UNIVERSITAT ROVIRA I VIRGILI

DEVELOPMENT OF A PROGRAM TO CONTROL RESPONSE BIASES AND ASSESSMENT OF ITS USEFULNESS IN TYPICAL PERFORMANCE MEASURES

David Navarro González

UNIVERSITAT ROVIRA i VIRGILI

Development of a program to control response biases and assessment of its usefulness in typical performance measures

David Navarro González

DOCTORAL THESIS
2018

Development of a program to control response biases and assessment of its usefulness in typical performance measures

DOCTORAL THESIS

Psychology department

2018

UNIVERSITAT ROVIRA i VIRGILI

David Navarro González

Directed by:

Urbano Lorenzo-Seva

Andreu Vigil-Colet

UNIVERSITAT
ROVIRA i VIRGILI

FAIG CONSTAR que aquest treball, titulat "Development of a program to control response biases and assessment of its usefulness in typical performance measures", que presenta David Navarro González per a l'obtenció del títol de Doctor, ha estat realitzat sota la meva direcció al Departament de Psicologia d'aquesta universitat.

HAGO CONSTAR que el presente trabajo, titulado "Development of a program to control response biases and assessment of its usefulness in typical performance measures", que presenta David Navarro González para la obtención del título de Doctor, ha sido realizado bajo mi dirección en el Departamento de Psicología de esta universidad.

I STATE that the present study, entitled "Development of a program to control response biases and assessment of its usefulness in typical performance measures", presented by David Navarro González for the award of the degree of Doctor, has been carried out under my supervision at the Department of Psychology of this university.

Tarragona, 5 de novembre / Tarragona, 5 de novembre / Tarragona, November 5th

El/s director/s de la tesi doctoral
El/los director/es de la tesis doctoral
Doctoral Thesis Supervisor/s

Urbano Lorenzo-Seva

Andreu Vigil-Colet

Acknowledgments

The present work would never have come into without the help and support of many mentors, friends and family. First, I would like to acknowledge all the members of the University of Groningen that made possible my research stay in this prestigious University, and treated me as one of them. I should make special mention to Dr. Jorge Tendeiro, who was my supervisor and an inestimable support during my time in Groningen. I learned a lot in my time at RUG, and I would like to thank everyone there for that.

Now, I would like to proceed with the acknowledgments in my natives languages, since it is easier for me to express gratitude in the language that I use to communicate with the people closer to me.

Agradecimientos

Agradecer a todas las personas con las que he coincidido el congresos y eventos durante estos años, en Mallorca (dos veces), Barcelona, Girona, Oviedo, Madrid y Nueva York. En especial agradecer a los miembros del área de metodología de la Universidad Autónoma de Madrid, por todos los momentos que hemos compartido. Siempre me he sentido arropado por vosotros en todos los congresos a los que iba solo. Agradecerles también su hospitalidad en nuestra visita a Madrid, y sus interesantes comentarios sobre nuestras aportaciones.

Agradecer también al Dr. José Muñiz su hospitalidad en nuestra visita a Oviedo, y por enseñarnos cómo sabe una buena sidra.

Agradecer a mis amigos y compañeros de clase durante todos estos años, habéis significado mucho para mí aunque nos hayamos distanciado. Agradecer especialmente a Rubén, por estar siempre dispuesto por mucho tiempo que pase.

Agraïments

M'agradaria començar agraint als meus dos directors de tesi, el Dr. Urbano Lorenzo i el Dr. Andreu Vigil per la seva dedicació i entrega que han permès que aquesta tesi sortís endavant. Agrair-los també les seves ensenyances durant aquests anys, han estat un pilar fonamental en la meva formació com a investigador. També m'agradaria agrair-los la paciència en determinats moments, especialment en la trepidant recta final d'aquesta tesis.

A tots els membres del departament de Psicologia, tots heu contribuït a que aquesta experiència hagi estat lo més satisfactòria possible. Al Dr. Pere Joan Ferrando, per ser una font inesgotable de coneixement psicomètric, i pel seu interès en col·laborar en tot tipus de projectes. A la Dra. Ivette Espinoza, pel seu recolzament i optimisme durant tot el procés. A la Dra. Fàbia Morales, pel seu suport i per ser la màxima font de feedback del Psychological Test Toolbox, el programa no seria el mateix sense els seus comentaris. A la Dra. Sandra Cosi, pel seu recolzament i per les bones estones que hem passat. A Toni Masip, per tot el suport informàtic i per les estones de cafè. A Joan Bernadet i Esther Carrasco per la seva disposició i suport en tots els tràmits i incidències. A les noies del despatx de NUTRISAM, pels seus ànims i per les estones que hem compartit. I als companys que segur que em deixo, gràcies per les estones plegats.

Per suposat, agrair a els meus pares, Mari Pau i Enric, pel seu suport en tot lo possible, no hauria arribat aquí sense ells. Als meus germans, Dani i Edi, per ser-hi sempre que ho he necessitat. A Andrés, pel seu optimisme i bon humor contagiós. A la Marta, per la seva energia i vitalitat. I a tots els meus tiets, cosins i nebots (que sou molts), agrair-vos també el vostre recolzament malgrat que no ens veiem sovint. Tots vosaltres heu fet possible que hagi arribat fins aquí, gràcies pel vostre recolzament incondicional.

Finalment, agrair a la Joana per tot el camí que hem recorregut junts. No hi ha paraules en cap dels idiomes que conec per definir el que ha suposat tenir-te al costat tots aquests anys.

Index

1. Introduction

The present work focuses on a major problem that psychologists have to deal with when trying to estimate personality: to what extent do the obtained measures reflects accurately the latent trait. Of all the possible sources of information, personality measures are usually obtained using self-reports, although there are some exceptions (for example, child assessment is usually hetero-reported).

In typical response measures, the participant's responses may reflect not only the content assessed but also other systematic effects. These effects are known as *response bias*, which Paulhus (1991, 2017) defined as a systematic tendency to answer the items on some basis other than the specific item content. The two best known response biases in typical response measures are Acquiescence (AC) and Social Desirability (SD).

Acquiescence can be defined as the tendency of responders to agree with a statement regardless of its content (Paulhus and Vazire, 2005), while *Social Desirability* can be defined as people's tendency to present themselves in a generally favorable fashion (Holden, 2010). Some authors distinguish between two primary SD factors: a cluster associated with *Alpha*, the general factor anxiety factor of MMPI (Block, 1965), which was defined as self-deception, an unconscious self-favorability; and a second cluster associated with another MMPI factor called *Gamma* (Wiggins, 1964), which was defined as impression management, the intentional distortion of self-descriptions.

A review of the literature shows that SD and AC have generated some controversy, particularly in the 1950s and 1960s. Basically, there are two main positions on the conceptualization of SD and AC as variables. The traditional position considers the response bias as nuisance variables or artifacts that should be controlled or suppressed with appropriate procedures since they are of no further substantive interest (Edwards, 1967; Hofstee, ten Berge, & Hendriks, 1998; Nunnally, 1978; Ray, 1979). On the other hand, other authors believe that SD and AC are meaningful variables that are potentially measurable and can provide information of interest about the respondent (Couch & Keniston, 1960; Eysenck & Eysenck, 1976; Morf & Jackson, 1972).

There is evidence to support the conception of response biases as measurable traits that are quite stable across different measures. SD seems to be a trait associated with the responder, while AC seems to depend more on the item characteristics. SD shows consistency across situations (Ellingson, Smith, & Sackett, 2001; Lönnqvist, Paunonen, Tuulio-Henriksson, Lönnqvist, & Verkasalo 2007), and stability over time (Paulhus & Reynolds, 1995) and across different cultures (Barrett, Petrides, Eysenck, & Eysenck, 1998).

Vigil-Colet (2014) showed that the correlations between the SD scores obtained through three different questionnaires correlate between .60 and .70, while AC scores only correlate .20 approximately. This indicates that SD is quite stable across measures, while AC depends more on each instrument. This result is consistent with the one reported by Stöber (2001), who found substantial correlations between different measures of SD.

As far as AC is concerned, Ferrando & Condon (2004) found that acquiescence presents low indices of reliability and convergent validity, even for research purposes. The certain degree of convergent validity suggests that if AC can be regarded as a trait, it is only a very weak and unreliable one. This is consistent with the studies by Ray (1979, 1983), which considered that the generalization of acquiescence between scales is dubious. In contrast, Couch & Keniston (1960) found a correlation of r=.64 between the overall agreement scale and the MMPI acquiescence scale. In response to these results, Nunnally (1978) conjectured that separate scales tend to share common variance, since they measure non-stylistic personality traits. However, when external, non-content-related acquiescence criteria are correlated, the correlations are far lower.

Theoretical discussions aside, both biases have a considerable impact on such aspects of a questionnaire as the factor structure or the participants' scores, as will be shown in the next chapter. Therefore, our opinion is that some sort of procedure needs to be applied to minimize or control AC and SD effects when a new inventory is designed.

1.1 Impact of response bias

Most research into response bias has focused on the impact of SD and AC on the validity of self-reports, and particularly on the effects of SD on employment selection processes (Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones, Viwesvaran, & Reiss, 1996; Salgado, 2005). These authors suggest that SD can distort participants' scores, and conclude that no single method can solve the problem entirely.

On the other hand, other researchers have focused on how self-reports can impact the factor structure and how SD and AC can distort the inter-item correlation matrix (Bentler, Jackson, & Messick, 1971; Rammstedt, Goldberg, & Borg, 2010; Soto, John, Gosling, & Potter, 2008). However, the impact depends on which bias is being measured: in the presence of AC, items worded in the same direction tend to show a positive relationship that cannot be attributed to the content measured, while items worded in different directions will tend to show a negative relationship. So, AC can potentially lead to an overestimation or underestimation of inter-item correlations depending on the direction of the items, which will obviously affect the factor structure of the measure.

In the case of SD, the situation is similar but there is one important difference: the set of items most affected by SD will be positively correlated but not because of the content measured. So, SD will also impact the correlation matrix by increasing the inter-item correlations, while AC can generate a stronger distortion because its effects work in both ways.

Some studies have analyzed the effects of AC on the factor structure of self-reports. For example, in samples with high levels of AC, the goodness-of-fit indices suggest that the fit of the five-factor model of personality is not always optimal. This has been investigated in an elderly population (Ross & Mirowsky, 1984; Vigil-Colet, Lorenzo-Seva, & Morales-Vives, 2015), samples with little education and in samples of adolescents or pre-adolescents (Morales-Vives, Lorenzo-Seva, & Vigil-Colet, 2017; Meisenberg & Williams, 2008; Rammstedt et al., 2010; Rammstedt & Kemper, 2011; Ross & Mirowsky, 1984; Soto et al., 2008). These studies suggest that AC levels vary throughout the lifespan, and seem to be higher in children, adolescents and the elderly. Therefore, the distortions produced by AC are greater in these populations, which often leads to compromised factor structures unless some sort of procedure is applied to control for the effects of AC.

1.2 Existing methods for controlling the impact of Social Desirability and Acquiescence

We should point out that, until the method that we will use in our software was developed, no model-based procedures had been proposed for simultaneously assessing the potential impact of SD and AC on responses to a questionnaire. For this reason, the methods described in this chapter are presented for the bias they are designed to deal with.

1.2.1 Methods for controlling Social Desirability

We should first point out that the impact of SD can be reduced by ensuring that items are appropriately designed. For example, neutrally worded item statements or an emphasis on confidentiality and anonymity during the test administration can reduce SD (Edwards, 1957), but cannot guarantee that all the variance attributable to SD is removed. In addition, researchers can administer a couple of items designed for detecting faking and SD, and use the scores on these items to exclude participants with high levels of distortion. However, the methods presented in this section aim to control distortions due to SD once they are present in a particular dataset, and researchers generally prefer not to eliminate participants from the sample. In addition, for some traits, it can be very difficult to formulate neutral items (for example, items for measuring aggression).

It should be noted that there are very few methods available that can analytically remove the distortions caused by SD. One option is to partial SD out of questionnaire scores. As McCrae and Costa (1983) pointed out, the main limitation of this approach is the real concern that some of the genuine content variance may be lost in the process. Furthermore, the procedure assumes that all items are parallel measures of the trait of interest, which is almost never true (Leite & Cooper, 2010).

Neill & Jackson (1970) and Paulhus (1981) proposed FA-based models that can remove the variance specifically attributable to SD from the data, and preserve the content variance.

More precisely, Paulhus (1981) proposed that in principal-factor analysis, the first factor to emerge in the unrotated solution usually represents SD, and he suggested that this factor, which essentially measures SD, should be suppressed. This method has several limitations: the assumption that the first factor to emerge represents SD is questionable, and the method can only be used with principal-components analysis. Furthermore, if the SD factor is eliminated from the analysis, SD factor scores cannot be provided.

Ziegler and Buehner (2009) proposed a structural equation model that models SD using a hybrid of within- (and between-) subject analysis. The most interesting feature of this approach is that it does not require a specific subset of items related to SD. However, it has some design requirements: the questionnaire is administered to two groups twice. A control group is asked to respond honestly both times and, on the second administration, an experimental group is specifically instructed to fake. People will fake the items that they believe are important and, as a result, the correlations between these items increases and influences the loading pattern for the common method factor. If there is a change in the common method factor on the second administration, it must be due to the specific instructions to fake, since this was the only difference between the groups. However, the method has an important limitation: the experimental design is complex, since it requires two equivalent groups and two administrations. There also seem to be some issues with the model fit, because some unexplained variance meant that the fit indices in the original test were questionable.

Burns & Christiansen (2011) used a three-factor classification system to classify the existing methods for measuring faking into nine categories. The first factor is the number of assessments required (that's to say, a single or a repeated measurement). The second one is whether the faking measure is computed from the responses to the content measures or whether additional data is required (internal or external). The final factor is based on the effect of faking on the measure, and whether this effect is portrayed as a mean shift or as an impact on the construct validity. The authors conclude that faking is a multidimensional phenomenon, and they suggest using multiple measures of faking aligned with the measures of each study since all faking measures have limitations. Furthermore, there are no measures of both effects (mean shift and changes in construct validity). The study, however, does not provide greater insight into any analytical methods; rather it focuses on overall approaches to faking. Particularly Factor Analysis (FA) based approaches are not studied, which are the focus of this thesis. The classification system is shown in Table 1, an adaptation of a table from Burns and Christiansen (2011).

Table 1. Classification of faking measures by Burns and Christiansen

Indicator	Repeated/Single measurement	Internal/External	Faking Effect
Response validity scales	Single	External	Mean shift
Bogus items and overclaiming	Single	External	Mean shift
Idiosyncratic item response	Single	Internal	Mean shift
Covariance Index	Single	Internal	Construct validity
Raw difference scores	Repeated	Internal	Mean shift
Regression adjusted different scores	Repeated	Internal	Mean shift
Percent agreement	Repeated	Internal	Construct validity
Within-person correlation	Repeated	Internal	Construct validity
Within-subject variances of differences	Repeated	Internal	Construct validity

Burns and Christiansen's classification (2011) cannot assign some methods to any existing categories. For example, the FA-based approaches that require the administration of external SD items should be regarded as "bogus items and overclaiming" methods, since it is an external measure, requiring a single administration and the mean shift being the main effect. In our opinion, this classification is questionable, since FA methods have little in common with the methods proposed in this category, suggesting that the classification proposed by the authors is not exhaustive.

Finally, Larson (2018) recently reviewed some of the options for identifying and mitigating SD impact in applied fields, and he concluded that not controlling SD could lead to inaccurate results. Larson suggests that an easy way to account for the variance due to SD is to administer an SD questionnaire and make a regression analysis of the SD scores on content scores. If this effect is significant then, it should be included in the final model, since SD seems to have a significant impact on the model's total explained variance.

1.2.2 Methods for controlling Acquiescence

Acquiescence is measured using two types of approach: (1) separate acquiescence scales, and (2) balanced scales. Below we review both approaches.

Separate acquiescence scales were developed on the assumption that acquiescence is a generalizable trait and essentially depends on the respondent. They are based on heterogeneous item pools, and are usually scored in the direction of agreement regardless of item content. The items on these scales are chosen to be neutral in terms of SD and to be of medium difficulty. A participant's tendency to agree across a heterogeneous and neutral pool can only be due to acquiescence. Acquiescence scales have been proposed by Bass (1956), Couch and Keniston (1960), and Hanley (1961), among others. However, these scales are unsuitable for clinical measurement because of their low reliabilities (Paulhus, 1991).

On the other hand, balanced scales were developed from the assumption that acquiescence essentially depends on the instrument that is used (Martin, 1964; Messick, 1967). There are basically two types of reversed items: items with an added negation, and items using an antonymic expression, which measures the same trait as the regular items, but in the opposite direction. Guidelines for item development generally recommend avoiding negative formulations (Haladyna & Rodríguez, 2013; Moreno, Martínez & Muñiz, 2015) because negatively worded items tend to have poor measuring properties (Barnette, 2000).

The use of reversed items is controversial, since some authors suggest that they do not reduce response bias, and can even reduce the model fit of some unidimensional instruments (Sonderen, Sanderman & Coyne, 2013; Suárez-Alvarez, Pedrosa, Lozano, García-Cueto, Cuesta & Muñiz, 2018; Woods, 2006). It should be pointed out that using reversed items to obtain a balanced scale without applying a control method does not guarantee that the impact of AC will be completely removed, since it assumes that all the items have the same AC loading. This is almost impossible, so the effect of AC may not be counterbalanced. As a result, for controlling variance to AC responding, a specific analytical method must be used to analyze the set of balanced items.

Some examples of these methods are ipsatizing (substracting the mean score of each individual from all the scores of that individual; Cattell, 1944), and partialling the mean component from the variables by linear regression, which removes the effects of AC from the inter-item correlation matrix (ten Berge, 1999). The main concern when ipsatizing is essentially removing the mean component (Clemans, 1966), which reduces the variance attributable to a general component. On the other hand, as in partialling SD variance, there is some concern that the content variance might be removed when the mean component is partialled.

Some FA-based procedures have also been proposed (Billiet & McClendon, 2000; Ferrando & Condon, 2006; Ferrando, Lorenzo-Seva & Chico, 2003; Lorenzo-Seva & Rodríguez-Fornells, 2006; Mirowsky & Ross, 1991; ten Berge, 1999). All of these methods assume that the items in the questionnaire are fully balanced: half of the items measure in one direction of the trait and the other half measure in the opposite direction. However, this condition is not always easy to achieve, especially in large questionnaires or when measuring a specific trait that can be difficult to assess in the reverse direction, so this is a potential limitation. The first method to assess acquiescence in partially balanced scales was the one proposed by Lorenzo-Seva & Ferrando (2009), which was the main inspiration for the acquiescence control method included in our final proposal.

It is worth mentioning the simulation study performed by Savalei (2014), which compared three methods for assessing AC. The three methods described were: (1) The CFA Method (Mirowsky & Ross, 1991; Maydeu-Olivares & Coffman, 2006), which proposes a model with all the AC loadings set to 1. (2) Chan & Bentler's (1993) approach (CB), which was developed for fitting factor models to ipsative data. (3) The EFA method (Ferrando, Lorenzo-Seva, & Chico, 2003), a precursor to the method proposed by Lorenzo-Seva & Ferrando (2009), in which two orthogonal factors are extracted and rotated to a partially specified target, and the main factor loadings are assumed to sum to zero. The results showed that CB and EFA methods performed badly when the substantive loadings were not balanced, while CFA performed quite well across all conditions.

The decision of the authors not to include the most recent proposal by Lorenzo-Seva & Ferrando (2009) is questionable, since it uses non-balanced scales, and would probably perform better than the methods selected by the authors under the condition of unbalanced loadings.

1.3 Our proposal for controlling Social Desirability and Acquiescence in a single analysis procedure

Of all the existing methods, the most complete FA-based method is the one proposed by Ferrando, Lorenzo-Seva, & Chico (2009), which allows SD and AC effects to be assessed and controlled simultaneously. In this method, the authors gather some of the proposed methods for controlling SD and AC and they implement them in a three-step unrestricted FA approach. However, it also allows for the possibility of applying the control procedure to only one of the response biases, if the methodological aspects of the questionnaire prevent both from being controlled or if the researcher is only interested in controlling one of them.

The method makes two assumptions: First, AC and SD are assumed to be independent from the content factor and also from each other. A priori, there is no reason why response biases should be related to most of the substantive traits (Billiet & McClendon, 2000; Edwards, 1967; Nunnally, 1978; Ray, 1979). However, some traits could be related to SD or AC (for example, agreeableness, sociability or external locus of control) (Bramble & Wiley, 1974; Krosnich & Fabrigar, 1998). When found, these relations seem to be quite weak (McCrae & Costa, 1983; Ones, Viswesvaran & Reiss, 1996). Regarding the relation between SD and AC, they are essentially uncorrelated (Greenwald & Clausen, 1970; Jackson & Messick, 1962; Stricker, 1961).

The second assumption is that acquiescence is not present in the items that will be used as pure measures of social desirability. Acquiescence tends to manifest in items that are neutral in SD and decrease as the SD level elicited by the item increases (Edwards, 1967; Jackson & Messick, 1962; Stricker, 1961). Therefore, it seems reasonable to assume that the effect of acquiescence on pure measures of SD is negligible.

The procedure does have some specific requirements. In order to control for SD, and in addition to the content items, the procedure needs at least four SD markers to be administered, which should be pure or almost pure measures of SD. Therefore, the questionnaire must consist of (a) a few SD markers, and (b) the items related to the content that the test aims to assess. The procedure can assess multiple independent or correlated latent variables.

The authors regard SD as impression management, which may be intentional or not. Some examples of good SD markers could be the following: "I always keep my word" for a positively worded item or "I have sometimes taken something that was not mine", for a negatively worded one. In these items, participants with high levels of SD responding will try to give a good impression of themselves by agreeing with the first statement and disagreeing with the second one.

A visual approach to the two assumptions is presented in Figure 1.

In order to control AC, the procedure assumes that the set of items are at least partially balanced, so a few items measure the latent variable in opposite directions. So, two subsets of items must be identified: (a) a balanced subset, in which the same items are both positively and negatively worded, and (b) an unbalanced set, containing all the remaining items that are worded in the same direction. Therefore, the procedure will identify acquiescence as a common style factor using the balanced set of items and then it will estimate the AC loadings of the unbalanced ones.

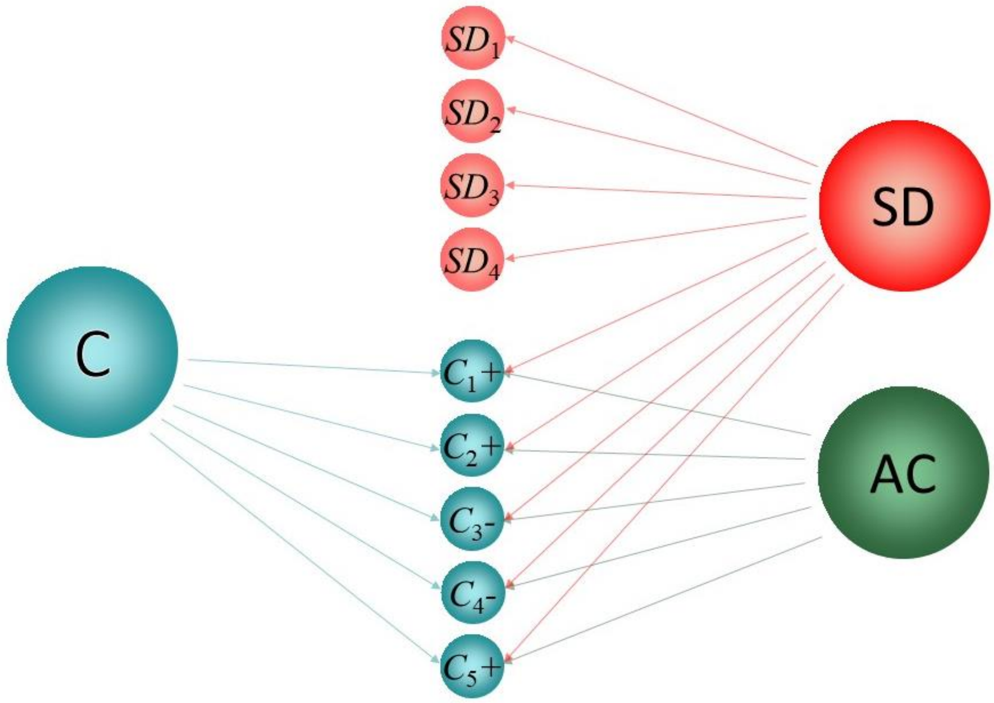

Figure 1. Visual representation of the relationship between items and factors

The summary of the procedure is presented in Figure 2.

As mentioned above, the procedure that we propose is a three-step procedure: (1) Control of Social Desirability, (2) Control of Acquiescence, and (3) Factor Analysis for the content items. In order to help the reader to understand our approach, we shall first present each step in a separate subsection of this document. As this first presentation is largely an intuitive explanation of the three steps, a final subsection explains the method from a mathematical point of view.

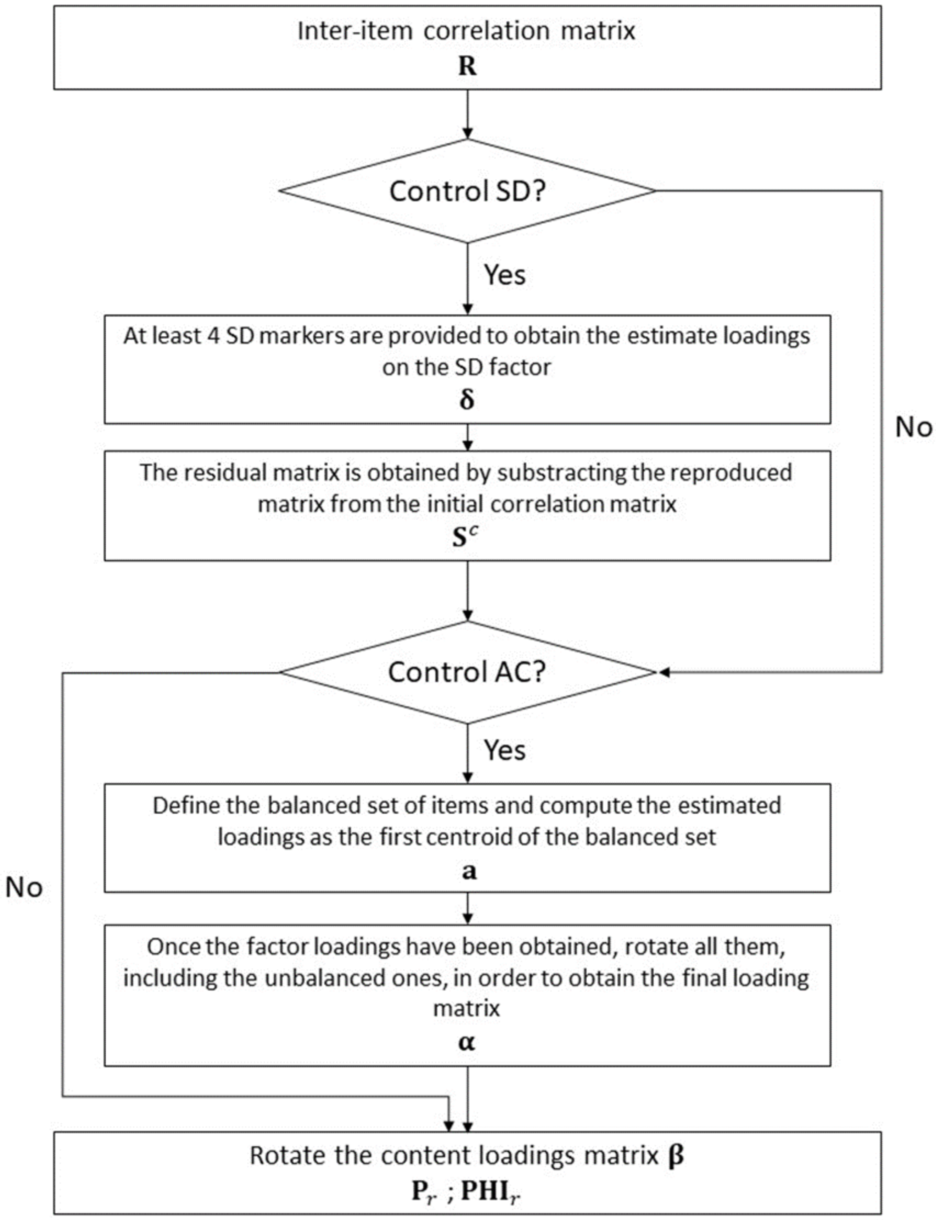

Figure 2. Summary of the response bias control procedure

1.3.1 Controlling Social Desirability

The first step is to estimate the SD factor, which is performed on the sole basis of the SD markers. This is the reason why a minimum number of SD items are required. In our studies we have used a minimum number of four items and obtained acceptable results. The estimation will be made using the *Instrumental Variables* technique (Hägglund, 1982), which will be detailed in the mathematical implementation section. It must be noted that this factor will be a good SD representation only if the selected markers are good and pure measures of the trait. Then, the method estimates the SD loadings of the content items from the cross-correlations between the SD markers and the content items.

Once the SD factor is available, a first reproduced correlation matrix is computed, which accounts for the variance due to SD. Finally, a reduced correlation matrix is obtained by subtracting the reproduced matrix from the initial correlation matrix.

This first residual matrix is expected to be free of SD variance and is the input matrix for the AC control procedure.

1.3.2 Controlling Acquiescence

The second step aims to control the acquiescence variance. Detecting Acquiescence may seem to be trivial, as it is essentially a systematic tendency of responders to agree that could easily be identified with a simple inspection of each person's responses. However, it is not so easy to identify the impact of AC. To illustrate this issue, figure 3 contains the scores of 13 participants on 4 hypothetical items, of which 1 and 4 are direct items (i.e., they measure in the positive direction of the trait), and 2 and 3 are reversed items (i.e., they measure in the opposite direction of the trait).

Participant	Item 1	Item 2	Item 3	Item 4
1	5	2	1	4
2	2	4	5	3
3	3	2	1	5
4	3	2	3	5
5	4	2	2	4
6	2	5	4	2
7	5	1	2	4
8	4	2	1	4
9	1	4	5	1
10	4	2	1	5
11	4	2	1	4
12	1	3	4	3
13	4	5	4	4

Figure 3. Distribution of the example item scores

As can be seen in Figure 3, a simple inspection of the raw scores does not help to distinguish which participants (or items) are most impacted by Acquiescence. An accurate inspection is needed to conclude that the responses of participant number 13 are the ones that are most impacted by AC. With a large set of items (not just four items as in our trivial example), it would not that evident when a responder has produced a pattern of acquiescent responses. As can be expected, distinguishing an AC response tendency is harder in real situations, in which there might be hundreds of participants and dozens of items. A trivial solution to detecting AC could be to compute the mean score of each item and then the deviation of each individual score from the mean of each item ($X_{ij} - \bar{X}_j$) (see Figure 4).

Participant	Item 1	Item 2	Item 3	Item 4	Mean
1	1.8	-0.8	-1.6	0.3	-0.1
2	-1.2	1.2	2.4	-0.7	0.4
3	-0.2	-0.8	-1.6	1.3	-0.3
4	-0.2	-0.8	0.4	1.3	0.2
5	0.8	-0.8	-0.6	0.3	-0.1
6	-1.2	2.2	1.4	-1.7	0.2
7	1.8	-1.8	-0.6	0.3	-0.1
8	0.8	-0.8	-1.6	0.3	-0.3
9	-2.2	1.2	2.4	-2.7	-0.3
10	0.8	-0.8	-1.6	1.3	-0.1
11	0.8	-0.8	-1.6	0.3	-0.3
12	-2.2	0.2	1.4	-0.7	-0.3
13	0.8	2.2	1.4	0.3	1.2

Figure 4. The difference between each score and the item mean, and the overall mean of each participant

In Figure 4 it is easier to see which participants deviate most from the item means, which is a simplistic way of obtaining an Acquiescence impact index. For example, participant number 13 presents a consistent positive deviation from all the item means, which could be due to the impact of Acquiescence. In addition, researchers should be interested in assessing which items are affected by acquiescent responding. This information could be of interest when developing a new test: items largely affected by acquiescent responses could turn out to be complex items, or items that the participants do not understand properly (perhaps because the vocabulary is too complex, or the wording is artificial). These items could be dropped from the test, and new items free of acquiescent responses proposed.

However, visual inspection of participants' raw responses will not always be sufficient to detect these items even if they are present in the set of items.

While the approximation illustrated in figure 4 is too simplistic for psychometric standards, the rationale behind our detection method is similar. A detailed mathematical explanation of the method can be found in subsection 1.5 (pages 33 to 38 of this document).

In order to compute our method, researchers must first identify which items define a balanced set of items: a set of items in which half of the items are worded in one direction, and the other half are worded in the opposite direction. Ideally, the items composing the balanced core should be the ones most impacted by AC, since this balanced core will define the AC factor. As proposed by ten Berge (1999), this factor is obtained on the basis of the first centroid of the residual inter-correlations between the balanced core of items. In other words, the acquiescence factor can be understood as a general factor that can identify a tendency of general agreement in each item and obtain a positive saturation for each one.

Once the centroid has been obtained, the unbalanced subset of items (i.e., the items that were not included in the set of balanced items) has to be projected on the centroid to obtain the loadings for all the items.

It should be noted that the centroid is not computed directly from the inter-items correlation matrix, but from the residual variance/covariance matrix previously obtained in Step 1 (explained in sub section 1.3.1, page 22 of this document). Finally, in order to obtain a variance/covariance matrix free of response biases effects, the reproduced correlation matrix that accounts for AC variance should be subtracted from the residual matrix obtained in step 1 (i.e., the one obtained in the SD control procedure) to obtain a second residual matrix variance/covariance matrix, which is expected to be free of both biases.

1.3.3. Obtaining content factors

The third step retains as many content factors as expected by performing an EFA with the residual variance/covariance matrix obtained after the variance of both biases has been removed. There are no restrictions on the methods for factor extraction and rotation of the content factors. In the Psychological Test Toolbox, we decided to implement MRFA as the method for factor extraction, since it computes the explained common variance accounted for each factor.

1.4 Practical applications of the method

The method has been considered of interest in applied research and it has already been successfully applied to develop several instruments that measure a wide variety of traits (Cupani & Lorenzo-Seva, 2016; Mas-Herrero, Marco-Pallares, Lorenzo-Seva, Zatorre, & Rodríguez-Fornells, 2013; Morales-Vives, Camps, & Lorenzo-Seva, 2013; Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, & Vigil-Colet, 2014; Saliba, Lorenzo-Seva, Marco-Pallares, Tillmann, Zeitouni, & Lehmann, 2016; Vigil-Colet, Morales-Vives, Camps, Tous, & Lorenzo-Seva, 2013).

The same procedure is also being used to develop other instruments such as the INventory of Callous-unemotional traits and Antisocial behavior (INCA, Morales-Vives, Cosi, Lorenzo-Seva, & Vigil-Colet, in revision) and the MAturity in Youth Assessment Scale (MAYAS, manuscript in preparation).

1.5 Technical implementation

Let's start with the first step. Consider a questionnaire composed of m content items, which was administered to n individuals. The m items are a set of items expected to be related to r content factors ($r < m$). The questionnaire is only partially balanced: a subset of p items is worded in one direction of the trait, and a subset of q items is worded in the opposite direction, where $p+q=m$. Also, a set of h SD items are administered together with the m content items. The $\mathbf{X}$ matrix containing scores of the n individuals (i.e., the responses of all the individuals to the test) can be partitioned as

$$\mathbf{X} = [\mathbf{X}^{sd}|\mathbf{X}^{\mathrm{c}}]$$

where $\mathbf{X}^{\mathrm{sd}}$contains the scores on the SD markers and $\mathbf{X}^{c}$ contains scores on all the m content items. $\mathbf{X}^{\mathrm{c}}$ can be partitioned as

$$\mathbf{X}^{\mathrm{c}} = [\mathbf{X}^{\mathrm{b}}|\mathbf{X}^{\mathrm{u}}]$$

where $\mathbf{X}^{\mathrm{b}}$ contains scores on the k balanced items, and $\mathbf{X}^{\mathrm{u}}$ contains scores on a set of l unbalanced items. The correlation between all the items included in X will be contained in R. Also, $\mathbf{R}^{\mathrm{c}}$ contains the correlation between $\mathbf{X}^{\mathrm{c}}$ items and $\mathbf{R}^{\mathrm{sd}}$contains the correlation between $\mathbf{X}^{\mathrm{sd}}$ items.

The structural model underlying the method assumes that each content item is a factorially complex measure, determined by: (a) the SD factor θ^{sd}, (b) an AC factor θ^{a}, and (c) the r content factors θ^{c}.

$$X_{ij} = \delta_j\theta_i^{sd} + \alpha_j\theta_i^{a} + \beta_{j1}\theta_{i1}^{c} + \beta_{j2}\theta_{i2}^{c} + \cdots + \beta_{jr}\theta_{ir}^{c} + \varepsilon_{ij}$$

For $i = 1 \dots n$ and j= $1 \dots m$, where δ is the SD factor loading, α is the AC factor loading, βs are the factor loadings for r content factors and εs are the residuals, with zero means and uncorrelated with the factors or one another. As mentioned above, the r factors are assumed to be uncorrelated with the response bias factors. Also, the SD factor and AC factor are expected to be uncorrelated with each other. However, the r content factors can be correlated with each other.

To simplify the model, let us suppose that all the content items only measure one latent trait θ^c, thus leading to a model such as

$$X_{ij} = \delta_j \theta_i^{sd} + \alpha_j \theta_i^a + \beta_j \theta_i^c + \varepsilon_{ij}$$

As presented above, an additional h set of SD markers was administered together with the content items. Their function is to provide factorial measures of SD, and the structural model for these items reduces to

$$X_{ih} = \delta_h \theta_h^{sd} + \varepsilon_{ih}$$

The h SD markers make it possible to estimate the loading of the content items on the SD factor using the Instrumental Variables (IV) technique, which was developed in the context of factor analysis by Hägglund (1982). First, one of the SD markers is taken as a proxy for θ^{sd} and the remaining $h-1$ markers are taken as instrumental variables. Without loss of generality we can take the first marker as a proxy. From correlation matrix $\mathbf{R}$, two vectors $\mathbf{r_h}$ and $\mathbf{r_j}$ can be defined. $\mathbf{r_h}$ is a column vector of order $((h-1) \times 1)$ that contains the covariance between the proxy and the other $h-1$ markers. $\mathbf{r_j}$ is a column vector of order $((h-1) \times 1)$ that contains the covariance between the content item j and the other $h-1$ markers. Then the loading of the m content items on the SD factor can be estimated as

$$\hat{\delta}_j = \mathbf{r}_\mathbf{j}'\mathbf{r}_\mathbf{h}(\mathbf{r}_\mathbf{h}'\mathbf{r}_\mathbf{h})^{-1}\delta_1$$

where $\hat{\delta}_j$ is the loading of the content item j, and the δ_1 is the factor loading of the proxy variable. The value of δ_1 can be computed from the correlation matrix of the h SD markers, or directly defined from a previous study.

This is how the estimate loadings of the SD factor for the m content items can be obtained. The loadings for the $h-1$ SD markers are estimated in the same way, and the loading for the first marker (proxy) can be estimated simply by choosing another pivot variable. Once the complete vector of m loading estimates $\boldsymbol{\delta}$ has been obtained, the reproduced correlation matrix is computed as $\boldsymbol{\delta}\boldsymbol{\delta}'$.

The first residual matrix $\mathbf{S}^\mathbf{c}$, which is free of SD impact, is obtained by subtracting the reproduced matrix from the initial correlation matrix between the content items $\mathbf{R}^\mathbf{c}$, defined as

$$\boldsymbol{S}^c = \boldsymbol{R}^c - \boldsymbol{\delta}\boldsymbol{\delta}'$$

Let us now go on to the second step. For fully balanced scales, the method is the following.

The first residual matrix obtained after the SD variance has been subtracted is used as the input in the second stage for estimating the loadings on the AC factor. As the influence of the SD factor has been partialled out, the structural model looks like this:

$$X_{ij} = \alpha_j \theta_i^a + \beta_j \theta_i^c + \varepsilon_{ij}$$

If $\mathbf{S^c}$ is the first residual matrix obtained after SD variance of the order $m \times m$ has been subtracted, $\mathbf{S^b}$ is the residual matrix between a set of balanced items. Then

$$\mathbf{a} = \mathbf{S^b 1}(\mathbf{1}'\boldsymbol{S^b}\mathbf{1})^{-1/2}$$

where **a** is the vector of correlations between the variables and their mean. Values of **a** show how much each variable is impacted by AC. A factor loading matrix $\mathbf{B_b}$of the order of $m \times (r+1)$ can be obtained by

$$\mathbf{S}^{\mathrm{b}} = \mathbf{B}_{\mathrm{b}}\mathbf{B}_{\mathrm{b}}' + \mathbf{M}_{\mathrm{b}}\mathbf{M}_{\mathrm{b}}' + \mathbf{\Psi}_{\mathrm{b}}^2$$

where $\mathbf{M_b}$ holds the loadings on those common factors that are discarded in the rank $(r+1)$ solution and $\mathbf{\Psi_b}$ is a diagonal matrix containing the unique factor standard deviations. Let the rotation matrix $\mathbf{W}$ be an orthonormal matrix of order $(r+1) \times (r+1)$. $\mathbf{W}$ must maximize the congruence between one column of the product $\mathbf{M_b W}$ and vector $\mathbf{a}$, so it is determined by the Korth and Tucker method (1976). Let $\mathbf{d}$ and $\mathbf{w}$ be vectors defined as

$$\mathbf{d} = (\boldsymbol{B}_b'\boldsymbol{B}_b)^{-1}\boldsymbol{B}_b'\boldsymbol{a}$$

and

$$\mathbf{w} = \mathbf{d}(\mathbf{d}'\mathbf{d})^{-1/2}$$

Given the eigendecomposition of the matrix

$$\mathbf{I} - \mathbf{ww}' = \mathbf{W\Delta W}'$$

where $\mathbf{I}$ is an identity matrix and Δ is a diagonal matrix with elements $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_{r+1} = 0$, the product

$$\mathbf{B}_b\mathbf{W} = [\boldsymbol{\beta}|\boldsymbol{\alpha}]$$

leads to a matrix whose last column $\boldsymbol{\alpha}$ contains the loading values of balanced items on the acquiescence factor, and $\boldsymbol{\beta}$ is a $k \times r$ matrix that can be rotated to show factor simplicity by any orthogonal or oblique rotation method. Note that $\boldsymbol{\beta}$ is a factor loading matrix that is free of variance caused by AC responding.

However, if the scale used is only partially balanced, some changes have to be made.

A factor loading matrix $\mathbf{L}$ of the order $m \times (r+1)$ can be obtained by

$$\boldsymbol{S}^c = \mathbf{LL}' + \mathbf{MM}' + \boldsymbol{\Psi}^2$$

where $\mathbf{S}^c$ is the covariance matrix obtained after SD variance has been subtracted, $\mathbf{M}$ holds the loadings on the common factors that are discarded in the rank$-(\mathbf{r}+\mathbf{1})$ solution and $\boldsymbol{\Psi}$ is a diagonal matrix containing the unique factor standard deviations. $\mathbf{L}$ can be portioned as

$$\mathbf{L} = [\frac{\mathbf{L}_b}{\mathbf{L}_u}]$$

where $\mathbf{L}_b$ contains the loading values related to the even set of balanced items and $\mathbf{L}_u$ contains the loading values related to the set of unbalanced items. Let the rotation matrix $\mathbf{U}$ be an orthonormal matrix of order $(r+1) \times (r+1)$. $\mathbf{U}$ must maximize the congruence between one column of the product $\mathbf{L}_b\mathbf{U}$ and a vector $\boldsymbol{a}$, so it is determined by the Korth and Tucker method (1976). Finally, $\mathbf{U}$ is used to rotate not only $\mathbf{L}_b$ but also the overall matrix $\mathbf{L}$, so that the product

$$\mathbf{LU} = [\boldsymbol{\beta}|\boldsymbol{\alpha}]$$

leads to a matrix whose last column $\boldsymbol{\alpha}$ contains the loading values of balanced and unbalanced items on the acquiescence factor and $\boldsymbol{\beta}$ is a $m \times r$ matrix that can be rotated to show factor simplicity by any orthogonal or oblique rotation method.

Finally, the third step is to factorize the content items using the final reproduced matrix once the impact of both biases have been subtracted. We decided to include Minimum Rank Factor Analysis (MRFA) in the Psychological Test Toolbox since it computes the explained common variance attributable to each factor. However, any other extraction method could be used instead.

1.6 Illustrative example

The aim of the analyses presented in this section is to provide a short, illustrative data example of how our method can be used in applied research to control SD and AC. A much more elaborate study can be found in Navarro-González, Vigil-Colet, Ferrando, and Lorenzo-Seva (in press), which is included in this thesis (see pages 68 and 69).

In this example, we are going to use the extraversion scale of the OPERAS questionnaire (Vigil-Colet, Morales-Vives, Camps, Tous, & Lorenzo-Seva, 2013), which contains seven items and the four SD markers. The questionnaire was administered to a sample of 500 responders (54.5% women) with ages ranging from 18 to 51.

The extraversion scale is partially balanced, with four positively worded items (labeled “+”) and three negatively worded items (labeled “-”).

We computed two exploratory factor analyses on the polychoric correlation matrix: (1) controlling both biases by applying the procedure described above, and (2) a regular exploratory factor analysis that retains only one content factor using the correlation between the seven extraversion items. In this way we compared the effects of the correction procedure on the participant’s factor scores. The Kaiser-Meyer-Olkin (KMO) test was .78, which is considered fair. We extracted a single content factor, and obtained a root mean squares of the residuals (RMSR) value of .048. For this dataset, Kelley’s criterion to consider the model as acceptable was an RMSR value of .049. The optimal implementation of parallel analysis (Timmerman & Lorenzo-Seva, 2011) suggested that two factors should be retained, corresponding with the variance attributable to the SD and extraversion factors. We believe that the parallel analysis failed to detect a third factor (which would be related to an AC factor) because AC accounted for very little variance in this data. Even so, we decided to apply our method, and to interpret a three dimensional model: two dimensions related to SD and AC, and a third dimension related to the content factor (i.e., Extraversion factor).

Table 2 shows the factor loadings for the response biases factors and the content factor.

Table 2. Rotated loading matrix

Item	SD	AC	Extraversion
1. SD -	-0.301	-	-
2. SD +	0.724	-	-
3. SD +	0.745	-	-
4. SD +	0.801	-	-
5. EX +	0.247	0.107	0.594
6. EX +	-0.074	0.419	0.651
7. EX -	0.156	0.110	-0.685
8. EX +	0.107	0.291	0.741
9. EX -	-0.107	0.258	-0.501
10. EX -	-0.091	0.226	-0.676
11. EX +	0.103	0.295	0.470

As expected, all the content items have relatively low loadings on the SD factor, since extraversion items are not usually related to SD, except item 5 (*"I am the life and soul of the party*). As far as the AC factor is concerned, the loadings are small–to-medium, suggesting that some items are influenced by AC to some extent. Finally, the loadings on the extraversion factor are all high and in the expected direction.

In regards to participants' item responses and factor scores, the outcomings are coherent with the expected results. For illustrative purposes, the answers to the items provided by three participants are shown in Table 3.

Table 3. Item answers of three participants

	Participant 1	**Participant 2**	**Participant 3**
1. DS -	1	3	2
2. DS +	4	2	4
3. DS +	5	3	3
4. DS +	5	3	3
5. EX +	4	5	2
6. EX +	2	4	3
7. EX -	3	5	5
8. EX+	1	5	2
9. EX -	4	4	4
10. EX -	2	5	4
11. EX +	4	5	3

At first glance, it is not easy to determine which bias, if any, has an impact on the scores of the participants. However, a closer look reveals some response patterns: Participant 1 seems to respond to the SD markers with an extreme response pattern, answering all positively worded SD items with a response of 5, and the negatively worded ones with a response of 1. Participant 2 tends to agree with the extraversion items regardless of the item content and direction, since the item responses range from 4 to 5. Finally, Participant 3 does not show any relevant response pattern in terms of response biases. These suppositions can be verified by inspecting participants' factor scores (see Table 4). Factor scores with no bias correction applied are presented in the first column of the table, and factor scores when controlling for SD and AC biases are presented in the last column. In addition, the two central columns show the scores on the SD and AC dimensions.

Table 4. Participants' scores for each factor, controlling and not controlling for biases

Participant	Standard FA	Controlling response biases		
	Extraversion	SD	AC	Extraversion
1	53.67	68.42	51.24	48.02
2	61.05	47.81	71.07	53.03
3	41.23	52.96	49.83	41.80

As shown, Participant 1 presents high levels of SD because of the answers to the SD items, but his scores on AC and the content factor are medium. However, when the impact of SD and AC is not controlled for, the Extraversion score (53.67) is higher than the one obtained when the variance is removed due to response biases.

Something similar happens with Participant 2. He/she presents high levels of AC due to the general tendency to agree shown in all content items. This tendency is also reflected in the extraversion score. Since the scale is almost balanced, the answers to the positively worded items were counterbalanced with the answers to the negatively worded ones, and a centered score was obtained (53.03). However, if the method is not applied, AC distorts and increases the participant's score on extraversion (61.05).

Finally, Participant 3 does not seem to be impacted by either SD or AC. He/she presented centered scores on both biases. In addition, he/she presented a low score on extraversion. Since there is no distortion in the responses of the participant, the factor score is almost the same in both analyses. This means that the method to control SD and AC does not affect the factor score of participants who did not produce biased responses. As we have already pointed out, in our paper Navarro-González et al. (in press, see pages 68 and 69 of the present document), we discuss a more elaborate study that provides greater insight into the practical consequences of applying our method to control SD and AC in applied research.

1.6 Psychological Test Toolbox

Psychological Test Toolbox is a program designed for performing Exploratory Factor Analysis while applying the aforementioned procedure for controlling SD and/or AC. The program is developed in MATLAB environment, and it can be downloaded in two ways: (a) as a stand-alone application (only for Windows-based computers), which requires the installation of the MATLAB runtime library, available from the Mathworks website; (b) as a MATLAB toolbox, which can be executed by MATLAB users from the code files. The Psychological Test Toolbox's front page is presented in Figure 5.

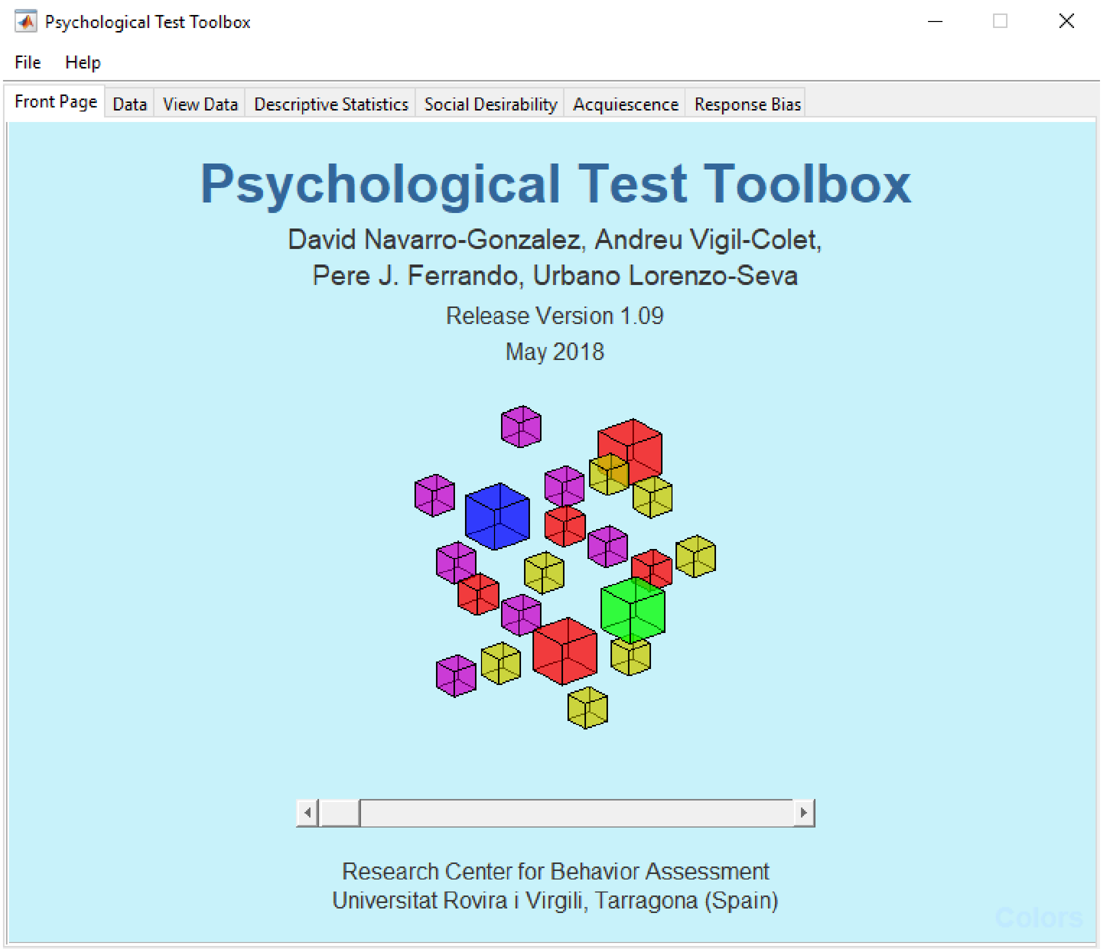

Figure 5. Front page of Psychological Test Toolbox

One of the main objectives of the Psychological Test Toolbox was to create a highly accessible program that does not require any expertise in statistical or programming languages like R or MATLAB. We developed a simple Graphical User Interface (GUI) using the coding tools available in MATLAB. We designed an application that was configurable through seven tabs that were organized in the logical order of the analytical process. A detailed description of the GUI application is given in the paper by Navarro-González (in press, pages 65 and 66 of this document).

The program is free, and it is distributed under General Public License version 3. Since the initial version in 2016, it has been continuously revised, new features have been added and bugs have been regularly fixed.

The program is available at the following website:

https://psico.fcep.urv.cat/utilitats/PsychologicalTestToolbox

On this website, researchers will find extensive documentation about the program and how to use it. As mentioned, the program is constantly being revised, so users are strongly advised to check the website regularly if they wish to stay up to date.

A detailed description of the usage of Psychological Test Toolbox (for example, the inputs required and the output information) can be found on pages 66 to 67 of this document. We have also developed tutorial videos that help researchers use our software. These videos can be found on the software's website.

1.6.1. Procedures and methods included in Psychological Test Toolbox

The program uses the following methods for each part of the analysis:

Univariate descriptive statistics of items:

- Univariate mean, variance, skewness, and kurtosis.
- Variable charts for ordinal variables, representing the distribution of the scores for each item.

Dispersion matrices:

- User defined type matrix.
- Covariance matrix.
- Pearson correlation matrix.
- Polychoric correlation matrix (Polychoric algorithm: Olsson ,1979a, 1979b; Tetrachoric algorithm: Bonett & Price, 2005). If the correlation matrix happens to be non-positive definite, the smoothing algorithm proposed by Devlin and colleagues needs to be computed (Devlin, Gnanadesikan, & Kettenring, 1975; Devlin, Gnanadesikan, & Kettenring, 1981).

As the analysis is exploratory, a procedure for determining the number of factors to be retained is computed:

- Optimal Parallel Analysis (PA). This is an implementation of Parallel Analysis (Horn, 1965) where it is computed based on Pearson or polychoric correlation matrices). We implemented Optimal implementation of PA as proposed by Timmerman and Lorenzo-Seva (2011).

Factor analysis:

- MRFA: Minimum Rank Factor Analysis (ten Berge, 1998; ten Berge, & Kiers, 1991, ten Berge, Snijders, & Zegers, 1981).
- Factor scores for continuous data (ten Berge, Krijnen, Wansbeek, & Shapiro, 1999), and expected a-posteriori (EAP) estimation of latent trait scores for ordinal data.

The rotation methods to obtain simplicity are:

- Varimax (Kaiser, 1958).
- Promin (Lorenzo-Seva, 1999).
- Semi-confirmatory factor analysis based on orthogonal and oblique rotation to a (partially) specified target (Browne, 1972a, 1972b).

Some of the indices used in the analysis are:

- Test on the dispersion matrix: Determinant, Bartlett's test and Kaiser-Meyer-Olkin (KMO).
- Goodness of fit statistics: Goodness of Fit Index (GFI) and Root Mean Square Error of Approximation (RMSEA).
- Simplicity indices: Bentler’s Simplicity index (1977) and Loading Simplicity index (Lorenzo-Seva, 2003).
- Mean, variance and histogram of fitted and standardized residuals. Automatic detection of large standardized residuals.
- Congruence index for assessing the congruence between the rotated loading matrix and the user provided target matrix (Tucker, 1951). The thresholds

proposed by Lorenzo-Seva and ten Berge (2006) are used to interpret the index.

1.7 Objectives and hypothesis

Considering the lack of software that allows users to analyze a dataset and control for the impact of response biases, the first objective was to create a user-friendly application that enables potential researchers to readily develop a questionnaire using the control bias method proposed by Ferrando et al. (2009). Since our intention was get as many people to use it as possible, the program was designed to be easy to use, and distributed under a freeware license.

The second objective was to show that the procedure implemented in the Psychological Test Toolbox was useful for two main purposes related to response biases.

The first of these was to determine whether the procedure can be useful for assessing the impact of response biases on the factor structure of self-assessed personality questionnaires, and provide more evidence about the distortions generated by SD and AC. We hypothesize that the factor structure of the self-assessed questionnaires will be heavily distorted by the impact of response biases, and especially by AC, since it generates a bigger distortion in the inter-correlation matrix than SD. In addition, controlling the impact of both response biases will improve the simplicity and coherence of the factor structure. Among other previous studies, there is a paper by Morales-Vives, Lorenzo-Seva and Vigil-Colet (2017) that analyzed the impact of response biases on the factor structure of a personality inventory. However, they did not distinguish between the impact of each bias, which is a limitation that we hope to overcome.

The second issue concerning the distortion of response biases is the relationship between response biases and the effects associated with the personality differentiation hypothesis across ability levels (PDH).

The PDH assumes that people with higher level of ability have a more differentiated personality structure (Brand, Egan, & Deary, 1994). Various explanations have been put forward to explain the differences in personality postulated by the PDH, one of which is the differential reliability associated with ability levels (DRAAL; Austin, Deary, & Gibson, 1997), which implies that people with a high level of ability present higher levels of reliability. However, the rationale behind the DRAAL itself is not entirely clear. Some authors suggest that DRAAL is caused by the high amount of cognitive processing required to answer items, so low-ability individuals may have trouble understanding some items (Austin et al. 1997, 2000). The same authors also proposed that high ability individuals have a “highly calibrated ruler” that enables them to give more meaningful responses. Other authors proposed that the DRAAL is related to differences in response styles and reflects the effects of SD or AC (Allik, Laidra, Realo, & Pullmann, 2004). The only results in this regard involve how acquiescence is associated with intelligence and low levels of education (Meisenberg & Williams, 2008), and the extent to which acquiescence distorts the factor structures of personality inventories (Rammstedt & Farmer, 2013; Soto et al., 2008). If low-ability individuals have higher levels of acquiescence, the factor structures for those individuals will be different, and this could provide an explanation for the DRAAL.

Taking into account that authors such as Allik et al. (2004) and Austin, Hofer, Deary, & Eber (2000) suggested that response biases are responsible for the differential reliability associated with ability levels (DRAAL), a phenomenon that has implications on the relationship between intelligence and personality differentiation, our objective was to investigate the relationship between response bias and intelligence measures in order to find evidence that supports or rejects this hypothesis. Considering that there is no evidence that SD and intelligence are related (De Fruyt, Aluja, García, Rolland, & Jung, 2006), and the relationship between intelligence and acquiescence is weak (Meisenberg & Williams, 2008), our hypothesis is that response biases will not be clearly responsible for DRAAL.

In our study, we also assessed the individual contribution to the reliability of each instrument using person fit indices (Ferrando, 2007, 2009, 2013) in an attempt to determine the relationship between individual consistency and ability.

2. Method

Each of the objectives was investigated separately and the results are presented in three separate papers. The method of each one is described in detail in the papers themselves, so here I will explain the general design of the three papers.

The first paper is the presentation of the Psychological Test Toolbox and the method for controlling response biases, so the method in this paper is unconventional. It is essentially a coding process for developing the application.

Once the program was ready, it could be used in applied fields, and this was the intention of the remaining two papers: to use the program to analyze data containing response bias effects.

For both papers, the participants were 532 volunteer students (52.6% women), from four different high schools in Tarragona province aged between 11 and 18 years old (M=14.75; SD=2.1).

Several instruments were administered, including:

- Overall Personality Assessment Scale (OPERAS; Vigil-Colet et al., 2013), which is a 40-item Big-Five Inventory.
- Indirect-Direct Aggression Questionnaire (IDAQ; Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, & Vigil-Colet, 2014), which measures Physical, Verbal and Indirect Aggression.
- Barratt Impulsiveness Scale-11 for children (BIS; Chahin, Cosi, Lorenzo-Seva, & Vigil-Colet, 2010; Cosi, Vigil-Colet, Canals, & Lorenzo-Seva, 2008), which is

a self-report questionnaire for assessing impulsivity in children and adolescents.

- Psychological Maturity Assessment Scale (PSYMAS; Morales-Vives, Camps, & Lorenzo-Seva, 2013), measuring three scales: work-orientation, self-reliance and identity.
- The Inventory of Callous Unemotional Traits (ICU; Frick, 2004; López-Romero, Gómez-Fraguela, & Romero, 2015), designed to evaluate the precursors of psychopathy in young populations.
- Thurstone's Primary Mental Abilities Test (Cordero, Seisdedos, González, & de la Cruz, 1989), which evaluates intelligence using five subscales (verbal, spatial, numerical, reasoning and word fluency).
- Raven's Progressive Matrices Test (Raven, 1996), measuring fluid intelligence free of cultural bias.
- The information scale of the WAIS intelligence test for adults (Wechsler, 2003), which is an indicator of crystallized intelligence.

The second paper uses the first two instruments to assess the impact of response biases on the factor structure of both tests. The third one uses all the measures except the OPERAS scores to assess the relationship between response bias and intelligence.

3. Results

3.1. Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias.

In press in *Journal of Statistical Software*. Uncorrected proof version is presented.

Impact index (2017): 22.737 (Statistics & Probability; 1 of 124 journals; Q1)

3.2. How response bias affects the factorial structure of personality self-reports.

Published in *Psicothema*, 28(4), 465-470. Full text is presented. (August, 2016)

Impact index (2016): 1.344 (Multidisciplinary, Psychology; 61 of 129 journals; Q2)

Cites in Scopus: 10.

3.3. Is general intelligence responsible for differences in individual reliability in personality measures? Published in *Personality and Individual Differences*, 130, 1-5. Due to authorship rights, the presented manuscript is the uncorrected proof version.

Impact index (2017): 1.967 (Psychology, Social; 20 of 62 journals; Q2).

3.1. Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias.

[JSS] Entered Post-Processing

jstatsoft admin

dom 31/12/2017 3:06

Para: David Navarro González <david.navarro@urv.cat>;

Cc: Andreu Vigil Colet <andreu.vigil@urv.cat>; Pere Joan Ferrando Piera <perejoan.ferrando@urv.cat>; Urbano Lorenzo Seva <urbano.lorenzo@urv.cat>;

Dear David Navarro-González:

Your submission

"Psychological Test Toolbox: A new tool to compute factor analysis controlling response bias"

has been conditionally accepted for publication in JSS. Your manuscript will now enter in the post-processing stage, which may take up to three weeks to complete. When this stage is complete you will be emailed with all necessary changes.

Thank you for considering JSS and contributing to free statistical software.

Best Regards,

Bettina Grün
Edzer Pebesma
Achim Zeileis

Journal of Statistical Software
http://www.jstatsoft.org/

Journal of Statistical Software

December 2016, Volume -, Issue -. *http://www.jstatsoft.org/*

Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias

David Navarro-Gonzalez
Universitat Rovira i Virgili

Andreu Vigil-Colet
Universitat Rovira i Virgili

Pere Joan Ferrando
Universitat Rovira i Virgili

Urbano Lorenzo-Seva
Universitat Rovira i Virgili

Abstract

The effects of response bias in psychological tests have been investigated for years, the two most common being Social Desirability (SD) and Acquiescence (AC). However, the traditional methods for controlling or eliminating the impact of those biases in participants' scores have several limitations. Some factor analysis-based methods can overcome some of these limitations, such as the procedure proposed by Ferrando, Lorenzo-Seva, and Chico (2009). Nevertheless, this method involves programming skills that are not common among applicate researchers or clinicians. Consequently, we have developed a stand-alone, user-friendly application that an easy way of using the aforementioned method to perform a factor analysis which controls for the effect of AC and SD. The program has been developed in a MATLAB environment and its distribution is entirely free.

Keywords: Response bias, Social Desirability, Acquiescence, Exploratory Factor Analysis, MATLAB.

1. Introduction

The present paper concerns the exploratory factor analysis of psychological tests. In a typical psychological test, the person being tested responds to a number of items by stating how well each item describes him/her. The responses to these kind of self-reports are susceptible to response bias, which is a systematic tendency to answer the items on some other basis than the specific item content (Paulhus 1991). The two best known response biases in questionnaires are Acquiescence (AC) and Social Desirability (SD). Acquiescence can be defined as the tendency of responders to agree with an statement regardless of its content (Paulhus and Vazire

2005), while social desirability can be defined as people's tendency to present themselves in a generally favorable fashion (Holden 2010).

A review of the literature on response biases indicates that AC and SD can impact both the individual scores and the factorial structure of typical response measures such as personality traits (Danner, Aichholzer, and Rammstedt 2015; Navarro-Gonzalez, Lorenzo-Seva, and Vigil-Colet 2016; Rammstedt and Kemper 2011; Rammstedt and Farmer 2013; Soto, John, Gosling, and Potter 2008) or aggression measures (Becker 2007; Harris 1997; Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, and Lorenzo-Seva 2012), among others. The distortions generated by these response biases can have a negative impact on personnel selection, individual assessment, research studies and clinical evaluation (Goffin and Christiansen 2003; Salgado 2005; Viswesvaran, Ones, and Hough 2001).

Given these findings, test developers often use some type of procedure to control or minimize the effect of AC and SD when designing questionnaires. However, most of these procedures are purely descriptive and have some shortcomings due to the ad hoc approaches inherent in those methods. In recent years, more sophisticated model-based procedures have been proposed.

Regarding AC, several authors have proposed procedures based on a factor analysis (FA) model that uses fully balanced sets of items, where half of the items measure in one direction of the trait and the other half measure in the opposite direction. Some procedures are based on restricted factor analysis (Billiet and McClendon 2000; Mirowsky and Ross 1991) and others are based on unrestricted factor analysis Ferrando, Lorenzo-Seva, and Chico 2003; Lorenzo-Seva and Rodríguez-Fornells 2006; Ten Berge 1999). However, in applied research it is usual to find scales that are only partially balanced, which makes difficult to apply the aforementioned procedures. To overcome this limitation, Lorenzo-Seva and Ferrando (2009) proposed a new method for controlling AC in partially balanced multidimensional scales.

In the case of SD, there have traditionally there are been two different approaches for dealing with its bias effects. Both approaches are based on administering an SD scale together with the content scales. The first method consists of using the SD scale to remove individuals with high scores in SD. This procedure has some limitations. First, removing participants with high scores in SD does not guarantee that the scores of the other participants are free of SD. Second, if the content that is being measured is related to SD, then individuals with high content scores may also be removed.

The second traditional method is known as "partialling", which is based on using the SD scale to partial out the SD effects on the content scale by regressing the SD scores onto the trait scales of interest and computing a residual score. This approach also has some limitations. First, it may remove meaningful variance from the relevant trait. Second, the procedure assumes that all items are parallel measures of the trait of interest, which is almost never true.

These limitations may be overcome by using methods based upon factor analysis. Some FA-based procedures for identifying an SD factor are those proposed by Paulhus (1981) or by Neill and Jackson (1970). In particular, in Neill and Jackson (1970) procedure SD is identified by using an SD scale as a marker variable. Ferrando (2005) developed a restricted FA model by expanding Jackson's idea to assess simultaneously the effects of AC and SD, thus allowing these biases to be modelled as additional factors that can be distinguished from the content factors (Ferrando *et al.* 2009). This procedure presents certain advantages. First, it removes the effect of both response biases from the factor structure, allowing the item structure to be analyzed once the distortion generated by SD and AC is removed. Second, it provides the estimated factor scores of the participants, which is a more precise and unbiased estimation of how the individuals stand with regard to the trait that is to be measured, and this can be very useful in individual assessment.

The main practical application of the procedure is to be applied in the development of new questionnaires, but can also be applied a posteriori if the analyzed questionnaire meets the characteristics described in Section 2.

The procedure has been considered of interest in applied research and it has been successfully used to develop different questionnaires (Cupani and Lorenzo-Seva 2016; Mas-Herrero, Marco-Pallares, Lorenzo-Seva, Zatorre, and Rodriguez-Fornells 2013; Morales-Vives, Camps, and Lorenzo-Seva 2013; Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, and Vigil-Colet 2014; Vigil-Colet, Morales-Vives, Camps, Tous, and Lorenzo-Seva 2013). However, only researchers with advanced knowledges of psychometrics and MATLAB programming can perform such an analysis, and this may hinder the wider application of the method.

Taking it into account we have developed a stand-alone application called **psychological test toolbox**, which is a user-friendly application that enables the implementation of the procedure described in Ferrando *et al.* 2009. It provides an easy way of performing factor analysis by controlling the effect of AC and SD, thus providing bias-free individual response scores . The program has been developed in a MATLAB environment (software propriety of MathWorks ®).

R is a popular software among researchers because it is freeware. We have implemented our software in MATLAB following the same philosophy as in R: any user can download and use it as a freeware software. We decided to use MATLAB because it is simpler to produce user-friendly software.

2. Characteristics of psychological tests in order to control SD and AC

The procedure proposed by Ferrando *et al.* (2009) cannot be applied in any typical response measure. In order to control SD, a number of items related to SD must be included in the psychological test. These items are known as SD markers. The greater the number of markers, the better the procedure is expected to work. However, in applied research the procedure has been successfully applied with as few as four SD markers. The psychological test (or questionnaire) must therefore be composed of (a) a short number of SD markers (at least four), and (b) the setoff items related to the psychological latent variables that the psychological test aims to assess. The procedure allows more than one latent variable to be assess and they can be correlated.

In order to deal with AC, the procedure assumes that it should be possible to identify acquiescence as a common style factor behind a set of content items that are semantically balanced (Mirowsky and Ross 1991). In a perfectly balanced scale, with respect to a psychological trait, half of the items are worded in one direction and the other half in the other. However, few questionnaires are designed so that exactly half of the items are worded in this way: most of the psychological tests are only partially balanced. Fortunately, the procedure by Lorenzo-Seva and Ferrando (2009) helps to handle partially balanced scales (i.e., where only a few items in the scale are worded in the opposite direction). In partially balanced scales, two subsets of content items must be identified: (a) a balanced subset (i.e., a subset of items where half of the items are worded in one direction and the other half in the other); and (b) an unbalanced subset (i.e., a subset of items where all the items are worded in the same direction). It must be noted that the procedure finally removes the variance caused by AC from all the items in the questionnaire (i.e., from the balanced subset of items, but also the unbalanced subsets of items).

An example of a psychological test that includes SD markers and partially balanced content items is OPERAS (Vigil-Colet *et al.* 2013). This test includes 4 SD markers, and 35 content items of which 18 are reversed. This psychological test assesses the individuals' scores for 5 latent variables. It must be noted that a psychological test that does not include SD markers but which does have (partially) balanced content items can only control for AC; whereas a psychological test that includes a number of SD markers, but with all the content items worded in the same direction, can only control SD. Finally, the procedure proposed by Ferrando *et al.* (2009) is based on two strong assumptions: (a) AC and SD measures are assumed to be uncorrelated from the content factors and from each other; and (b) AC is assumed not to operate in pure SD markers. As a consequence, SD and AC can be controlled in consecutive and separate steps. In Section 3.1 we present how SD can be controlled using the SD markers, and in Section 3.2 and Section 3.3 we describe how AC can be controlled in partially balanced scales. In Section 4 we discuss some of the existent similar software. In Section 5 we present our stand-alone package for computing the procedure.

3. Model overview

3.1. Controlling social desirability

Let's consider a questionnaire administrated to n individuals and composed of m content items. The m items are a set of items expected to be related to r latent content variables ($r < m$). The questionnaire is partially balanced: a subset of k items is worded in one direction of the trait, and subset of l items is worded in the opposite direction, where $k+l = m$. Additionally, a set of h SD markers are administrated together with the content items. The $\mathbf{X}$ matrix containing scores of the n individuals (i.e., the responses of individuals to the test) can be partitioned as

$$\mathbf{X} = [\mathbf{X}^{sd}|\mathbf{X}^{c}] \tag{1}$$

where $\mathbf{X}^{sd}$ contains the scores in the SD markers and $\mathbf{X}^{c}$ contains scores related to all the m content items. $\mathbf{X}^{c}$ can be partitioned as

$$\mathbf{X}^{c} = [\mathbf{X}^{b}|\mathbf{X}^{u}] \tag{2}$$

where $\mathbf{X}^{b}$ contains scores related to an even set of k balanced items, and where $\mathbf{X}^{u}$ contains scores related to a set of l unbalanced items. The correlation between all the items included in $\mathbf{X}$ will be contained in $\mathbf{R}$. Also, $\mathbf{R}^{c}$ contains the correlation between $\mathbf{X}^{c}$ items and $\mathbf{R}^{sd}$ contains the correlation between $\mathbf{X}^{sd}$ items

The structural model assumes that each content item to be factorially complex measure, determined by: (a) the SD factor θ^{sd}, (b) an AC factor θ^{a}, and (c) the r content factors θ^{c}

$$X_{ij} = \delta_j\theta_i^{sd} + \alpha_j\theta_i^{a} + \beta_{j1}\theta_{i1}^{c} + \beta_{j2}\theta_{i2}^{c} + \ldots + \beta_{jr}\theta_{ir}^{c} + \varepsilon_{ij} \tag{3}$$

for $i = 1 \dots n$ and $j = 1 \dots m$, where δ is the SD factor loading, α is the AC factor loading, βs are the factor loading for r content factors and the εs are the residuals, with zero means and uncorrelated with the factors or one another. As mentioned above, the r factors are assumed to be uncorrelated with the response bias factors. Also, the SD factor and AC factor are also expected to be uncorrelated with each other. However, the r content factors can be correlated among themselves.

To simplify the model, let us suppose that all content items in the questionnaire are measuring a one-dimensional trait θ^c, thus leading to a model such as

$$X_{ij} = \delta_j \theta_i^{sd} + \alpha_j \theta_i^a + \beta_j \theta_i^c + \varepsilon_{ij}. \tag{4}$$

Consider now the additional set of h items designed to be pure measures of SD, which are administrated together with the content items. Their function is provide factorially simple measures of SD, and the structural model for these items reduces to:

$$X_{ih} = \delta_h \theta_i^{sd} + \varepsilon_{ih}. \tag{5}$$

The h SD markers allow the loading of the content items on the SD factor to be estimated using the Instrumental Variables (IV) technique. This technique was developed in the context of factor analysis by Hägglund (1982). First, one of the SD markers is taken as a proxy for θ^{sd} and the remaining $h-1$ markers are taken as instrumental variables. Without loss of generality we can take the first marker as proxy. From correlation matrix $\mathbf{R}$, two vectors $\mathbf{r_h}$ and $\mathbf{r_j}$ can be defined. $\mathbf{r_h}$ is a column vector of order $(h-1) \times 1)$ that contains the covariance between the proxy and the other $h-1$ markers. $\mathbf{r_j}$ is a column vector of order $(h-1) \times 1$ that contains the covariance between the content item j and the other $h-1$ markers. Then the loading of the m content items on the SD factor can be estimated as,

$$\hat{\delta}_j = \mathbf{r_j}^\top \mathbf{r_h} (\mathbf{r_h}^\top \mathbf{r_h})^{-1} \delta_1. \tag{6}$$

where $\hat{\delta}_j$ is the loading of the content item j, and the δ_1 is the factor loading of the proxy variable. The value of δ_1 can be computed from the correlation matrix of the h SD markers, or directly defined from a previous study.

This is how the estimate loadings of the SD factor for the m content items can be obtained. The loadings for the $h-1$ SD markers are estimated in the same way, and the loading for the first marker (proxy) can be estimated simply by choosing another pivot variable. Once the complete vector of m loading estimates $\boldsymbol{\delta}$ have been obtained, the reproduced correlation matrix is computed as $\boldsymbol{\delta}\boldsymbol{\delta}^\top$.

The first residual matrix $\mathbf{S^c}$, which is free of SD impact, is obtained by substracting the reproduced matrix from the initial correlation matrix between the content items $\mathbf{R^c}$, defined as

$$\mathbf{S^c} = \mathbf{R^c} - \boldsymbol{\delta}\boldsymbol{\delta}^\top. \tag{7}$$

3.2. Controlling acquiescence: Method for fully balanced scales

This first residual matrix obtained after subtracting SD variance is used as the input in the second stage for estimating the loadings on the AC factor. As the influence of the SD factor has been partialled out, the structural model looks like this:

$$X_{ij} = \alpha_j \theta_i^a + \beta_j \theta_i^c + \varepsilon_{ij}. \tag{8}$$

If $\mathbf{S}^c$ is the first residual matrix obtained after substracting SD variance of the order $m \times m$, $\mathbf{S}^b$ is the residual matrix between a set of balanced items. Then

$$\mathbf{a} = \mathbf{S}^b\mathbf{1}(\mathbf{1}^\top\mathbf{S}^b\mathbf{1})^{-1/2} \tag{9}$$

where $\mathbf{a}$ is the vector of correlations between the variables and their mean. Values of $\mathbf{a}$ show how much each variable is impacted by AC. A factor loading matrix $\mathbf{B}_b$ of the order of $m \times (r+1)$ can be obtained by

$$\mathbf{S}^b = \mathbf{B}_b\mathbf{B}_b^\top + \mathbf{M}_b\mathbf{M}_b^\top + \Psi_b^2 \tag{10}$$

where $\mathbf{M}_b$ holds the loadings on those common factors that are discarded in the rank$-(r+1)$ solution and Ψ_b is a diagonal matrix containing the unique factor standard deviations. Let the rotation matrix $\mathbf{W}$ be an orthonormal matrix of order $(r+1)\times(r+1)$. $\mathbf{W}$ must maximize the congruence between one column of the product $\mathbf{M}_b\mathbf{W}$ and vector $\mathbf{a}$, so it is determined by the method of Korth and Tucker (1976). Let $\mathbf{d}$ and $\mathbf{w}$ be vectors defined as

$$\mathbf{d} = (\mathbf{B}_b^\top\mathbf{B}_b)^{-1}\mathbf{B}_b^\top\mathbf{a} \tag{11}$$

and

$$\mathbf{w} = \mathbf{d}(\mathbf{d}^\top\mathbf{d})^{-1/2}. \tag{12}$$

Given the eigendecomposition of the matrix

$$\mathbf{I} - \mathbf{w}\mathbf{w}^\top = \mathbf{W}\Delta\mathbf{W}^\top \tag{13}$$

where $\mathbf{I}$ is an identity matrix and Δ is a diagonal matrix with elements $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_{r+1} = 0$, the product

$$\mathbf{B_b W} = [\boldsymbol{\beta}|\boldsymbol{\alpha}] \tag{14}$$

leads to a matrix whose last column $\boldsymbol{\alpha}$ contains the loading values of balanced items on the acquiescence factor, and $\boldsymbol{\beta}$ is a $k \times r$ matrix that can be rotated to show factor simplicity by any orthogonal or oblique rotation method. Note that $\boldsymbol{\beta}$ is a factor loading matrix that is free of variance caused by AC responding.

3.3. Controlling acquiescence: Method for partially balanced scales

A factor loading matrix $\mathbf{L}$ of the order of $m \times (r+1)$ can be obtained by

$$\mathbf{S}^c = \mathbf{L}\mathbf{L}^\top + \mathbf{M}\mathbf{M}^\top + \Psi^2 \tag{15}$$

where $\mathbf{S}^c$ is the covariance matrix obtained after substracting SD variance, $\mathbf{M}$ holds the loadings on the common factors that are discarded in the rank$-(r+1)$ solution and Ψ is a diagonal matrix containing that unique factor standard deviations. $\mathbf{L}$ can be portioned as

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_b \\ \hline \mathbf{L}_u \end{bmatrix} \tag{16}$$

where $\mathbf{L}_b$ contains the loading values related to the even set of balanced items and $\mathbf{L}_u$ contains the loading values related to the set of unbalanced items. Let the rotation matrix $\mathbf{U}$ be an orthonormal matrix of order $(r+1)\times(r+1)$. $\mathbf{U}$ must maximize the congruence between one column of the product $\mathbf{L}_b\mathbf{U}$ and vector $\mathbf{a}$, so it is determined by the method of Korth and Tucker (1976). Finally, $\mathbf{U}$ is used to rotate not only $\mathbf{L}_b$ but also the overall matrix $\mathbf{L}$, so that the product

$$\mathbf{LU} = [\boldsymbol{\beta}|\boldsymbol{\alpha}] \tag{17}$$

leads to a matrix whose last column $\boldsymbol{\alpha}$ contains the loading values of balanced and unbalanced items on the acquiescence factor and $\boldsymbol{\beta}$ is a $m \times r$ matrix that can be rotated to show factor simplicity by any orthogonal or oblique rotation method. If $\mathbf{T}_r$ is a $r \times r$ rotation matrix, the rotated loading matrix related to the content factors is obtained by

$$\mathbf{P}_r = \boldsymbol{\beta}\mathbf{T}_r \tag{18}$$

while the correlation matrix between factor scores is obtained by

$$\boldsymbol{\Phi}_r = \mathbf{T}_r^{-1}(\mathbf{T}_r^{-1})^\top. \tag{19}$$

The hull procedure is summarized in Figure 1.

4. Software packages available to compute FA controlling response bias

Factor analysis is implemented in most software packages. As a stand-alone package, among the most widely used freeware is **FACTOR** (Lorenzo-Seva and Ferrando 2013), which implements several methods for computing factor analysis including some of the most recent methodological developments. The most common R distribution package for computing factor analysis is probably the **psych** package (Revelle 2016), which also contains several configuration options, and is up to date in methodological developments. Both options are really good tools for computing FA, and are clearly more configurable than **psychological test toolbox** in terms of the number of procedures available for the user to choose. However, none of them are able to control response biases in their procedures, which is the main reason that we created our tool.

Regarding the response bias function, there are certain other factor analysis procedures for assessing the impact of Acquiescence or Social Desirability. However, to the best of our knowledge, none of those methods are available for distribution via R package or any other software. The only way some of the aforementioned methods can be used is by manually calculating all the computing steps with the equations provided in the articles.

The only similar tools for controlling response bias are correctors in specific instruments, which are designed solely to provide the participant's factor scores for the specific version of a given questionnaire.

5. Psychological test toolbox: Computing FA controlling response bias

5.1. Overall description of psychological test toolbox

As mentioned previously, **psychological test toolbox** is a program designed for performing factor analysis while controlling the effect of both AC and SD or only one of these biases. The program was developed in a MATLAB environment (software propriety of MathWorks

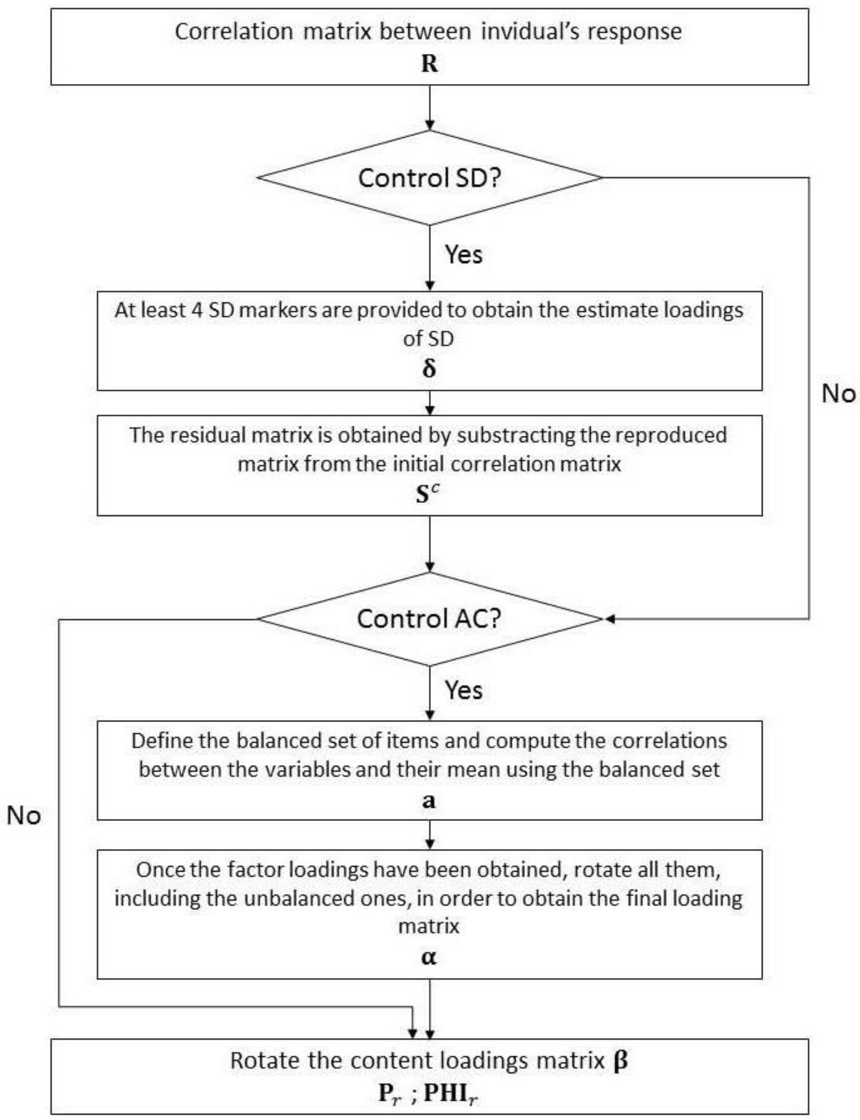

Figure 1: Procedure for controlling social desirability, acquiescence or both.

®), and it is released in two formats: as stand-alone application (only for Windows-based computers) and also as a MATLAB toolbox, which can be executed by MATLAB users on any operating system which supports MATLAB.

The program is free, it only requires the installation of MATLAB Compiler Runtime (MCR) for R2017a, which is also free and is available from MathWorks ® website.

5.2. Procedures implemented in psychological test toolbox

It is important to mention that this project contains more than one hundred of computing functions, including the principals and the secondary ones (invoked by the primaries), so it is not practical to list them all in this article. However, we make an effort to comment each one in the code, especially the primaries, to describe their usability. Also, in the principal function (PsychologicalTestToolbox.m), all the objects and the functions embedded have a comment line to guide the MATLAB user during the calculation process.

To obtain a summary of the functions used in the program, the MATLAB user can introduce the following command line in the MATLAB prompt:

```
MATLAB> addpath lib
MATLAB> help lib
```

Regarding the authorship of the functions used in the program, the vast majority of them are entirely written by members of our research team. The only exception are some internal computing functions present in the polychoric matrix calculation, which were originally created by Beasley and Springer (1977), Brown (1977) and Donnelly (1973). Also, if the code is based on a method proposed by a certain author, this is mentioned in the code itself or in the reference section if the contribution to the calculation is significant

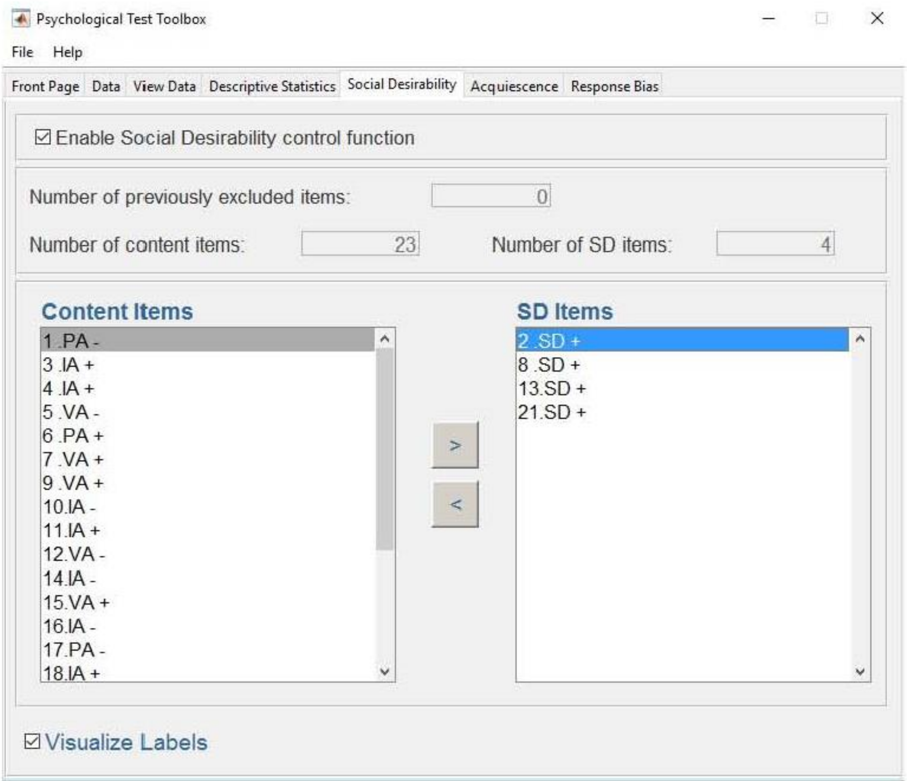

Figure 2: Determining the social desirability markers.

The program allows to compute factor analysis to be computed using different kinds of dispersion matrices, including covariance matrices, Pearson correlation matrices and tetrachoric/polychoric correlation matrices. The suitability of the dispersion matrix is assessed by three tests: the determinant of the matrix, Barlett's test, and the Kaiser-Meyer-Olkin index.

The number of factors to be retained has to be specified, and the Optimal Implementation of Parallel Analysis (PA) (Timmerman and Lorenzo-Seva 2011) can be computed to assess which number of factors is suitable. The eigenvalues of the dispersion matrices and Cattell's scree test are also generated.

For factor analysis, the program uses two procedures: unweighted least squares (ULS) and minimum rank factor analysis (MRFA) (Ten Berge and Kiers 1991). For assessing the model's goodness of fit, the program provides the Goodness of Fit Index (GFI), the Root Mean Square of Residuals (RMSR), and descriptive of statistics of the distribution of residuals.

Regarding rotation methods, the program includes Varimax (Kaiser 1958) and Promin (Lorenzo Seva 1999), which is an special case of simplimax (Kiers 1994), for assessing orthogonal and oblique solutions respectively. For assessing semi-specified target rotation it provides the methods developed by Browne (1972a) and Browne (1972b).

After the rotation phase, **psychological test toolbox** provides Bentler's simplicity index (Bentler 1977) and Lorenzo's-Seva simplicity (Lorenzo-Seva 2003) for assessing the level of simplicity attained in the rotated solution. Also, if the target matrix is provided, the congruence indices between the rotated solution and the expected solution are given, thus providing the congruence for each item and for each factor as well as the overall congruence.

Factor scores are computed by an improved implementation of Bayes EAP estimation described in Ferrando and Lorenzo-Seva (2016), which also provides the standard error of prediction for all responders.

The missing values were processed using the Multiple Imputation of missing values described in Lorenzo-Seva and Van Ginkel (2016).

5.3. Design of the user interface

The design of the graphic user interface (GUI) was one of the most important phases in the development process because one of the main objectives of **psychological test toolbox** was to create a very accessible program. We tried to develop a simple GUI using the tools that provided by MATLAB language and divided the hull application into 7 tabs which are organized according to the logical order of the analyzing process. The name and description of each tab are as follows:

1. Front Page: This is the first tab displayed and indicates the name of the program, the authors and the current version of the program.

2. Data: The user uses this tab to import the data that will be used in the analysis. Once it is imported, the user can exclude certain items.

3. View Data: This tab displays the imported data, that could be useful for doing some checking without having to open the file externally.

4. Descriptive Statistics: The user can make certain changes to the configuration of the analysis such as changing the dispersion matrix that will be used or setting up Parallel Analysis. Also, the descriptive statistics section can be computed and displayed in this

tab, which could be useful in certain cases if the user is not interested in computing a factor analysis at that particular moment.

5. Social Desirability: In this tab the user can enable the SD control function. If enabled, the application requires the user to select at least three items as SD markers.

6. Acquiescence: In this tab the user can enable the AC control function. If enabled, the application provides the user with the option of excluding certain items from the balanced core of items. As explained in Section 2, the questionnaire has to be at least partially balanced in order to control AC.

7. Response bias: This is the final tab, which includes a complete report of the data to be analyzed including the items excluded from the analysis and those selected as SD markers. In this tab the user has to specify the number of factors to be retained; least 3 items are required per factor. Also, the user can switch between the rotation methods available, require the participant's factor scores to be computed and require all possible bias combinations to be computed. Once the analysis is complete, the output is displayed in the embedded sub-window and can be saved for external viewing.

Finally, the program has a "Help" menu and a "File" Menu, the latter featuring certain options such as importing data, saving matrices generated during the analysis or exiting. One of the functions available in the File menu is saving the current configuration of the program, including the data imported and the output generated (if any), and allowing the user to close the program and resume the analysis at a later date by clicking on "Save analysis" and "Open analysis". This could be a useful tool for replicating certain results using the same exact configuration, or for doing a complex analysis at two different times.

All the GUI objects, figures and graphics are generated by code, which gave us more flexibility to handle and structure them. The main figure where the application is embedded cannot be resized to prevent distortions of the objects from being viewed. The figures containing plots and the output section are fully resizable.

5.4. Input and output

To run the stand-alone program, **psychological test toolbox** must be executed in Windows operative system. To run the user-friendly interface in MATLAB context as a toolbox, the following command line must be executed in the MATLAB prompt

```
MATLAB> PsychologicalTestToolbox
```

Once the main window is in execution, the program requires some input data to work with, which can be a raw data file or a dispersion matrix. The program can import files in different formats (.dat, .txt, .xls, .xlsx), and is able to identify variable labels at the header of the file. There are some optional input files, such as a text file containing the variable labels (in .txt) and a file containing the semi-specified rotation target matrix.

If the data contains missing values, the user has to assign a unique value to these (for example: 999), specify the option "The data file contains missing values" and define the previously determined value.

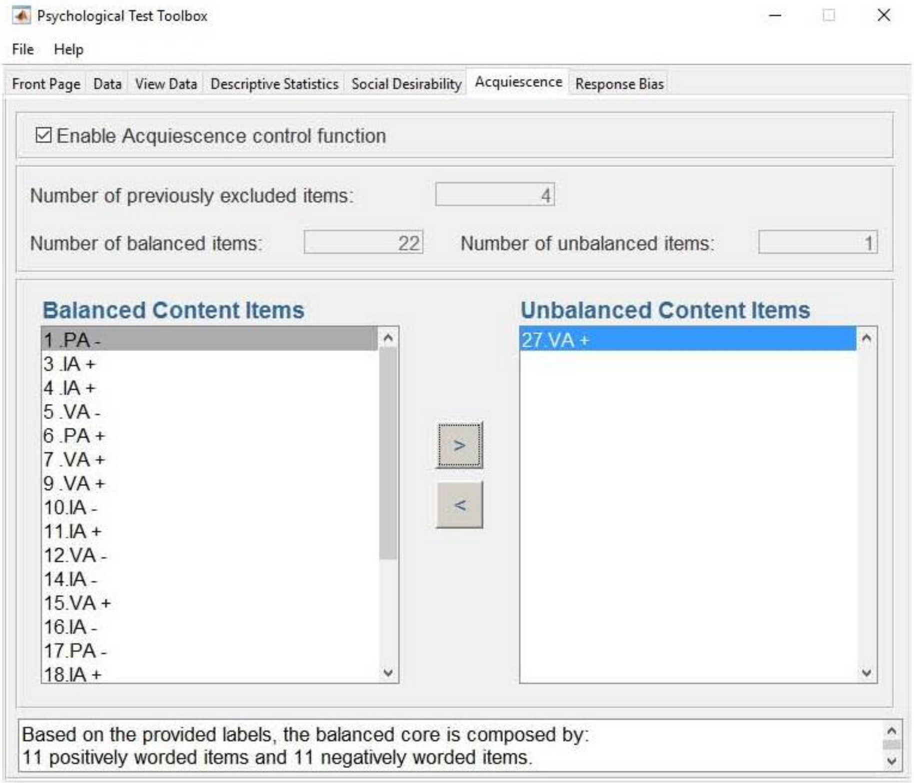

Figure 3: Selecting the balanced core of items.

An extensive output is provided, depending on the selected options selected. The output is divided into two parts: the first output section contains the descriptive statistics of the items, including:

- A summary of the analysis.
- Univariate item descriptive statistics.
- Dispersion matrix.
- Parallel analysis output (if requested).
- Indices of adequacy of the dispersion matrix.
- Scree test.
- Descriptive statistics related to missing data (if applicable).
- References for this section.

And the second output section presents the factor analysis output, including:

- Goodness of Fit Index.

- Target loading matrix (if provided).
- Rotated loading matrix.
- Correlation between content factors.
- Indices of factor simplicity.
- Congruence indices between the rotated loading matrix and target matrix (if a procrustean rotation is selected).
- Distribution of residuals.
- EAP scores of the participants and the reliability of these scores (if required).
- References for this section.

Note that the factor analysis output will be generated for the desired option (controlling for only one response bias, or both or neither) but there is an option for computing and displaying all the possible bias analysis combinations.

The output can be saved in three different formats: in plain text (.txt), in Rich Text Format (.rtf) which is fully compatible with Microsoft Word and presents all the tables in a proper format, and also in **LaTeX** (.tex) format which generates a complete report that presents all the information in a clean manner.

6. An illustrative example

In this example we are going to use the indirect and direct aggression questionnaire (I-DAQ, Ruiz-Pamies *et al.* 2014), which was one of the first questionnaires developed using this procedure to control for response biases. This questionnaire was administered to a sample of 1479 respondents (536 men and 943 women) with ages ranging from 18 to 96.

The questionnaire measures 3 aggression dimensions: physical aggression (PA), verbal aggression (VA) and indirect aggression (IA). The questionnaire consists of 27 Likert items, i.e., 23 items measuring the 3 content dimensions and 4 SD markers for applying the procedure described in this article. For clarity, we used labels indicating which dimensions were being measured and their direction; the positively keyed items are labeled "+" and the negatively worded items are labeled "-". The selection of the SD markers in the program is presented in Figure 2.

The content items are only partially balanced and consist of 12 positively worded items and only 11 negatively worded items. In this example, we will exclude item number 27 from the balanced core (see Figure 3).

Finally, Figure 4 is presents the hull configuration, including the full list of content items and the configuration options used in this analysis.

We computed an exploratory factor analysis, controlling both biases based on the polychoric interitem correlation matrix, because polychoric correlation is advised when the univariate distributions of ordinal items are asymmetric or with excess of kurtosis, which is the case. If both indices are lower than one in absolute value, then Pearson correlation is advised. The root mean squares of the residuals (RMSR) was .037. An optimal implementation of parallel

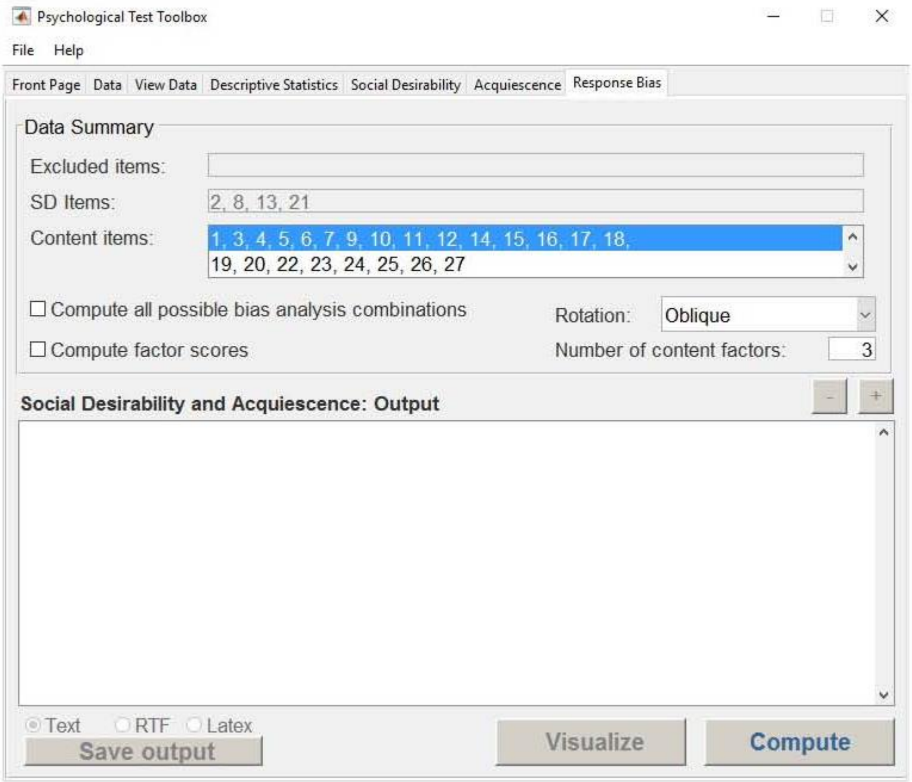

Figure 4: The configuration screen which displays the list of content items and SD markers.

analysis was computed and the results are shown in Table 1, showing that the advised number of factors to retain are 3.

Table 2 shows the factor solution obtained by controlling for response bias using the procedure described above. When applying the procedure to control for the effect of SD and AC, the factor structure becomes congruent with the expected solution. All the items have their salient loading on the expected factor, without any factorially complex item, thus resulting in a simple solution.

Another advantage of using this procedure is the ability to look at the loadings on the biases factors in order todetermine which items are more impacted by which bias. For example, we can see that some items have high loadings on the SD factor, such as item 27 (.449), item 3 (.417) and item 4 (.353). Also, there are other items with high loadings on the AC factor such as item 18 (.334) and item number 23 (.334).

7. Program limitations

The program's options for configuring the analysis have been simplified to make the process easier for applied researchers who may be less familiar with some of the psychometric and statistical concepts. However, this decision has left the program with only a few configuration options, and some advanced users may consider them too limited.

Item	Real-data % of variance	Mean of random % of variance	95 percentile of random % of variance
1	29.09**	7.33	7.84
2	8.78**	6.89	7.31
3	8.25**	6.55	6.91
4	5.84	6.25	6.61
5	5.38	5.96	6.27
6	4.00	5.70	6.02
7	3.59	5.45	5.74
8	3.40	5.20	5.45
9	3.24	4.96	5.19
10	2.81	4.72	4.96
11	2.75	4.48	4.67
12	2.67	4.24	4.44
13	2.48	4.00	4.23
14	2.45	3.76	3.98
15	2.26	3.50	3.71
16	2.15	3.26	3.48
17	1.97	3.00	3.25
18	1.85	2.75	3.00
19	1.61	2.49	2.75
20	1.57	2.22	2.51
21	1.43	1.94	2.22
22	1.04	1.67	1.97
23	0.66	1.38	1.69
24	0.36	1.08	1.42
25	0.21	0.76	1.07
26	0.15	0.45	0.70
27	0.00	0.00	0.00

Table 1: Parallel analysis output. ** advised number of factors

The stand-alone version is only available for Windows-based computers. However, the code version of the program can be executed from any OS if the user has a MATLAB license.

There are some analyses can take several minutes or even hours. This is not the standard, but in some configurations the computing time can increase exponentially, from a few seconds to several minutes. The options that can further increase computation time are:

- Selecting the polychoric matrix as the dispersion matrix for the analysis.
- Requesting optimal implementation of parallel analysis.
- Requesting factor scores for each participant.

If the user only wishes to use one of these options the increase in computing time will be

Item	SD	AC	Physical	Indirect	Verbal
2 .SD +	0.6468	0.0000	0.0000	0.0000	0.0000
8 .SD +	0.4668	0.0000	0.0000	0.0000	0.0000
13.SD +	0.7851	0.0000	0.0000	0.0000	0.0000
21.SD +	0.7706	0.0000	0.0000	0.0000	0.0000
1 .PA -	-0.1806	0.1098	**-0.3863**	-0.1306	-0.2143
6 .PA +	0.1495	0.3287	**0.5937**	-0.0771	-0.0380
17.PA -	-0.1381	0.2308	**-0.7347**	0.0121	0.0292
19.PA -	-0.1685	0.2648	**-0.7004**	0.0589	0.0007
20.PA +	0.2985	0.2528	**0.6365**	0.0720	-0.1066
25.PA +	0.2347	0.2109	**0.5231**	0.0566	0.1719
3 .IA +	0.4166	0.3004	0.0328	**0.4692**	-0.0080
4 .IA +	0.3528	0.2982	0.0875	**0.3860**	-0.0072
10.IA -	-0.0896	0.2345	0.0306	**-0.4247**	0.0408
11.IA +	0.1982	0.3219	-0.0809	**0.5189**	0.0334
14.IA -	-0.2053	0.1864	0.0010	**-0.4552**	-0.1040
16.IA -	-0.0921	0.1572	0.1003	**-0.6288**	0.0242
18.IA +	0.2984	0.3344	0.0876	**0.6144**	-0.1073
23.IA +	0.2689	0.3341	0.0879	**0.5271**	-0.0164
24.IA -	-0.0486	0.1366	0.0460	**-0.4031**	-0.0742
26.IA -	-0.1850	0.1497	-0.0031	**-0.5082**	-0.0271
5 .VA -	-0.1620	0.1060	0.0094	0.0047	**-0.6618**
7 .VA +	0.1709	0.1797	-0.0731	0.0851	**0.5408**
9 .VA +	0.2881	0.2258	0.0394	0.0786	**0.5314**
12.VA -	-0.1913	0.1410	-0.1039	-0.0147	**-0.5356**
15.VA +	0.2508	0.1170	-0.0654	-0.1663	**0.7512**
22.VA -	-0.1609	0.1016	-0.1278	-0.0748	**-0.2232**
27.VA +	0.4490	0.1538	0.1885	0.0767	**0.2739**

Table 2: Rotated loading matrix.

acceptable. However, the computing time begins to increase significantly when some of these options are requested in combination, for example if the user selects the polychoric matrix and requests the factor scores for each participant.

Furthermore, computing can also be increased by certain sample characteristics such as having a large amount of items or participants or including missing data.

8. Program availability

As mentioned previously, the **psychological test toolbox** is a freeware program and can be downloaded from the website of our department:

http://psico.fcep.urv.cat/utilitats/PsychologicalTestToolbox/

In the website, the user will find extensive documentation, including tutorial videos organized in sections depending on which functionalities is the user more interested. We strongly recommend to visit the site for stay up to date regarding the current version of the program, as well as knowing possible new features introduced a posteriori.

The library required for executing the stand-alone application can be downloaded from the MathWorks website. This is not necessary if the user has a current license for MATLAB 2017a.

http://www.mathworks.com/supportfiles/downloads/R2016a/deployment_files/R2017a/installers/win64/MCR_R2016a_win64_installer.exe

Acknowledgements

The research was partially supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2014SGR73) and by a grant from the Spanish Ministry of Economy and Competitivity (PSI2014-52884-P).

References

Beasley JD, Springer SG (1977). "Algorithm AS 111: The Percentage Points of the Normal Distribution." *Applied Statistics*, **26**(1), 118–121.

Becker G (2007). "The Buss–Perry Aggression Questionnaire: Some Unfinished Business." *Journal of Research in Personality*, **41**(2), 434–452. doi:10.1016/j.jrp.2006.05.004.

Bentler PM (1977). "Factor Simplicity Index and Transformations." *Psychometrika*, **42**(2), 277–295. doi:10.1007/bf02294054.

Billiet JB, McClendon MJ (2000). "Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items." *Structural Equation Modeling: a Multidisciplinary Journal*, **7**(4), 608–628. doi:10.1207/s15328007sem0704_5.

Brown MB (1977). "Algorithm AS 116: The Tetrachoric Correlation and its Asymptotic Standard Error." *Applied Statistics*, **26**(3), 343–351.

Browne MW (1972a). "Oblique Rotation to a Partially Specified Target." *British Journal of Mathematical and Statistical Psychology*, **25**(2), 207–212. doi:10.1111/j.2044-8317.1972.tb00492.x.

Browne MW (1972b). "Orthogonal Rotation to a Partially Specified Target." *British Journal of Mathematical and Statistical Psychology*, **25**(1), 115–120. doi:10.1111/j.2044-8317.1972.tb00482.x.

Cupani M, Lorenzo-Seva U (2016). "The Development of an Alternative IPIP Inventory Measuring the Big-Five Factor Markers in an Argentine Sample." *Personality and Individual Differences*, **91**, 40–46. doi:10.1016/j.paid.2015.11.051.

Danner D, Aichholzer J, Rammstedt B (2015). "Acquiescence in Personality Questionnaires: Relevance, Domain Specificity, and Stability." *Journal of Research in Personality*, **57**, 119–130. doi:10.1016/j.jrp.2015.05.004.

Donnelly T (1973). "Algorithm 462: Bivariate Normal Distribution." *Comunications of the ACM*, **16**(10), 638. doi:10.1145/362375.362414.

Ferrando PJ (2005). "Factor Analytic Procedures for Assessing Social Desirability in Binary Items." *Multivariate Behavioral Research*, **40**(3), 331–349. doi:10.1207/s15327906mbr4003_3.

Ferrando PJ, Lorenzo-Seva U (2016). "A Note on Improving EAP Trait Estimation in Oblique Factor-Analytic and Item Response Theory Models." *Psicológica*, **37**, 235–247.

Ferrando PJ, 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. doi:10.1207/s15327906mbr3803_04.

Ferrando PJ, 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**(2), 364–381. doi:10.1080/10705510902751374.

Goffin R, Christiansen N (2003). "Correcting Personality Tests for Faking: A Review of Popular Personality Tests and an Initial Survey of Researchers." *International Journal of Selection & Assessment*, **11**(4), 340–344. doi:10.1111/j.0965-075x.2003.00256.x.

Hägglund G (1982). "Factor analysis by instrumental variables methods." *Psychometrika*, **47**(2), 209–222. doi:10.1007/BF02296276.

Harris JA (1997). "A Further Evaluation of The Aggression Questionnaire: Issues of Validity and Reliability." *Behaviour Research and Therapy*, **35**(11), 1047–1053. doi:10.1016/s0005-7967(97)00064-8.

Holden R (2010). "Social Desirability." In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons.

Kaiser HF (1958). "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, **23**(3), 187–200. doi:10.1007/bf02289233.

Kiers HAL (1994). "Simplimax: Oblique Rotation to an Optimal Target with Simple Structure." *Psychometrika*, **59**(4), 567–579. doi:10.1007/bf02294392.

Korth B, Tucker LR (1976). "Procrustes Matching by Congruence Coefficients." *Psychometrika*, **41**(4), 531–535. doi:10.1007/bf02296973.

Lorenzo-Seva U (1999). "Promin: A Method for Oblique Factor Rotation." *Multivariate Behavioral Research*, **34**(3), 347–365. doi:10.1207/s15327906mbr3403_3.

Lorenzo-Seva U (2003). "A Factor Simplicity Index." *Psychometrika*, **68**(1), 49–60. doi:10.1007/bf02296652.

Lorenzo-Seva U, Ferrando PJ (2009). "A General Factor-Analytic Procedure for Assessing Response Bias in Questionnaire Measures." *Structural Equation Modeling: a Multidisciplinary Journal*, **16**(2), 364–381. doi:10.1080/10705510902751374.

Lorenzo-Seva U, Ferrando PJ (2013). "FACTOR 9.2: A Comprehensive Program for Fitting Exploratory and Semiconfirmatory Factor Analysis and IRT Models." *Applied Psychological Measurement*, **37**(6), 497–498. doi:10.1177/0146621613487794.

Lorenzo-Seva U, Rodríguez-Fornells A (2006). "Acquiescent Responding in Balanced Multidimensional Scales and Exploratory Factor Analysis." *Psychometrika*, **71**(4), 769–777. doi:10.1007/s11336-004-1207-4.

Lorenzo-Seva U, Van Ginkel JR (2016). "Multiple Imputation of Missing Values in Exploratory Factor Analysis of Multidimensional Scales: Estimating Latent Trait Scores." *Anales de Psicología*, **32**(2), 596–608. doi:10.6018/analesps.32.2.215161.

Mas-Herrero E, Marco-Pallares J, Lorenzo-Seva U, Zatorre RJ, Rodriguez-Fornells A (2013). "Individual Differences in Music Reward Experiences." *Music Perception: an Interdisciplinary Journal*, **31**(2), 118–138. doi:10.1525/mp.2013.31.2.118.

Mirowsky J, Ross C (1991). "Eliminating Defense and Agreement Bias from Measures of the Sense of Control: A 2 X 2 Index." *Social Psychology Quarterly*, **54**(2), 127–145. doi:10.2307/2786931.

Morales-Vives F, Camps E, Lorenzo-Seva U (2013). "Development and Validation of the Psychological Maturity Assessment Scale (PSYMAS)." *European Journal of Psychological Assessment*, **29**, 12–18. doi:10.1027/1015-5759/a000115.

Navarro-Gonzalez D, Lorenzo-Seva U, Vigil-Colet A (2016). "How Response Bias Affects the Factorial Structure of Personality Self-Reports." *Psicothema*, **28**(4), 465–470. doi:10.7334/psicothema2016.113.

Neill JA, Jackson DN (1970). "An Evaluation of Item Selection Strategies in Personality Scale Construction." *Journal of Research in Personality*, **30**(3), 647–661.

Paulhus DL (1981). "Control of Social Desirability in Personality Inventories: Principal-Factor Deletion." *Journal of Research in Personality*, **15**(3), 383–388. doi:10.1016/0092-6566(81)90035-0.

Paulhus DL (1991). "Measurement and Control of Response Bias." In JP Robinson, PR Shaver, LS Wrightsman (eds.), *Measures of Personality and Social Psychological Attitudes*. Academic Press. doi:10.1016/B978-0-12-590241-0.50006-X.

Paulhus DL, Vazire S (2005). "The Self-Report Method." In R Robins, RC Fraley, RF Krueger (eds.), *Handbook of Research Methods in Personality Psychology*. Guilford Press.

Rammstedt B, Farmer R (2013). "The Impact of Acquiescence on the Evaluation of Personality Structure." *Psychological Assessment*, **25**(4), 1137–1145. doi:10.1037/a0033323.

Rammstedt B, Kemper B (2011). "Measurement Equivalence of the Big Five: Shedding Further Light on Potential Causes of the Educational Bias." *Journal of Research in Personality*, **45**(1), 121–125. doi:10.1016/j.jrp.2010.11.006.

Revelle W (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.6.9, URL https://CRAN.R-project.org/package=psych.

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**, 310–315. doi:10.1017/sjp.2014.43.

Salgado JF (2005). "Personalidad y Deseabilidad Social en Contextos Organizacionales: Implicaciones para la Práctica de la Psicología del Trabajo y las Organizaciones." *Papeles del Psicólogo*, **26**(092), 310–315.

Soto CJ, John OP, Gosling SD, 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. doi:10.1037/0022-3514.94.4.718.

Ten Berge JM (1999). "A Legitimate Case of Component Analysis of Ipsative Measures, and Partialling the Mean as an Alternative to Ipsatization." *Multivariate Behavioral Research*, **34**(1), 89–102. doi:10.1207/s15327906mbr3401_4.

Ten Berge JM, Kiers HAL (1991). "A Numerical Approach to the Approximate and the Exact Minimum Rank of a Covariance Matrix." *Psychometrika*, **56**(2), 309–315. doi:10.1007/bf02294464.

Timmerman ME, Lorenzo-Seva U (2011). "Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis." *Psychological Methods*, **16**, 385–408. doi:10.1037/a0023353.

Vigil-Colet A, Morales-Vives F, Camps E, Tous J, Lorenzo-Seva U (2013). "Development and Validation of the Overall Personality Assessment Scale (OPERAS)." *Psicothema*, **25**(1), 100–106. doi:10.7334/psicothema2011.411.

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), 115–128.

Viswesvaran C, Ones DS, Hough LM (2001). "Do Impression Management Scales in Personality Inventories Predict Managerial Job Performance Ratings?" *International Journal of Selection & Assessment*, **9**(4), 277–289. doi:10.1111/1468-2389.00180.

Affiliation:

David Navarro-Gonzalez
Department of Psychology
Universitat Rovira i Virgili
Ctra de Valls s/n, 43007 - Tarragona, Spain
E-mail: david.navarro@urv.cat

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume -, Issue - *Submitted:* 2016-12-7
December 2016 *Accepted:* –

3.2. How response bias affects the factorial structure of personality self-reports.

Psicothema

Psicothema 2016, Vol. 28, No. 4, 465-470
doi: 10.7334/psicothema2016.113

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

How response bias affects the factorial structure of personality self-reports

David Navarro-González, Urbano Lorenzo-Seva and Andreu Vigil-Colet
Universitat Rovira i Virgili

Abstract

Background: Various studies have shown that acquiescence can distort the factor structure of personality questionnaires based on the five-factor model. In the present study, we analysed how acquiescence and social desirability affect the factor structure of a measure based on this personality model and a measure of aggression. **Method:** We analysed the factor structures of both tests before and after removing both biases in a sample of 532 adolescents aged between 11 and 18 (M= 14.75, SD= 2.1). **Results:** The factor structure of both tests presented a worse fit to the expected model when response bias was not controlled, and the congruence indexes for the personality and aggression measures showed a moderate (from C= .948 to C= .872) or great (from C= .931 to C= .475) decrease, respectively. Furthermore, acquiescence was largely responsible for these effects, and social desirability effects were only shown on the aggression measure. **Conclusions:** Response bias, and especially acquiescence, should be controlled during the development of personality measures to avoid distorting them, especially with samples of people with a high level of acquiescence (for example, those with little education, the young or the elderly). Furthermore, the use of response bias loadings as a criterion for choosing the items minimizes those distortions.

Keywords: Response bias, personality, factor structure.

Resumen

Efectos de los sesgos de respuesta en la estructura factorial de los autoinformes de personalidad. **Antecedentes:** diversos estudios han mostrado que la aquiescencia genera distorsiones en la estructura factorial de los cuestionarios de personalidad. En este estudio analizamos los efectos tanto de la aquiescencia como de la deseabilidad social en la estructura factorial de dos cuestionarios. **Método:** se analizó la estructura factorial de ambos con y sin sesgos de respuesta en una muestra de 532 adolescentes con edades entre los 11 y los 18 años (M= 14.75 SD= 2.1). **Resultados:** cuando no se eliminó el efecto de los sesgos de respuesta, el ajuste de ambos tests en relación al modelo esperado empeoró, disminuyendo la congruencia factorial moderadamente (desde C= .948 hasta C= .872) o notablemente (desde C= .931 a C= .475) para las medidas de personalidad y agresividad, respectivamente. Además, la aquiescencia fue la principal responsable de estos efectos, mientras que la deseabilidad social tan solo afectó la medida de agresividad. **Conclusiones:** es necesario controlar los sesgos de respuesta para evitar estructuras factoriales distorsionadas, especialmente en muestras con elevados niveles de aquiescencia, como poblaciones con bajo nivel educativo, adolescentes o en la tercera edad. Además, la minimización de los sesgos de respuesta durante el proceso de elección de ítems parece reducir dichas distorsiones.

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

The impact of response bias on typical response measures is an issue that has generated a great deal of research in recent decades. Although most of the research has focused on questionnaires within the framework of the five-factor model (FFM) of personality (i.e., Holden & Passey, 2010; Ones, Viswesvaran, & Reiss, 1996), the impact of response biases has also been assessed in other typical response measures such as impulsivity (Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, & Lorenzo-Seva, 2012), aggression and violence (Becker, 2007; Bell & Naugle, 2007), psychological maturity (Morales-Vives, Camps, & Lorenzo-Seva, 2013), mood states (Soubelet & Salthouse, 2011) or well-being (Kozma & Stones, 2012).

Received: April 12, 2016 • Accepted: July 28, 2016
Corresponding author: Andreu Vigil-Colet
Research Center for Behavior Assessment
Universitat Rovira i Virgili
43007 Tarragona (Spain)
e-mail: andreu.vigil@urv.cat

A glance at the scientific literature on the issue reveals that the two most important response biases are social desirability (SD), defined as the tendency for people to present themselves in a generally favourable fashion (Holden, 2010), and acquiescence (AC), defined as the tendency of respondents to agree with statements without regard to their content (Paulhus & Vazire, 2005).

Most of the research in this field has focused on how response bias affects the validity of self-reports. For instance, a great deal of research has analysed the effects of SD on test scores in employment selection processes (Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones et al., 1996; Salgado, 2005). Other issues, however, such as the effects of response bias on the factor structure of questionnaires, especially in the case of SD, have received less attention.

Response bias can affect the factor structure of questionnaires because it distorts the inter-item correlation matrix pattern (Bentler, Jackson, & Messick, 1971; Rammstedt, Goldberg, & Borg, 2010; Soto, John, Gosling, & Potter, 2008). For instance, in the presence of AC, items worded in the same direction tend to show a positive relationship that is not due to the content measured, while items worded in different directions will tend to show a negative relationship. In this case, AC will lead to the overestimation or underestimation of correlations in terms of the items' direction. A similar effect may be expected in the case of SD: that is, items most affected by SD will show a positive correlation independently of their content. As a consequence, the distortion of the inter-item correlation matrix may have a considerable impact on the resulting factor structure.

Some studies have shown these effects for AC. In this respect, tests administered to samples of low educational levels, low intelligence or adolescents and pre-adolescents have the worst fits to the five factor model of personality (Meisenberg & Williams, 2008; Rammstedt et al., 2010; Rammstedt & Kemper, 2011; Soto et al., 2008). Therefore, the validity of personality measures may be affected in subpopulations with high levels of AC such as those described above or others who also have high levels of AC, such as the elderly (Ross & Mirowsky, 1984; Vigil-Colet, Lorenzo-Seva, & Morales-Vives, 2015).

These studies have analysed the effects of AC because such methods as ipsatizing allow the effects of AC to be removed from the inter-item correlation matrix (Ten Berge, 1999). Nevertheless, there are fewer methods available for removing the distortions due to SD.

Ferrando, Lorenzo-Seva, & Chico (2009) developed a general method for controlling both biases simultaneously. The first step in the method identifies a factor related to SD by using items that are taken as markers of SD.

Then the loadings of the content items on this SD factor are used to compute a residual inter-item correlation matrix free of SD. Subsequently the residual correlation matrix is analysed by applying the method developed by Lorenzo-Seva & Ferrando (2009), which removes from the content those items of the variance that are due to acquiescent responding. This process makes it possible to analyse a residual inter-item correlation matrix that is free of the distortions caused by SD and AC, and can be used in classical exploratory factor analysis (EFA) to determine the factor structure of the questionnaire.

As can be seen the main objective of the method is to remove the effects of both biases and compare the factor structures obtained with and without bias. Nevertheless, an EFA can be run on each residual matrix and the factor structures can be compared without controlling for bias, controlling only for AC, controlling only for SD and controlling for both biases. Therefore, the impact of the distortion caused by AC and SD on the factor structure of questionnaires can be assessed.

This is the main objective of the present research, which focuses on the conjoint and individual effects of response bias on the factor structure. Although previous research has shown the effects of AC, less is known about the effects of SD and the conjoint effect of both biases. This is relevant because their effects seem to depend upon the nature of the item. In this regard, the effects of AC seem to be weak and the effects of SD stronger when the items reflect highly desirable or undesirable behaviours. In neutral items, however, the effects of AC are stronger (Ferrando & Anguiano-Carrasco, 2010). From this viewpoint, the effects of both kinds of response bias on the structure of questionnaires may depend upon the level of social acceptation or rejection of the content measured. Taking this into account, we will analyse the impact of biases on a personality questionnaire based on the FFM and an aggression questionnaire because these kinds of measure are highly affected by SD (Becker, 2007; Morren & Meesters, 2002; Vigil-Colet et al., 2012). Furthermore, the tests were administered to adolescents, who usually show higher levels of AC than adults (Soto et al., 2008).

Method

Participants

A total of 532 volunteer students (252 men and 280 women) from 4 different high schools from the Tarragona province with ages ranging from 11 to 18 years old (M=14.75 SD=2.1) participated. A total of 29.2% of the sample was aged between 11 and 13 years old, 52.5% between 14 and 16, and 18.3% between 17 and 18. Two high schools were in small cities and two in Tarragona.

Instruments

Overall Personality Assessment Scale –OPERAS– (Vigil-Colet, Morales-Vives, Camps, Tous, & Lorenzo-Seva, 2013). This is a 40-item instrument which gives scores for the factors: Extraversion (EX), Agreeableness (AG), Conscientiousness (CO), Emotional stability (ES), and Openness to experience (OE). Item responses are produced using a 5-point Likert scale. The test has suitable psychometric properties, with the following factorial consistencies: $r_{\theta\theta}$=.86, $r_{\theta\theta}$=.71, $r_{\theta\theta}$=.77, $r_{\theta\theta}$=.86, and $r_{\theta\theta}$=.81 for EX, AG, CO, ES and OE, respectively. This questionnaire contains four items of SD, the aim of which is to control this response bias. In addition, some of the items are content balanced so that the acquiescent responding bias can be controlled.

The indirect-direct aggression questionnaire –IDAQ– (Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, & Vigil-Colet, 2014). This test gives 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 I-DAQ have appropriate factor reliabilities: $r_{\theta\theta}$=.83, $r_{\theta\theta}$=.77 and $r_{\theta\theta}$=.78 for PA, VA and IA, respectively.

Procedure

School approval and parental written informed consent were obtained before participation in the study. Children's participation was voluntary, and no incentives were given for their participation. About 96% of the children who were invited to participate in the study eventually did so.

A professional psychologist administered the tests collectively. The participants were asked to volunteer to answer the inventories in their classroom. The questionnaires were anonymous, and respondents had to provide only their gender and age.

Data analysis

We computed four EFA for each questionnaire, which took into account its three- (IDAQ) or five- (OPERAS) factor structure. These EFA were performed on the polychoric inter-item correlation matrix and on the residual matrixes obtained after SD, AC and both biases had been removed, using the method developed by Ferrando et al. (2009). To assess the fit of each loading matrix to the expected factorial solutions, the congruence index developed by Tucker (1951) was computed between the rotated loading matrix and the ideal loading matrix. Data was analysed using MatLab (MathWorks, 2012).

Results

Table 1 shows the loading matrix for IDAQ with and without removing bias. As can be seen, when both SD and AC were removed, all the items except item 10 had their salient loading on the expected factor. On the other hand, when they were not removed, 9 of the 23 content items did not load on the expected factor. Congruence indexes showed clear improvement when response biases were removed. It should be taken into account that indexes higher than .85 imply 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 & Berge, 2006). When response biases were removed, all the congruence indexes were over .85, and most of them were around .95. But when they were not removed, most of the indexes showed a bad fit. It is worth mentioning that when biases were not removed, item number ten seems to load on the IA scale but, when they were, the loading seems to reflect the distortion caused by bias.

As can be seen, the questionnaire items showed that both biases had a considerable effect, the mean loading on SD (excluding the items used as markers) and AC being λ_m=.232 and λ_m=.222, respectively.

Table 2 shows the improvement when only one of the two response biases was removed. When SD was controlled, there was an improvement in the factor structure. Only three items did not show their salient loading on the expected factor, and the congruencies were clearly better than the ones obtained without any control, although not good enough. We found a greater improvement when AC was controlled. In this case, the results were quite similar to those obtained when both biases were removed, and the congruence indexes were similar.

Table 1
Loading matrix obtained with and without controlling response bias and factorial congruence with expected solution for I-DAQ. In bold face salient loadings on content factors

		Controlling bias					With bias		
	Item	**SD**	**AC**	**PHY**	**IND**	**VER**	**PHY**	**IND**	**VER**
	13	.608							
	2	.647							
	8	.391							
	21	.747							
PHYSICAL	1	-.209	.235	**-.496**	-.117	-.076	**.594**	.040	.180
	6	.294	.217	**.688**	-.090	-.153	**-.495**	.101	.128
	17	-.195	.314	**-.729**	-.081	-.052	**.794**	.156	.209
	19	-.213	.195	**-.785**	.077	.056	**.692**	.092	.013
	20	.324	.254	**.639**	.018	.026	**-.548**	.214	.091
	25	.272	.078	**.466**	.033	.287	**-.612**	.121	.012
INDIRECT	3	.386	.269	.036	**.463**	-.075	.101	**.691**	-.152
	4	.281	.283	.194	**.234**	-.071	-.066	**.453**	.018
	10	.212	.305	-.038	.020	-.033	.106	**.360**	.229
	11	.401	.343	.030	**.396**	.008	.038	**.645**	-.055
	14	-.001	.296	.142	**-.597**	-.076	-.047	-.145	**.573**
	16	-.119	.328	.057	**-.546**	-.076	.067	-.149	**.516**
	18	.114	.197	-.030	**.721**	-.022	.091	**.506**	-.333
	23	.330	.175	.078	**.323**	.020	-.081	**.402**	-.113
	24	.101	.310	-.027	**-.153**	.023	.073	.213	**.326**
	26	-.034	.325	-.085	**-.504**	.095	.081	-.025	**.593**
VERBAL	5	-.278	.117	.020	.010	**-.660**	**.410**	-.136	-.015
	7	.056	.198	.021	-.083	**.381**	**-.210**	.028	.073
	9	.459	.183	.010	.101	**.489**	-.320	**.374**	.051
	12	-.138	.135	-.128	-.045	**-.450**	**.419**	-.034	.069
	15	.331	.028	-.016	-.104	**.583**	**-.367**	.142	.101
	22	-.062	.197	-.047	-.080	**-.224**	**.219**	.037	.152
	27	.513	.105	.118	.035	**.278**	-.305	**.403**	.142
	λ_m	.232	.222						
	Congruence:			.953	.884	.955	.851	.534	.003
	Overall congruence:				.931			.475	

Note: SD: social desirability, AC: acquiescence PHY: physical aggression, VER: Verbal aggression, IND: Indirect aggression. λ_m: mean of loadings

Table 2
Loading matrix controlling social desirability or acquiescence and factorial congruence with expected solution. In bold face salient loadings on content factors

		Controlling SD			Controlling AC		
	Item	**PHY**	**IND**	**VER**	**PHY**	**IND**	**VER**
PHYSICAL	1	**.528**	.178	-.028	**-.508**	-.126	-.146
	6	**-.602**	.201	-.023	**.747**	-.084	-.155
	17	**.774**	.187	.066	**-.750**	-.082	-.112
	19	**.782**	-.040	.093	**-.794**	.069	-.024
	20	**-.560**	.108	.185	**.654**	.016	.035
	25	**-.464**	-.017	.310	**.456**	.021	.314
INDIRECT	3	.066	**-.240**	.147	.035	**.524**	-.028
	4	-.091	-.036	**.146**	.199	**.256**	-.049
	10	.138	**.180**	.167	-.070	.035	.052
	11	.077	-.158	**.273**	.039	**.417**	.047
	14	-.076	**.674**	-.033	.158	**-.630**	-.053
	16	.005	**.619**	.015	.055	**-.577**	-.101
	18	.123	**-.529**	.203	-.046	**.697**	-.084
	23	-.026	**-.184**	.117	.078	**.347**	.083
	24	.100	**.296**	.146	-.056	**-.162**	.073
	26	.154	**.601**	.148	-.101	**-.514**	.098
VERBAL	5	.053	.057	**-.547**	.049	.021	**-.747**
	7	-.038	.076	**.367**	.018	-.123	**.312**
	9	-.019	-.039	**.585**	.036	.116	**.517**
	12	.183	.116	**-.363**	-.108	-.030	**-.470**
	15	-.050	.059	**.522**	-.043	-.112	**.695**
	22	.103	**.166**	-.152	-.045	-.080	**-.255**
	27	-.092	.068	**.386**	.128	.072	**.417**
Congruence		.952	.808	.837	.944	.884	.940
Overall Congruence			.864			.922	

Note: SD: social desirability, AC: acquiescence PHY: physical aggression, VER: Verbal aggression, IND: Indirect aggression

We performed the same analysis for OPERAS. Table 3 showed that, when biases were removed, all the items loaded on the expected dimension, with appropriate or excellent congruence indexes. When biases were not removed, the fit was worse, especially in the case of AG, and the congruence indexes decreased. Furthermore, five items did not load on the expected dimension. EX and ES were the least affected dimensions and, to reduce the size of the tables, their loadings are not included in the table. As can be seen, the impact of response biases was lower than for IDAQ, especially for SD, which showed a mean loading of λ_m=.177.

Table 4 shows the results when only one bias was controlled. When SD effects were removed, the improvement in the congruencies was almost negligible but, when only AC was controlled, we obtained nearly the same structure as the one reported in Table 3 when both biases were removed.

Discussion

Various authors have pointed out that AC is a source of bias in typical response measures which may distort their factor structure, but less is known about the effects of SD. The present study analysed the impact of both response biases on the factor structure of personality questionnaires and showed that both AC and SD have effects on factor structures, but of different magnitudes, and apparently related to the content measured.

The results discussed above seem to show that response bias has a considerable impact on the factor structure of questionnaires, and can cause a great deal of distortion. In fact, in the case of IDAQ, the structure obtained when bias is not controlled is absolutely incongruent with the expected structure, while in the case of OPERAS, the fit is worse. It is worth mentioning that during the development of OPERAS, the items were chosen by taking their loadings on AC and SD into account so as to minimize their impact on the test. It is logical, then, that controlling for bias had less effect on its structure than in the case of IDAQ. The results reported above seem to show the importance of controlling response bias when the factor structure is assessed.

Our results also show that the distortion due to AC is clearly bigger than the distortion due to SD. This result has been reported previously by several authors who assessed the effects of AC on the

Table 3

Loading matrix obtained with and without controlling response bias. In bold face salient loadings on content factors for OPERAS' scales conscientiousness, agreeableness and openness to experience

		Controlling bias							With bias				
	item	SD	AC	EX	ES	CO	AG	OP	EX	ES	CO	AG	OP
	5	-.310											
	11	.731											
	19	.722											
	26	.792											
CONSCIENTIOUSNESS	4	-.251	.228	.074	-.296	**-.327**	.174	.095	-.183	.101	-.300	**.429**	.042
	10	.311	.063	.064	.155	**.599**	-.170	-.120	.166	.044	**.463**	-.335	-.116
	16	.325	.059	-.080	.115	**.588**	-.065	-.151	.153	-.054	**.556**	-.238	-.157
	22	.271	.155	-.089	.016	**.432**	-.139	.135	.156	.019	**.441**	-.082	.079
	28	-.009	.188	-.010	.106	**-.342**	-.267	.198	.209	.099	**-.236**	.017	.230
	33	.379	.183	-.175	.172	**.526**	-.018	-.060	.262	-.081	**.628**	-.118	-.072
	38	-.128	.357	.098	-.144	**-.481**	-.061	.036	.027	.213	-.250	**.342**	.042
AGREEABLENESS	6	-.272	.319	.003	-.095	-.157	**.350**	.095	.004	-.022	-.195	**.513**	-.004
	12	-.371	.301	.006	-.110	-.220	**.512**	.120	-.042	-.060	-.275	**.637**	-.003
	17	-.207	.204	.168	-.188	.060	**.332**	-.017	-.131	.096	-.076	**.330**	-.091
	23	.390	.065	-.092	.036	.018	**-.468**	.097	.168	.103	.216	**-.303**	.157
	29	.486	.126	-.042	.173	.130	**-.390**	-.070	.291	.150	**.364**	-.351	.019
	34	-.276	.362	.043	-.207	-.120	**.426**	.248	-.103	.026	-.061	**.687**	.119
	39	.444	.159	.106	-.112	-.078	-.176	**-.223**	.051	**.311**	.273	-.099	-.159
OPENNESS	7	.026	.080	.105	-.101	.018	-.058	**-.640**	-.035	.124	-.011	-.100	**-.599**
	13	-.125	.296	-.015	-.128	-.179	-.107	**.462**	.043	.080	-.141	.282	**.382**
	18	.041	.077	.046	.064	-.037	.102	**-.240**	.065	.041	-.029	.003	**-.207**
	24	.013	.061	-.151	.088	-.041	.045	**.694**	.087	-.127	.001	.140	**.686**
	30	-.114	.263	.208	-.056	-.302	.282	**.373**	.021	.234	-.147	**.508**	.280
	35	-.070	.451	.198	-.167	-.109	.140	**.510**	.033	.287	.084	**.569**	.362
	40	.109	.161	-.026	-.070	.047	.016	**-.761**	.026	.039	.148	-.042	**-.689**
	λ_m	.177	.206										
	Congruence			.968	.965	.959	.902	.942	.930	.892	.864	.760	.914
	Overall Congruence					.948					.872		

Note: SD: social desirability AC: Acquiescence EX: extraversion; AG: agreeableness; CO: conscientiousness; ES: emotional stability; OE: openness to experience. λ_m: mean of loadings structure of personality measures based on the FFM (Meisenberg & Williams, 2008; Rammstedt et al., 2010; Soto et al., 2008) but there have been no studies about its effects on other personality dimensions. In this regard, our results show that AC has a similar effect on aggression measures.

Two important questions are raised by the results reported above. The first is that AC has a greater effect on factor structures than SD does, and the second is the different effect that controlling SD has on the two tests analysed.

To answer the first question, we hypothesise that the effects of SD and AC on the inter-item correlation matrix are slightly different. When a pair of items is affected by SD, their correlation may increase in two circumstances. If they are measuring the same content, SD overestimates their relationship and, if their contents are not related, SD may generate a correlation where none is expected. It is noted that both distortions always give rise to a positive relationship. The effect of AC is more complex because, if the items share the same content, AC may over- or underestimate their relationship depending on whether the items are keyed in the same or in opposite directions. On the other hand, if the contents are not related, AC generates positive or negative relationships between items where no relationship is expected. Therefore, AC affects not only the magnitude of the relationships but also their sign; it involves a greater distortion of the inter-item correlation matrix and has a deeper impact on the factor structure.

The second question refers to the differences observed after controlling SD in both tests. In the case of OPERAS, removing SD effects led to negligible improvement on the congruence of the factor solution (from C=.872 to C=.874) while in the case of IDAQ, the improvement was considerable (from C=.475 to C=.864). These differences may be because aggression measures are usually highly affected by SD, because this kind of behaviour is socially rejected (Vigil-Colet et al., 2012). In fact, the mean loadings on SD for IDAQ (λ_m=.232) were greater than for OPERAS (λ_m=.177). Therefore, it is possible that SD distorts aggression measures more than overall personality measures, which may explain the differences observed.

To summarize, the main conclusion of the present study is that response bias, and especially AC, should be controlled when assessing the dimensionality of a personality measure because, if it is not, the researcher may find a distorted dimensionality or a dimensionality that is not the one expected from the theoretical model underlying the measure. This point is even more important when the sample comprises a group of people with a high level of acquiescence such as people with a low level of education, the young or the elderly (Meisenberg & Williams, 2008; Soto et al., 2008; Vigil-Colet, Morales-Vives, & Lorenzo-Seva, 2013). Taking into account that even low levels of AC have effects on item covariation (Rammstedt & Farmer, 2013), further research should apply the method used in the present study in samples with moderate or low levels of AC to assess the degree of distortion of the factor structures in those populations.

Table 4

Loading matrix controlling social desirability or acquiescence. In bold face salient loadings on content factors for conscientiousness, agreeableness and openness to experience

		Controlling SD					Controlling AC				
	item	EX	ES	CO	AG	OP	EX	ES	CO	AG	OP
CONSCIENTIOUSNESS	4	.083	-.194	-.278	**.393**	-.013	-.279	.040	**-.373**	.310	.092
	10	.092	.186	**.543**	-.259	.063	.170	.093	**.604**	-.221	-.138
	16	-.067	.151	**.569**	-.149	.138	.113	-.040	**.627**	-.194	-.153
	22	-.039	.147	**.464**	.055	-.092	.016	-.058	**.433**	-.236	.121
	28	.060	.217	**-.345**	.034	-.211	.123	-.020	**-.400**	-.203	.186
	33	-.161	.260	**.545**	.011	.090	.143	-.130	**.601**	-.239	-.043
	38	.158	.051	**-.422**	.357	.046	-.140	.072	**-.543**	.034	.034
AGREEABLENESS	6	-.026	-.001	-.099	**.489**	.019	-.079	-.037	-.195	**.490**	.089
	12	-.056	-.050	-.147	**.596**	.020	-.098	-.046	-.257	**.658**	.119
	17	.134	-.133	.090	**.351**	.103	-.173	.128	.003	**.413**	-.021
	23	.012	.154	.002	**-.199**	-.145	.034	-.042	.057	**-.598**	.089
	29	.037	.282	.114	-.206	.029	.153	.019	.211	**-.598**	-.062
	34	.011	-.093	-.034	**.631**	-.100	-.208	.003	-.158	**.497**	.259
	39	.160	.025	-.036	.062	**.264**	-.147	.164	.023	**-.441**	-.201
OPENNESS	7	.126	-.039	.002	-.108	**.635**	-.090	.107	.007	-.019	**-.632**
	13	.054	.045	-.137	.315	**-.402**	-.103	-.041	-.246	.052	**.442**
	18	.026	.062	-.049	.016	**.242**	.064	.054	-.015	.063	**-.245**
	24	-.150	.081	-.037	.172	**-.689**	.091	-.148	-.029	.022	**.684**
	30	.179	.023	-.217	**.536**	-.247	-.072	.188	-.294	.274	**.393**
	35	.228	.057	.015	**.660**	-.338	-.187	.176	-.123	.186	**.544**
	40	-.014	.020	.056	-.032	**.785**	-.087	-.016	.075	-.052	**-.756**
Congruence		.965	.890	.945	.658	.909	.936	.853	.856	.858	.846
Overall Cong.				.874					.944		

Note: EX: extraversion; AG: agreeableness; CO: conscientiousness; ES: emotional stability; OE: openness to experience

One possible limitation of the study is that the factorial invariance across the age groups was not tested. Although the age range is small, the quick development of personality during adolescence suggests that further research with larger samples should analyse the factorial invariance across age groups for both tests.

Acknowledgments

The research was partially supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2014SGR73) and by a grant from the Spanish Ministry of Economy and Competitivity (PSI2014-52884-P).

References

Becker, G. (2007). The Buss-Perry aggression questionnaire: Some unfinished business. *Journal of Research in Personality, 41*, 434-452.

Bell, K. M., & Naugle, A. E. (2007). Students' self-Reporting of partner abuse perpetration and victimization. *Violence and Victims, 22*, 243-257.

Bentler, P., Jackson, D., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*, 186-204.

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). Acquiescence and social desirability as item response determinants: An IRT-based study with the Marlowe-Crowne and the EPQ Lie scales. *Personality and Individual Differences, 48*, 596-600.

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A General factor-Analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling, 16*, 364-381.

Holden, R. (2010). Social desirability. In *Corsini Encyclopedia of Psychology*. London: John Wiley & Sons, Inc.

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49*, 446-450.

Kozma, A., Stones, M. J. (1988). Social desirability in measures of subjective well-being: Age comparisons. *Social Indicators Research, 20*, 1-14.

Lorenzo-Seva, U., & Berge, J. Ten. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*, 57-64.

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *The British Journal of Mathematical and Statistical Psychology, 62*, 319-326.

MathWorks, M. (2012). *The language of technical computing*. Mathlab Incorporated. Natick, Massachusetts, USA.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539-1550.

Morales-Vives, F., Camps, E., & Lorenzo-Seva, U. (2013). Development and validation of the psychological maturity assessment scale (PSYMAS). *European Journal of Psychological Assessment, 29*, 12-18.

Morren, M., & Meesters, C. (2002). Validation of the Dutch version of the aggression questionnaire in adolescent male offenders. *Aggressive Behavior, 28*, 87-96.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Paulhus, D. L., & Vazire, S. (2005). The Self-Report Method. In R. W. Robins, R. Fraley & R. F. Krueger (Eds.). *Handbook of research methods in personality psychology* (pp. 224-239). New York: Guilford Press.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*, 1137-1145.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 44*, 53-61.

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.

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health, *25*, 189-197.

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.

Salgado, J. F. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y las organizaciones [Personality and social desirability in organizational contexts: Implications for practice of psychology of work and organizations]. *Papeles del Psicólogo, 26*, 115-128.

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.

Soubelet, A., & Salthouse, T. (2011). Influence of social desirability on age differences in self-reports of mood and personality. *Journal of Personality, 79*, 741-762.

Ten Berge, J. M. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research, 34*, 89-102.

Tucker, L. (1951). A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984). *Washington D.C. Department of the Army*.

Vigil-Colet, A., Lorenzo-Seva, U., & Morales-Vives, F. (2015). The effects of ageing on self-reported aggression measures are partly explained by response bias. *Psicothema, 27*, 209-215.

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, U. (2013). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema, 25*, 100-106.

Vigil-Colet, A., Morales-Vives, F., & Lorenzo-Seva, U. (2013). How social desirability and acquiescence affect the age-personality relationship. *Psicothema, 25*, 342-348.

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*, 310-315.

3.3. Is general intelligence responsible for differences in individual reliability in personality measures?

Personality and Individual Differences xxx (2018) xxx-xxx

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com

ELSEVIER

Is general intelligence responsible for differences in individual reliability in personality measures?

David Navarro-González, Pere Joan Ferrando, Andreu Vigil-Colet*

Universitat Rovira i Virgili, Research Center for Behavior Assessment, Spain

ARTICLE INFO

Article history:
Received 31 January 2018
Received in revised form 16 March 2018
Accepted 17 March 2018
Available online xxx

Keywords:
Person reliability
Response bias
Personality differentiation hypotheses

ABSTRACT

One possible hypothesis for personality differentiation is the higher reliability of high-ability individuals in typical response measures. This differential reliability has been explained as resulting from different verbal abilities as a consequence of the difficulties that low-ability individuals have in understanding items, or as the effect of response bias, or due to higher precision in the answers of high-ability individuals. The lack of an estimation of individual reliability has made it difficult to test these hypotheses. However, recent psychometric advances have made it possible to measure person reliability and thus address the issue. The present study analyses the relationships between person reliability measures and the response bias of different personality measures in measurements of intelligence in a sample of 532 adolescents. The results show that person reliability is more closely related to general intelligence than to specific abilities and that the results for low-ability individuals cannot be explained by verbal deficits or by higher levels of acquiescence or social desirability. The differential reliability of measures across ability levels therefore seems to be related to higher levels of traitedness in high-ability individuals, i.e. traits are represented in them with greater strength and clarity.

1. Introduction

The potential interactions between intelligence and personality measures are a subject that has generated considerable controversy for many decades. These interactions do not refer directly to the relationships between personality and intelligence, but rather to a series of problems related to (a) the extent to which intelligence levels affect the factorial structure of personality measures or the relationships between personality dimensions, and (b) the possibility that the level of differentiation of abilities may depend on certain personality dimensions.

The issue summarized above was first reported by Shure and Rogers (1963), who found that the factor structure of personality scales differed as a function of individual levels of intelligence, and Eysenck and White (1964), who found a different factor structure of intelligence depending on individual levels of neuroticism. These types of result were later integrated into the personality differentiation hypothesis (PDH) framework developed by Brand, Egan, and Deary (1994). The PDH suggests that people with a higher level of ability have a more differentiated personality structure because they have more freedom to develop their personality, and this, results in greater distinction between them. If this hypothesis is true, then certain outcomes can be predicted when analysing the interactions between measures of personality and measures of ability. First we can expect a lack of factorial invariance when assessing the structure of personality measures across different intelligence levels, insofar as fewer dimensions will be needed to describe the personality structure of less intelligent individuals. Second, high-ability individuals will show greater variability in personality measures than low-ability individuals. Finally, we can expect a lack of invariance of ability measures across levels of personality due to different relationships between ability measures across levels of different personality dimensions such as neuroticism.

The above predictions have generated a considerable amount of research over the last 30 years, but so far the evidence in favour of the PDH is inconsistent. With respect to the first issue mentioned, certain studies have detected a lack of invariance in personality measures across intelligence levels (Allik, Laidra, Realo, & Pullmann, 2004; Mclarnon & Carswell, 2013) or different correlations between personality measures across ability levels (Austin et al., 2002). Others, however, have reported that personality remains essentially invariant (De Fruyt, Aluja, García, Rolland, & Jung, 2006; Waiyavutti, Johnson, & Deary, 2012) or that the correlations between personality measures were equal across ability levels (Austin, Deary, & Gibson, 1997).

With regard to the second prediction, different authors have reported an increased variance of personality scores among high-ability individuals, but only for some of the personality dimensions analysed. Austin et al. (1997), for instance, reported this effect only for openness and neuroticism, while Harris, Vernon, and Jang (2005) found an increased variance for three of the twenty dimensions of personality and De Fruyt et al. (2006) found increased variance only for neuroticism and extraversion. However, other studies have reported no differences in any dimension (Allik et al., 2004; Escorial, García, Cuevas, & Juan-Espinosa, 2006).

* Corresponding author at: Universitat Rovira i Virgili, Departament de Psicologia, Crtra. Valls s/n, 43007 Tarragona, Spain.
Email address: andreu.vigil@urv.cat (A. Vigil-Colet)

https://doi.org/10.1016/j.paid.2018.03.034
0191-8869/ © 2017.

Finally, regarding the lack of invariance of ability measures across levels of personality, Austin et al. (1997) and Austin, Hofer, Deary, and Eber (2000) found that the correlation between two intelligence measures increased as neuroticism increased, while Austin et al. (2002) found that the correlation between fluid and crystallized intelligence increased with the level of neuroticism. Nevertheless, Escorial et al. (2006) found no difference between the eigenvalues of the *g* factor across different levels of personality dimensions, and Bonaccio and Reeve (2006) reported that the structure of cognitive abilities remained invariant across neuroticism levels.

Overall, the results so far summarized suggest that, despite the inconsistencies, there is partial support for the predictions deriving from the PDH. However, a clear and univocal rationale for the results obtained is still lacking. Although the PDH suggests that more intelligent individuals have more differentiated personalities, there are other explanations that may account for these results. Different authors have reported that personality measures have varying amounts of reliability depending upon individual levels of ability and education, with high-ability groups showing higher reliability (Allik et al., 2004; Austin et al., 1997; McFarland & Sparks, 1985). This increase in turn is expected to result in both higher variability (because of the increase in true variance) and higher score correlations (because they become less attenuated by measurement error). In the end these stronger correlations are expected to impact the factor structure of the measures analysed. Taking this alternative explanation into account, Austin et al. (1997) suggested that the results associated with the PDH may be reflecting (a) a true personality differentiation, (b) a simple effect of differential reliability, or (c) a mixture of the two.

Different explanations have been put forward regarding the differential reliability associated with ability levels (DRAAL). These explanations mainly derive from the fact that the process of answering items requires a considerable amount of cognitive processing, and have therefore focused on issues such as the difficulties that low-ability individuals have in understanding certain items, differences at verbal ability level, and the presence of a "highly calibrated ruler" in high-ability subjects that enables them to give more meaningful responses (Austin et al., 1997, 2000). Other authors have suggested that the DRAAL may ultimately be related to differences in response styles between high and low-ability groups, i.e. groups may show different levels of faking, self-enhancement and/or acquiescence which may be the cause of differential reliability (Allik et al., 2004; Austin et al., 2000). So far, however, there has been little research relating response styles and intelligence. De Fruyt et al. (2006) found no relationship between intelligence and self-enhancement. Meanwhile acquiescence has been related to intelligence and low levels of education (Meisenberg & Williams, 2008) and has been proved to have a considerable impact on the factor structures of personality insofar as the number of factors extracted in a personality test varies depending upon whether or not acquiescence effects are removed (Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016; Rammstedt & Farmer, 2013; Soto, John, Gosling, & Potter, 2008). These results may partly explain the effects described in the PDH because, if low ability individuals have higher levels of acquiescence and these effects are not removed, then different factor structures for these individuals are expected to arise.

Overall, as pointed out by Austin et al. (2000), the main problem is that it is difficult to disentangle which of the effects associated with the PDH are due to changes in personality structure across ability levels and which are due to other problems such as differential reliabilities on the sole basis of self-report results. At a group level, it is quite straightforward to assess whether the marginal reliability of personality scores is lower for the low-ability groups. However, assessing (a) the individual contributions to reliability, and (b) further potential relations to response bias indexes, verbal ability measures, etc. is not so simple. It is submitted here that a more finely-graded analysis that would enable points (a) and (b) above to be assessed would, in turn, enable the different explanations given for the DRAAL to be better investigated. This type of analysis, which is based on the concept of person reliability, is already feasible and is summarized below.

1.1. Person reliability

Conventional psychometric models for personality consider only a single parameter for each respondent: his/her level of the trait being measured. Implicitly, therefore, this modelling assumes that all individuals respond to the test with the same degree of consistency and accuracy. This view has been challenged for over 70 years (Coombs, 1948; Mosier, 1942) and the evidence in personality is also against it; some individuals respond to personality items with very high consistency, almost deterministically, whereas the responses of others are much more random. This differential degree of consistency has been labelled "person fluctuation", "person reliability" or "person discrimination" (Ferrando, 2007, 2009). Person reliability is the term we shall use here.

Ferrando (2007, 2008, 2013) proposes a comprehensive item response theory (IRT) model for assessing person reliability under a variety of response formats. Essentially, the proposal consists of a series of extended conventional IRT models with an extra parameter that functions as an individual slope or discrimination index, and which models the degree of response consistency. This parameter is bounded below by zero and has no upper bound. Values near zero imply that the way the individual responds is almost random, i.e. totally insensitive to the normative item ordering, whereas very high values imply an almost deterministic, Guttman-type responding.

Following Tellegen (1988), Ferrando (2007, 2009, 2013) conceptualized person reliability as a relevant individual-differences dimension to partly explain the behaviour of the individual responding to a test. Furthermore, and also in line with previous proposals (Markus, 1977; Tellegen, 1988), Ferrando hypothesized that this dimension was related to the degree of clarity and strength with which the trait was organized in the individual. Recent empirical evidence suggests that this interpretation is tenable, with person reliability measures being indicators of traitedness (LaHuis, Barnes, Hakoyama, Blackmore, & Hartman, 2017). More generally, applied research results suggest that person reliability estimates have certain relevance in personality assessments. They are directly related to measures of conscientiousness and impulsivity (Austin, Deary, Gibson, McGregor, & Dent, 1998; Ferrando, 2007; Ferrando, 2009) and they have also been shown to function as moderator variables in validity assessments, in the sense that stronger relevant validity relations have been found for the most reliable individuals (Ferrando, 2015).

1.2. Aims of the study

The feasibility of obtaining reliability estimates at individual level might enable us to answer some of the questions discussed above. Thus if the low marginal (i.e. mean) reliability values found in low-ability groups is due to a poor understanding of the item content by low-ability individuals, then we can expect the person reliability estimates to be more closely related to measures of verbal ability than to measures of fluid or general intelligence. In the present research we shall use different personality measures, some of which have been developed using a method proposed by Ferrando, Lorenzo-Seva, and Chico (2009), enabling not only content but also acquiescence (AQ) and social desirability (SD) scores to be obtained for each individual. Hence relationships between intelligence measures and response bias measures can also be directly assessed. If, as authors such as Allik et al. (2004) and Austin et al. (2000) have suggested, response biases are responsible for the DRAAL results, then substantial relations between intelligence measures and response biases should be expected.

2. Method

2.1. Participants

The sample consisted of 532 student volunteers (252 men and 280 women) from 8 different high schools in the province of Tarragona, with ages ranging from 11 to 18 years old ($M=14.75$ $SD=2.1$). The same sample was the community sample used as a control in a study comparing the personality and abilities of juvenile offenders and community adolescents (Duran-Bonavila, Vigil-Colet, Cosi, & Morales-Vives, 2017).

2.2. Measures

The Indirect-Direct Aggression Questionnaire (IDAQ) (Ruiz-Pamies, Lorenzo-Seva, Morales-Vives, Cosi, & Vigil-Colet, 2014). This test gives scores for physical aggression (PA), verbal aggression (VA) and indirect aggression (IA) factors and an overall aggression score. The items were chosen from an initial pool selected by a panel of judges from the best of existing aggression measures refined after two studies using exploratory factor analysis. The test was developed using a method to control social desirability and acquiescence and has a considerable effect on the scores and factor structure of aggressive behaviour self-reports (Navarro-González et al., 2016; Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, & Lorenzo-Seva, 2012). In addition to the three content factors, therefore, the test also gives scores for SD and AQ. The reliabilities of the factor score estimates derived from the IDAQ are appropriate: $r_{\theta\theta}=0.83$, $r_{\theta\theta}=0.77$ and $r_{\theta\theta}=0.78$ for PA, VA and IA respectively.

The Barratt Impulsiveness Scale-11 for children (Chahin, Cosi, Lorenzo-Seva, & Vigil-Colet, 2010; Cosi, Vigil-Colet, Canals, & Lorenzo-Seva, 2008). This is a self-report questionnaire for assessing impulsivity that is specifically designed for children and adolescents. The test gives scores for motor impulsivity (MI), non-planning impulsivity (N-PI) and cognitive impulsivity (CI). MI is related to a lack of inhibition and delay, N-PI is related to planning abilities and CI to the tendency to make quick cognitive decisions.

The Psychological Maturity Assessment Scale (PSYMAS; Morales-Vives, Camps, & Lorenzo-Seva, 2013). This questionnaire consists of three scales: work-orientation (WO), self-reliance (SR) and identity (ID). It is made up of 25 items: seven items for each scale and four social desirability items. The reliability of the factor score estimates derived from the total scale is $r_{\theta\theta}=0.82$, while the reliability of the subscale score estimates are $r_{\theta\theta}=0.71$ for WO, $r_{\theta\theta}=0.78$ for SR and $r_{\theta\theta}=0.77$ for ID. In addition to these content factors, the test also gives factor score estimates for SD and AQ.

The Inventory of Callous Unemotional Traits (ICU; Frick, 2004). This is a questionnaire specifically designed to evaluate the precursors of psychopathy in youth populations. We use the Spanish adaptation developed by López-Romero, Gómez-Fraguela, and Romero (2015), which consists of 24 items with a 4-point response format (0=never/almost never; 3=always/almost always) with reliabilities of $\alpha=0.76$, $\alpha=0.82$ and $\alpha=0.78$ for CA, UC and UE respectively.

Thurstone's Primary Mental Abilities Test (Cordero, Seisdedos, González, & de la Cruz, 1989). The subscales of Thurstone's test were verbal (PMA-V, $\alpha=0.91$), spatial (PMA-S, $\alpha=0.73$), numerical (PMA-N, $\alpha=0.95$), reasoning (PMA-R, $\alpha=0.92$) and word fluency (PMA-WF, $\alpha=0.73$). The test comprises fluid and crystallized intelligence scales.

Raven's Progressive Matrices Test (Raven, 1996). This can be regarded as a measure of fluid intelligence free of cultural bias. The test has a reliability of $\alpha=0.91$.

The information scale of the WAIS intelligence test for adults (Wechsler, 2003). This scale is an indicator of crystallized intelligence. The test has a reliability of $\alpha=0.84$.

2.3. Procedure

School approval and parental written informed consent were obtained before participation in the study. Participation was voluntary and no incentives were given. About 96% of the participants who were invited to take part in the study eventually did so. A professional psychologist administered the tests collectively. The participants were asked to volunteer to answer the inventories in their classroom. The questionnaires were anonymous and respondents had to provide only their gender and age.

2.4. Data analysis

Individual scores on general intelligence were maximum likelihood (ML) factor score estimates obtained from the first principal factor based on all the intelligence measures. As for the personality measures, person reliability score estimates for each individual were obtained separately for each measure. These were ML estimates obtained as proposed in Ferrando (2013). We should note that standard errors and confidence intervals for these estimates can be obtained analytically and are important when the aim of the study is individual assessment. Person reliability estimates were not computed for each scale but only for the overall measures, since they need a minimum of 20 items to reach stability. However, it should be pointed out that all the personality measures used here enable an overall score to be used, which implies that the assumption of a general factor running through all the items is tenable. SD and AC were computed using the program Psychological Test Toolbox (Navarro-González, Vigil-Colet, Ferrando, & Lorenzo-Seva, in press). Finally, the relationships between intelligence and personality measures were analysed using product-moment correlations.

3. Results

Scores for the ability measures were factor-analysed using maximum-likelihood extraction. Sampling adequacy was good (KMO=0.83). Only one factor had an eigenvalue greater than one, accounting for 43% of the variance. (See Table 1.)

Table 2 shows the relations between levels of ability and person reliability on the one hand, and the reliabilities of the personality measures on the other. In both 'g' factor and person reliability, high and low groups were obtained by using a median cut-off value. It can be seen that there is a slight tendency of high-ability individuals to show greater reliabilities, in the same sense as reported by Austin et al. (1997) and Allik et al. (2004). We observe a more noticeable ef-

Table 1
Loadings of ability scales on the first factor extracted by maximum likelihood used to estimate "g" factor scores.

Scale	Loading
WAIS (information)	0.601
PMA verbal	0.581
PMA spatial	0.524
PMA reasoning	0.698
PMA numeric	0.518
PMA word fluency	0.605
Raven	0.564

Table 2
Reliability of typical response measures for high and low-ability groups and person consistency groups. Cronbach's alpha for BIS and CFI and factorial reliability for IDAQ and PSYMAS. Between brackets 95% confidence interval for reliabilities.

	Ability		Person consistency	
	Low	High	Low	High
IDAQ	0.705 (0.652–0.753)	0.806 (0.772–0.838)	0.721 (0.671–0.767)	0.797 (0.761–0.830)
BIS	0.814 (0.780–0.844)	0.825 (0.794–0.854)	0.789 (0.751–0.823)	0.848 (0.820–0.873)
PSYMAS	0.725 (0.676–0.770)	0.750 (0.705–0.791)	0.693 (0.638–0.743)	0.786 (0.747–0.821)
ICU	0.709 (0.657–0.753)	0.739 (0.692–0.781)	0.705 (0.652–0.753)	0.747 (0.701–0.788)

fect when we compare the high and low person-reliability groups. In this case, as expected, the more inconsistent individuals have lower reliability coefficients.

Table 3 shows the product-moment correlations between ability measures and response bias as derived from the IDAQ and PSYMAS measures. As can be seen, SD seems to be mostly unrelated to ability measures in the sense reported by De Fruyt et al. (2006). However, there seems to be a noticeable inverse relationship between AC and ability, with about half of the relationships between ability measures and AC being significant, although the magnitude of the relationships is quite low. Despite the low magnitude of the correlation coefficients, this result may partly support the hypothesis that AC may be responsible for the lower reliability of the low-abilities group.

Person reliability was calculated from each of the personality measures. Table 4 shows the relationships between these person reliabilities and the various ability measures. There is a clear positive pattern of relationship between person reliability and ability found in all personality measures. This consistent result suggests that individuals with higher ability are more consistent when answering items in those

Table 3
Correlations of response bias in IDAQ and PSYMAS with ability measures.

	Acquiescence		Social desirability	
	IDAQ	PSYMAS	IDAQ	PSYMAS
WAIS (information)	**−0.157**	**−0.145**	**0.112**	−0.044
PMA verbal	**−0.160**	**−0.116**	0.077	−0.015
PMA spatial	0.040	−0.048	0.028	−0.035
PMA reasoning	−0.088	0.008	0.065	−0.004
PMA numeric	0.043	0.024	−0.035	−0.063
PMA word fluency	**−0.151**	**−0.123**	0.039	0.023
PMA total	**−0.145**	*−0.098*	0.055	−0.022
Raven	**−0.121**	*−0.087*	0.067	−0.034
G estimate	**−0.164**	**−0.114**	*0.091*	−0.023

p < .01 *p < .05.*

Table 4
Correlations between ability measures and person reliabilities computed from each personality measure ranked by the average correlation across personality measures.

	IDAQ	PSYMAS	BIS	ICU	Average
PMA verbal	**0.172**	**0.106**	0.04	**0.124**	**0.110**
PMA word fluency	**0.166**	**0.147**	0.068	0.08	**0.115**
PMA spatial	**0.143**	**0.147**	**0.104**	**0.154**	**0.137**
PMA numeric	**0.218**	**0.176**	**0.144**	**0.18**	**0.180**
WISC (information)	**0.284**	**0.202**	*0.095*	**0.144**	**0.181**
PMA reasoning	**0.202**	**0.167**	**0.166**	**0.218**	**0.188**
PMA total	**0.253**	**0.211**	**0.147**	**0.216**	**0.207**
Raven	**0.266**	**0.210**	**0.178**	**0.25**	**0.226**
G estimate	**0.308**	**0.245**	**0.171**	**0.246**	**0.243**

p < .01 *p < .05.*

questionnaires. The same table shows the mean correlation of personality measures and the different ability measures, ranking ability measures from the lowest correlation to the highest. It can be seen that the lowest correlations between person reliability and abilities were found for specific abilities such as verbal, word fluency, etc., while the highest correlations were found for overall measures of ability such as Raven's test or the "g" factor estimate. Finally, taking into account the potential relationships between AC and abilities as discussed above, we also computed the correlations shown in Table 4 controlling for AC only for IDAQ and PSYMAS (i.e. the tests that allow for this control). The magnitude of the correlation coefficients, however, remained mainly unaffected.

4. Discussion

Since Austin et al. (1997) suggested that results relating to the PDH may be partly or totally due to a differential reliability effect, little research has been conducted to test the possible reasons for this effect. Three possible causes that may either individually or jointly explain the effect have been proposed: differences in social desirability and/or acquiescence related to ability, difficulties for low-ability individuals at the verbal processing level, and the possibility that high-ability individuals are more accurate when self-assessing.

The results reported in this article tend to discard some of these "a priori" explanations and basically suggest that high-ability individuals tend to provide self-assessments that are more accurate. As for the response bias conjectures, almost none of the SD measures showed relationships with either specific ability measures or general intelligence, confirming the results found by De Fruyt et al. (2006). However, a low but consistent relation between AC and intelligence measures was obtained. This is consistent with results reported by Meisenberg and Williams (2008) and suggests that AC may, at least in part, be a candidate for explaining the DRAAL.

Perhaps the most relevant results are those relating to the person reliability estimates. As Austin et al. (2000) pointed out, the main problem in testing the possible causes of DRAAL is the difficulty in determining the contribution of each individual to the overall reliability of a personality measure. The person reliability indexes used here, however, enable the response consistency to be estimated at the individual level and the relations between reliability and relevant individual-difference measures to be assessed. In this study the results point to a clear relationship between individual consistency and ability, with high-ability individuals showing greater person response consistency. Furthermore, these relations were found for both specific abilities and general intelligence. An inspection of their magnitude suggests that consistency is more closely related to general intelligence than to specific abilities. The lowest relationships were found for verbal abilities, which would seem to discard the conjecture that DRAAL is reflecting difficulties in understanding the items at this level. Overall, our results seem to favour the hypothesis by Austin et al. (1997) in that high-ability individuals are able to give more meaningful responses to the items and provide more accurate self-assessment. More generally, high-ability individuals appear to have higher levels of traitedness, i.e. it seems that personality traits are more meaningful for them, with the traits being represented with greater strength and clarity.

The results reported here therefore also have consequences in the domain of traitedness. Previous studies have shown that individual differences in traitedness were related to personality dimensions such as conscientiousness or impulsivity (Austin et al., 1998; Ferrando, 2007; Ferrando, 2009). Our results suggest that traitedness is related not only to personality but also to ability levels, which have consequences in the development of personality measures. Perhaps the most relevant consequence is the importance of avoiding biased samples such as university students when developing these types of measures because their higher intelligence, and therefore their higher person reliability, may falsely inflate the overall test reliability.

Although the results reported above may help to clarify the causes of DRAAL, they cannot answer the question as to whether the PDH is mainly due to the DRAAL or whether it is only partly explained by this phenomenon. Further research is needed to answer this. One possibility may be to apply the same approach used in this paper to bigger samples and then analyse the structure of personality at different levels, equating the individuals within each ability level in person reliability. The expected result, if the PDH can indeed be explained by differential reliability effects, is that we should find a lack of factorial invariance of personality across ability levels when individuals are not equated, and factorial invariance when the analysis is performed with individuals equated in person reliability.

Acknowledgements

This work was supported by a grant from the Spanish Ministry of Economy and Competitivity (PSI2017-82307-P).

References

Allik, J., Laidra, K., Realo, A., Pullmann, H., 2004. Personality development from 12 to 18 years of age: Changes in mean levels and structure of traits. European Journal of Personality 18, 445–462.
Austin, E.J., Deary, I.J., Gibson, G.J., 1997. Relationships between ability and personality: Three hypotheses tested. Intelligence 25, 49–70.
Austin, E.J., Deary, I.J., Gibson, G.J., McGregor, M.J., Dent, J.B., 1998. Individual response spread in self-report scales: Personality correlations and consequences. Personality and Individual Differences 24, 421–438.
Austin, E.J., Deary, I.J., Whiteman, M.C., Fowkes, F.G.R., Pedersen, N.L., Rabbitt, P., … McInnes, L., 2002. Relationships between ability and personality: Does intelligence contribute positively to personal and social adjustment?. Personality and Individual Differences 32, 1391–1411.
Austin, E.J., Hofer, S.M., Deary, I.J., Eber, H.W., 2000. Interactions between intelligence and personality: Results from two large samples. Personality and Individual Differences 29, 405–427.
Bonaccio, S., Reeve, C.L., 2006. Differentiation of cognitive abilities as a function of neuroticism level: A measurement equivalence/invariance analysis. Intelligence 34, 403–417.
Brand, C., Egan, V., Deary, I.J., 1994. Intelligence, personality, and society: "Constructivist" versus "essentialist" possibilities. In: Detterman, D.K. (Ed.), Current topics in human intelligence. Vol. 4, Ablex, New Jersey, pp. 29–42.
Chahin, N., Cosi, S., Lorenzo-Seva, U., Vigil-Colet, A., 2010. Stability of the factor structure of Barrat's Impulsivity Scales for children across cultures: A comparison of Spain and Colombia. Psicothema 22, 983–989.
Coombs, C.H., 1948. A rationale for the measurement of traits in individuals. Psychometrika 13, 59–68.
Cordero, A., Seisdedos, N., González, M., de la Cruz, V., 1989. PMA. Primary mental abilities. TEA Ediciones, Madrid.
Cosi, S., Vigil-Colet, A., Canals, J., Lorenzo-Seva, U., 2008. Psychometric properties of the Spanish adaptation of the Barrat mpulsiveness scale 11-a for children. Psychological Reports 103, 336–346.
De Fruyt, F., Aluja, A., García, L.F., Rolland, J.P., Jung, S.C., 2006. Positive presentation management and intelligence and the personality differentiation by intelligence hypothesis in job applicants. International Journal of Selection and Assessment 14, 101–112.
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.
Escorial, S., García, L.F., Cuevas, L., Juan-Espinosa, M., 2006. Personality level on the big five and the structure of intelligence. Personality and Individual Differences 40, 909–917.
Eysenck, H.J., White, P.O., 1964. Personality and the measurement of intelligence. British Journal of Educational Psychology 24, 197–201.
Ferrando, P.J., 2007. A Pearson-type-VII item response model for assessing person fluctuation. Psychometrika 72, 25–41.
Ferrando, P.J., 2009. A graded response model for measuring person reliability. The British Journal of Mathematical and Statistical Psychology 62, 641–662.
Ferrando, P.J., 2013. A general linear framework for modeling continuous responses with error in persons and items. Methodology 9, 150–161.
Ferrando, P.J., 2015. The relevance of measuring individual discrimination in personality: A study with the test anxiety scale for children. Personality and Individual Differences 82, 253–257.
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.
Frick, P.J., 2004. The inventory of callous-unemotional traits. University of New Orleans, New Orleans, LA, Unpublished rating scale.
Harris, J.A., Vernon, P.A., Jang, K.L., 2005. Testing the differentiation of personality by intelligence hypothesis. Personality and Individual Differences 38, 277–286.
LaHuis, D.M., Barnes, T., Hakoyama, S., Blackmore, C., Hartman, M.J., 2017. Measuring traitedness with person reliabilities parameters. Personality and Individual Differences 109, 111–116.
López-Romero, L., Gómez-Fraguela, J.A., Romero, E., 2015. Assessing callous-unemotional traits in a Spanish sample of institutionalized youths: The inventory of callous-unemotional traits. Journal of Psychopathology and Behavioral Assessment 37, 392–406.
Markus, H., 1977. Self-schemata and processing information about the self. Journal of Personality and Social Psychology 35, 63–78.
McFarland, S.G., Sparks, C.M., 1985. Age, education, and the internal consistency of personality scales. Journal of Personality and Social Psychology 49, 1692–1702.
Mclarnon, M.J.W., Carswell, J.J., 2013. The personality differentiation by intelligence hypothesis: A measurement invariance investigation. Personality and Individual Differences 54, 557–561.
Meisenberg, G., Williams, A., 2008. Are acquiescent and extreme response styles related to low intelligence and education?. Personality and Individual Differences 44, 1539–1550.
Mosier, C.I., 1942. Psychophysics and mental test theory II: The constant process. Psychological Review 48, 235–249.
Navarro-González, D., Lorenzo-Seva, U., Vigil-Colet, A., 2016. How response bias affects the factorial structure of personality self-reports. 28, 465–470.
Navarro-González, D., Vigil-Colet, A., Ferrando, P.J., Lorenzo-Seva, U., 2018. Psychological test toolbox: A new tool to compute factor analysis controlling response bias. Journal of Statistical Software(in press).
Rammstedt, B., Farmer, R.F., 2013. The impact of acquiescence on the evaluation of personality structure. Psychological Assessment 25, 1137–1145.
Raven, J.C., 1996. Matrices progresivas. Escalas CPM Color y SPM General. TEA Ediciones, Madrid.
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. Spanish Journal of Psychology 17.
Shure, G.H., Rogers, M.S., 1963. Personality factor stability for three ability levels. The Journal of Psychology 55, 445–456.
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.
Tellegen, A., 1988. The analysis of consistency in personality assessment. Journal of Personality 56, 622–663.
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, 310–315.
Waiyavutti, C., Johnson, W., Deary, I.J., 2012. Do personality scale items function differently in people with high and low IQ?. Psychological Assessment 24, 545–555.
Wechsler, D., 2003. Wechsler intelligence scale for children-WISC-IV. Psychological Corporation, New York, NY.

4. Discussion

In order to achieve our first objective, we created the Psychological Test Toolbox program. The program was released in 2016, so it has been available on our website for more than two years, and it has been updated and modified to improve its functionalities. To analyze the impact and usage of the program, we cannot show the number of citations, since the paper is still in press. However, we have the metrics of the website traffic, which may be an indicator of researchers' interest. Since it was released in 2016, there have been 823 visits to the website, with a mean of 275 per year. Overall, we conclude that the first objective has been achieved, since we developed a user-friendly application that enables applied researchers to use our method to control biases.

To generate interest, we have shown the program to the scientific community on several occasions, and we have received feedback from colleagues, which has led to improvements. The program was presented on the 7th European Congress of Methodology, organized by the European Association of Methodology (Navarro-González, Vigil-Colet, Ferrando, & Lorenzo-Seva, 2016), in a software demonstration entitled: *"Psychological Test Toolbox software demonstration: A new tool to compute factor analysis controlling response bias"*. It was also presented as a workshop at the *"Cátedra de modelos y aplicaciones psicométricos"*, organized by the Universidad Autónoma de Madrid in 2016 and at the Spanish National Congress of Psychology in July 2017 as part of a symposium about new methods for controlling response biases in self-assessment (Navarro-González, 2017). Finally, the development of the method and the program was summarized last summer at the

International Meeting of the Psychometric Society (IMPS) at Columbia University in New York with the title: *"Controlling the impact of response biases: an FA-Based approach"* (Navarro-González, Vigil-Colet, Ferrando, & Lorenzo-Seva, 2018).

Since the Psychological Test Toolbox has been released, some researchers have been developing new instruments with the program: for example, the INventory of Callous-unemotional traits and Antisocial behavior (INCA, Morales-Vives, Cosi, Lorenzo-Seva, & Vigil-Colet, in revision) and MAturity in Youth Assessment Scale (MAYAS, manuscript in preparation).

We hypothesize that once the paper is finally published (Navarro-González, Vigil-Colet, Ferrando, & Lorenzo-Seva, in press), the overall interest in the program will grow and enable more researchers to benefit from the potential advantages of the proposed method.

Our second objective was to illustrate the the two main uses of the response bias control procedure implemented in the Psychological Test Toolbox. The first use (Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016) is that it can determine the impact that SD and AC have on the factor structure of two questionnaires. The results of our investigations demonstrated that they both have effects on the factor structures, but that these effects are of different magnitudes. In both questionnaires, the impact of AC was clearly bigger than the impact of SD, which is congruent with the findings of other studies that have analyzed the impact of AC on some personality inventories based on the Big Five Model (Meisenberg & Williams, 2008; Rammstedt et al. 2010; Soto et al., 2008). However, this issue has not been tested in other personality dimensions. In this regard, our results showed that AC has a similar impact on aggression measures.

Regarding the difference of the impact AC and SD have on the factor structure, we have a hypothesis that could explain why AC leads to bigger distortion: When two items are impacted by SD, their correlation may increase artificially because they share common variance attributable to SD, regardless of their content. So, if the content of both items is different but SD affects both, a correlation could be generated where none is expected, and if the content of both are related, their correlation could increase due to the impact of SD.

However, the impact of AC is more complex. AC can generate over- or underestimations in the relationships between items, which will alter the correlation between these items in a different way than SD. Furthermore, when two items are impacted by AC, AC can overestimate this correlation if they are keyed in the same direction or underestimate it if they are keyed in the opposite direction, regardless of their content. Therefore, AC affects not only the magnitude of the correlations between items but also their sign, which involves a greater distortion of the inter-item correlation matrix and will have a deeper impact on the factor structure, since FA is performed using this matrix.

We also showed that the impact of both biases is different for each test, especially when controlling for SD. When analyzing a Big Five personality test like the OPERAS, removing SD effects led to negligible improvement in the congruence of the factor solution (congruence values between .872 and .874), which indicates that SD hardly distorted the factor structure at all. However, in the case of IDAQ, the improvement was considerable (congruence values between .475 and .864), which can be explained by the content being measured: IDAQ measures aggressive behavior, which is expected to be widely rejected in society (Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, & Lorenzo-Seva, 2012). This hypothesis was also supported by the mean loadings on the SD factor for both questionnaires ($\lambda_m = .232$ for IDAQ and $\lambda_m = .177$ for OPERAS). Therefore, it is possible that SD distorts aggression measures more than overall personality measures because they are more impacted by SD, which could explain the differences observed.

Our findings support the hypothesis that AC is the main source of distortion and that, when controlling response biases, the interpretability and simplicity of the structure of the questionnaires is improved. Furthermore, the distortion due to AC appears to be even greater than expected, because in the case of IDAQ, when AC is not controlled, the factor structure is completely incongruent, and even causes the verbal aggression factor to disappear.

Our last objective was to investigate the possible role of response biases in the DRAAL, which is a phenomenon that could explain the PDH. The results of our investigation (Navarro-González, Ferrando, & Vigil-Colet, 2017) showed that none of the SD measures showed a relationship with either specific ability or intelligence measures, which is consistent with the findings of De Fruyt et al. (2006). This is especially significant if it is borne in mind that in our study we used several measures of SD and AC from different questionnaires. However, a small but consistent inverse relation between AC and intelligence measures was found, which is consistent with the results reported by Meisenberg and Williams (2008). Of the

18 relationships between ability measures and AC 12 were significant, although the mean correlation was r=-.125, which shows a small-to-moderate effect. Regarding our hypothesis, only AC showed any relationship with intelligence, so we cannot conclude that response biases explain the DRAAL, but the results suggest that AC may partially explain the DRAAL. The main finding of this study was the relationship between ability and individual consistency, obtained using person reliability indices (Ferrando, 2007, 2009, 2013). These indices are a measure of the individual's consistency when answering a group of items. The lowest relationships were found for verbal abilities, which suggests that the conjecture that DRAAL reflects difficulties in understanding the items should be discarded. Furthermore, the results suggest that consistency is more related with general intelligence than specific abilities.

Recent studies have shown that person reliability indices are related to traitedness, which is consistent with Austin et al. (1997), who propose that high-ability individuals appear to have better internal rule, and are more able to assess their own personality, so personality traits are more meaningful for them, with traits being represented with greater strength and clarity.

Nevertheless, the sample size does not enable us to separate individuals with high and low levels of ability and equivalent levels of person reliability so we cannot test whether PDH effects disappear in these individuals. In this regard, we conducted another study with N=8000, which is currently under review. The results showed that the factorial structure of the NEO-PI-R test showed that dimensionality was a function of ability levels. However, this effect was not found when the analyses were performed on individuals with the same person reliabilities. The results related to the factorial structure are consistent with the hypotheses that personality differentiation across ability levels may reflect a question of differential reliability.

Considering the impact of person reliability on personality research, we are seriously considering implementing this kind of index in a new version of the Psychological Test Toolbox: our software would allow researchers to control not only the impact of SD and AC, but also to assess the reliability of individuals, which seems to be a great source of variance.

To sum up, the scientific community now has a tool to apply the response bias control method proposed by Ferrando, Lorenzo-Seva and Chico (2009), and we also provide more evidence about the impact of SD and AC in two specific situations. We encourage researchers to consider the impact of these biases, especially in the development phase of any self-assessed inventory and, for those interested, the Psychological Test Toolbox could be an easy way to control their impact.

5. References

Allik, J., Laidra, K., Realo, A., & Pullmann, H. (2004). Personality development from 12 to 18 years of age: Changes in mean levels and structure of traits. *European Journal of Personality*, 18, 445-462.

Austin, E. J., Deary, I. J., & Gibson, G. J. (1997). Relationship between ability and personality: Three hypothesis tested. *Intelligence, 25*, 49-70.

Austin, E. J., Hofer, S. M., Deary, I. J., & Eber, H. W. (2000). Interaction between intelligence and personality: Results from two large samples. *Personality and Individual Differences*, 29, 405-427.

Barnette, J, J. (2000). Effects of stem and likert response option reversals on survey internal consistency: if you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361-370.

Barrett, P.T., Petrides, K.V., Eysenck, S.B.G., & Eysenck, H.J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25*, 805–819.

Bass, B. M. (1956). Development and evaluation of a scale for measuring social acquiescence. *Journal of Abnormal and social Psychological Measurement*, 26, 291-309.

Bentler, P. M. (1977). Factor simplicity index and transformations. *Psychometrika, 42*(2), 277-295.

Bentler, P., Jackson, D., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*, 186-204.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 6*, 431-444.

Block, J. (1965). The challenge of response sets. New York: Appleton-Century Crofts.

Boada-Grau, J., Prizmic-Kuzmica, A. J., Serrano-Fernández, M. J., & Vigil-Colet, A. (2013). Estructura factorial, fiabilidad y validez de la escala de adicción al trabajo (WorkBAT): Versión española. *Anales de psicología, 29*(3), 923-933.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*(2), 213-225.

Bramble, W. J., Wiley, D. E. (1974). Estimating content-acquiescence correlation by covariance structure analysis. *Multivariate Behavioral Research, 9*, 179-190.

Brand, C., Egan, V., & Deary, I. J. (1994). Intelligence, personality, and society: “Constructivist” versus “essentialist” possibilities. In D. K. Detterman (Ed.), Current topics in human intelligence, 29-42. New Jersey: Ablex.

Browne, M. (1972a). orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, *25*, 115-120.

Browne, M. (1972b). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, *25*, 207-212.

Burns, G. N. & Christiansen, N. D. (2011). Methods of Measuring Faking Behavior. *Human Performance, 24*, 358-372.

Cattell, R. B. (1944). Psychological measurement: normative, ipsative, interactive. Psychol. Rev., 51, 292-303.

Chahin, N., Cosi, S., Lorenzo-Seva, U., Vigil-Colet, A. (2010). Stability of the factor structure of Barrat's Impulsivity Scales for children across cultures: A comparison of Spain and Colombia. *Psicothema, 22*, 983-989.

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research, 22*, 214-247.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ispative measures. *Psychometric Monographs, 14*.

Cordero, A., Seisdedos, N., González, M., & de la Cruz, V. (1989). PMA. Primary mental abilities. Madrid: TEA Ediciones.

Cosi, S., Vigil-Colet, A., Canals, J., & Lorenzo-Seva, U. (2008). Psychometric properties of the Spanish adaptation of the Barrat impulsiveness scale 11-a for children. *Psychological Reports, 103*, 336-346.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 60, 151–174.

Cupani, M., & Lorenzo-Seva, U. (2016). The development of an alternative IPIP inventory measuring the Big-Five factor markers in an Argentine sample. *Personality and Individual Differences, 91*, 40-46.

De Fruyt, F., Aluja, A., García, L. F., Rolland, J. P., & Jung, S. C. (2006). Positive presentation management and intelligence and the personality differentiation by intelligence hypothesis in job applicants. *International Journal of Selection and Assessment, 14*, 101-112.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531-545.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, *76*, 354-362.

Edwards, A. L. (1957). The Social Desirability Variable in Personality Assessment and Research. New York: Dryden Press.

Edwards, A. L. (1967). The social desirability variable: A broad statement. In I. A. Berg (Ed.), Response set in personality assessment (pp. 32-47). Chicago: Aldine.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*, 122–133.

Eysenck, H. J., & Eysenck, S. B. G. (1976). Psychoticism as a dimension of personality. New York: Crane-Russak

Ferrando, P. J. (2007). A Pearson-type-VII item response model for assessing person fluctuation. *Psychometrika, 72*, 25–41.

Ferrando, P. J. (2009). A graded response model for measuring person reliability. *The British Journal of Mathematical and Statistical Psychology, 62*, 641–662.

Ferrando, P. J. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology, 9*, 150–161.

Ferrando, P. J., & Condon, L. (2004). The convergent validity of acquiescence: an empirical study relating balanced scales and separate acquiescence scales. *Personality and Individual Differences, 37*, 1331-1340.

Ferrando, P. J., & Condon, L. (2006). Assessing acquiescence in binary responses: IRT-related ítem-factor analytic procedures. *Structural Equation Modeling, 13*, 96-115.

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

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling, 16*, 364-381.

Frick, P. J. (2004). The inventory of callous-unemotional traits. New Orleans, LA: University of New Orleans Unpublished rating scale.

Greenwald, H. J., & Clausen, J. D. (1970). Test of relationship between yeasaying and social desirability. *Psychological Measurement, 4*, 43-55.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. *Psychometrika, 47*(2). 209-222.

Haladyna, T. M., & Rodríguez, M. C. (2013). Developing and validating test ítems. New Tork, NY: Routledge.

Hanley, C. (1961). Social desirability and response bias in the MMPI. *Journal of Consulting Psychology*, 25, 13–20.

Hofstee, W. K. B., ten Berge, J. M. F., & Hendriks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences, 25*, 897-909.

Holden, R. (2010). Social Desirability. In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 17

Jackson, D. N., & Messick, S. (1962). Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition* (pp. 129-155). New York: Wiley.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Korth, B., & Tucker, L. R. (1976). “Procrustes matching by congruence coefficients. *Psychometrika, 41*(4), 531-535.

Krosnich, J. A., & Fabrigar, L. R. (1998). *Designing good questionnaires: Insights from cognitive and social psychology.* New York: Oxford University Press.

Larson, R. B. (2018). Controlling social desirability bias. *International Journal of Market Research*, 1-14.

Leite, W. L., & Cooper, L. A. (2010). Detecting Social Desirability Bias Using Factor Mixture Models. *Multivariate Behavioral Research*, 45, 271-293.

Lönnqvist, J. E., Paunonen, S. V., Tuulio-Henriksson, A., Lönnqvist, J., & Verkasalo, M. (2007). Substance and style in socially desirable responding. *Journal of Personality, 75*, 291–322.

López-Romero, L., Gómez-Fraguela, J. A., & Romero, E. (2015). Assessing callous-unemotional traits in a Spanish simple of institutionalized youths: The inventory of callous-unemotional traits. *Journal of Psychopathology and Behavioral Assessment, 37*, 392-406.

Lorenzo-Seva, U. (1999). Promin: a method for oblique factor rotation. *Multivariate Behavioral Research*, *34*, 347-356.

Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika*, 68, 49-60.

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*, 319-326.

Lorenzo-Seva, U., & Rodríguez-Fornells, A. (2006). Acquiescent responding in balanced multidimensional scales and exploratory factor analysis. *Psychometrika, 71*, 769-777.

Lorenzo-Seva, U. & ten Berge, J.M.F. (2006). Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity. *Methodology*, *2*, 57-64.

Martin, J. (1964). Acquiescence-measurement and theory. *British Journal of Social and Clinical Psychology, 3*, 216-225.

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception: An Interdisciplinary Journal, 31*(2), 118-138.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11, 344-363.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882-888.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539-1550.

Messick, S. J. (1967). The psychology of acquiescence: an interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 115-145). Chicago: Aldine.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense agreement bias from measures of the sense of control. A 2 x 2 index. *Psychology Quarterly, 54*, 127-145.

Morales-Vives, F., Camps, E., & Lorenzo-Seva, U. (2013). Development and validation of the psychological maturity assessment scale (PSYMAS). *European Journal of Psychological Assessment, 29*, 12-18.

Morales-Vives, F., Cosi, S., Lorenzo-Seva, U., & Vigil-Colet, U. (under review). The Inventory of Callous-unemotional traits and Antisocial behaviour (INCA) for young people: Development and validation in a community sample. *Frontiers in Psychology.*

Morales-Vives, F., & Dueñas, J. M. (2018). Predicting suicidal ideation in adolescent boys and girls: the role of psychological maturity, personality traits, depression and life satisfaction. *The Spanish journal of psychology, 21*, 1-12.

Morales-Vives, F., Lorenzo-Seva, U., & A. Vigil-Colet (2017). How response biases affect the factor structure of Big Five personality questionnaires. *Anales de Psicología*, 33 (3), 589-596.

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

Morf, M. E., & Jackson, D. N. (1972). An analysis of two response styles: True responding and item endorsement. *Educational and Psychological Measurement, 30*, 647-661.

Navarro-González, D. (2017, July). El programa Psychological Test Toolbox en el desarrollo de cuestionarios autoinformados libres de sesgos. In A. Vigil-Colet (Chair). Aplicando nuevos métodos de control de sesgos de respuesta en autoinformes. Symposium conducted at the III Congreso Nacional de Psicología, Oviedo, Spain.

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.

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.

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva. (in press). Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias. *Journal of Statistical Software*.

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva. (2016, July). Psychological Test Toolbox: a new tool to compute factor analysis controlling response bias. Paper presented at the European Congress of Methodology, Palma de Mallorca, Spain.

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva. (2018, July). Controlling the impact of response biases: an FA-Based approach. Paper presented at the International Meeting of the Psychometric Society, New York, NY.

Neill, J. A., & Jackson, D. N. (1970). An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement, 30*, 647-661.

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

Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443-460.

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485-500.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Paulhus, D. L. (1981). Control of social desirability in personality inventories: Principal factor deletion. *Journal of Research in Personality, 15*, 383-388.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (vol. 1, pp. 17–60). New York: Academic Press.

Paulhus, D. L. (2017). Socially desirable responding on self-reports. Encyclopedia of Personality and Individual Differences, 1-5.

Paulhus, D.L., & Reynolds, S.R. (1995). Enhancing target variance in personality impressions: Highlighting the person in person-perception. *Journal of Personality and Social Psychology, 69*, 344–354.

Paulhus, D. L., & Vazire, S. (2005). The self-report method. In R. Robins, R.C. Fraley, R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology*. Guildford Press.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 45*, 121-125.

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.

Raven, J. C. (1996). Matrices progresivas. Escalas CPM Color y SPM General. Madrid: TEA Ediciones.

Ray, J. J. (1979). Is the acquiescent response style not so mythical after all? Some results from a successful balanced F scale. *Journal of Personality Assessment, 43*, 638-643.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology, 121*, 81-96.

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior, 25*, 189-197.

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

Saliba, J., Lorenzo-Seva, U., Marco-Pallares, J., Tillmann, B., Zeitouni, A., & Lehmann, A. (2016). French validation of the Barcelona Music Reward Questionnaire. *PeerJ, 4*, e1760.

Salgado, J. F. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y las organizaciones [Personality and social desirability in organizational contexts: Implications for practice of psychology of work and organizations]. *Papeles del Psicólogo, 26*, 115-128.

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

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.

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17). Convergent validity, discriminant validity and relationship with age. *European Journal of Psychological Assessment, 17*(3), 222-232.

Stricker, L. J. (1961). Some item characteristics that evoke acquiescence and social desirability response sets on psychological scales. Unpublished doctoral dissertation, New York University, New York.

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*(2), 149-158

ten Berge, J. M. F. (1998). Some recent developments in factor analysis and the search for proper communalities. In A. Rizzi, M. Vichi & H.-H. Bock (Eds.), Advances in data science and classification (pp.325-334). Heidelberg: Springer.

ten Berge, J. M. F. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative of ipsatization. *Multivariate Behavioral Research, 34*, 89-102.

ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. Psychometrika, 56, 309-315.

Ten Berge, J., Krijnen, W., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications, 289*, 311-318.

ten Berge, J. M. F., Snijders, T. A. B. & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. *Psychometrika, 46*, 201-213.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. *Psychological Methods*, *16*, 209-220.

Torrent-Sellens, J., Ficapal-Cusí, P., & Boada-Grau, J. (2016). Dispositional Employability and Online Training Purchase. Evidence from Employees' Behavior in Spain. *Frontiers in psychology*, 7, 831.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

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).

Vigil-Colet, A. (2014). ¿Es deseable controlar los sesgos de respuesta en los autoinformes de personalidad? Invited conference at the III Jornadas de la Asociación Iberoamericana de las diferencias individuales, Barbastro, Spain.

Vigil-Colet, A., Lorenzo-Seva, U., & Morales-Vives, F. (2015). The effects of ageing on self-reported aggression measures are partly explained by response bias. *Psicothema, 27*(3), 209-215.

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, U. (2013). Desarrollo y validacion de las Escalas de Evaluacion Global de la Personalidad (OPERAS*). Psicothema, 25*(1), 100-107.

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*, 310-315.

Wechsler, D. (2003). Wechsler intelligence scale for children-WISC-IV. New York, NY: Psychological Corporation.

Wiggins, J. S. (1964). Convergences Among Stylistic Response Measures from Objective Personality Tests. *Educational and Psychological Measurement, 24*(3).

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioural Assessment, 28*(3), 189-194.

Ziegler, M., & Buehner, M. (2009). Modeling Socially Desirable Responding and Its Effects. *Educational and Psychological Measurement, 69*(4), 548-56

UNIVERSITAT
ROVIRA i VIRGILI