The Spanish Journal of Psychology (2022), **25**, e24, 1–9.

doi:10.1017/SJP.2022.20

To be Direct or not: Reversing Likert Response Format Items

Jaime García-Fernández, Álvaro Postigo, Marcelino Cuesta, Covadonga González-Nuevo, Álvaro Menéndez-Aller and Eduardo García-Cueto

Universidad de Oviedo (Spain)

Abstract. Likert items are often used in social and health sciences. However, the format is strongly affected by acquiescence and reversed items have traditionally been used to control this response bias, a controversial practice. This paper aims to examine how reversed items affect the psychometric properties of a scale. Different versions of the Grit-s scale were applied to an adult sample ($N = 1,419$). The versions of the scale had either all items in positive or negative forms, or a mix of positive and negative items. The psychometric properties of the different versions (item analysis, dimensionality and reliability) were analyzed. Both negative and positive versions demonstrated better functioning than mixed versions. However, the mean total scores did not vary, which is an example of how similar means could mask other significant differences. Therefore, we advise against using mixed scales, and consider the use of positive or negative versions preferable.

Received 24 January 2022; Revised 19 September 2022; Accepted 21 September 2022

Keywords: Acquiescence, Grit-S, Likert scales, reversed items

Likert-type items (Likert, 1932) are one of the most widely-used multiple-response question formats for assessing no cognitive variables. In this type of item, the participant selects an option from a group of alternatives ordered by the level of agreement with the item statement. Positive forms of items (also called direct or non-reversed items) give high scores when the participant has a high level in the assessed trait. Negative forms of items (reversed) give low scores when the participant has a high level in the trait. Items can be reversed either by adding a negation to the item statement, a technique known as reverse orientation (e.g., from "I consider myself a good person" to "I do not consider myself a good person") or by using reverse wording using antonyms (e.g., from "I consider myself a good person" to "I consider myself a bad person"; Suárez-Álvarez et al., 2018; van Sonderen et al., 2013). In applied research, one question to ask when constructing a scale is whether the scale should include reversed items.

Reversed items aim to control one of the main response biases in self-report measures: Acquiescence (Navarro-González et al., 2016). Acquiescence is defined as the tendency to agree with an item statement, disregarding its content (Paulhus & Vazire, 2005). It is not a response set bias (like social desirability) but a response style bias (like inattention; van Sonderen et al., 2013). Despite widespread use, psychometric research generally does not advise this practice (Podsakoff et al., 2012; Vigil-Colet et al., 2020), although some authors do defend it, declaring that a small number of negative items may cause slower, more careful reading of items (Józsa & Morgan, 2017).

Reversed items complicate cognitive processing of item statements (Marsh, 1986; Suárez-Álvarez et al., 2018; van Sonderen et al., 2013), hence they are not considered advisable (Irwing, 2018; Lane et al., 2016; Moreno et al., 2015; Muñiz & Fonseca-Pedrero, 2019). Furthermore, reversed items have a differential effect on participants depending on their cultures (Wong et al., 2003), personality traits (DiStefano & Motl, 2009), intelligence, and linguistic performance (Suárez-Álvarez et al., 2018). In addition, reversed items complicate

Correspondence concerning this article should be addressed to Jaime García-Fernández. Universidad de Oviedo. Facultad de Psicología. Plaza de Feijoo, S/N. 33003 Oviedo (Spain).

E-mail: garciafernandezj@uniovi.es. Phone: +34–985104140.

Funding statement: This investigation was supported by a predoctoral grant from the Universidad de Oviedo (PAPI–21–PF–24).

Conflicts of Interest: None.

How to cite this article:
García-Fernández, J., Postigo, Á., Cuesta, M., González-Nuevo, C., Menéndez-Aller, Á., & García-Cueto, E. (2022). To be direct or not: Reversing likert response format items. *The Spanish Journal of Psychology*, *25*. e24. Doi:10.1017/SJP.2022.20

inter-item correlation estimations (Navarro-González et al., 2016), diminish items' discriminatory power (Chiavaroli, 2017; Józsa & Morgan, 2017), reduce scale reliability (Carlson et al., 2011), and produce different scores in positive and negative items. With regard to the latter, inverted items usually lead to higher scores once their scores are redirected (Suárez-Álvarez et al., 2018; Vigil-Colet, 2020), as people tend to disagree more with negative items than with direct ones (i.e., people may doubt whether they "*finish* every task they start", but will probably disagree with the idea of "*not finishing* every task they start"). However, Solís Salazar (2015) found higher scores for positive items, even when negative items are redirected.

Another problem caused by the use of reversed items is having worse dimensionality indexes in essentially-unidimensional constructs. In fact, a psychological construct could even move from being unidimensional to having two method factors when positive and negative items are mixed—one factor for positive and another for negative items—(Essau et al., 2012; Horan et al., 2003; van Sonderen et al., 2013; Woods, 2006). The grit construct is an example of this issue. Grit is a trait based on perseverance combined with passion for accomplishing long-term goals (Duckworth, 2016; Duckworth & Quinn, 2009). The best-known scale for grit assessment is Grit-S (Duckworth & Quinn, 2009), which is supposed to assess two dimensions (perseverance of effort and consistency of interest). Here, negative items make up the first factor, while the second is made up of positive ones. Recent research has shown that grit has a unidimensional structure, with the bidimensional model being caused by reversed items (Areepattamannil & Khine, 2018; Gonzalez et al., 2020; Morell et al., 2021; Postigo et al., 2021; Vazsonyi et al., 2019). Therefore, some grit scales have been developed following the unidimensional hypothesis, such as the Oviedo Grit Scale (EGO; Postigo et al., 2021).

Research on item redirection usually uses unidimensional scales to show the effects of inverse items (Solís Salazar, 2015; Suárez-Álvarez et al., 2018; Vigil-Colet, 2020). However, reversed items in the Grit-S scale produced a method factor that had serious consequences in terms of the substantive conceptualization of the construct. Given this, we believe that demonstrating what effects reversed items have on the Grit-s scale may be interesting for grit researchers. Applied researchers may also benefit from a clear example of how item reversal may affect scales in terms of item properties, total scores, factor structures, and reliability. It is important to analyze all of these differences, because although some properties may not vary between groups, this does not mean that the remaining properties will behave in the same way.

Another interesting point is the effect that reversed items might have when the scale is related to other variables. Although there is much research about how item reversal affects internal consistency, reliability, and even total scores (as previously explained), we have not found any studies mentioning the effects negative items can have in correlations with other psychological constructs. Previous research on grit has reported that high levels of grit are related to low levels of neurotic disorders, such as anxiety or depression (Datu et al., 2019; Musumari et al., 2018). It would be interesting to see how this relationship (grit-neuroticism) may be affected by item reversal.

The present study examines whether item reversal in Likert response format items influences the psychometric properties of a grit scale (Grit-S) and the relationship with another variable (Neuroticism).

First, we aim to determine how item reversal affects the factorial structure of the scale. As a consequence of the methodological artifact, we would expect scales that mix both types of items to have a bidimensional structure (caused by a methodological artifact), and the positive and negative versions to have a unidimensional structure. The second objective is to analyze possible changes in the total score due to using reversed items. If negative items tend to have higher scores, the more negative items in a scale, the higher the total scores. Thus, we would expect the negative version to have higher total scores than the mixed or original versions, which would both also have higher scores than the positive version. Third, we aim to show how reliability is affected by item reversal. As negative items usually correlate between each other more than positive ones (Solís Salazar, 2015), and because the Cronbach's alpha (α) coefficient is based on these correlations, negative scales should have higher reliability coefficients than positive scales. In addition, mixing the two types of items can force a scale from being unidimensional to being bidimensional. This would worsen the reliability coefficients, which are conceived to be estimated on unidimensional scales. Finally, the fourth objective is to analyze how correlations with another variable are affected by the use of reversed items. As explained above, grit has an inverse relation with Neuroticism, so negative correlations with the Neuroticism subscale of the NEO Five Factor Inventory (NEO-FFI) are expected, and this relationship should be stronger for the more reliable scales.

Method

Participants

The study sample comprised 991 Spaniards who completed an online questionnaire. 103 participants were excluded because they demonstrated suspicious response behavior (i.e., taking too much or too little time to answer

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

the questionnaire or leaving some items unanswered). This sample was complemented by another 531 participants from the same population who took part in a previous study where Grit-S scales were applied.

The final sample consisted of 1,419 participants divided into five groups (Table 1). As the table shows, the different groups had similar mean ages, sex ratios, and levels of educational qualifications. Most of the sample had completed university (66.8%), followed by those who finished high school (19.0%), vocational training (10.2%), and secondary/primary school (4.0%).

The sample size is adequate for Exploratory Factor Analysis as each group contains over 200 participants and the scales have no more than 10 five-point Likert items (Ferrando & Anguiano-Carrasco, 2010).

Instruments

Grit-S. Grit-S (Duckworth & Quinn, 2009) is a scale with eight items assessing two dimensions (four items for each dimension): Perseverance of effort and consistency of interest. The items use a five-point Likert response format. We used the Spanish version by Arco-Tirado et al. (2018), in which Cronbach's alpha = .77 for the consistency of interest dimension, Cronbach's alpha = .48 for the perseverance of effort dimension and Cronbach's alpha = .75 for the total score. This version of the scale has five inverted items (the original English scale has four), four of which are in the consistency of interest dimension. Another three versions of the scale were developed (positive, negative, and mixed—explained below). The reversal process was as follows: a group of seven experts in Psychometrics and Psychological Assessment created several alternative versions for each original item (positives or negatives depending on the original item) using the reversed wording technique. The main reason for using reversed wording instead of reversed orientation is that the second one is not recommended by previous research (Haladyma & Rodríguez, 2013; Irwing, 2018; Muñiz & Fonseca-Pedrero, 2019). Afterwards, the representativeness of each alternative version was discussed. The versions with a minimum consensus of six out of seven (86%) experts were selected for developing the different scale versions. Hence, we created the Grit-S positive (all items in direct form), Grit-S negative (all items reversed) and Grit-S mixed (half of the items were randomly selected and inverted, disregarding their dimension). Although the original Grit-S scale is already a mixed scale, the reversed items in the Grit-S mixed version were randomly selected, and the consistency of interest dimension contains more than solely reversed items.

The four Grit-S scale versions are shown in Table A1 (see Appendix). The structure of each scale is given in Table A2 (see Appendix).

Neuroticism subscale, NEO-FFI test. The NEO-FFI test (Costa & McCrae, 1985) is an inventory for assessing personality following the Big Five personality model. The Neuroticism subscale is composed of 12 Likert-type items with five response categories from *completely disagree* to *completely agree.* It was adapted to Spanish by Cordero et al., (2008). The original Cronbach's alpha coefficient for the scale was .90. In this study, we found a Cronbach's alpha coefficient of .86.

Procedure

Each group completed one scale in an online survey platform. The participants were found through non-probabilistic convenience sampling. Data collection lasted 5 months. Participants completed the scale anonymously and voluntarily without any compensation. All participants gave their informed consent, and their anonymity was ensured according to Spanish data protection legislation, Organic Law 3/2018, de 5th December, on Individual Data Protection and the Guarantee of Digital Rights (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales).

Data Analysis

Dimensionality

Several Exploratory Factor Analyses (EFA) were conducted in order to assess the dimensionality of the

Table 1. *Sample Groups Regarding the Answered Scale*

Group	*n*	% Women	*M (SD)* Age	% Studies
1 – Grit-S positive	302	63.2	44.42 (14.46)	72.8/14.9/8.9/3.4
2 – Grit-S mixed	289	68.2	41.24 (16.93)	61.6/27.7/8.0/2.7
3 – Grit-S negative	297	68.7	38.10 (16.86)	51.2/31.6/12.8/4.4
4 – Grit-S original	531	62.0	38.60 (14.90)	78.3/9.8/7.9/4.0
Total	1,419	64.9	40.26 (15.84)	68.1/19.1/9.2/3.7

Note. M = mean; *SD* = standard deviation; % studies = university studies/high school/vocational training/secondary or primary studies.

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

scales. When items have five or more response alternatives, and skewness and kurtosis are less than one, a Pearson correlation matrix is advised for factorial analysis (Lloret-Segura et al., 2014). The suitability of the matrix for factorial analysis was assessed using the Kaiser-Meyer-Olkin (KMO) and the Bartlett statistic. KMO should be greater than .80 to ensure a feasible analysis (Kaiser & Rice, 1974). Robust Unweighted Least Squares (RULS) was used as an estimation method. To decide on the number of extracted factors, we used an optimal implementation of Parallel Analysis (PA; Timmerman & Lorenzo-Seva, 2011). The feasibility of the factorial structure was assessed using total explained variance and the Comparative Fit Index (CFI). More precisely, to assess the suitability of the unidimensional structure, we estimated Explained Common Variance (ECV; Ferrando & Lorenzo-Seva, 2017). CFI should be greater than .95 (Hu & Bentler, 1999), and ECV greater than .80 (Calderón Garrido et al., 2019). The factor loadings of the different versions were compared using the Wrigley and Neuhous congruence coefficient (García-Cueto, 1994).

Descriptive Statistics, Item Analysis and Differences in Scores

We calculated descriptive statistics (mean, standard deviation, skewness, kurtosis and discrimination index) for each item in each grit scale. The discrimination index should be higher than .20 to consider an item a good measure of the trait (Muñiz & Fonseca-Pedrero, 2019). To verify if reversed items had significantly affected the total Grit-S scores, ANOVA between scale versions (original, positive, negative, mixed) was performed.

Reliability

Scale reliability was assessed using Cronbach's alpha. We computed Feldt's *w* statistic (Feldt, 1969) to assess whether there were significant differences between the reliability of the scales.

Descriptive statistics, ANOVA and the *t*-test were estimated using IBM SPSS Statistics (Version 24). Reliability and EFAs were assessed using FACTOR 12.01.01 (Lorenzo-Seva & Ferrando, 2013).

Correlations with Other Variables

Three versions of the Grit-S scale (positive, negative, mixed) were correlated with the Emotional Stability score of the NEO-FFI test. We could not estimate the correlation for the original version of the Grit-s as this sample did not complete the Emotional Stability scale.

Results

Dimensionality

A total of four EFAs were conducted, one for each version of the scale. Optimal Implementation of Parallel Analysis recommended one dimension in all versions of the scale (see Figure 1). Table 2 shows the KMO, Bartlett significance level, percentage of total explained variance, ECV and CFI for each version of the scale. Table 3 shows the comparisons between factorial loadings of the four Grit-S scales.

The Grit-S negative version gave the best fit, followed by the positive, original, and mixed versions. The original and mixed versions did not reach the requirement established for KMO and ECV, thus indicating a bad fit to a unidimensional structure.

Descriptive Statistics, Item Analysis and Differences in Scores

Descriptive statistics for the items are shown in Table 4. The items from the versions of the Grit-S scale had means between 2.58–4.26 and standard deviations between 0.79–1.27. Apart from the kurtosis value for Item 2 (–1.00) and the skewness value for Item 5 (–1.28) —both from the negative Grit-S scale—all skewness and kurtosis indexes were between ±1.

Discrimination indexes were generally lower in the mixed versions and higher in the negative versions than the positive versions. Item 5 of the Grit-S scale demonstrated no discriminatory power (.00)

The ANOVA for the four versions of the Grit-S scale showed no significant differences between the total scores for the original, mixed, positive, and negative versions ($F = 0.972$; $df = 3$; $p = .405$).

Reliability

The reliability for each version of the scale is shown in Table 2. The original and mixed versions demonstrated the worst reliability. Reliability comparisons are shown in Table 5. The negative version of the Grit-S negative version had significantly better reliability than the other versions, and the positive version had better reliability than the original or mixed versions.

Correlations with Other Variables

The Pearson correlations between Neuroticism and the grit scales were –.26 for the positive version, –.38 for the mixed version and –.53 for the negative version.

Discussion

Reversed items have been questioned by previous research for various reasons (Carlson et al., 2011;

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

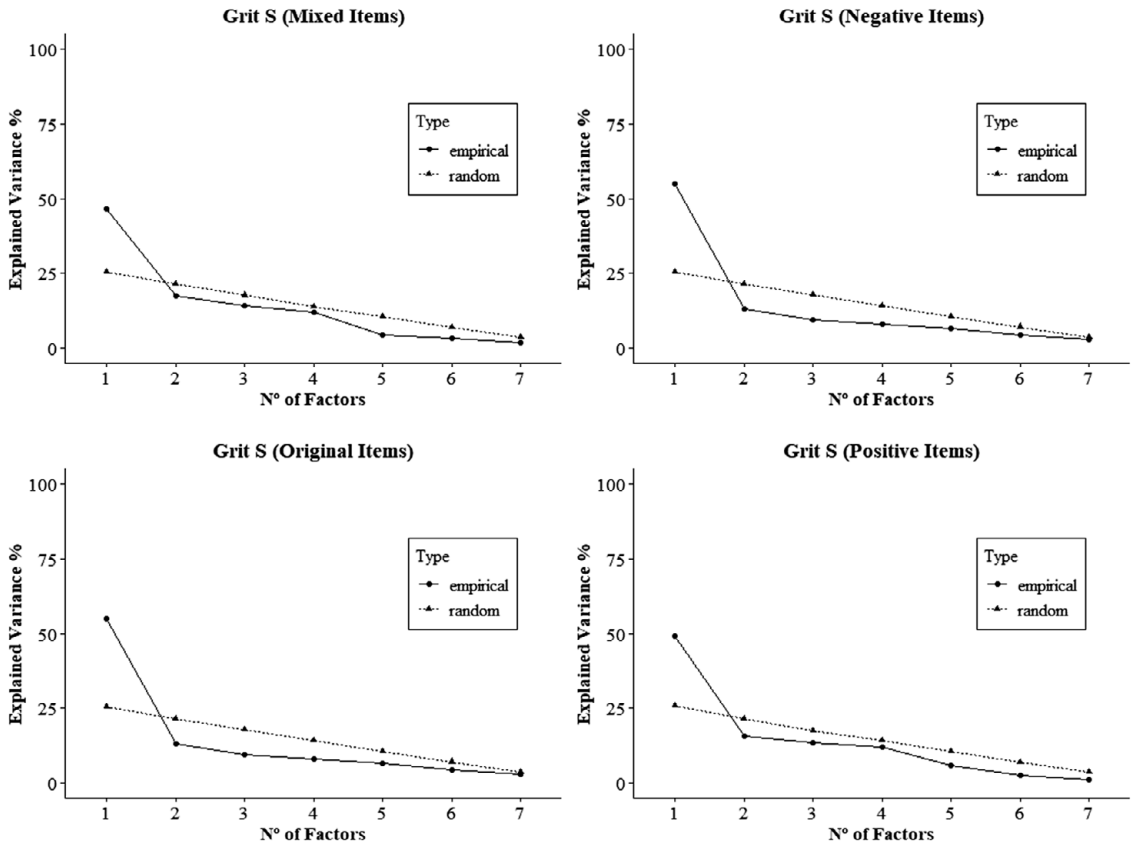

Figure 1. Results of the Optimal Implementation of Parallel Analysis

Table 2. *Fit Indices of Exploratory Factor Analysis for Grit Scale Versions*

	Sphericity					
	KMO	(*p* Bartlett)	Explained Variance	ECV	CFI	α
Grit-S (−)	.868	(< .001)	47.24	.820	.965	.83
Grit-S (+)	.816	(< .001)	41.91	.820	.936	.77
Grit-S (M)	.773	(< .001)	39.00	.729	.944	.72
Grit-S (O)	.788	(< .001)	35.63	.768	.893	.73

Note. KMO = Kaiser-Meyer-Olkin statistic. ECV = Explained Common Variance. CFI = Comparative Fix Index. α = Cronbach's α.

Table 3. *Factorial Loadings Comparison of Grit Scale Versions*

Comparison	r_c	Significance
Grit-S (−) — Grit-S (M)	.954	$p > .05$
Grit-S (−) — Grit-S (O)	.954	$p > .05$
Grit-S (−) — Grit-S (+)	.932	$p > .05$
Grit-S (O) — Grit-S (+)	.957	$p > .05$
Grit-S (O) — Grit-S (M)	.937	$p > .05$
Grit-S (+) — Grit-S (M)	.982	$p > .05$

Note. (−) = negative; (+) = positive; (M) = mixed; (O) = original; r_c = Congruence coefficient.

Chiavaroli, 2017; Essau et al., 2012; Navarro-González et al., 2016). The present study examined the effect of item reversion on a grit scale, as well as any potential consequences of that when relating the scale to other variables.

Looking at the dimensionality of the versions of the scale, EFA points to a unidimensional structure, similar to previous results (Areepattamannil & Khine, 2018; Gonzalez et al., 2020; Postigo et al., 2021), meaning that the hypothesis of a two-factor structure for mixed versions is refuted. However, the best fit indexes were found for the negative and positive versions, while the mixed versions (both mixed and original Grit-S scales)

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

Table 4. *Descriptive Statistics of the Items*

	Grit-S original version							Grit-S positive version					
	M	*SD*	*sk*	*k*	*DI*	*FL*		*M*	*SD*	*sk*	*k*	*DI*	*FL*
[a] i–1	2.68	1.02	0.20	–0.50	.37	.44	i–1	2.83	1.00	0.07	–0.43	.31	.31
[a] i–2	3.20	1.14	–0.08	–0.85	.45	.53	i–2	3.45	1.09	–0.45	–0.44	.32	.36
[a] i–3	3.61	1.02	–0.51	–0.32	.59	.70	i–3	3.79	0.97	–0.51	–0.28	.71	.81
[a] i–4	3.66	1.13	–0.60	–0.40	.45	.53	i–4	3.70	1.01	–0.60	–0.21	.63	.73
[a] i–5	3.22	1.10	–0.17	–0.64	.26	.31	i–5	2.80	1.09	0.21	–0.70	.16	.16
i–6	4.26	0.79	–0.94	0.68	.35	.43	i–6	4.09	0.88	–0.85	0.51	.53	.67
i–7	3.66	0.96	–0.52	–0.20	.59	.71	i–7	3.62	1.01	–0.54	–0.19	.57	.69
i–8	4.01	0.89	–0.89	0.77	.59	.41	i–8	4.01	0.89	–0.89	0.77	.59	.68
Total	28.3	4.73	–0.26	–0.08	-		Total	28.3	4.89	–0.26	–0.09	-	-

	Grit-S negative version							Grit-S mixed version					
	M	*SD*	*sk*	*k*	*DI*	*FL*		*M*	*SD*	*sk*	*k*	*DI*	*FL*
[a] i–1	2.99	1.13	0.03	–0.79	.54	.60	i–1[a]	2.86	1.09	0.18	–0.74	.43	.51
[a] i–2	3.17	1.23	–0.11	–1.00	.60	.43	i–2	3.45	1.11	–0.49	–0.41	.15	.22
[a] i–3	3.30	1.20	–0.20	–0.97	.63	.48	i–3	3.68	0.99	–0.56	–0.19	.66	.76
[a] i–4	3.54	1.27	–0.44	–0.97	.71	.62	i–4[a]	3.66	1.22	–0.66	–0.59	.68	.77
[a] i–5	2.97	1.12	0.04	–0.79	.33	.13	i–5	2.58	1.12	0.33	–0.64	.00	.02
[a] i–6	4.22	1.02	–1.28	0.87	.54	.35	i–6	3.96	0.94	–0.73	0.08	.49	.57
[a] i–7	3.75	1.10	–0.64	–0.36	.58	.42	i–7[a]	3.74	1.12	–0.68	–0.36	.48	.63
[a] i–8	3.86	1.20	–0.86	–0.25	.59	.43	i–8[a]	3.90	1.09	–0.80	–0.20	.50	.61
Total	27.8	6.34	–0.35	–0.43	-	-	Total	27.8	5.06	–0.46	–0.08	-	-

Note. *M* = mean; *SD* = standard deviation; *sk* = skewness; *k*= kurtosis; *DI* = Discrimination Index; *FL* = Factorial Loading; [a] = negative items.

exhibited the worst unidimensional fit. In other words, the use of both positive and negative items promotes the multidimensionality of the scale (Essau et al., 2012; Horan et al., 2003; Woods, 2006). This is not only a problem for the scale's internal consistency, but can have serious consequences for the theoretical framework that researchers are developing, for example, conceptualizing more factors than necessary because of the method factor that negative items may produce. Continuing with factorial structure, the items' factor loadings did not exhibit statistically significant differences between versions. This indicates that the factorial structure did not differ due to the use of reversed items, although this structure is less clear when using mixed scales (as they had worse fit indexes).

In the Grit-S scale, the negative version demonstrated greater reliability (α = .83) than the positive version (α = .77). This can be explained as due to the higher correlations between the negative items than between the positive items (Solís Salazar, 2015). The positive version exhibited a higher reliability coefficient than the mixed and original versions. Finally, there were no statistically significant differences in reliability between the mixed and original versions, which was expected as both of these scales mix positive and negative items. This confirms previous findings about the reduced reliability coefficients when using mixed scales (Carlson et al., 2011).

There were no statistically significant differences between the versions with regard to the total scores. This refutes our second hypothesis, as our data did not replicate the results of previous findings (Suárez-Álvarez et al., 2018; Vigil-Colet, 2020). This could be seen as the grit scale being a "special case" due to its items (people tend to agree or disagree in the same way with negative and positive items when asked about their grit levels) or the length of the questionnaire, as previous research has shown these differences with questionnaires that are at least twice as long as the Grit-S scale. One might think that the scales could be used interchangeably, given that there were no mean differences between versions. We advise against this interpretation, as having the same mean does not imply that an individual would have the same score in both versions. As we mentioned previously, the quality of factorial scores worsens with mixed versions, as do the reliability coefficients, and these differences are statistically significant.

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

Table 5. *Reliability Comparison of Grit Scale Versions*

Comparison	*w*	Significance
Grit-S (−) — Grit-S (M)	0.600	$p < .001$
Grit-S (−) — Grit-S (O)	1.636	$p < .001$
Grit-S (−) — Grit-S (+)	1.370	$p < .005$
Grit-S (O) — Grit-S (+)	1.195	$p < .001$
Grit-S (O) — Grit-S (M)	0.982	$p < .360$
Grit-S (+) — Grit-S (M)	0.822	$p < .047$

Note. (−) = negative; (+) = positive; (M) = mixed; (O) = original

Another example of what might be masked by similar total mean scores is the change in the correlation coefficients with Neuroticism. By redirecting just half of the items, the correlation goes from –.26 to –.38 (a difference of 7.6 in the percentage of explained variance). If all items are redirected, that produces a correlation of –.53 (the percentage of explained variance grows by 21 points). This proves that redirecting items can have a powerful effect on the relationship with other variables. We believe that the reason for this difference is the increase in the variance of the total scores produced by negative items, which affects the correlation coefficient (Amón Hortelano, 1990). This may vary depending on the psychological construct being assessed (positive items may exhibit more variance than negative items for a different variable).

The results of this study should be assessed in light of some limitations. First, using a cross-sectional design, with different samples responding to each scale, could have biased the results, although the groups did have similar sociodemographic characteristics. In this regard, future studies should apply longitudinal designs. Secondly, the possibility of developing a "perfect-inverted item" is unclear, given semantic, grammatical and/or expressive issues. Some reversed expressions may sound 'weird' to a native speaker, leading to grammatical changes that make the sentence clearer but further from being a precise reversed version of the original item. This is not only a limitation for the present study, but also another argument against the use of reversed items in scale development.

Applied researchers should avoid developing mixed scales. Note that the problems with negative items come when they are included in a scale along with positive items (i.e., mixed scales). Having an entirely negative scale—with properly constructed items—cannot be considered bad practice, as this study shows. Thus, researchers should select which form (positive or negative) they prefer considering the theoretical framework of the construct. It is also important to note that having the same mean total scores does not mean that the compared scales are equivalent, as the factorial structure, reliability, and the relationship with other variables may differ significantly.

References

Amón Hortelano, J. (1990). *Estadística para psicólogos: Estadística descriptiva* [Statistics for psychologists: Descriptive statistics] (12th Ed.). Pirámide.

Arco-Tirado, J. L., Fernández-Martín, F. D., & Hoyle, R. H. (2018). Development and validation of a Spanish version of the Grit-S scale. *Frontiers in Psychology*, *9*, Article 96. https://doi.org/10.3389/fpsyg.2018.00096

Areepattamannil, S., & Khine, M. S. (2018). Evaluating the psychometric properties of the original Grit Scale using rasch analysis in an Arab adolescent sample. *Journal of Psychoeducational Assessment*, *36*(8), 856–862. http://doi.org/10.1177/0734282917719976

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

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research and Evaluation*, *22*, Article 3. https://doi.org/10.7275/ca7y-mm27

Cordero, A., Pamos, A., & Seisdedos, N. (2008). *Inventario de Personalidad Neo Revisado (NEO PI-R), Inventario Neo Reducido de Cinco Factores (NEO-FFI): Manual profesional* [The Neo Personality Inventory Revised (NEO-PI-R), the Reduced Five Factor Personality Inventory (NEO-PI-R): Professional manual] (3rd Ed.). TEA.

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Psychological Assessment Resources.

Datu, J. A. D., King, R. B., Valdez, J. P. M., & Eala, M. S. M. (2019). Grit is associated with lower depression via meaning in life among Filipino high school students. *Youth & Society*, *51*(6), 865–876. https://doi.org/10.1177/0044118x18760402

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*(3), 309–313. https://doi.org/10.1016/j.paid.2008.10.020

Duckworth, A. (2016). *Grit: The power of passion and perseverance*. Scribner/Simon & Schuster.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Essau, C. A., Olaya, B., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., O'Callaghan, J., & 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. https://doi.org/10.1002/mpr.1364

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, *34*(3), 363–373. https://doi.org/10.1007/BF02289364

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología [Factor analysis as a research technique in psychology]. *Papeles del Psicólogo*, *31*(1), 18–33.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*(5), 762–780. https://doi.org/10.1177/0013164417719308

García-Cueto, E. (1994). Coeficiente de congruencia [Congruence coefficient]. *Psicothema*, *6*(3), 465–468.

Calderón Garrido, C., Navarro González, D., Lorenzo Seva, U., & Ferrando Piera, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Gonzalez, O., Canning, J. R., Smyth, H., & MacKinnon, D. P. (2020). A psychometric evaluation of the Short Grit Scale. *European Journal of Psychological Assessment*, *36*(4), 646–657. https://doi.org/10.1027/1015-5759/a000535

Haladyma, T. M., & Rodríguez, M. C. (2013). *Developing and validating test items*. Taylor & Francis. https://doi.org/10.4324/9780203850381

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Irwing, P., Booth, T., & Hughes, D. J. (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. John Wiley & Sons. https://doi.org/10.1002/9781118489772

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.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Lane, S., Raymond, M. R., & Haladyma, T. M. (2016). Handbook of test development (2nd Rd.). *Routledge*. https://doi.org/10.4324/9780203102961

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 55.

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada [Exploratory Item Factor Analysis: A practical guide revised and updated]. *Anales de Psicología*, *30*(3). https://doi.org/10.6018/analesps.30.3.199361

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2. *Applied Psychological Measurement*, *37*(6), 497–498. https://doi.org/10.1177/0146621613487794

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

Morell, M., Yang, J. S., Gladstone, J. R., Turci Faust, L., Ponnock, A. R., Lim, H. J., & Wigfield, A. (2021). Grit: The long and short of it. *Journal of Educational Psychology*, *113*(5), 1038–1058. https://doi.org/10.1037/edu0000594

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

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema*, *31*(1), 7–16. https://doi.org/10.7334/psicothema2018.291

Musumari, P. M., Tangmunkongvorakul, A., Srithanaviboonchai, K., Techasrivichien, T., Suguimoto, S. P., Ono-Kihara, M., & Kihara, M. (2018). Grit is associated with lower level of depression and anxiety among university students in Chiang Mai, Thailand: A cross-sectional study. *PLOS ONE*, *13*(12), Article e0209121. https://doi.org/10.1371/journal.pone.0209121

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). Efectos de los sesgos de respuesta en la estructura factorial de los autoinformes de personalidad [How response bias affects the factorial structure of personality self-reports]. *Psicothema*, *28*(4), 465–470. https://doi.org/10.7334/psicothema2016.113

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales [Organic Law 3/2019, of December 5, on the protection of personal data and guarantee of digital rights] (2018, December 6th). *Boletín Oficial del Estado*, 294, Sec. I, pp. 119788–119857. https://www.boe.es/eli/es/lo/2018/12/05/3/dof/spa/pdf

Paulhus, D. L., & Vazire, S. (2005). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guildford Press.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Postigo, Á., Cuesta, M., García-Cueto, E., Menéndez-Aller, Á., González-Nuevo, C., & Muñiz, J. (2021). Grit assessment: Is one dimension enough? *Journal of Personality Assessment*, *103*(6), 786–796. https://doi.org/10.1080/00223891.2020.1848853

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

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press

Likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

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), Article e68967. https://doi.org/10.1371/journal.pone.0068967

Vazsonyi, A. T., Ksinan, A. J., Ksinan Jiskrova, G., Mikuška, J., Javakhishvili, M., & Cui, G. (2019). To grit or not to grit, that is the question! *Journal of Research in Personality*, *78*, 215–226. https://doi.org/10.1016/j.jrp.2018.12.006

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

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

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Appendix

Table A1. *Positive and Negative Items for Grit-s Scales*

Grit-S (positive items)	Grit-S (negative items)
1. New ideas and projects never distract me from previous ones.	1. New ideas and projects sometimes distract me from previous ones.[a]
2. I have been obsessed with some idea or project for a long time without losing interest.	2. I have been obsessed with a certain idea or project for a short time but later lost interest.[a]
3. I am constant in my goals.	3. I often set a goal but later choose to pursue a different one.[a]
4. I maintain my attention on projects that take a long time to complete (more than a few months).	4. I have difficulty maintaining my focus on projects that take more than a few months to complete.[a]
5. Setbacks do not discourage me.	5. Setbacks discourage me. [a]
6. I am a hard worker. [a]	6. I am a little worker.
7. I finish whatever I begin. [a]	7. I never finish everything I start.
8. I am diligent. [a]	8. I am lazy.

Note. [a] = original Grits-S item.

Table A2. *Item Direction (Positive or Negative in Each Scale Versions)*

Scale (version)	Items
Grit-S (O)	1(–), 2(–), 3(–), 4(–), 5(–), 6, 7, 8
Grit-S (M)	1(–), 2, 3, 4(–), 5, 6, 7(–), 8(–)
Grit-S (+)	1, 2, 3, 4, 5, 6, 7, 8
Grit-S (–)	1(–), 2(–), 3(–), 4(–), 5(–), 6(–), 7(–), 8(–)

Note. (O) = original version; (M) = mixed version; (+) = positive version; (–) = negative version.

https://doi.org/10.1017/SJP.2022.20 Published online by Cambridge University Press