Psicothema

Psicothema 2021, Vol. 33, No. 4, 639-646
doi: 10.7334/psicothema2021.131

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2021 Psicothema
www.psicothema.com

Article

Is Correcting for Acquiescence Increasing the External Validity of Personality Test Scores?

Ana Hernández-Dorado, Andreu Vigil-Colet, Urbano Lorenzo-Seva, and Pere J. Ferrando
Universitat Rovira i Virgili

Abstract

Background: Balanced scales control for acquiescence (ACQ) because the tendency of the respondent to agree with the positive items is cancelled out by the tendency to agree with opposite-pole items. When full balance is achieved, ACQ is not expected to affect external validity. Otherwise, attenuated estimates are expected to appear if no control methods such as Lorenzo-Seva & Ferrando's (2009) are used. **Method:** Expected results were derived analytically. Subsequently, a simulation was carried out to assess (a) how ACQ impacted external validity and (b) how validity estimates behaved when ACQ was corrected. Two illustrative examples are provided. **Results:** A sizable number of items and/or high content loadings tended to decrease ACQ's impact on validity estimates, making the empirical coefficient closer to its structural value. Furthermore, when scales were well balanced, the controlled and uncorrected scores were close to each other, and led to unbiased validity estimates. When the scales were unbalanced and no corrections were used, attenuated empirical validity coefficients inevitably appeared. **Conclusions:** Designing a well-balanced test or correcting for ACQ are the best ways to minimize attenuation in external validity estimation.

Keywords: Response biases; external validity; measurement applications.

Resumen

¿La Corrección por Aquiescencia Aumenta la Validez Externa de las Puntuaciones en Personalidad? **Antecedentes:** construir escalas balanceadas permite controlar la aquiescencia (ACQ), haciendo que la tendencia del encuestado a estar de acuerdo con los ítems positivos se cancele con la tendencia a estar de acuerdo con los ítems del polo opuesto. En caso contrario, se esperarán estimaciones atenuadas de los coeficientes de validez externa en caso de no utilizar algún método de control (Lorenzo-Seva & Ferrando, 2009). **Método:** se llevó a cabo (a) un desarrollo analítico (b) una simulación para evaluar (a) el impacto de ACQ en la validez externa y (b) el comportamiento de las estimaciones de validez cuando se corrige por ACQ. Incluyendo finalmente dos ejemplos ilustrativos. **Resultados:** número alto de ítems y/o cargas altas en el factor de contenido tienden a disminuir el impacto de ACQ en las estimaciones de validez. Además, con escalas balanceadas por diseño, las diferencias entre las puntuaciones corregidas y no corregidas son menores, llevando a estimaciones de validez insesgadas. En escalas no balanceadas ni corregidas aparece una atenuación en el coeficiente de validez empírico. **Conclusiones:** diseñar pruebas balanceadas o corregir ACQ son las mejores maneras de minimizar la atenuación en la estimación de la validez externa.

Palabras clave: sesgos de respuesta; validez externa; metodología aplicada.

Believing that the answer to an item is an accurate reflection of the trait to be measured is highly optimistic. Item responses may be affected by several factors other than the intended content, such as social desirability, extreme response and acquiescence (ACQ) (e.g. Bentler et al., 1971). It has been estimated that ACQ causes 3-5% of the variance in personality or attitude scales, and it can spuriously inflate inter-item correlations and, therefore, reliability estimates (Lechner et al., 2019). And as has been shown by some studies of scales based on the Five-Factor Model, ACQ can also lead to an unrealistic factor structure (Soto et al., 2008; Danner et al., 2015; Morales-Vives et al., 2017).

Over the years, ACQ has been defined (see Baumgartner & Steenkamp, 2001; Ferrando et al., 2016) and studied (Ray, 1983; Wetzel et al., 2016) from various points of view. In all cases, however, control has been the objective and, of the different forms of control, balancing the scale is one of the most classical and effective. However, balancing scales is by no means easy. Neither is there any guarantee that it will control for ACQ properly inasmuch as it can alter the latent structure of the data and therefore affect the method (Ferrando et al., 2003). On the other hand, reverse items tend to be more complex, they can only be understood by respondents with good language skills and so they tend to have lower factorial weights than direct items (Condon et al., 2006; Suárez-Álvarez et al., 2018). Likewise, it is not always clear that using positive and reversed items in the same test reduces response biases (Suárez-Álvarez et al., 2018).

There are several "a posteriori" methods in which ACQ is allowed to occur but is then eliminated using statistical procedures. Most of these procedures are based on fully balanced scales (Ferrando et al., 2003; Billiet & McClendon, 2000), but some, such as Lorenzo-Seva & Ferrando (2009), also allow ACQ to be corrected on quasi-balanced and unbalanced scales. In applied research, quasi-balanced scales (same number of positive and negative items but

Received: April 8, 2021 • Accepted: July 17, 2021
Corresponding author: Ana Hernández-Dorado
Research Center for Behavior Assessment. Department of Psychology
Universitat Rovira i Virgili
43007 Tarragona (Spain)
e-mail: ana.hernandez@urv.cat

with unequal saturations) and partially balanced scales (different number of positive and negative items) are relatively common.

ACQ needs to be controlled and new forms of ACQ-control, such as Maydeu-Olivares & Coffman (2006), are still being investigated. The RIFA method, tested by De la Fuente & Abad (2020), could be a good alternative to the EFA-based method because it is easier to implement and is robust to the violation of the assumption of tau-equivalence in ACQ factor loadings. However, this approach is less accurate to highly heterogeneous ACQ loadings patterns because its estimate of the loadings of ACQ is the calculation of the average of these loadings. And, therefore, if we assume that the items will be affected differently by the ACQ and there is interest in the study of the ACQ factor, the Lorenzo-Seva & Ferrando (2009) model will be more appropriate. For more information on the development of ACQ control methods that have been developed, see Primi, Santos et al. (2019)

Although numerous studies have been made of ACQ, very few focus on validity, which might be partly because the concept is sometimes abstract and unfathomable. Primi, De Fruyt et al. (2019) observed differences in the criterion-related validity of the direct scores of the ACQ-uncorrected and corrected tests, and found "false-keyed" items to be more valid than "true-keyed" items.

Model-based Predicted Results

Basic Results and Validity Coefficients

We shall consider a general bi-dimensional model in which each item response is a measure of a content factor (θ_c) and an acquiescence factor (θ_a). The model has two parts: a measurement sub-model (1), and an extended structural sub-model (2) in which an external variable (a criterion) is regressed on the latent constructs defined in (1)

$$x_j = \lambda_{jc}\theta_c + \lambda_{ja}\theta_a + \varepsilon_j \tag{1}$$

$$y = \lambda_y\theta_c + \varepsilon_y \tag{2}$$

where λ_{jc} is the loading on content, λ_{ja} is the loading on ACQ and λ_y is the "true" validity coefficient. Both factors in (1) as well as the external variable (y) in (2) are assumed to be scaled which mean 0, and unit variance. In the simplest case: (a) ACQ and content are assumed not to be correlated, and (b) the criterion is an objective variable, and so uncorrelated with the ACQ factor.

This article will now go on to deal with the effects of ACQ on the coefficient of validity when the test scores are factor score estimates. Information about the corresponding effects when scores are raw or unit-weight sum scores will be provided by the authors on request.

By extending the definitions in Lord & Novick (1968, sect. 12.1), we now define the theoretical validity coefficient as the correlation $\rho\theta_c y$ which, given our adopted scaling and the fact that (2) is a single-regressor equation, is simply λ_y. Next, we define the empirical validity coefficient as the correlation between the content factor score estimates (based on model (1)) and y, denoted by "$\rho\hat{\theta}_{c'y}$". The relation between both coefficients is given by:

$$\rho\hat{\theta}_{c'}y = \frac{\rho\theta_c y}{\sqrt{1+\dfrac{1}{\sum_j \dfrac{\lambda_{jc}^2}{\sigma_{\varepsilon j}^2}}}} \tag{3}$$

where $\sigma_{\varepsilon j}^2$ is the error variance of the *jth* item. Equation 3 predicts that, if the factor estimates are corrected for ACQ, the empirical validity coefficient is still an attenuated measure of the "true" relationship between the test content and the criterion y (i.e. the theoretical validity). Clearly, attenuation is mitigated when test length and magnitude of the loadings increase, a result noted in previous studies (e.g. Soto & John, 2019).

Gone with the ACQ

Consider now that the measurement model (1) holds, and the data is fitted by the unidimensional model, so assuming that all the common variance is due to the content and that, therefore the presence of ACQ is ignored. Denote by $\hat{\theta}_g$ the maximum likelihood (ML) factor score estimates of the general factor. In this case, the validity relations are:

$$\rho_{\hat{\theta}_g{}'y=}\frac{\delta_c\rho\theta_c y}{\sqrt{1+\dfrac{1}{\sum_j \dfrac{\lambda_{jg}^2}{1-\lambda_{jg}^2}}}} \tag{4}$$

where δ_c is the covariance between the general factor scores estimates ($\hat{\theta}_g$) and the level of content (θ_c). The expression for δ_c is provided in Ferrando (2010, equation 18) and can be also estimated by regression. The main point here, however, is that, being almost a correlation (θ_c is standardized and $\hat{\theta}_g$ almost is), its value is always smaller than 1. Again, the empirical validity in (4) is an attenuated estimate of the theoretical validity. However, the attenuation is stronger here because of the additional term δ_c as well as the different amounts of information in the denominator: In effect, the error variance based on the single general factor is larger than the error variance in (3) obtained after two factors have been extracted.

The validity relation in (4) could be questioned because it is based on a wrong unidimensional model that is fitted to bidimensional data. Indeed, fitting the wrong model is expected to result in biased parameter estimates, somewhat larger item residual variances (as noted above), and, to some extent, bad model-data fit (see Ferrando, 2010). However, as long as model in (1) and (2) holds, prediction (4) continues to be correct even though the estimates provided by the unidimensional model are biased or the model does not fits well the data.

The Cost of not Correcting

The cost of not correcting for acquiescence in terms of validity, when validity is based on factor score estimates, can now be operationalized by relating the empirical validity based on general score estimates $\rho\hat{\theta}_{g'}y$ to the empirical validity based on score estimates corrected for ACQ $\rho\hat{\theta}_{c'}y$. The theoretical relation is

$$\rho\hat{\theta}_{g'}y = \rho\hat{\theta}_{c'}y\left[\delta_c\frac{\sqrt{1+\dfrac{1}{\sum\dfrac{\lambda_{jc}^2}{\sigma_{\varepsilon j}^2}}}}{\sqrt{1+\dfrac{1}{\sum\dfrac{\lambda_{jg}^2}{1-\lambda_{jg}^2}}}}\right] \tag{5}$$

Although equation (5) has the general form of a correction-for-attenuation formula, strictly speaking it is an attenuation relation between two already attenuated estimates, but of a different order. The general-factor-based validity estimate is an attenuated estimate of the corrected-factor-based estimate, and the attenuating factors are: the strength of the relation between the general and the content factor, the number of items, and the amount of variance due to ACQ. The impact of test length has already been discussed and the impact of the remaining two sources is only to be expected. When the impact of ACQ is low, (a) the general factor is close to (or mostly reflects) the content factor, (b) the variance due to ACQ is small, and, (c) the two empirical validity coefficients become closer one to another.

Simulation Study

Goals

The present simulation aims to assess (a) how the internal characteristics of the test (test length, content factor loadings, and the balancing of content factor and ACQ factor) impact the criterion-related validity; and (b) what happens to the estimated validity when the variance due to ACQ is removed from biased test scores.

Method

The factorial design of this first study was 2 × 3 × 2 × 2 × 5 × 3 = 360 conditions with 100 replicas per condition. Previous trials indicated that a higher number of replicas only increased the estimated time, rather than vary results. The independent variables were: (1) correcting (C) versus not correcting (NC) ACQ; (2) degree of balance in the factor loadings pattern: B, balanced, Q quasi-balanced, and U unbalanced; (3) high (H) versus low (L) loadings on content factor; (4) balanced versus unbalanced ACQ pattern; (5) number of items (4, 10, 16, 30, 50) and; (6) "theoretical" correlation between the external variable (criterion) and the content factor (.70, .50, .30). Table 1 summarises the independent variables and levels.

The dependent variable was the estimated validity coefficient, which, in order to be compared, was centered by calculating the difference between the estimated coefficient and the theoretical (i.e. true) correlation. The quality of the validity estimates was assessed through an ANOVA. The measure of effect size was eta square, and Cohen's interpretation criteria were: values close to .01 would have little effect, .06 moderate and .14 or greater high.

Table 1
Summary of Variables

Variable	Levels	Name of levels
Correcting	2	Correcting ACQ (C) vs Not Correcting (NC)
Balanced	3	Balanced (B; equal number of positive and reversed items), Quasi-Balanced (Q; different mean loadings) Unbalanced (U; 75% of positive items)
Loadings	2	(in content factor) High (H; .70) vs Low (L; .50)
ACQ pattern	2	Equal (E; .20) vs Not Equal (NE; between .10 - .30)
Number of items	5	4, 10, 16,30, 50 items
Theoretical correlation	3	.70, .50, .30

Note: For quasi-balanced and unbalanced scales the simulated patterns were: at "High Loadings" .75 in positive items and -.65 in reverted items; and at "Low Loadings" .55 in positive items and -.45 in reverted

The *psych* package (Revelle, 2021) (for factor analysis) and the *vampyr* package (Navarro-González et al., 2020) for the ACQ controlling condition were used. Finally, ANOVA was implemented in the R *stats* package. The code used in this simulation will be available to the reader on request.

Results

At a general level, there is a slight attenuation effect in all correlations in both the NC and the C condition. However, as expected, attenuation tends to be stronger in the uncorrected condition (see Figure 1, graphic A).

The ANOVA results are summarised in Table 2. As a measure of attenuation (dependent variable), we used the difference between "true" (.7, .5 and .3) and estimated correlations. Only those variables with an effect size greater than .01 have been included. Effect sizes are large for the number of items (η^2= .383) and loading size (η^2= .136); and medium for the magnitude of the "true" correlation (η^2= .085). It is also noted that high loadings generate correlations closer to the "true" values at all levels of the variable "number of items" (η^2= .054) with less intragroup difference than in the test with low loadings (Figure 1, graph C)**.** Finally, in the case of low "true" correlations (true cor. = 0.3), the differences are smaller (both in high and low loadings condition) (Figure 1, graph B), and this effect is present at all levels of the variable "number of items" (Figure 1, graph D).

Figure 2 compares the differences between correcting and not correcting, the number of items, and the amount of balance (whose effect size is η^2= .072). Here, there is hardly any discrepancy in the balanced conditions B and Q, while in condition U the difference in C is visibly greater than in NC. However, the effect size is medium. This may be due to the lower level of the variable "number of items".

At first sight, the three graphs in Figure 2 do not explain the effect size obtained in the analysis. There is no discrepancy between conditions B, Q and U. The differences tend to be greater in the NC condition, and this trend is repeated in the three balanced levels and at all levels of the variable number of items, (except the "4 items" level). The effect size is moderate due to the fact that in the lowest condition of the "number of items", in the corrected ACQ condition, high differences only arise when the tests are unbalanced.

Table 2
Summary of ANOVA results

		F	sig	effect size
principal effects	C vs NC	153.516	< .001	.001
	nº items	53877.846	< .001	**.383**
	balanced	69.930	< .001	.000
	true cor.	12010.547	< .001	**.085**
	Loadings	19073.277	< .001	**.136**
interaction double	nº items * loadings	4281.508	< .001	.054
	true cor. * loadings	2223.746	< .001	.015
	nº items * true cor.	1736.925	< .001	.041

Note: Only significant results have been included. Highlights in bold are those values of effect size that are higher according to Cohen

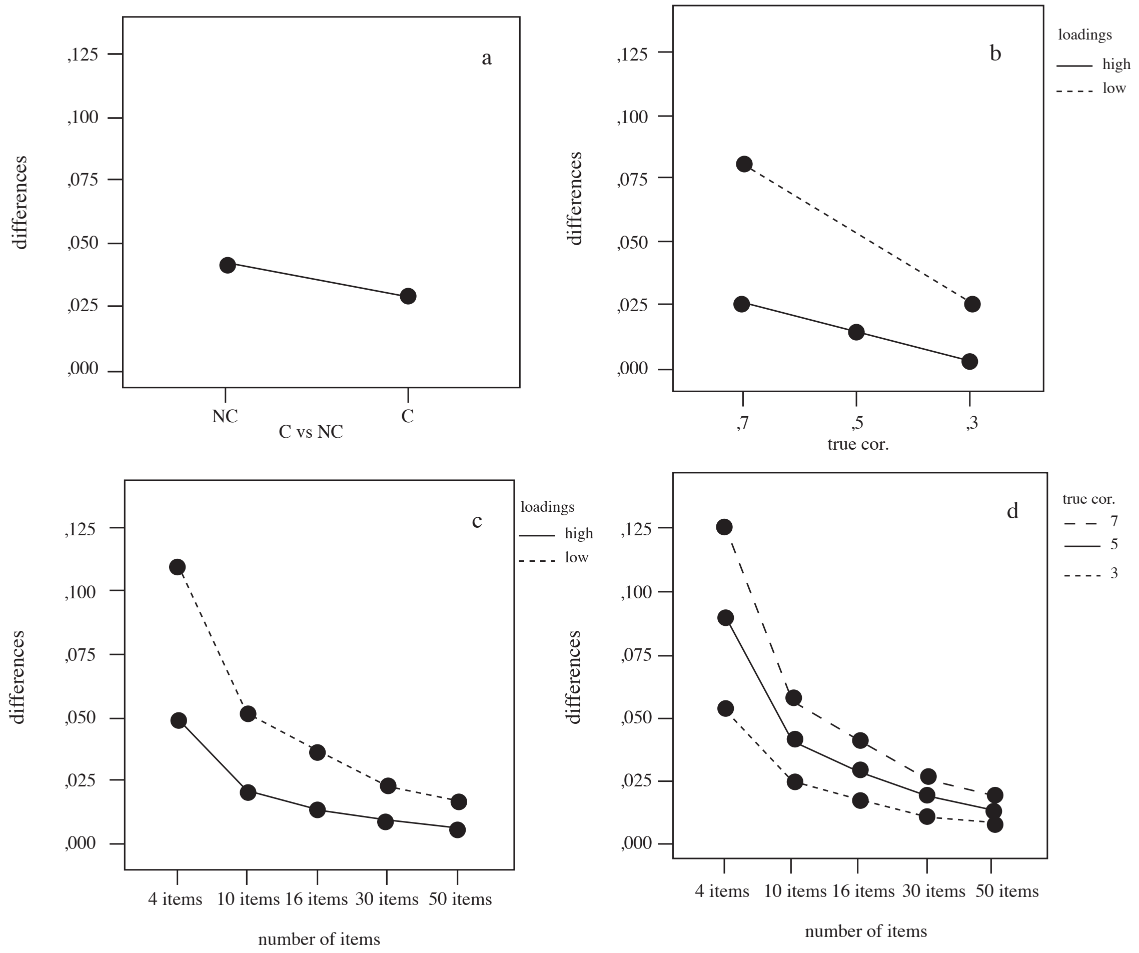

Figure 1. Graphs of ANOVA
Note: graph A: Top left; B: Top right; C: bottom left; D: bottom right

Illustrative Example 1

This first example assesses the relationship between the scores on an anxiety test, and the marks on an exam. So, the external variable can be properly considered as an objective, non-test, criterion measure.

Method

Participants

The sample consisted of 299 Psychology undergraduates (54 men and 245 women; that is, 82% of the sample were women) from six universities, with ages ranging from 18 to 60 years old (*M*=20.9 *SD*=4).

Instruments

The Examination Anxiety subscale of SAS (Statistical Anxiety Scales was used for this purpose (Vigil-Colet et al., 2008). It is a fully balanced measure made up of six 5-point Likert-type items. The numerical qualification on an exam was also recorded using a scale that ranged between 0 and 10 points (being 10 the maximum score).

Procedure

The questionnaire was administered collectively during class hours, in groups of 65 students, by a psychologist. Participation was voluntary and the protection of personal data was ensured. For one of the groups of 65 students, the teacher of the statistical subject of the group was actually the psychologist collecting the questionnaires: for this group an identification number was used so that qualifications in the subject could be added to participants' questionnaire responses. Once the data of interest was collected, any participant's identification was deleted from the study.

Data Analysis

All the computations were carried out with FACTOR (Ferrando & Lorenzo-Seva, 2017) and the Psychological Test Toolbox software (Navarro-González et al., 2019).

Results

Adequacy of correlation matrix to be factor analyzed

Data sampling adequacy was first checked and found acceptable: KMO=.810. Essential unidimensionality was assessed using Explained Common Variance (ECV), and Item Residual Absolute Loadings (I-REAL) (for details, see Ferrando & Lorenzo-Seva, 2018). The ECV and I-REAL values and the corresponding 95% confidence intervals were .861 [.828, .912)], and .215 [.171, .237]. Both satisfy the thresholds of unidimensionality. Optimal Implementation of Parallel Analysis (Timmerman & Lorenzo-Seva, 2011) also recommended retaining a single factor.

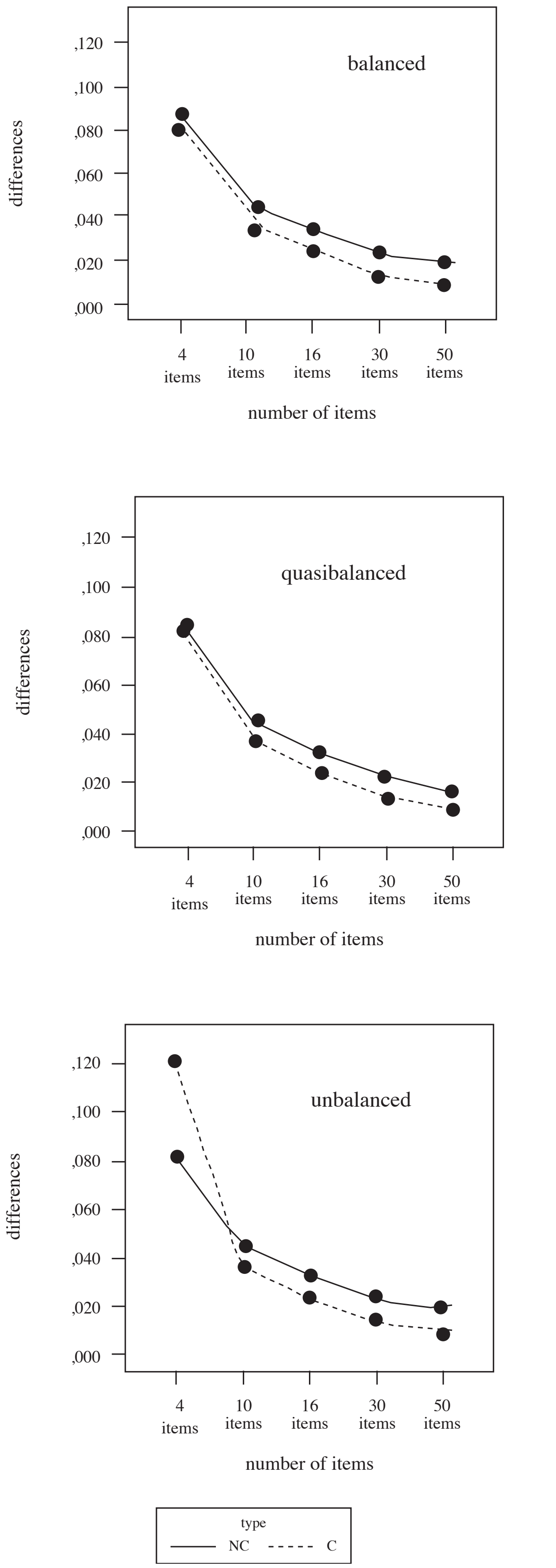

Figure 2. Triple interaction between balanced, C vs NC and number of items
Notes: graph A: balanced; B: quasi-balanced and C: unbalanced

Factor model fit

Robust linear exploratory factor analysis was computed using the ML criterion and specifying a single factor. Goodness of model–data fit was assessed by using both the conventional approach and the recent proposal by Yuan et al. (2017) based on equivalence testing. Goodness-of-fit measures were based on the second-order (mean and variance) corrected chi-square statistic proposed by Asparouhov & Muthen (2010) and the results are in the column labeled *Unidimensional Model* in Table 3. Overall, the results suggest that the fit of the unidimensional model is acceptable. The corresponding loading values for the unidimensional solution are in the column *Non-controlled AQ variance* (NC-ACQ) in Table 4.

Study of the impact of acquiescent variance in the factor structure

As the set of six items was fully worded balanced (i.e., half of them items are worded in the opposite direction to the other half of the items), the Ferrando et al. (2003) procedure for controlling for variance due to acquiescent responding was computed. As described in the paper by Ferrando at al, the procedure is based on an exploratory factor analytic approach: (a) it assesses the

Table 3
Goodness-of-fit values for illustrative example 1

Index	Unidimensional model	Bidimensional model
CFI	.964	.961
95% CI CFI	(.948; ,975)	(.894; .978)
T-size CFI	.919 (close)	.914 (close)
GFI	.990	.997
95% CI GFI	(.982; .997)	(.992; .999)
Z-RMSR	.069	.031
95% CI Z-RMSR	(.038; .087)	(.013; .046)

Table 4
Factor solutions for non-controlled ACQ variance (NC-ACQ) and for controlled AQ variance (C-ACQ) for illustrative example 1

Item	NC-ACQ	C-ACQ	
	EA	ACQ	EA
Direct	.916	.266	.914
Reversed	-.820	.185	-.854
Direct	.692	.373	.684
Reversed	-.422	.333	-.468
Direct	.501	.425	.503
Reversed	-.686	.315	-.779

Note: EA; Examination Anxiety

dimensionality and structure of a balanced personality scale taking into account the potential effects of acquiescent responding, and (b) it corrects the individual trait estimates for acquiescence. Goodness of fit results are in the column labeled Bidimensional Model in Table 3. Unsurprisingly, the fit of the bidimensional model is also acceptable, but is not a significant improvement on the unidimensional model. The corresponding loading matrix after controlling for ACQ is in the column *Controlled ACQ variance* (C-ACQ) in Table 4. Note that the last four items show a salient loading on the ACQ dimension (column ACQ), and that the loading of the fifth is the largest. Substantive loadings on *Examination Anxiety* when controlling for ACQ were quite similar to the ones obtained when there was no control (i.e., the unidimensional model). In fact, the value of the congruence index between the two columns is .998. The threshold value for considering two factor loading solutions to be equal is .95 (Lorenzo-Seva & ten Berge, 2006).

Validity study: Analysis of participants' acquiescent responses

EAP factor score estimates (see Ferrando & Lorenzo-Seva, 2016, for further details) on the unidimensional and the bidimensional factor solutions were computed for the 65 participants for whom qualification marks were available. For the unidimensional model, and as expected, the correlation between EA score estimates and the criterion was negative (-.368). The corresponding correlation between the "cleaned" score estimates and the criterion was -.386.

Overall, we note that the EA scale is characterized by high substantive factor weights, a relatively low decompensation between positive and negative items (see Table 4) and a small but sufficient number of items (Soto & John, 2019). In conclusion, as the predictions made here suggest, the empirical validity estimate is only slightly better when it is based on the factor score estimates corrected for ACQ.

Illustrative Study 2

In this example, we analyse the relationships between aggressiveness and intelligence, and how controlling for ACQ affects this relationship. Intelligence has often been related to violent behaviour (Ayduk et al., 2007), and this relationship has also been found using tests which control for response biases (Vigil-Colet et al., 2012; Duran-Bonavila, Morales-Vives et al., 2017). In the present study we analyse whether, as predicted by the model, the relationship increases when ACQ is controlled. It should be taken into account that all intelligence measures are maximal performance measures and, in consequence, they are not contamined by ACQ. So, we analyse the effects of removing ACQ on validity when the criterion is free from this type of bias.

Method

Participants

The sample consisted of a total of 532 students (252 men and 280 women) from 8 public high schools in the province of Tarragona, with ages ranging from 11 to 18 years old (*M*=14.75 *SD*=2.1) (see Duran-Bonavila, Vigil-Colet et al., 2017 for further details)**.**

Instruments

The Indirect-Direct Aggression Questionnaire (IDAQ) (Ruiz-Pamies et al., 2014) provides scores for physical aggression (PA), verbal aggression (VA) and indirect aggression (IA) factors as well as an overall aggression score. Although the questionnaire has a correlated-factors structure in three dimensions, we used the overall score for two reasons. First, the tri-dimensional structure of IDAQ items depends on whether ACQ is removed or not (Navarro-González et al., 2016; Vigil-Colet et al., 2020). As a consequence, if we analyze the effects of removing acquiescence at the multidimensional level, the items comprising the solutions with and without controlling acquiescence may be different. Second, the fit of the unidimensional model after controlling for ACQ is quite acceptable (CFI=.98, RMSR=.04, RMSEA=.07), which supports the idea that the IDAQ scores measure a general factor of indirect aggression.

Three tests were used as intelligence measures: *Thurstone's Primary Mental Abilities Test* (Cordero et al., 1989), which contains scales of fluid and crystallized intelligence; *Raven's Progressive Matrices Test* (Raven, 1996), an indicator of crystallized intelligence; and *the information scale of the WAIS intelligence test for adults* (Wechsler, 2003) which is an indicator of crystallized intelligence. Intelligence measures were used as objective criterion variables, as they are all maximum performance measures and are therefore not ACQ-biased.

Procedure

School approval and parental written informed consent were obtained before participation in the study. Participation was voluntary and no incentives were given. The questionnaires were anonymous, and respondents had to provide only their gender and age.

Data Analysis

We analyzed the data reported by Duran-Bonavila, Morales-Vives et al. (2017) estimating new factor scores with and without controlling for ACQ using all the IDAQ items. Data was analyzed using the Psychological Test Toolbox (Navarro-González et al., 2019) and SPSS 25.

Results

Table 5 shows the correlations between all intelligence measures and IDAQ's overall aggression scores with and without removing ACQ effects. In all cases, the correlation between the intelligence measures and aggression is negative, a result that has consistently been obtained in previous studies (Kavish et al., 2018; González-Moraga et al., 2019). More relevant here, for all intelligence measures the correlations between aggression and intelligence were slightly larger when ACQ effects were removed. The critical threshold here appears to be -.1: when ACQ was removed, most of the correlations shown were over -.1, while they were under -.1 when no correction was used. As for differential effects, both RAVEN and WAIS had an approximate increase of .05. Correlation with RAVEN and WAIS corrected (*r*= -.147 and r_r= -.182) and not corrected were respectively (*r* = -.096 and *r*= -.135).

Table 5
Product moment correlations between intelligence measures and overall aggression with and without controlling acquiescence

	With Bias	Controlling ACQ
WISC information	-.135*	-.182*
PMA verbal	-.058	-.102**
PMA spatial	-.074	-.086**
PMA reasoning	-.158*	-.193*
PMA numeric	-.100**	-.102**
PMA word fluency	-.055	-.077
PMA Total	-.130*	-.165*
Raven	-.097**	-.147*
G estimate	-.105**	-.148*

Note: * p < .01; **p < .05

Discussion

Despite the considerable interest in biases and response styles, their effect on validity has hardly been studied. This state of affairs justifies the main aim of our proposal: to “quantify” the effects of the internal characteristics of the test and the correction of ACQ on external validity. Three studies were carried out for this purpose: a simulation and two illustrative examples based on real data.

Studies such as the one by Soto & John (2019) allowed us to make initial predictions: that the use of balanced scales with a sufficient number of items and high loadings, would effectively correct for the impact of ACQ and improve validity estimates. Furthermore, a general starting point is that empirical validity is a biased estimate of true validity. These initial assumptions raised a number of questions that we have tried to answer throughout the study.

Evidence from analytical development as from simulated and empirical results suggests that the first prediction above was right:

Validity decreases when there are fewer than 10 items, when the loadings of the substantive factor are low, and when the scale is unbalanced. These are important benchmarks to be considered when designing scales. That is, attenuation is mitigated when both the length of the test and the magnitude of the pattern loadings increase, a result observed in previous studies (Soto & John, 2019), and derived from our analytical approach.

As expected, the data strongly supported that empirical validity is a biased estimate of ‘true’ validity. Furthermore, the amount of bias (downward bias or attenuation, to be more specific) seems to largely depend on the internal characteristics of the test and the “true” validity. Again, unfavorable internal characteristics will increase attenuation. On the other hand, attenuation is less pronounced when “true” validity is low, regardless of the number of items or loadings.

Finally, the third hypothesis raised the question of the expected gain in validity when correcting for ACQ. The theoretical results, the results from the simulation study, and those from the two proposed examples suggest that validity generally improves when the scale is corrected for acquiescence. This improvement is, in some cases, very subtle but nontrivial, and appears even in the case of almost fully balanced scales. That is, when the impact of ACQ is low, the general factor is close to the content factor and the two validity coefficients get closer to one another.

Correcting for ACQ is not expected to improve validity in scenarios in which there is already a pre-balancing correction and in tests where, because of the conditions, it is difficult to “extract” the ACQ factor. This can be seen in the decrease in the difference between correcting and not correcting.

Now, in the light of the results obtained, how should we proceed? First, they open the possibility of further investigating the effect of ACQ correction on validity, including variables or levels of variables that have been omitted here and that imply a limitation in the present study, such as a condition of no balance (without any reverted item) or including the correlation between the criterion and the ACQ. On the other hand, the results support the need for using a good design and not relying (or solely relying) on post-hoc corrections. An appropriate number of items with good internal characteristics in terms of both content loadings and balance of positive and negative items would go a long way to avoiding further validity biases. Therefore, we strongly suggest that great care be taken when designing the measuring instrument. On the other hand, in cases where the test does not have the required positive features and the items are believed to be affected by ACQ it is strongly recommended to use a correction method, since it is expected to lead to improvements in the estimated structure of the test, the individual score estimates derived from this structure, and (the point of this article) the external validity estimate.

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

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Mplus technical appendix,* 1-8. https://www.statmodel.com/download/ WLSMV_new_chi21.pdf

Ayduk, O., Rodríguez, M. L., Mischel,W., Shoda, Y., &Wright, J. (2007). Verbal intelligence and self-regulatory competencies: Joint predictors of boys’ aggression. *Journal of Research in Personality, 41*, 374-388 https://doi.org/10.1016/j.jrp.2006.04.008

Baumgartner, H., & Steenkamp, J. E. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156, https://doi.org/10.1509/jmkr.38.2.143.18840

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76*(3), 186-204. https://doi.org/10.1037/h0031474

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608-628, https://doi.org/10.1207/S15328007SEM0704_5

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality and Individual Differences*, *40*(3), 403-407. https://doi.org/10.1016/j.paid.2005.07.019

Cordero, A., Seisdedos, N., González, M., & de la Cruz, V. (1989). *PMA. Aptitudes Primarias Mentales* [Primary Mental Abilities]. TEA Ediciones.

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

de la Fuente, J., & Abad, F. J. (2020). Comparing Methods for Modeling Acquiescence in Multidimensional Partially Balanced Scales. *Psicothema*, *32*(4), 590-597. http://10.7334/psicothema2020.96

Duran-Bonavila, S., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2017). How impulsivity and intelligence are related to different forms of aggression. *Personality and Individual Differences, 117,* 66-70. https://doi.org/10.1016/j.paid.2017.05.033

Duran-Bonavila, S., Vigil-Colet, A., Cosi, S., & Morales-Vives, F. (2017). How Individual and Contextual Factors Affects Antisocial and Delinquent Behaviors: A Comparison between Young Offenders, Adolescents at Risk of Social Exclusion, and a Community Sample. *Frontiers in Psychology*, *8*, 1-12, https://doi.org/10.3389/fpsyg.2017.01825

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*(2), 427-448. https://doi.org/10.1348/000711009X470740

Ferrando, P. J., & Lorenzo-Seva, U. (2016). A note on improving EAP trait estimation in oblique factor-analytic and item response theory models. *Psicológica, 37*(2), 235-247. https://www.redalyc.org/pdf/169/16946248007.pdf

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*(2), 236-240 https://doi.org/ 10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva, U. (2018). 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

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*(3), 353-374, https://doi.org/10.1207/S15327906MBR3803_04

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 713-725. https://doi.org/10.1080/10705511.2016.1185723

González Moraga, F. R., García, D., Billstedt, E., & Wallinius, M. (2019). Facets of Psychopathy, Intelligence and Aggressive Antisocial Behaviors in Young Violent Offenders. *Frontiers in Psychology*, *10*, 984. https://doi.org/10.3389/fpsyg.2019.00984

Kavish, N., Bailey, C., Sharp, C., & Venta, A. (2018). On the relation between general intelligence and psychopathic traits: An examination of inpatient adolescents. *Child Psychiatry & Human Development*, *49*(3), 341-351. https://doi.org/10.1007/s10578-017-0754-8

Lechner, C. M., Partsch, M. V., Danner, D., & Rammstedt, B. (2019). Individual, situational, and cultural correlates of acquiescent responding: Towards a unified conceptual framework. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 426-446. https://doi.org/10.1111/bmsp.12164

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores, Reading*. Addison-Wesley.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. https://doi.org/10.1027/1614-2241.2.2.57

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Morales-Vives, F., Lorenzo-Seva, U., & Vigil-Colet, A. (2017). Cómo afectan los sesgos de respuesta a la estructura factorial de los tests basados en el modelo de los Cinco Grandes factores de personalidad [How response biases affect the factor structure of Big Five personality questionnaires]. *Anales de Psicología/Annals of Psychology, 33*(3), 589-596. https://doi.org/10.6018/analesps.33.3.254841

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*(4), 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), 1-21. https://doi.org/10.18637/jss.v091.i06

Navarro-González, D., Vigil-Colet, A., Ferrando, P.J., Lorenzo-Seva, U., & Tendeiro, J.N. (2020). *vampyr: Factor Analysis Controlling the Effects of Response Bias* (version 1.1.1) [R package]. https://cran.rstudio.com/web/packages/vampyr/index.html

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John O. P. (2019). True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement. *International Journal of Testing 20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 447-465. https://doi.org/10.1111/bmsp.12168

R Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.R-project.org/

Raven, J. C. (1996). *Matrices progresivas. Escalas CPM Color y SPM General* [Raven Progressive Matrices]. TEA Ediciones.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology, 121*(1), 81-96, http://doi.org/10.1080/00224545.1983.9924470

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research* (version 2.1.6) [R package]. https://cran.rstudio.org/web/packages/psych/psych.pdf

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

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*(4), 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*(4), 718-737. https://doi.org/10.1037/0022-3514.94.4.718

Suárez Álvarez, J., Pedrosa, I., Lozano, L. M., García Cueto, E., Cuesta Izquierdo, M., & Muñiz Fernández, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*(2), 149-158. http://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

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema, 20*(1), 174-180. http://www.psicothema.com/pdf/3444.pdf

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

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema*, *24*(2), 310-315. https://www.redalyc.org/pdf/727/72723578021.pdf

Wechsler, D. (2003). *Escala de inteligencia de Wechsler para niños-IV (WISC-IV)* [Wechsler Intelligence Scale for Children-WISC-IV]. Psychological Corporation.

Wetzel, E., Böhnke, J. R., & Brown, A., (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 349-363). Oxford University Press.

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, *23*(3), 319-330. https://doi.org/10.1080/10705511.2015.1065414