in the three forms of the self-efficacy test were analyzed graphically to examine the fit of the data to the model.

In addition, a multivariate analysis of variance (MANOVA) was performed to examine the effect of the order of application. Then, to study acquiescence bias, a Levene test of homogeneity of variance was performed. A test of repeated measures ANOVA was done to study the difference in the means of the participants in the three self-efficacy measures (intrasubject factor). Finally, the scores in abstract reasoning and verbal comprehension were added as covariates to control the influence that aptitude variables may have on the participants' scores (ANCOVA). The effect size was estimated using partial eta-squared (Trigo & Martínez, 2016). The data were analyzed with SPSS 20 (IBM, 2011), FACTOR 9.2 (Lorenzo-Seva & Ferrando, 2013; Ferrando & Lorenzo-Seva, 2017), TAP 12 (Brooks & Johanson, 2003), MPLUS 7.3 (Muthén & Muthén, 2012), FlexMIRT 2 (Cai, 2013) and ResidPlots-2 (Liang, Han, & Hambleton, 2009).

Table 1
Discrimination indices of regular, reversed and combined items

Items	Form A: Regular		Form B: Reversed		Form C: Combined	
	r_{ix}	a	r_{ix}	a	r_{ix}	a
1	.740	3.01	.569[1]	1.85[1]	.469[1]	1.42[1]
2	.615	1.83	.569[2]	1.58[2]	.508	1.42
3	.733	2.78	.653[1]	1.98[1]	.430[1]	1.05[1]
4	.719	2.59	.610[1]	1.66[1]	.500[1]	1.46[1]
5	.643	2.01	.676[1]	2.01[1]	.460	1.24
6	.608	1.50	.554[1]	1.39[1]	.379	0.88
7	.569	1.58	.556[1]	1.66[1]	.316[1]	0.82[1]
8	.659	2.00	.648[2]	1.99[2]	.541	1.59
9	.492	1.13	.258[1]	0.55[1]	.420	0.98
10	.662	1.89	.658[1]	1.97[1]	.574[1]	1.62[1]
11	.619	1.71	.534[1]	1.34[1]	.529	1.36
12	.455	0.99	.674[1]	2.22[1]	.607[1]	1.90[1]
13	.672	2.07	.575[2]	1.87[2]	.576	1.78
14	.647	2.09	.705[1]	2.33[1]	.544[1]	1.56[1]
15	.646	1.81	.591[1]	1.58[1]	.500[1]	1.35[1]
16	.721	2.58	.731[2]	2.54[2]	.599	1.99
17	.511	1.24	.509[1]	1.16[1]	.485[1]	1.18[1]
18	.515	1.21	.526[1]	1.27[1]	.556	1.45
19	.547	1.26	.545[1]	1.43[1]	.431[1]	1.13[1]
20	.644	1.95	.606[1]	1.82[1]	.358	0.97
Cronbach's Alpha	.932		.921		.879	

Note: [1] Reversed items – antonyms- ; [2] Reversed items – negations -; r_{ix} = item-test correlation corrected; *a* = IRT discrimination parameter
The IRT's Information Function are presented in Figure 1

Results

Reliability and Item Discrimination

Table 1 shows the discrimination indices of the regular, reversed and combined forms of the items in the self-efficacy test. As can be seen, the discrimination indices are substantially lower in the combined form compared to the regular. The difference in the discrimination indices between the regular and combined forms vary between .26 and .30. The standard error of the estimate was .10 (CI = ± .196), confidence intervals of the items did not overlap in 14 out of the 20 items when comparing regular and combined forms, so statistically significant differences were found in these 14 items (α = .05; Cumming & Finch, 2006). No pattern was observed in loss of discriminative power in terms of the strategy used to reverse the items (i.e. negation vs antonymic expression).

The reliability coefficients are also reduced, reaching statistical significance when the regular and combined forms are compared (p < .001*)* and when the reversed and combined are compared (p < .001) but not between the regular and reversed forms (p = .074). The Spearman-Brown formula was used to predict the test-retest reliability if in place of 10 common items there had been 20. The test-retest reliability coefficient for the regular items (2, 5, 6, 8, 9, 11, 13, 16, 18, and 20) was .77, for the reversed items (1, 3, 4, 7, 10, 12, 14, 15, 17, and 19) it was .80. The data were also analyzed using Samejima's Graded Response Model. Looking at the standardized residual errors of the items in each of the three forms of the self-efficacy test (i.e. regular, reversed, and combined) shows that more than 90% of the residuals are found between ±2 standard deviations, which indicates an adequate fit of the data to the model (Liang, Han, & Hambleton, 2008). Table 1 gives the parameter a for the items of the regular, reversed, and combined forms as from a classical approach. Parameter a varies between .99 and 3.01 for the regular form (a_{mean} = 1.86), between 0.55 and 2.54 for the reversed form (a_{mean} = 1.71), and between 0.82 and 1.99 for the combined form (a_{mean} = 1.36). It is clear that the items with worse discriminatory power are in the combined form. In the regular form, 13 of the 20 items may be considered highly discriminatory (a > 1.7; Baker, 2001), while in the reversed form there are 10 and in the combined, 3. As may be seen in Figure 1, the combined form would be the least accurate whereas the most accurate would be the regular form, with the reversed form being very close to it. These results explain, to a large extent, the differences found between the information functions of the tests, and converge with the results from CTT.

Dimensionality and Measurement Invariance

Table 2 shows the psychometric properties of the regular, reversed and combined forms of the self-efficacy test. Regarding the CFA, a worse fit of the data to the model was seen in the combined form compared to the regular and reversed forms. Furthermore, a nested two-factorial model with all regular items on the first factor, all reverse items on the second factor, and covariance of the two factors shows a clearly improved fit (χ^2= 233.907, *df*=169; CFI = .942; TLI = .935; RMSEA = 0.032; SRMR = 0.084; AIC = 18,859; BIC = 19,098) in comparison to the single factor model (Table 2).

In order to evaluate the measurement invariance, a one trait–three form single CFA model that constrained the factor loadings to be equal across forms A, B, and C, decreasing the model fit substantially, was performed (χ^2= 2914.259, *df* =1750; CFI = .604; TLI = .599; RMSEA = 0.042; SRMR = 0.095; AIC = 54,786; BIC = 55,335).

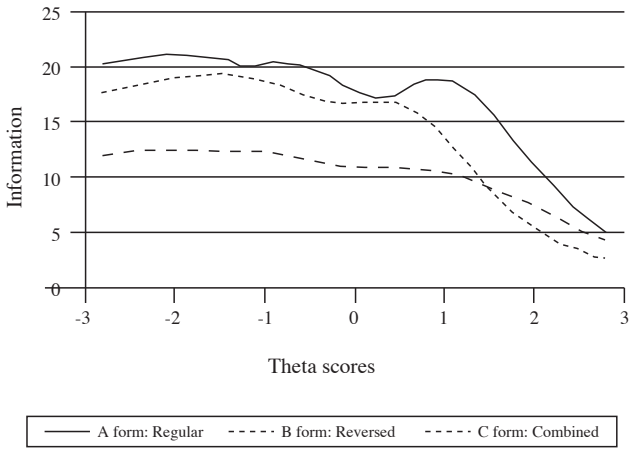

Figure 1. Tests information function of the regular, reversed, and combined forms of the self-efficacy test

Table 2
Psychometric properties of the regular, reversed, and combined forms of the self-efficacy test

	χ^2(df)	CFI	TLI	RMSEA [CI 90%]	SRMR	AIC	BIC
Form A: Regular	341.624 (170)*	.903	.892	0.052 [0.044 - 0.060]	0.049	16.125	16.360
Form B: Reversed	298.821 (170)*	.911	.901	0.045 [0.036 - 0.053]	0.048	18.249	18.484
Form C: Combined	455.306 (170)*	.746	.716	0.067 [0.060 - 0.074]	0.088	19.300	19.535

Note: χ^2=Chi-square test; df =degrees of freedom; CFI = Comparative Fit Index; TLI= Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR= Standardized Root Mean Square Residual; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria
* p < .001
The percentage of explained variance – using the CFA approach – was 48.19 for the regular form, 46.89 for the reversed form and 35.87 for the combined form

Means, Variances, and Correlations

In order to check the effect of the application order, an MANOVA was performed to study the differences between the total scores in the regular, reversed, and combined form by the six application order. The multivariate tests for the order effect using the Pillai' trace test were not statistically significant (F = 1.165; df = 15; p = .294; η^2= 0.021) using both Wilks' Lambda (F = 1.167; df = 15; p = .293; η^2= 0.021) and Hotelling's Trace (F = 1.168; df = 15; p = .291; η^2= 0.021). Examination of the differences in the three forms of the self-efficacy test in terms of the order of application (tests of between-subjects effect) demonstrated that there are no statistically significant differences in either the regular form ($F_{(5, 276)}$ = 0.914; p = .472) or the reversed form ($F_{(5, 276)}$ = 1.536; p = .179). Although statistically significant differences were found in the combined form ($F_{(5, 276)}$ = 2.447; p = .034), the size of the effect estimated via partial eta squared (η^2 = 0.042) indicates that the differences are small. Examination of the differences in the three forms of the self-efficacy test in terms of order of application (tests of between-subjects effect) demonstrated that there are no statistically significant differences either in the regular form ($F_{(5, 276)}$ = 0.914; p = .472) or the reversed form ($F_{(5, 276)}$ = 1.536; p = .179). Although in the combined form statistically significant differences were found ($F_{(5, 276)}$ = 2.447; p = .034), the effect size estimated via partial eta squared (η^2 = 0.042) indicates that the differences are small. These results indicate that the order of application has no effect on the results.

In order to check the possible effects of the acquiescence bias, the standard deviations of the three forms (regular, reversed, and combined) were calculated (Table 3). The variance of the scores in the regular form was 127.69, in the reversed form 152.80, and in the combined form 123.62. There are no statistically significant differences (Levene Test = 1.445; df_1 = 2; df_2 = 843; p = .236) between the variances of the regular, reversed, and combined forms.

Table 3 shows the results of the repeated measures ANOVA to examine the difference in means between the regular, reversed, and combined forms of the self-efficacy test. *Mauchly's* sphericity test was statistically significant (p = .004) and the *Greenhouse-Geisser* correction was used in the interpretation of the results. As can be seen in Table 3, the differences between the means were statistically significant (p < .001) and the effect size moderate (η^2 > 0.10). Furthermore, the *Bonferroni* test was statistically significant in all the pairwise comparisons (p < .001). The highest mean self-efficacy scores were seen in the reversed form, followed by the combined and the regular forms. Three repeated measures ANCOVAS were carried out including scores obtained by the participants in the verbal comprehension and abstract reasoning tests as covariates. The previously seen differences in means from the regular reversed, and combined forms disappear when controlling for the effect of verbal comprehension (sayings and antonyms). In the case of abstract reasoning, the differences are statistically significant but the effect size indicates that the differences found are practically zero.

Table 3
Descriptive statistics and significance tests: Regular, reversed, and combined forms of the self-efficacy test

Separately by order	Order of administration					
	A	B	C			
	M (SD)	M (SD)	M (SD)	F	p	η^2
A-B-C	73.10(10.25)	80.07(11.75)	77.42(10.15)	16.58	<.001	0.222
A-C-B	70.00(10.41)	75.13(12.37)	70.54 (9.77)	4.98	.016	0.096
B-A-C	71.04(12.30)	76.48(11.75)	73.69(12.87)	8.94	<.001	0.160
B-C-A	72.08(13.39)	74.46(14.18)	72.24(10.83)	.849	.432	0.023
C-A-B	71.93(10.21)	77.56(12.35)	74.95(10.63)	7.49	<.001	0.151
C-B-A	74.45(11.59)	79.09(11.82)	75.13 (11.48)	5.74	.004	0.111
	A	B	C			
Collapsed across order	M (SD)	M (SD)	M (SD)	F	P	η^2
	72.13 (11.30)*	77.33 (12.36)*	74.18 (11.12)*	35.57	<.001	0.112
Adjusted means by covariates						
Saying	71.79	77.38	74.30	0.19	.831	0.001
Antonyms	71.89	77.19	74.27	0.91	.403	0.004
PMA	72.16	77.36	74.23	3.39	.036	0.012

Notes: A = Form A (regular items); B = Form B (reversed items); C = Form C (regular and reversed items combined). *M* = Global average between-subjects. Mean, *SD* = Standard Deviation
* Note that the *F* tests reported in this table are for within-subjects differences, the global averages of A, B, and C are statistically significant in all cases. Post-hoc test: B > C > A, p < .001

Table 4
Correlations between verbal comprehension and discrepancies between a pair of scores

Absolute difference	Sayings	Antonyms	PMA	Composite score[1]
|A-B|	-.181**	-.104	-.157**	-.172*
|A-C|	-.187**	-.045	-.153**	-.094
|B-C|	.015	-.117	-.025	-.089

[1] Composite scores are created by converting each raw score of Sayings, Antonyms, and PMA into a z-score and calculating the average z-score
** p < .01; * p < .05

In order to highlight the relationship between verbal comprehension and the item wording, further analysis was performed. For each case, the absolute mean difference between a pair of scores: |A-B|, |A-C|, and |B-C| was calculated, and each of the three absolute mean difference scores was correlated with each of the covariates. As Table 4 shows, individuals with high verbal comprehension produce less discrepancy between any pair of scores. Although the results show a modest relationship, they seem to be strong enough to make the statistically significant differences between averages found in the ANCOVA disappear (Table 3).

The correlation between the scores in the regular form and the reversed form was .62 ($p < .001$), between the regular and combined it was .62 ($p < .001$), and between the reversed and the combined forms it was .66 ($p < .001$). After applying the correction for attenuation, the correlations were .67, .63, and .73, respectively. The correlation between direct empirical scores and estimates from the IRT model (θ) of each of the forms was above .97 in each of the three cases.

Discussion

The use of both regular and reversed items in tests was introduced with the aim of reducing response bias (Nunnally, 1978; Paulhus, 1991). Currently, a significant number of measurement instruments continue to use this strategy and there are researchers that recommend its use (Weijters et al., 2013). Those who advocate combining regular and reversed items in the same test argue that when all the items are in the same direction, acquiescence bias and other response bias may be present. The reason they give to justify the use of this strategy is that method effects produced by these mechanisms are completely masked, and can be undetectable unless a direct measure of method effects is used (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003); which means combining regular and reversed items to get such a measure. The logic these authors follow is that it is better to try to correct these effects, despite that meaning the combination of regular and reversed items, rather than to ignore them completely (Weijters et al., 2013). Several authors have proposed different methods to detect response style bias (Ferrando & Lorenzo-Seva, 2010; Ferrando, Lorenzo-Seva, & Chico, 2003; Savalei & Falk, 2014). One the most novel approaches is the decomposing of rating data into multiple response processes based on a multinomial processing tree (Böckenholt, 2012; Khorramdel & von Davier, 2014). The pros of the use of these methods are clear; they allow detecting response style bias and controlling the effect that they have on the scores. Despite that most of these methods present empirical evidence for its use, there are also a number of cons that can be listed.

Based on the results of the present study there are four fundamental reasons to discourage the combination of regular and reversed items in the same test. The first reason is that the cognitive process used by respondents for each type of item is not necessary the same, according to previous research (Marsh, 1986, 1996; Mestre, 1988). Although this study does not provide specific empirical evidence on this issue, the results obtained are in line with this hypothesis. From a psychological point of view, the comprehension of a reversed item needs better linguistic skills, so these items favor those examinees with better verbal ability. The problem is aggravated when examinees have to alternate between processing regular and reversed items, as is recommended to control acquiescence bias. The results of this study show that to be the case, as the method effect disappears when controlling for verbal comprehension. Therefore, combining regular and reversed items in the same test should, at the very least, be accompanied by a justification for the possible bias introduced by differences in the participants' cognitive processes.

The second reason is that, in contrast to the prevailing view, combining regular and reversed items in the same test decreases the variability in the responses. It is notable that the lowest variance of test scores was found in the combined form, which included both regular and reversed items. These results are more understandable bearing in mind that various researchers suggest that acquiescence is stable over time (Alessandri et al., 2010; Weijters, Geuens, & Schillewaert, 2010). Consequently, it is not reasonable to think that examinees change their response style at the moment of a specific evaluation. Furthermore, the presentation of tests to the participants was random. Therefore, the results presented here do not confirm that the strategy of using regular and reversed items in the same test reduces response bias, which is in line with previous findings by van Sonderen et al. (2013).

The third reason is that the test's psychometric properties are substantially worse when regular and reversed items are combined in the same test. The results show that the precision of the test and the discriminatory power of the items diminish when regular and reversed items are included in the same test. These results are in line with those found by other authors (Bourque & Shen, 2005; Carlson et al., 2011; Chiavaroli, 2017; Ebesutani et al., 2012; Hughes, 2009; Józsa & Morgan, 2017; Solís-Salazar, 2015). It is also worth noting that the test-retest reliability is around .80, however, the correlations between the regular, reversed, and combined forms vary between .62 and .66. This data indicates that despite the fact that the forms evaluate the same construct, they cannot be considered parallel forms of the same test (Evers et al., 2013). The evaluation of factorial invariance provides evidence that the items are not measuring with the same precision in each group (Dimitrov, 2010). The same happens with dimensionality, as the fit of the data to the model is also affected, making it more difficult to support the idea of unidimensionality. These results converge with other results in the scientific literature, and support the idea that including this type of items especially affects unidimensionality, making the fit worse, and increasing the rejection of unifactorial in favor of multidimensional models (Dunbar et al., 2000; Essau et al., 2012; Horan et al., 2003; Woods, 2006). These results seem to be in line with the idea that, when regular and reversed items are combined in the same test, it benefits those examinees with better verbal abilities. As a consequence, the construct being measured may be contaminated by other variables which have little relation to the objective of the evaluation. Some researchers have suggested that personality may also be involved in the manner of answering regular and reversed items in the same test (DiStefano & Motl, 2009; Horan et al., 2003). Future research will shed more light on what is actually being evaluated when items are combined and on how the construct supposedly being measured is masked.

The fourth reason is that there are statistically significant differences in the average scores in terms of whether the items of the test are regular, reversed, or combined. Specifically, the highest scores are seen in the reversed form, followed by the combined, and the regular. These results may be related to confirmation bias, the tendency to activate beliefs which are consistent with the

sense in which the item is written (Davies, 2003). Previous studies suggested that if a reversed item is presented first, scores in reversed items are higher (Weijters et al., 2013). However, the results of the present research show that the highest scores are obtained when all of the items are reversed. Regarding these results, it is worth remembering that examinees tend to disagree with reversed items more than they agree with regular items (Solís-Salazar, 2015). For example, with the item "*I am able to organize my own work*" 38% of participants responded, "*Totally agree*". When the item was presented to the same examinees as "*I am incapable of organizing my own work*", 48.2% responded, "*Totally disagree*". Note that both of these items would have the maximum score in self-efficacy once the reversed item is redirected.

In short, according to the results obtained, the strategy of using regular and reversed items combined in a single test has significant negative consequences: a) the measurement precision of the instrument is flawed; b) the interpretation of instrument unidimensionality is jeopardized by secondary sources of variance; c) the variance of the combined form is reduced; d) examinees's scores differ significantly from those obtained in tests where all of the items are of a similar form; e) verbal skills influence examinees' responses. These conclusions are worthy of consideration for several reasons. Firstly, a repeated measures design was used, which has not been used before with these aims. This allows a much more thorough, rigorous investigation, the reduction of sources of error, and the attribution of differences in characteristics of the measurement instrument, avoiding confounding effects. Secondly, comparing the results when all items are regular, all are reversed, and when both types are combined emphasizes that the problem is not with regular items, but rather with the combination of regular and reversed items in the same test. Thirdly, the evaluation of verbal skills leads to a better understanding of the consequences of combining items on the participants' psychological processes when responding, which confirm previous research (van Sonderen, 2013; Weijters et al., 2013). In conclusion, the results of this study present a trade-off between a potential acquiescence bias when items are positively worded and a potential different understanding when combining regular and reversed items in the same test. The specialized literature recommends combining regular and reversed items for controlling for response style bias, but these results caution researchers in using them as well after accounting for the potential effect of linguistic skills and the findings presented in this study.

Certain limitations must be borne in mind when interpreting the results. Most importantly, it would be advisable to improve both the representativeness of the sample and to use other samples to check the validity of the results at a transcultural level (Byrne & van de Vijver, 2017; Essau et al., 2012; Muñiz, Elosua, Padilla, & Hambleton, 2016). This would improve the robustness of the results related to the goodness of fit evaluation, precision of parameter estimates, and measurement invariance evaluation. The results of this research focus mainly on acquiescence bias, in the future it would be useful to look more deeply at other response bias such as careless responding (Kam & Meyer, 2015). In such cases, it would be advisable to use scales of infrequency (where the response to the item is previously known) which would allow the detection of people who respond randomly or dishonestly (Muñiz et al., 2014). For an estimation of the effect of acquiescence on responses, post-hoc controls are recommended via explicit measures of acquiescence (Baumgartner & Steenkamp, 2001; Weijters et al., 2013). The importance of acquiescence bias when using a computerized adaptive test administration has to be investigated (Pedrosa, Suárez-Álvarez, García-Cueto, & Muñiz, 2016). Finally, one of the more promising alternatives for controlling response styles is the use of anchoring vignettes (Bolt, Lu, & Kim, 2014) despite that they have also presented limitations when assumptions are violated (von Davier, Sim, Khorramdel, & Stankov, 2017).

Acknowledgements

The views expressed in the paper represent the views of the individual authors and do not represent an official position of the Organisation for Economic Co-operation and Development. This research was funded by the Spanish Association of Methodology of Behavioral Sciences and Health (AEMCCO), member of the European Association of Methodology (EAM), and by the FPI programme from the Ministry of Economy and Competitiveness of the Government of Spain (PSI2014-56114-P, BES2012-053488, and PSI2017-85724-P).

References

Abad, F.J., Olea, J., & Ponsoda, V. (2011). *Medición en ciencias sociales y de la salud* [Measurement in social sciences and health]. Madrid: Síntesis.

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P.M., Barbaranelli, C., Medda, E., …, & Caprara, G.V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the Life Orientation Test Revised. *Structural Equation Modeling, 17*, 642-653. doi:10.1080/10705511.2010.510064

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

Baumgartner, H., & Steenkamp, J.B.E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156. doi:10.1509/jmkr.38.2.143.18840

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*, 665-678.

Bolt, D.M., Lu, Y., & Kim, J.S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528-541. doi: 10.1037/met0000016

Bourque, L.B., & Shen, H. (2005). Psychometric characteristics of Spanish and English versions of the Civilian Mississippi scale. *Journal of Traumatic Stress*, *18*(6), 719-728. doi:10.1002/jts.20080

Brooks, G.P., & Johanson, G.A. (2003). Test analysis program. *Applied Psychological Measurement, 27,* 305-306.

Brown, A. (2015). Item response models for forced-choice questionnaires: A Common framework. *Psychometrika, 81*(1), 135-160. doi: 10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2012). How IRT can solve problems of ipsative data in forced-choice questionnaire. *Psychological Methods, 18*(1), 36-52.

Byrne, B., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29*, 539-551.

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Carlson, M., Wilcox, R., Chou, C-P., Chang, M., Yang, F., Blanchard, J., …, & 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. doi:10.1037/a0022484.

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research and Evaluation*, *22*(3), 1-14.

Chiorri, C., Anselmi, P., & Robusto, E. (2009). Reverse items are not opposites of straightforward items. In U. Savardi (Ed.), *The Perception and Cognition of Contraries* (pp. 295-328). Milano: McGraw-Hill.

Cronbach, L.J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475-494. doi:10.1177/001316444600600405

Cronbach, L.J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3-31. doi:10.1177/001316445001000101

Cumming, G., & Finch, S. (2006). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170-180. doi:10.1037/0003-066X.60.2.170

Davies, M.F. (2003). Confirmatory bias in the evaluation of personality descriptions: Positive test strategies and output interference. *Journal of Personality and Social Psychology, 85*, 736-744. doi:10.1037/0022-3514.85.4.736

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

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149. doi: 10.1177/0748175610373459

DiStefano, C., & Motl, R.W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg self-esteem scale. *Personality and Individual Differences, 46*, 309-313. doi:10.1016/j.paid.2008.10.020

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. doi: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. doi:10.1080/00223891.2012.662188.

Elosua, P., & Zumbo, B.D. (2008). Reliability coefficients for ordinal response scales. *Psicothema, 20*(4), 896-901.

Essau, C. A., Guzmán, B.O., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., ..., & Ollendick, T.H. (2012). Psychometric properties of the Strength and Difficulties Questionnaire from five European countries. *International Journal of Methods in Psychiatric Research*, *21*(3), 232-245. doi:10.1002/mpr.1364

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283-291. doi:10.7334/psicothema2013.97

Feldt, L.S. (1969). A test of the Hypothesis that Cronbach's alpha or Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Ferrando, P.J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology, 63*, 427-448.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*, 236-240.

Ferrando, P.J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research, 38*, 353-374.

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2012). Imputation methods for missing data in educational diagnostic evaluation. *Psicothema, 24*(1), 167-175.

Fonseca-Pedrero, E., & Debbané, M. (2017). Schizotypal traits and psychotic-like experiences during adolescence: An update. *Psicothema, 29*, 5-17.

García-Cueto, E., Muñiz, J., & Yela, M. (1984). Estructura factorial de la comprensión verbal [Factorial structure of verbal comprehension]. *Investigaciones Psicológicas, 2*(2), 59-75.

Haladyna, T.M., Downing, S.M., & Rodríguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Haladyna, T.M., & Rodríguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Horan, P. M., DiStefano, C., & Motl, R.W. (2003). Wording effects in self esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*, 444-455.

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

IBM (2011). *IBM SPSS Statistics for Windows, Version 20* [Computer software]. Armonk, NY: IBM Corp.

Józsa, K., & Morgan, G.A. (2017). Reversed items in Likert scales: Filtering out invalid responders. *Journal of Psychological and Educational Research*, *25*(1), 7-25.

Kam, C.C.S., & Meyer, J.P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*(3), 512-541. doi:10.1177/1094428115571894

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(29), 161-177. doi:10.1080/00273171.2013.866536.

Kline, R.B. (2010). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Lane, S., Raymond, M. R., & Haladyna, T. M. (2016). *Handbook of test development (2nd edition)*. New York, NY: Routledge.

Liang, T., Han, K.T., & Hambleton, R.K. (2008). *User's guide for ResidPlots-2: Computer software for IRT graphical residual analyses, Version 2.0* (Center for Educational Assessment Research Report No. 688). Amherst: Center for Educational Assessment, University of Massachusetts.

Liang,T., Han, K.T., & Hambleton, R.K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement, 33*(5), 411-412.

Lorenzo-Seva, U., & Ferrando, P.J. (2013). *Manual of the program FACTOR v. 9.20*. Retrieved from: http:// psico.fcep.urv.es/utilitats/factor/documentation/Manual-of-the-Factor-Program-v92.pdf

Marsh, H.W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37-49. doi: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*, 810-819. doi:10.1037/0022-3514.70.4.810

Mestre, J.P. (1988). The role of language comprehension in mathematics and problem solving. In R.R. Cocking & J.P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200-220). Hillsdale, NJ: Lawrence Erlbaum Associates.

Moreno, R., Martínez, R., & Muñiz, J. (2004). Guidelines for the construction of multiple choice test items. *Psicothema, 16*(3), 490-497.

Moreno, R., Martínez, R., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*(2), 65-72.

Moreno, R., Martínez, R., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394. doi:10.7334/psicothema2015.110

Muñiz, J., Elosua, P., Padilla, J. L., & Hambleton, R. K. (2016). Test adaptation standards for cross-lingual assessment. In C. S. Wells & M. Faulkner-Bond (Eds.), Educational measurement. From foundations to future (pp. 291-304). New York: The Guilford Press.

Muñiz, J., Sánchez, P., & Yela, M. (1986). Comprensión verbal en monolingües y bilingües [Verbal comprehension on monolingual and bilingual]. *Informes de Psicología, 5*, 139-153.

Muñiz, J., Suárez-Álvarez, J., Pedrosa, I., Fonseca-Pedrero, E., & García-Cueto, E. (2014). Enterprising personality profile in youth: Components and assessment. *Psicothema, 26*(4), 545-553. doi:10.7334/psicothema2014.182

Muthén, L.K., & Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*, 465-470.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York, NY: McGraw-Hill.

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

Pedrosa, I., Suárez-Álvarez, J., García-Cueto, E., & Muñiz, J. (2016). A computerized adaptive test for enterprising personality assessment in youth. *Psicothema, 28*, 471-478.

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

Prieto, G., & Delgado, A.R. (1996). Construcción de los ítems [Item development]. In J. Muñiz (Ed.), *Psicometría* (pp. 105-135). Madrid: Universitas.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Savalei, V., & Falk, C.F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49*(5), 407-424. doi:10.1080/00273171.2014.931800

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

Suárez-Álvarez, J., Pedrosa, I., García-Cueto, E., & Muñiz, J. (2014). Screening enterprising personality in youth: An empirical model. *Spanish Journal of Psychology, 17*(E60). doi: 10.1017/sjp.2014.61

Swain, S. D., Weathers, D., & Niedrich, R.W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45,* 116-131.

Thurstone, L. (1996). *Test de Aptitudes Primarias* [Primary Mental Abilities]. Madrid: TEA Ediciones (Orig. 1938).

Trigo, M. E., & Martínez, R. J. (2016). Generalized ETA square for multiple comparisons on between-groups designs. *Psicothema, 28*, 340-345.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

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), e68967. doi:10.1371/journal.pone.0068967

von Davier, M., Shin, H-J., Khorramdel, L., & Stankov, L. (2017). The effects of vignette scoring on reliability and validity of Self-Reports. *Applied Psychological Measurement*. Advance online publication. doi: 10.1177/0146621617730389

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated Items in surveys: A Review. *Journal of Marketing Research, 49,* 737-747.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reverse item bias: An integrative model. *Psychological Methods, 18,* 320-334.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, *26*(1), 2-12. doi:10.1016/j.ijresmar.2008.09.003

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*, 96-110.

Wilson, M. (2005). *Constructing measures: An item response modelling approach*. Mahwah, NJ: Erlbaum.

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

Yela, M. (1987). *Estudios sobre inteligencia y lenguaje* [Studies on intelligence and language]. Madrid: Pirámide.