Psicothema

Psicothema 2020, Vol. 32, No. 1, 108-114
doi: 10.7334/psicothema2019.286

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

To reverse or to not reverse Likert-type items: That is the question

Andreu Vigil-Colet, David Navarro-González, and Fabia Morales-Vives
Universitat Rovira i Virgili (CRAMC)

Abstract

Background: The suitability of using reversed items in typical response measures has been a matter of controversy for many years. While some authors recommend their use, others reject them due to their undesirable effects on tests' psychometric properties. The present research intends to analyse a third alternative based on the use of reversed items plus a procedure to control response bias effects. **Method:** We analysed two forms of the same test, one with direct and reversed items and another composed only of direct items, and compared them both before and after applying a procedure to control response biases. **Results:** The factorial structure and factorial reliability of both versions was almost equivalent after controlling response biases. When no effect biases were controlled, the version with both types of items exhibited less acceptable psychometric properties. **Conclusions:** The use of reversed items is not advisable without the application of a procedure to control response bias effects. When such effects are mitigated, the results are equivalent to those obtained with only direct items, but with the added value of controlling for acquiescence effects.

Keywords: Response bias, personality, factor structure.

Resumen

Invertir o no invertir ítems tipo Likert: esa es la cuestión. **Antecedentes:** la utilización de ítems invertidos en medidas de respuesta típica ha sido durante mucho tiempo una cuestión controvertida. Mientras algunos autores aconsejan su utilización, otros la rechazan debido a sus efectos indeseables en las propiedades psicométricas de las medidas. El presente estudio pretende analizar una tercera vía, basada en el uso de ítems invertidos juntamente con un método para eliminar los efectos de los sesgos de respuesta. **Método:** se analizaron dos versiones de una misma prueba, una incorporando ítems directos e invertidos y otra compuesta únicamente de ítems directos. Posteriormente se compararon ambas versiones antes y después de controlar los efectos de los sesgos de respuesta. **Resultados:** la estructura factorial y la fiabilidad de las puntuaciones factoriales de ambas versiones tras eliminar los efectos de los sesgos de respuesta fue equivalente, mientras que la versión con ambos tipos de ítems sin control de sesgos mostró peores propiedades psicométricas. **Conclusiones:** la utilización de ítems revertidos sin la aplicación de un método de control de sesgos está claramente desaconsejada. Cuando dichos métodos se utilizan los resultados de ambas versiones son equivalentes con el añadido que en la versión con ítems revertidos se controlan los efectos de aquiescencia.

Palabras clave: sesgos de respuesta, personalidad, estructura factorial.

The use of reversed items in typical performance measures has been a controversial issue in recent decades. Reversed items may be defined as those which must be recoded so that all the items of a scale have the same directional relationship with the construct of interest (Weijters, Baumgartner, & Schillewaert, 2013). The use of these types of items was explicitly addressed many years ago by authors such as Nunnally (1978) or Ray (1983) as a means of controlling for the effects of acquiescence (AC). Acquiescence is one of the most common response biases, and may be defined as an individual's tendency to agree with a statement regardless of its content (Paulhus & Vazire, 2005). One of the consequences of this tendency is that it may be difficult to disentangle whether an individual with a high score on a questionnaire has a high trait level or whether the high score reflects a high tendency to agree with questionnaire items (Ferrando & Lorenzo-Seva, 2010; Weijters et al., 2013).

Received: September 19, 2019 • Accepted: November 27, 2019
Corresponding author: Andreu Vigil-Colet
Facultad de Ciencias de la Educación y Psicología
Universitat Rovira i Virgili (CRAMC)
43007 Tarragona (Spain)
e-mail: andreu.vigil@urv.cat

In order to control AC effects, the most commonly used method is the use of balanced scales, that is, scales with all of the items positively worded (with no negations to change the meaning of the item) but with half of the items measuring the dimension of interest in one direction and the other half measuring in the opposite direction. Some authors advocate the use of reversed items and cite many positive effects resulting from their inclusion, but others have argued that reversed items also have negative consequences. Therefore, both the pros and the cons of reversed items must be considered before deciding whether to include them or not.

In addition to controlling for AC effects in balanced scales, the use of reversed items is expected to offer other advantages. For instance, it may increase the validity of the scales by providing a more complete representation of the underlying construct to be measured and by increasing the accuracy in the prediction of other constructs (Józsa & Morgan, 2017; Weijters & Baumgartner, 2012). Furthermore, reversed items may act as "speed bumps" and lead to slower and more careful reading of the items (Józsa & Morgan, 2017; Kam & Meyer, 2015; Weijters et al., 2013).

Nevertheless, introducing reversed items in typical response measures can also give rise to several negative effects. Firstly, the introduction of reversed items usually implies a reduction in the

scales' internal consistency due to a lower item-total correlation (Ebesutani et al., 2012; Paulhus & Vazire, 2005; Salazar, 2015; Suárez-Alvarez et al., 2018). Secondly, the use of reversed items may affect the factorial structure of measures and usually leads to a poor fit to the expected model (Danner, Aichholzer, & Rammstedt, 2015; Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016; Soto, John, Gosling, & Potter, 2008), and in many cases their inclusion may result in a two-dimensional structure containing positive and negative items in different factors when measuring unidimensional constructs (Brown, 2003; Dunbar, Ford, Hunt, & Der, 2000; Kam & Meyer, 2015; Paulhus & Vazire, 2005; Weijters & Baumgartner, 2012; Woods, 2006). Lastly, reversed items usually yield lower mean item responses than direct items (Salazar, 2015; Suárez-Alvarez et al., 2018; Weems, Onwuegbuzie, & Colins, 2006).

Most of the above negative effects might be considered to be due to the impact of AC in the inter-item correlation matrix. On an scale comprising direct and reversed items, AC distorts the correlation between items, overestimating the correlations when items are polarised in the same direction and underestimating them when they are polarised in opposite directions, resulting in a worse model fit and a decrease in scale reliability (Rammstedt & Kemper, 2011; Rammstedt & Farmer, 2013; Ray, 1983). On the other hand, in a scale consisting only of direct polarised items, AC tends to overestimate the correlation between items, artificially improving model fit and reliability (Danner et al., 2015; Salazar, 2015).

Considering the discussion above, there are two main positions regarding the use of reversed items. Taking into account the positive and negative consequences of their use, some authors continue to defend balanced scales in the belief that the benefit is worth the cost (Ferrando & Lorenzo-Seva, 2010; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Weijters et al., 2013). Other authors argue for avoiding the use of reversed items and using scales consisting only of direct items (DeVellis, 2003; Suárez-Alvarez et al., 2018). Nevertheless, there is also a third way to approach this problem: to use reversed items and apply a procedure that controls their undesirable effects while retaining their positive effects, such as controlling for AC (Dunbar et al., 2000; Kam & Meyer, 2015; Woods, 2006). Several studies have shown that modelling AC as a method factor and removing its effects seem to mitigate at least the poor fit to the expected model in personality measures which combine direct and reversed items (Morales-Vives, Lorenzo-Seva, & Vigil-Colet, 2017; Navarro-González et al., 2016; Rammstedt & Farmer, 2013; Soto et al., 2008). Nevertheless, these studies have not analysed whether the obtained fit would be equivalent to that obtained from a test using only direct items or whether these methods can prevent the decrease in reliability due to the use of reversed items.

Following upon the discussion above, the main objective of this paper is to compare two versions of a typical response measure: one consisting of a mixture of direct and reversed items, and another version of the same measure consisting only of direct items. This comparison allows us to study how the use of reversed items affects different properties of the measure, such as factorial structure, reliability, and the mean and variability of the scores. It also enables us to assess how the use of a procedure to control AC effects attenuates the undesirable effects of reversed items or even removes them completely, reaching factorial structures and reliability coefficients equivalent to those in measures consisting only of direct items.

To control AC effects, we used the procedure developed by (Ferrando, Lorenzo-Seva, & Chico, 2009), who developed a general procedure for controlling not only for AC but also for social desirability effects (SD), defined as the tendency to give answers that make the person assessed look good (Paulhus, 1991). The first step in the procedure identifies a factor related to SD using items taken as markers of SD. These are used to compute the SD loadings of the content factors and to compute a residual inter-item correlation matrix free of SD. In a second step, the residual correlation matrix is analysed by applying the procedure developed by Lorenzo-Seva & Ferrando (2009), which subtracts the items of variance due to acquiescent responses from the content. More precisely, the procedure assumes that it should be possible to identify acquiescence as a common style factor present in a set of content items that are at least partially balanced (i.e., where only a few items in the scale are worded in the opposite direction). A balanced subset of items must be identified (i.e., a subset of items in which half of the items are worded in one direction and the other half in the other), and this balanced core of items is used to obtain a centroid solution, which estimates the AC loadings of this subset. Finally, the remaining items (the ones left from the balanced subset) are projected on to the centroid to obtain their corresponding AC loadings. This procedure differs from other EFA and IRT acquiescence control procedures in that: (a) items are allowed to have their own unique acquiescence loadings on the acquiescence factor, which captures the idea that each item may elicit acquiescence to a different extent, (b) the set of items does not need to be fully balanced, (c) no acquiescence items or scale need to be used, and (d) it can be used alongside the procedure for controlling SD described above.

The most important difference was the one discussed in point (a), because other procedures assume that all the items have the same loading on acquiescence while the EFA procedure proposed by Ferrando et al. (2009) can estimate the specific acquiescence loading of each item (Savalei & Falk, 2014).

In the final step, the residual inter-item correlation matrix free of the distortions caused by SD and AC can be used in a classical exploratory factor analysis (EFA) to determine the factor structure of the questionnaire. Furthermore, it is possible to compare the factorial structures when both biases are corrected for, when only one is corrected for, and when no bias correction has been implemented. We took advantage of the fact that the procedure used can also analyse the effects of SD because, although it seems that the AC response bias has the greatest impact on factorial structures, it is possible that SD may affect them as well (Morales-Vives et al., 2017; Navarro-Gonzalez et al., 2016).

Method

Participants

A total of 619 volunteer university students (30.7% male) participated from four different degree programmes (psychology, education, pedagogy and social work) at Universitat Rovira i Virgili in Tarragona, Spain, with ages ranging from 18 to 57 years old (M=21.27 SD=4.5).

Instruments

The indirect-direct aggression questionnaire (I-DAQ) (Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, & Vigil-Colet, 2014). This test yields scores for the factors physical aggression (PA),

verbal aggression (VA) and indirect aggression (IA), as well as SD and AC scores for each individual. The factors measured by the I-DAQ have appropriate factor-score reliabilities: $r_{\theta\theta}$=.83, $r_{\theta\theta}$=.77 and $r_{\theta\theta}$=.78 for PA, VA and IA respectively.

Two versions of the questionnaire were used: the standard version, consisting of 12 direct items and 11 reversed items plus 4 items used as SD markers, and the direct version. In the direct version, all the reversed items were transformed into direct items, for example, "Aunque esté enfadado, mi manera de hablar es poco agresiva." (Even though I am angry, my way of speaking is not very aggressive) was modified to "Cuando estoy enfadado, mi manera de hablar es muy agresiva" (When I am angry, my way of speaking is very aggressive).

Procedure

The tests were administered collectively in the students' classrooms. The participants were asked to volunteer to respond to the inventories. The questionnaires were anonymous, and respondents provided only their gender and age. The test version (standard or direct) administered was assigned randomly to each participant.

Data analysis

We computed different EFAs for each version of the questionnaire using the procedure developed by Ferrando et al. (2009). For the direct version, we computed one EFA without any bias correction, and another that controlled the SD effects. For the standard version, we computed EFAs without any correction and with corrections for SD, AC and both biases simultaneously. These EFAs were performed on the polychoric inter-item correlation matrix. To assess the fit of each loading matrix to the expected factorial solutions, the congruence index developed by L. Tucker was computed between the rotated loading matrix and the ideal loading matrix. Indexes higher than *C*=.85 constitute a fair congruence between the rotated loading matrix and the ideal loading matrix, while indexes of .95 or higher imply that the rotated loading matrix and the ideal loading matrix are essentially equal (Lorenzo-Seva & ten Berge, 2006). For each EFA, we computed the reliability of the derived factor score estimates. We also computed the raw scores for each scale and for the subscales containing direct or reversed items for both versions of the test without removing biases. Data was analysed using the Psychological Test Toolbox (Navarro-González, Vigil-Colet, Ferrando, & Lorenzo-Seva, 2019) and SPSS 25. The Psychological Test Toolbox is freeware software developed in MATLAB. It was designed to perform EFA by applying the procedure described in (Ferrando et al., 2009) for assessing response biases impact. It is available as a stand-alone program at the following link: https://psico.fcep.urv.cat/utilitats/PsychologicalTestToolbox

Results

Table 1 contains the descriptive statistics for the raw scores of the three scales of the I-DAQ, and shows that the standard version (containing direct and reversed items) always has higher item means than the direct version, while Levene's tests did not reveal any differences in variances. The effect size for these differences was low for PA and IA and medium for IA. In order to determine whether these differences were due to the items' directionality, we computed item means for subscales containing only direct or reversed items. As the data show, the subscales consisting of items which are direct in both versions showed no differences in either version, indicating that reversed items do not affect the response pattern for positive items when combined in the same scale. However, the mean of the items belonging to subscales consisting of only reversed items was always lower than their direct counterparts, even when the effects were large as in the case of VA, indicating a possible AC effect in I-DAQ items. It should be taken into account that for the reversed versions, the items were recoded to measure the trait in the direction of direct items, therefore, the effects of AC are reflected in a lower mean item score.

The presence of AC in the I-DAQ was also reflected in the relationships between the subscales consisting of direct or reversed items. Table 2 contains Pearson correlations between these subscales, and shows that the correlations between direct and

Table 1
Descriptive statistics, Student's test and Cohen's d for the raw scores of the full scales in the direct and standard version and for the subscales consisting of direct and reversed items

Scale	Version	Mean	SD	t	d
Physical	standard	1.70	0.76	*2.34*	0.19
	direct	1.81	0.60		
Indirect	standard	3.39	1.15	*2.33*	0.18
	direct	3.61	1.19		
Verbal	standard	2.33	0.72	**5.59**	0.44
	direct	2.66	0.74		
Physical	standard	1.72	0.76	-0.30	
(Direct items)	direct	1.74	0.71		
Physical	standard	1.70	0.95	**8.1**	0.64
(Reversed items)	direct	2.31	0.87		
Indirect	standard	1.81	0.61	0.99	
(Direct items)	direct	1.76	0.61		
Indirect	standard	1.58	0.89	**4.1**	0.33
(Reversed items)	direct	1.86	0.75		
Verbal	standard	2.75	0.82	0.57	
(Direct items)	direct	2.71	0.83		
Verbal	standard	2.33	0.74	**10.4**	0.83
(Reversed items)	direct	2.67	0.74		

p<.01; *p<.05*

Table 2
Correlation matrix between subscales consisting of direct (d) and reversed (r) items

	Physical d	Physical r	Indirect d	Indirect r	Verbal d	Verbal r
Physical d	–					
Physical r	**.357**	–				
Indirect d	**.390**	**.267**	–			
Indirect r	**.154**	**.285**	**.155**	–		
Verbal d	**.300**	**.280**	**.363**	.067	–	
Verbal r	**.242**	**.388**	**.212**	**.287**	**.540**	–

p<.01; *p<.05*

reversed subscales measuring the same trait are low to moderate. Furthermore, and more relevantly, for PA and IA, correlations with the other two scales with items in the same direction are even higher than the correlation with the same scale in the opposite direction. For instance, IA direct has a r=.155 with IA reversed, but a r=.390 with PA direct. As a consequence, in certain cases, the direction effect is even higher than the effect of the trait measured, and the inter-item correlation matrix will be distorted by this effect.

Table 3 contains the loading matrix for the direct version with and without controlling for SD effects, and shows that, with the exception of items 22 and 23, all items have their salient loading in the expected dimension. Furthermore, controlling for SD does not seem to improve the factorial structure of the questionnaire. A quite different result was obtained for the standard version of the I-DAQ, as shown in table 4. When biases were not controlled for, the VA scale was mainly comprised of the expected items, but the first factor, labelled PA, is a mixture of direct PA and IA items, while the third factor, labelled IA, is a mixture of reversed items from the same scales.

Removing SD effects improved the structure of the PA factor, but the IA factor was comprised only of reversed IA items. When only AC was controlled, there was a clear improvement in the structure of all the items. With the exception of items 27 and 23, all items had their salient loadings on the expected factor and, when both biases were controlled, only item 27 loaded on a different factor.

Table 3
Loading matrix with and without bias for the direct version of the I-DAQ. Salient loadings on content factors in bold

		With bias			Controlling social desirability			
	Item	**PHY**	**VER**	**IND**	**SD**	**PHY**	**VER**	**IND**
Social Des.	2sd				-.61	.00	.00	.00
	8sd				-.32	.00	.00	.00
	13sd				-.65	.00	.00	.00
	21sd				-.72	.00	.00	.00
Physical Aggression	1f-	**.68**	.08	-.14	.07	**.74**	.08	-.08
	6f+	**.51**	.12	-.02	-.24	**.63**	.07	-.15
	17f-	**.68**	-.02	.17	-.18	**.66**	.00	.19
	19f-	**.55**	-.02	.05	.01	**.56**	-.01	.12
	20f+	**.68**	-.05	.13	-.16	**.71**	-.06	.10
	25f+	**.44**	.40	-.14	-.06	**.44**	.42	-.12
Verbal Aggression	5v-	.06	**.66**	-.06	.05	.00	**.74**	.04
	7v+	.05	**.53**	-.04	-.06	.05	**.54**	-.02
	9v+	-.10	**.72**	.18	-.22	-.05	**.70**	.12
	12v-	.13	**.60**	.09	-.14	.12	**.64**	.09
	15v+	-.05	**.95**	-.16	-.11	-.04	**.92**	-.15
	22v-	.01	.15	**.28**	-.31	.02	.16	**.19**
	27v+	.04	**.50**	.05	-.25	.07	**.52**	-.07
Indirect Aggression	3i+	.12	.05	**.50**	-.35	.21	.02	**.36**
	4i+	.19	.06	**.46**	-.36	**.33**	.00	.31
	10i-	-.02	.06	**.67**	-.37	-.03	.07	**.61**
	11i+	-.08	-.01	**.66**	-.32	-.01	-.04	**.57**
	14i-	-.02	-.01	**.73**	-.34	-.02	.01	**.68**
	16i-	-.11	.06	**.82**	-.32	-.10	.09	**.77**
	18i+	.11	-.10	**.70**	-.25	.15	-.11	**.68**
	23i+	-.03	**.39**	.29	-.20	-.05	**.43**	.27
	24i-	.02	.06	**.62**	-.39	.06	.06	**.50**
	26i-	.09	-.05	**.58**	-.02	.00	.00	**.75**

It is worth mentioning that an inspection of the AC loadings of direct and reversed items showed that the mean of the loadings for reversed items (λ=.34) was much greater than the loadings of direct items (λ=.18, $t_{(21)}$= 3,7 p<.01, d=1.5), suggesting that reversed items generated almost twice the AC as direct items. There were also differences in the acquiescence loadings across scales: VA had a mean loading of λ=.18, while PA and IA showed greater loadings in AC with values of λ=.24 and λ=.32 respectively.

Table 4 shows a comparison of the factorial congruence and reliability of all the analyses performed on the direct and standard versions. The direct version showed fair (VA) or good (PA and IA) congruences with the expected structure with negligible differences when SD was controlled for. Factorial-based reliabilities were also quite good in both cases. A different scenario was found for the standard version. When biases were not controlled, none of the scales reached acceptable congruence, and PA factor scores had unacceptable reliability (in fact, this factor and the IA factor were a mixture of PA and IA items). When SD was controlled, there were slight improvements in congruences, two of which were greater than C=.85, but the greatest improvement was found when AC effects were controlled. In this case, all congruences were fair (VA) or good (PA and IA) and reliabilities were also good. No noticeable increase was found for either congruences or for reliabilities when SD was controlled in addition to AC. As the table indicates, when AC or AC and SD effects were controlled, the congruences and reliabilities were equivalent to the ones reported for the direct version of the I-DAQ.

Discussion

The results of the present study illustrate the typical problems associated with combining direct and reversed items in a questionnaire: a decrease in reliability, a poor fit to the expected factorial structure, lower scores for reversed items, and higher levels of AC in reversed items. On the other hand, the same scale consisting only of direct items exhibited the expected advantages of this format: a better fit to the expected factorial structure and higher reliability.

Therefore, it seems that the least advisable procedure is to use tests comprising direct and reversed items without using any procedure to control AC effects. This issue is especially relevant because most of the typical response measures administered in today's context combine both types of items and do not make use of a procedure to control AC effects.

Based on our results, it seems clear that if AC effects are not controlled, using only direct items is more desirable than using scales combining both types of items. However, this approach must be viewed with caution for several reasons. The first is that AC effects are not controlled for in direct-item-only measures, and this affects the scores of different individuals, making them a mixture of their trait level and their tendency to agree with the items and thereby preventing their unequivocal interpretation. A similar effect may be found for the case of careless responding and confirmation bias, the effects of which can be confounded with content variance (Józsa & Morgan, 2017; Weijters & Baumgartner, 2012; Weijters et al., 2013). Furthermore, AC is expected to affect the inter-item correlation matrix, whose values may be inflated,

Table 4
Loading matrix with and without bias for the I-DAQ. Salient loadings on content factors in bold

		With bias			Controlling for acquiescence				Controlling for social desirability				Controlling for social desirability and acquiescence				
	ITEM	PHY	VER	IND	ACQ	PHY	VER	IND	SD	PHY	VER	IND	SD	ACQ	PHY	VER	IND
Social Des.	2sd								-.81	.00	.00	.00	-.81	.00	.00	.00	.00
	8sd								-.36	.00	.00	.00	-.36	.00	.00	.00	.00
	13sd								-.65	.00	.00	.00	-.65	.00	.00	.00	.00
	21sd								-.71	.00	.00	.00	-.71	.00	.00	.00	.00
Physical Aggression	1f-	-.22	**-.30**	.15	.17	**-.75**	.10	.17	.17	**-.34**	-.29	.06	.17	.24	**-.68**	.11	.18
	6f+	**.47**	-.05	-.06	.26	**.52**	.22	.17	-.34	**.48**	-.11	.01	-.34	.17	**.52**	.25	.12
	17f-	-.22	-.25	**.36**	.26	**-.80**	.04	.03	.07	**-.45**	-.23	.27	.07	.31	**-.81**	.03	.02
	19f-	-.14	-.28	**.34**	.30	**-.69**	.11	.07	.08	**-.31**	-.27	.25	.08	.33	**-.64**	.11	.06
	20f+	**.46**	.17	-.05	.29	**.50**	-.03	.16	-.28	**.49**	.15	.02	-.28	.20	**.49**	-.01	.16
	25f+	**.46**	.28	-.02	.28	**.53**	-.15	.13	-.28	**.51**	.28	.08	-.28	.19	**.51**	-.14	.13
Verbal Aggression	5v-	.12	**-.72**	.08	.07	-.06	**.73**	.08	.11	.13	**-.75**	.06	.11	.12	-.04	**.73**	.09
	7v+	.02	**.66**	.09	.15	-.03	**-.68**	-.01	-.08	.01	**.68**	.11	-.08	.14	-.03	**-.68**	.03
	9v+	.20	**.58**	.11	.30	.04	**-.63**	.08	-.23	.12	**.65**	.15	-.23	.24	.03	**-.64**	.08
	12v-	-.02	**-.58**	.26	.12	-.17	**.57**	-.12	.15	-.05	**-.59**	.22	.15	.18	-.16	**.57**	-.10
	15v+	-.06	**.84**	.10	.20	-.01	**-.86**	-.11	-.15	-.12	**.88**	.12	-.15	.17	-.02	**-.86**	-.09
	22v-	.03	**-.27**	.17	.16	-.13	**.30**	-.01	.11	.08	**-.33**	.19	.11	.22	-.12	**.30**	.04
	27v+	**.38**	.19	.10	.29	.15	-.16	**.18**	-.33	**.25**	.19	.11	-.33	.20	.13	**-.16**	.13
Indirect Aggression	3i+	**.77**	-.08	.00	.34	.11	.04	**.64**	-.42	**.64**	-.08	.03	-.42	.21	.09	.04	**.59**
	4i+	**.58**	-.08	.03	.28	.15	.09	**.42**	-.31	**.47**	-.08	.05	-.31	.18	.14	.08	**.36**
	10i-	.19	.01	**.68**	.48	.01	.09	**-.33**	.00	.13	-.02	**.69**	.00	.52	.00	.08	**-.28**
	11i+	**.63**	-.16	-.03	.26	.18	.18	**.52**	-.43	**.48**	-.17	-.01	-.43	.11	.13	.16	**.40**
	14i-	.00	.00	**.59**	.40	.08	.14	**-.52**	-.04	-.08	-.02	**.68**	-.04	.44	.07	.13	**-.56**
	16i-	-.18	.19	**.64**	.34	.10	-.03	**-.71**	.17	-.15	.17	**.69**	.17	.43	.10	-.04	**-.65**
	18i+	**.62**	-.02	-.06	.30	.16	.01	**.53**	-.33	**.58**	-.04	-.02	-.33	.22	.16	.03	**.53**
	23i+	**.30**	.27	-.02	.25	-.01	**-.35**	.31	-.21	.22	**.31**	-.01	-.21	.16	.02	**-.32**	.27
	24i-	-.01	.05	**.60**	.36	.08	.08	**-.52**	.01	-.06	.02	**.61**	.01	.38	.04	.07	**-.48**
	26i-	.11	-.04	**.68**	.47	.17	.22	**-.53**	.09	.14	-.08	**.75**	.09	.53	.16	.21	**-.44**

thus artificially improving model fit and overestimating internal consistency based reliability (Danner et al., 2015; Ferrando & Lorenzo-Seva, 2010; Salazar, 2015; Weijters et al., 2013).

The present paper proposes a third approach to this issue that attempts to overcome some of the problems associated with the use of reversed items: employing some type of procedure to control the undesirable effects of these items while maintaining their advantages, such as controlling for AC effects. As we have shown, the reliability and congruence with the expected structure after controlling for AC effects is equivalent to that reported in the direct version of the questionnaire. As also found in previous research, the distortions in the factorial structure due to response biases are more related to AC than to SD (Navarro-González et al., 2016; Soto et al., 2008). A clear example of the effects of removing biases can be seen in Table 4, which shows that when the effects of response biases were not controlled, the only scale that retains its factorial structure reasonably well is VA, whose items had lower AC loadings, while the PA and IA items, which are more affected by AC, were split into two method factors associated with direct and reversed items.

In summary, it seems that if the test administrator is not concerned about possible AC effects, the best option is the use of only direct items. However, if the items used in a questionnaire have a noticeable AC impact, the alternative of mixing both types of items with a procedure to control AC effects may be of

Table 5
Factorial reliabilities and congruences with expected solution for the direct and original versions of the I-DAQ with and without controlling for biases

	DIRECT VERSION					
	With bias			Controlling for social desirability		
	PHY	VER	IND	PHY	VER	IND
$r_{\theta\theta}$	.82	.89	.89	.85	.90	.86
C	.95	.86	.94	.95	.87	.93
	ORIGINAL VERSION					
	With bias			Controlling for social desirability		
$r_{\theta\theta}$	.83	.78	.78	.78	.88	.85
C	.52	.79	.79	.79	.87	.85
	Controlling for acquiescence			Controlling for both bias		
$r_{\theta\theta}$	.87	.88	.85	.85	.88	.80
C	.95	.87	.95	.95	.88	.94

interest, because it controls for AC effects without consequential undesirable effects on reliability or factorial structure.

Despite the advantages, some problems are not solved by using a procedure to control for AQ. For instance, the assumption that participants both interpret and respond to the items in the same manner independently of item direction does not seem tenable. As this study shows, reversed items have both greater item means (lower means after keying the item in the same direction of direct items) and greater loadings on AC than direct items. It seems logical to think that the difference in item means may be due to these differential AC effects, but in our case the greatest difference in means was found in VA, which was the scale least impacted by AC. Therefore, it seems that AC cannot completely explain the lack of equivalence between direct and reversed items. This may be because, as different authors have proposed, reversed items may not elicit the same cognitive demands as direct items, and this effect may partially explain the mean differences between the two types of items due to factors such as verbal skills (Suárez-Alvarez et al., 2018; Weems, Onwuegbuzie, & Collins, 2006).

It is worth mentioning that one limitation of this study is that the effects of acquiescence on the factorial structure of the I-DAQ reported above was found in a university sample and only for one measure. Nevertheless, finding these effects in this sample is relevant because as the cognitive level of individuals increases, their person reliability increases, and their AC level decreases (Escorial, Navarro-González, Ferrando, & Vigil-Colet, 2019; Meisenberg & Williams, 2008; Navarro-González, Ferrando, & Vigil-Colet, 2018). So, our results show that AC has a noteworthy impact in the factorial structure of typical response measures, even in samples which have low AC levels and high individual consistency. Furthermore, the effects reported above are quite similar to those reported in heterogeneous samples for the same measure (Navarro-González et al., 2016). Further research is, however, needed in samples that usually display higher levels of AC, such as low ability and elderly groups, and in other measures of typical performance, and to determine the effects of scales composed of direct or direct and reversed items on predictive validity. This last issue is relevant because acquiescence may under or overestimate validity depending on whether the trait and criterion are imbalanced in the same or in different directions (Danner et al., 2015; Soto & John, 2019).

Acknowledgments

This project has been possible with the support of a grant from the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behaviour Research and Therapy*, *41*, 1411-1426. https://doi.org/10.1016/S0005-7967(03)00059-7

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. https://doi.org/10.1016/j.jrp.2015.05.004

DeVellis, R.F. (2003). Scale development: Theory and applications (2nd ed), Newbury Park, CA. Sage.

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*, 13-19. https://doi.org/10.1027//1015-5759.16.1.13

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, 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*, 427-437. https://doi.org/10.1080/00223891.2012.662188

Escorial, S., Navarro-González, D., Ferrando, P. J., & Vigil-Colet, A. (2019). Is individual reliability responsible for the differences in personality differentiation across ability levels? *Personality and Individual Differences*, *139*, 331-336. https://doi.org/10.1016/j.paid.2018.12.004

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *The British Journal of Mathematical and Statistical Psychology*, *63*, 427-448. https://doi.org/10.1348/000711009X470740

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 364-381. https://doi.org/10.1080/10705510902751374

Józsa, K., & Morgan, G. A. (2017). Reversed items in likert scales: Filtering out invalid responders. *Journal of Psychological and Educational Research*, *25*, 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*, 512-541. https://doi.org/10.1177/1094428115571894

Lorenzo-Seva, U., & Berge, J. Ten. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*, 57-64. https://doi.org/10.1027/1614-2241.2.2.57

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010

Morales-Vives, F., Lorenzo-Seva, U., & Vigil-Colet, A. (2017). How response biases affect the factor structure of big five personality questionnaires. *Anales de Psicologia*, *33,* 589-596. https://doi.org/10.6018/analesps.33.2.254841

Navarro-González, D., Ferrando, P. J., & Vigil-Colet, A. (2018). Is general intelligence responsible for differences in individual reliability in personality measures? *Personality and Individual Differences*, *130*, 1-5. https://doi.org/S0191886918301600

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports, *28*, 465-470. https://doi.org/10.7334/psicothema2016.113

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva, U. (2019). Psychological Test Toolbox: A new tool to compute factor analysis controlling response bias, *Journal of Statistical Software*, *91*(6). https://doi.org/10.18637/jss.v091.i06

Nunnally, J.C. (1978). Psychometric theory, New York: McGraw-Hill.

Paulhus, D.L., & Vazire, S. (2005). The self-report method. In R.W. Robins & R. C. Fraley (Ed.), *Handbook of Research Methods in Personality Psychology* (pp. 224-239). New York: Guilford Press.

Paulhus, D. (1991). Measurement and control of response bias. *Measures of Personality and Social Psychological Attitudes* (October), 17-59. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

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

Rammstedt, B., & Kemper, C. J. (2011). Measurement equivalence of the Big Five: Shedding further light on potential causes of the educational bias. *Journal of Research in Personality*, *45*, 121-125. https://doi.org/10.1016/j.jrp.2010.11.006

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*, 1137-1145. https://doi.org/10.1037/a0033323

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology*, *121*, 81-96. https://doi.org/10.1080/00224545.1983.9924470

Ruiz-Pamies, M., Lorenzo-Seva, U., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2014). I-DAQ: A new test to assess direct and indirect aggression free of response bias. *The Spanish Journal of Psychology*, *17*, E41. https://doi.org/10.1017/sjp.2014.43

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

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*, 407-424. https://doi.org/10.1080/00273171.2014.931800

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics Affect external validity? *Psychological Assessment*, *31*, 444-459. https://doi.org/10.1037/pas0000586

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*, 718-737. https://doi.org/10.1037/0022-3514.94.4.718

Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema*, *30*, 149-158. https://doi.org/10.7334/psicothema2018.33

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

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

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

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