*Psicológica (2011), 32*, 87-105.

# The impact of acquiescence on forced-choice responses: A model-based analysis

Pere J. Ferrando*, Cristina Anguiano-Carrasco, & Eliseo Chico

*'Rovira i Virgili' University, Spain*

The general aim of the present study is to assess the potential usefulness of the normative Forced Choice (FC) format for reducing the impact of acquiescent responding (AR). To this end it makes two types of contributions: methodological and substantive. Methodologically, it proposes a model-based procedure, derived from a basic response mechanism, for assessing the impact of AR. Substantively, it applies the procedure in three large datasets, which use well-known normative FC personality questionnaires: Rotter's Locus of Control Scale, the Sensation Seeking Scale (form V), and Vando's Reducer-Augmenter Scale. The results suggest that the impact of AR is minimal except for the Locus of Control Scale, so FC-based questionnaires are a good alternative for controlling acquiescence. The different results obtained with the LOC Scale might be explained for some particularities of this scale which are discussed in the paper.

In recent decades in Industrial and Organizational Psychology there has been renewed interest in the use of personality measures as predictors (e.g. Barrick & Ryan, 2003). This interest, in turn, has given a new lease of life to certain research topics which had been quite active during the 1950s, 60s and 70s, but which had then been consigned to oblivion. It is widely acknowledged that the main weakness of personality measures is the potential distortive effects of response biases, mainly social desirability (SD) and acquiescent responding (AR), on the test scores (e.g. Paulhus, 1991). So, most of the research in applied personality measurement, both

* **Acknowledgments:** This research was partially supported by a grant from the Spanish Ministry of Science and Technology (SEJ2005-09170-C04-04/PSIC) with the collaboration of the European Fund for the Development of Regions, and by a grant from the Catalan Ministry of Universities, Research and the Information Society (2005SGR00017). **Correspondence**: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicología. Carretera Valls s/n. 43007 Tarragona (Spain). E-mail: perejoan.ferrando@urv.cat

past and present, has focused on assessing the effects of response biases on real data, and on the procedures for controlling them or, at least, minimizing their effects.

The forced-choice (FC) format was introduced at the beginning of the 1950s precisely as an attempt to reduce SD and AR (Guilford, 1954; Gordon, 1951; Zavala, 1965). Generally speaking, a binary FC item consists of a pair of statements of which the respondent is asked to select one. However, this general type of item can be used within two very distinct measurement modelling frameworks: ipsative and normative. In the ipsative framework, every item consists of two statements, with matching SD values, each of which measures a different trait or dimension (e.g. Edwards, 1970). The type of measurement provided by these items is not amenable to conventional item or factor analysis, and requires a specific methodology which is not free from problems (Edwards & Abbot, 1973; Johnson, Wood, & Blinkhorn, 1988). We shall not consider ipsative FC items in this paper.

In the normative measurement framework, both statements measure the same trait or dimension. Furthermore, they are usually intended to have different locations on the continuum of the trait that is measured (Hicks, 1970). Finally, as far as possible, both statements are also matched on SD. However, matching on SD must be more difficult in this case because, in a well-designed item, the statements will be at the positive and negative trait extremes. And one of the extremes might be more desirable than the other. The Myers-Briggs Type Indicator (Myers, 1962), Zuckerman's Sensation Seeking Scales (1996) and Rotter's Locus of Control Scale (LOC) (1966) are examples of widely used personality measures that use the normative FC framework. This is the framework we shall consider in this paper.

Empirical evidence on the success of FC items at reducing response bias has mainly focused on SD, and the results based on the normative framework are not conclusive. Some studies found that SD is not well controlled by the FC format (Feldman & Corah, 1960; Christiansen, Burns & Montgomery, 2005), whereas in others the SD influence was clearly minimized (Saltz, Reece & Ager, 1962; Jackson, Wroblewski & Ashton, 2000). As far as AR is concerned, the literature is very scarce, and empirical evidence is almost nonexistent (but see Berkowitz & Wolkon, 1964, and Mukherjee, 1969). Even so, two main positions can be distinguished from the literature review. On the one hand, some authors (Ford, 1964) consider that the FC items are intrinsically free from acquiescence because the respondents choose between two statements and not between "true/agree" and "false/disagree". On the other hand, Ray (1989), and Schuman and Presser, (1981) suggested that, with FC items, the tendency to agree might simply become a tendency to choose or agree with whatever statement is

presented first. At present, there seems to be no clear empirical evidence for or against either position so they should be regarded mainly as conjectures. This lack of evidence may be due to the lack of an appropriate modelling framework that allows a clear empirical assessment to be made.

In our opinion, assessing the potential impact of AR on FC items is an issue of clear practical relevance. It has been recently suggested (Morgeson et al. 2007, Smyth, Dillman, Christian & Sterns, 2006) that FC is a promising alternative for developing personality measures which are more resistant to biases and, ultimately, better predictors. However, before efforts are taken to develop FC measures, it should first be empirically verified that they are superior to traditional measures in this respect. This point is even more relevant if it is taken into account that developing good FC scales is usually far harder than developing conventional binary or Likert scales (Ray, 1989). Second, if FC items are resistant to AR there would be no need to balance FC scales. Generally achieving a well balanced scale is a complex and difficult task (e.g. Ray, 1989).

The present paper proposes a model-based procedure for empirically assessing the appropriateness of the conjectures described above, and, more generally, for assessing the impact of AR on FC items. Furthermore, the procedures are illustrated with three real-data studies based on well known FC personality scales.

The next section describes the modelling framework and the rationale for the procedure we propose. It has often been claimed that dominance-based psychometric models such as factor-analytic (FA) models, conventional item-response-theory (IRT) models and classical-test-theory models are not suitable for analyzing FC scales (e.g. Guilford, 1954; Nunnally, 1978). In our opinion, most of this criticism is unjustified, and is the result of confusing FC with ipsative measurement. Even so, we believe that we should first provide a clear modelling rationale based on a response mechanism, which will serve as a basis for the procedure we propose. We also note that several FA and IRT applications with normative FC items give acceptable fits and meaningful results (e.g. Harvey & Murry, 1994, Steinberg & Thissen, 1995).

## The Model and Rationale

Consider a test made up of $n$ normative FC items. We shall start by assuming that the behaviour, belief or feeling evoked by item $j$ elicits a value of an underlying variable (UV) of response strength, with a standard normal distribution, and denoted by $X_j^*$. For the reasons discussed below, we assume that this response variable is governed by two common factors:

$\theta_1$ and $\theta_2$. Let $X^*_{ij}$ be the response-strength value elicited when participant $i$ responds to item $j$. The UV FA model is:

$$X^*_{ij} = \alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2} + \varepsilon_{ij} \qquad (1)$$

where the $\alpha$s are the factor loadings, and the $\varepsilon$s are the measurement errors, with zero means, and uncorrelated with the factors. Both factors are scaled in a $z$-score metric (mean 0 and variance 1). For fixed $\theta_1$ and $\theta_2$ values, the error is normally distributed, with constant variance $\sigma^2_{\varepsilon j}$. Also, for fixed $\theta_1$ and $\theta_2$ and for any pair of items $j$, $k$, the $X^*_j$ and $X^*_k$ responses are independent (i.e. local independence). We note that the factor-analytic model (1) is a general bidimensional UV model that can be used with different kinds of items. The specific application to FC items depends on the particular threshold formulation that is described below. In more detail, it depends on the relation between the UV and the observed responses.

Each statement of item $j$, say $A_j$ and $B_j$, is characterised by a fixed location or threshold on the continuum of $X^*_j$, denoted by $\beta j_A$ and $\beta j_B$. In fact, from the discussion above, we can assume that in a well-designed FC item both locations will be clearly separated. Denote now by $\beta j=(\beta_{jA}+\beta_{jB})/2$, the midpoint between the two locations (see Figure 1).

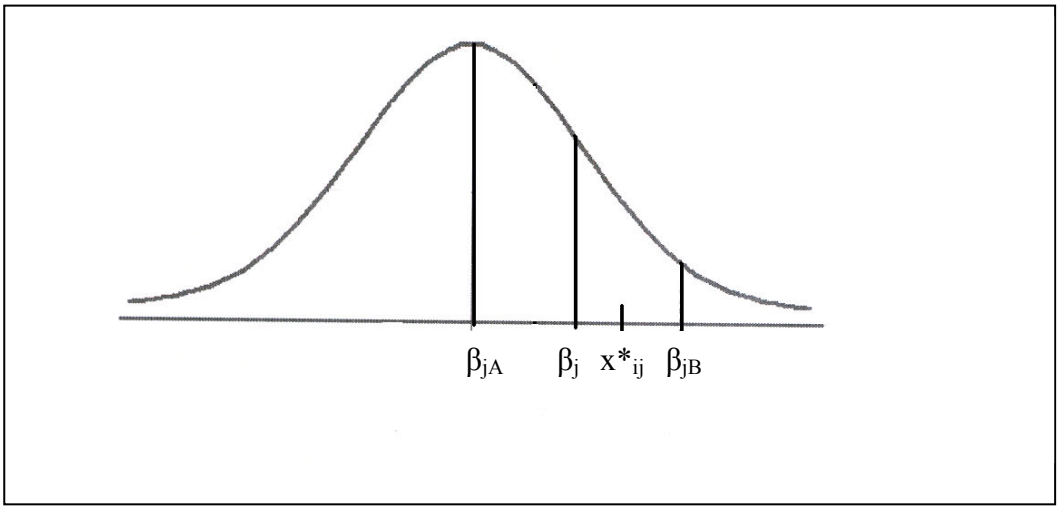


**Figure 1: the response mechanism for a normative FC item.**

The response mechanism assumed here was initially proposed by Coombs (1948). Respondent $i$ compares his or her elicited response-strength value $X^*_{ij}$ to both statements simultaneously, and chooses the statement whose threshold is nearest this value. Overall, according to all our assumptions, the conditional probability of choosing statement $\beta_j$ for fixed $\theta_1$ and $\theta_2$ values is given by (see figure 1):

$$P(X_{ij} = B_j \mid \boldsymbol{\theta}_i, \sigma_{\varepsilon j}, \beta_j) = P(X^*_{ij} > \beta_j \mid \boldsymbol{\theta}_i, \sigma_{\varepsilon j}, \beta_j) = \Phi\left(\frac{\alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2} - \beta_j}{\sigma_{\varepsilon j}}\right) \quad (2)$$

where the notation $\Phi(X)$ is used for the cumulative distribution function of a standard normal variable. Equation (2) defines the Item Characteristic Surface of the bidimensional two-parameter normal-ogive model (2PNOM). The model can be defined using the FA parameterization (2) or an IRT parameterization (see Takane & de Leeuw, 1987). It also can be closely approximated by the corresponding logistic model.

The developments discussed so far, can be considered as a bidimensional extension of a model proposed by Ferrando (2006; Ferrando used a Thurstonian framework, while here we use a FA framework). They provide a rationale for the calibration of normative FC items with the bidimensional 2PNOM or with its logistic counterpart. Note that in this calibration the item parameter $\beta_j$ would be the midpoint between the locations of the statements.

The model in (1) and (2) can be fitted by a variety of IRT and FA based procedures. The simplest and oldest procedure, which Bock and Lieberman (1971) called the heuristic approach, consists of fitting the inter-item tetrachoric correlation matrix using the centroid method of FA (Lawley, 1955). This is the approach we shall use here to derive the main results needed for our procedure. Indeed, we acknowledge that modern estimation procedures are superior in many respects to the simple centroid. In particular, they are more efficient, and provide statistics that allow goodness of model-data fit to be assessed. However, modern procedures are all iterative and complex, and deriving direct and simple results from them can be a cumbersome and practically impossible task. In a related series of analyses of conventional items, Ferrando and Lorenzo-Seva (2009) used the simple centroid approach to derive the main results that were required, and then showed that modern estimation procedures arrived at virtually the same results predicted by the centroid approach. We shall use here the same strategy. First, we use the simple centroid approach to derive the main

results that are required, and next, we show that modern estimation procedures should arrive at virtually the same results that are predicted from the centroid approach.

Consider now that the $n$ items ($n$ even) form a balanced scale which aims to measure a single dimensional personality trait with two ends or poles (e.g. extravert-introvert or high vs. low stability). In half of the items, the first statement (say A) measures the upper end of the trait and the second statement (B) measures the lower end. In the other half, the roles of A and B are reversed. Regardless of the keying, each item is scored 1 if the respondent chooses statement A, and 0 if he/she chooses B. The analysis of this scale will be based on the bidimensional model (2). We shall assume that factor $\theta_1$ is the "content" dimension that the scale aims to measure, and that $\theta_2$ is an acquiescence factor, understood as in Ray's (1989) hypothesis, and conceptualized here as an individual-differences variable of propensity to choose the first statement regardless of the item content. Given this conceptualization, we shall assume that $\theta_1$ and $\theta_2$ are independent (i.e. orthogonal factors).

Under the scoring schema described above, half of the loadings corresponding to $\theta_1$ in equation (1) are expected to be positive, and the other half negative. On the other hand, all of the loadings on $\theta_2$ are expected to be positive. This is because, for all of the items, agreement with the first statement will lead to a higher item score. Furthermore, we shall adopt the so called "weak assumption of balance" (Ferrando & Condon, 2006, Miller & Cleary 1993). In this case, this means that the sum of the loadings on the "content" factor is expected to be zero (i.e. that the sum of the positive and negative loadings cancels each other out).

As discussed above, to derive the results which are needed for the procedure we propose, we shall fit model (2) by using the centroid analysis of the reduced tetrachoric-correlation matrix with communalities in the main diagonal (as in Lawley, 1955). Also, to keep them simple, the results that follow will be derived in the population, and sampling issues will not be considered.

From equation (1) it follows that, for a fixed item $j$, the sum of the tetrachoric correlations with all of the test items is

$$\sum_{k=1}^{n} \sigma_{jk} = \alpha_{j1} \sum_{k=1}^{n} \alpha_{k1} + \alpha_{j2} \sum_{k=1}^{n} \alpha_{k2} \ . \qquad (3)$$

It is noted that in (3) we assume that the correlation of item *j* with itself is its communality. If the weak assumption of balance is met, the first sum on the right hand side of the equal sign vanishes, and equation (3) will reduce to

$$\sum_{k=1}^{n} \sigma_{jk} = \alpha_{j2} \sum_{k=1}^{n} \alpha_{k2} \, . \qquad (4)$$

So

$$\sum_{j=1}^{n} \sum_{k=1}^{n} \sigma_{jk} = (\sum_{j=1}^{n} \alpha_{j2})^2 . \qquad (5)$$

Now, the centroid loading $\alpha_j$ for the first common factor is obtained as (e.g. Lawley, 1955)

$$\alpha_j = \frac{\sum_{k=1}^{n} \sigma_{jk}}{\sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} \sigma_{jk}}} \qquad (6)$$

By using results (4) and (5), in equation (6), it follows that

$$\alpha_j = \alpha_{j2} \, . \qquad (7)$$

Thus, under the assumptions considered here, the factor which is estimated by the first centroid would be the acquiescence factor in our proposed bidimensional model. And the item loadings on this factor would be unbiased estimates of the 'true' loadings on this acquiescence factor.

From equation (1) and (7) it follows that the expected values of the elements in the residual correlation matrix after the first centroid has been extracted would be

$$res_{jk} = \alpha_{j1}\alpha_{k1}. \qquad (8)$$

A second centroid could not be directly fitted to the residual correlation matrix (8) because the expected values of the sums in (6) would be zero. So, the usual practice of reversing the signs of the negative correlations and restoring the signs in the final loading estimates should be used. As well as this, the same rationale we used in equations (3) to (7) shows that the second centroid would correspond to the "content" factor in our proposed model, and that the loadings on this factor would be unbiased estimates of the corresponding "true" loadings.

Overall, the rationale of the developments described so far can be summarized as follows. First, the basis is a sequential two-step FA. Second, under certain conditions (unidimensionality and weak balance) the first-step centroid factor is an estimate of the (probably secondary) acquiescence factor. Third, the acquiescence factor is partialled-out from the correlation matrix and the second-step centroid is extracted from the residual correlation matrix. This second factor is an estimate of the (probably dominant) content factor.

The sequential approach is didactic and useful for predicting results. However, the same results can be obtained in a single step by estimating a bidimensional solution in the canonical form (i.e. each successive factor accounts for as much variance as possible, see Ferrando & Lorenzo-Seva, 2009). Furthermore, the results derived from the centroid are expected to hold essentially for modern FA estimation procedures (particularly unweighted and weighted least squares and maximum likelihood). This is for two reasons. First, least squares and maximum likelihood methods are all based on the general criterion of minimizing the (unweighted or weighted) sum of the squared discrepancies between the observed and reproduced inter-item correlation. This criterion, in turn is equivalent to maximizing the sum of the squared loadings on each successive factor, and the centroid approximates this criterion because it maximizes the sum of the absolute loadings (Choulakian, 2003). Second, the centroid loading estimates are consistent (Lawley, 1955). Given that the least-squares, ML, and centroid procedures optimize essentially the same function and are consistent, it follows that in a reasonably large sample they should essentially converge to the same solution (Lawley, 1955). This was the result obtained by Ferrando and Lorenzo-Seva (2009) with standard items.

Provided that a good balanced scale is available, the results discussed so far allow the impact of AR on FC items to be empirically assessed. More

specifically, they provide tools for testing the two main conjectures discussed above. If AR has a negligible impact on the item responses, a unidimensional model should show a good fit to the data. Furthermore, the goodness of fit is not significantly improved if a bidimensional model is fitted to this data. After fitting the unidimensional model, the resulting factor should be the "content" factor, and the pattern should be bipolar. The positively keyed items should show positive loadings, the negatively keyed items negative loadings, and the sum of the loadings should be near zero (i.e. the weak condition of balance).

If the impact of AR is nontrivial: (a) the unidimensional model should not fit the data that well, and (b) fitting the bidimensional model should noticeably improve the fit to the extent that it becomes acceptable. One of the canonical factors (possibly the dominant factor) should be the "content" factor, and its pattern should exhibit the bipolar characteristics discussed above. The other canonical factor should be the acquiescence factor, and its pattern should exhibit positive manifold (using Spearman's terms). The loadings should all be positive and generally significantly different from zero.

## EMPIRICAL STUDIES

To increase the generalizability of the results, we used our procedure with three personality measures which had been administered to different samples. However, to avoid redundancies, the three studies will be presented and discussed together.

### Measures

If it is to work properly, our procedure requires two basic conditions to be met. The item set must be: (a) essentially unidimensional and (b) well balanced. Nowadays, acquiescence is considered to be a secondary factor with respect to the main "content" factor that the items attempt to measure (see e.g. Ferrando & Condon, 2006). So it might be difficult to identify this secondary factor if the items are already impacted by several "content" factors or are incompletely balanced. These considerations guided the choice of the item sets used in the present study.

The instrument used in the first study was Rotter's Locus of Control Scale (LOC, Rotter, 1966). Even today, the LOC scale is probably the best known and most widely used measure of general locus of control. It consists of 23 content items whose statements refer mainly to beliefs, thoughts or attitudes. A typical item from the LOC scale is given below.

A. Many of the unhappy things in people's lives are partly due to bad luck.
B. People's misfortunes result from the mistakes they make.

Rotter (1966) developed his scale as a unidimensional instrument that was intended to measure a broad, general dimension of locus of control. However, the numerous factor analyses of the scale have repeatedly suggested that the scale is multidimensional. A meta-analytical proposal based on studies from different cultures and languages (Berndt, 1978), which also agrees with our previous studies, is a two-factor model with a more general 'Personal Control' dimension, and a second, more specific "Socio-Political" dimension. For the reasons discussed above, we worked only with the Personal Control items. Furthermore, because the items of interest were not completely balanced, we paired items according to their content, and produced a 14-item measure in which half of the items had option A measuring externality and B internality and the other half of the items the other way round.

The second study used the Sensation Seeking Scale form V (SSS-V; Zuckerman, 1996), which is also an instrument that is widely used in personality measurement. Unlike the case mentioned above, the SSS-V is explicitly intended to be multidimensional, and consists of four subscales, each of which has 10 items. Of these subscales, the Thrill and Adventure Seeking (TAS) scale appears to be the best defined and the most reliable both in the original version (Zuckerman, 1996; Eysenck & Haapasalo, 1989), and in the Spanish adaptation (Ferrando & Chico, 2001). It is completely balanced, with five items keyed in each direction. So, in this case we used the scale without any item selection. Below is a typical TAS item:

A. A sensible person avoids activities that are dangerous.
B. I sometimes like to do things that are a little frightening.

The last measure considered was the Vando Reducer-Augmenter Scale (RAS; Vando, 1974, Clapper, 1990). The content of the RAS items refers to preferences for situations which involve higher or lower levels of stimulation intensity. A typical RAS item is

A. Climb a mountain.

B. Read about a dangerous adventure.

The same as the LOC, the RAS was designed to be essentially unidimensional. Nevertheless, the numerous factor analyses of the scale (including our own previous analyses, see Ferrando, Vigil-Colet, Tous, & Lorenzo-Seva, 1993), tend to systematically obtain a tridimensional structure, with a more general dimension that has been named "General Lifestyle RA" and two more specific dimensions which refer respectively to musical and artistic preferences, and to situations of thrill and danger (Kohn et al., 1986). For reasons discussed above, we choose to work only with the General Lifestyle items. Furthermore, because the number of positively keyed and negatively keyed items was quite unequal, we paired items according to their content, and produced a fully balanced scale made up of 14 items.

To close this section we would like to point out that even when the three chosen measures are made up of normative FC items, the types of items are quite different from one another. In the LOC scale, the statements refer to beliefs or attitudes, they are generally rather long, and the respondents have to decide which statement they most agree with. In the SSS-V the participants have to choose between two behaviours that are mostly externally observable, and the statements tend to be of average length. Finally in the RAS the respondents have to choose between two options which are either internal feelings or observable behaviours, and the statements are usually rather short.

### Participants

In the three studies, we analysed rather large samples which were mostly made up of undergraduates from different faculties in our University. All the samples, then, consisted mainly of young people (mean age about 21) with a relatively high cultural level. The proportion of genders was about 70% female. The sample sizes were: N=1035 (LOC study), N=448 (SSS-V-TAS study) and N=904 (RAS study).

## ANALYSES AND RESULTS

The sequence of analyses was two-step, and followed the rationale discussed above. In each data set the unidimensional model was fitted first. The bidimensional model was fitted next, and the arbitrary initial solution

was put in canonical form so as to make it comparable to the centroid solution on which our procedure is based. Given that in all the studies the item sets were of medium-short size and the samples were rather large, the tetrachoric matrices were fitted by using weighted-least-squares (WLS) estimation. WLS estimation enables the goodness-of-fit to be rigorously assessed and the improvement of fit in the nested models to be strictly compared. This last point is particularly relevant in our case in which we need to compare the improvement of fit when going from the one-factor to the two-factor model. The program used for fitting the model was Lisrel 8.80 (Jöreskog & Sörbom, 1996).

Table 1 shows the goodness-of-fit results obtained in the three studies. It seems clear that there are common trends in all of them. First, the fit of the unidimensional model is more than acceptable in the three datasets. In fact, according to present standards it is good or very good. (RMSEA below 0.05 and goodness-of-fit indices above 0.95, see Hu & Bentler, 1999). Second, according to the chi-square difference test, in the three cases the fit improves noticeably when going from the one-factor model to the two-factor model. However, the high power of the test in these large samples must be taken into account (in particular in studies 1 and 3 the samples are of about 1000 participants). When the improvement of fit is judged by the increments in the goodness of fit indexes, which are theoretically less affected by sample size, the conclusions are not so clear. The RMSEA-based measure of improvement of fit, the root deterioration per restriction (RDR, Browne & du Toit 1992) has the same scaling as the overall RMSEA. So values between 0.05 and 0.08 would indicate that there are no significant changes in the degree of fit (Browne & Cudeck 1993). According to this criterion, none of the studies would show a clear improvement when going from one to two factors. The conclusion would be different if the study were to be based on the GFI increment ($\Delta$GFI). Cheung and Rensvold (2002) proposed that a reference value of $\Delta GFI < 0.001$ means that the improvement of fit is negligible. And the increments are above this cut-off value in all three studies.

Overall, the goodness-of-fit results are not as clear as the proposals we made in the previous section except in the second study. In the SSS-V-TAS data set, the fit of the unidimensional model was already excellent, and imposing a second factor even made it too good (virtually, RMSEA=0 and GFI=1). In this case it seems clear that forcing a second factor leads to overfitting. In studies 1 and 3, however, we found that the fit of the unidimensional model was already good, but that it could be improved by adding a second factor. It seems clear, then, that the solutions obtained must be examined before conclusions can be reached.

**Table 1. Goodness-of-fit results in the three studies.**

| a) LOC dataset | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Model* | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | *RMSEA* | *RDR* | *GFI* | $\Delta GFI$ |
| 1factor | 200 | 77 | 61 | 13 | 0.039 | | 0.987 | 0.003 |
| 2 factor | 139 | 64 | | | 0.034 | 0.05 | 0.990 | |
| b) SSS-V-TAS dataset | | | | | | | | |
| *Model* | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | *RMSEA* | *RDR* | *GFI* | $\Delta GFI$ |
| 1 factor | 43.07 | 34 | 21.17 | 12 | 0.020 | | 0.993 | 0.003 |
| 2 factor | 21.90 | 22 | | | 0.000 | 0.041 | 0.996 | |
| c) RAS dataset | | | | | | | | |
| *Model* | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | *RMSEA* | *RDR* | *GFI* | $\Delta GFI$ |
| 1 factor | 238.30 | 77 | 88.76 | 13 | 0.048 | | 0.983 | 0.006 |
| 2 factor | 149.54 | 64 | | | 0.040 | 0.080 | 0.989 | |

The factor pattern solutions were first assessed by comparing the solution obtained with the unidimensional model to the dominant canonical solution obtained with the bidimensional model. The results were clear. In each of the three datasets both patterns were virtually identical (as expected) and showed the perfect bipolar structure that would be expected from the theory if this factor was the 'content' factor to be measured.

To assess the second canonical factor in the bidimensional solution (supposedly the acquiescence factor), we considered two criteria. The first was the comparison of the point-estimated loadings with their corresponding standard errors. The second was a heuristic criterion widely used in applied settings (e.g. McDonald, 1985): a minimum of 3-4 variables with loadings larger than 0.30 are needed if a factor is to be adequately defined. The results can be summarized as follows. In the SSS-V-TAS and the RAS most of the loadings did not reach statistical significance and fluctuated randomly around 0. And only one item in each scale had a loading well above 0.30. So it appears that for these two measures, the second factor is clearly residual and does not comply with the minimum requirement that it can be interpreted in any way (in our case as the acquiescence factor). The second LOC factor did have more substantial loads, so it will be assessed below in more detail. In accordance with the results described so far, table 2 shows only the dominant (content) canonical pattern for the SSS-V-TAS and the RAS, and the complete bidimensional solution for the LOC. For a clearer interpretation, the items on each scale that were expected to have a positive content loading are marked with a 'P' and the items with an expected negative loading with an 'N'. Finally the centroid bidimensional solution for the LOC is presented

together with the canonical WLS solution. This would serve to illustrate the expected degree of equivalence between both approaches that was discussed above.

**Table 2. Factorial solutions for the three studies.**

| *Item number* | *LOC (WLS solution)* | | *LOC (Centroid solution)* | | *SSS-V-TAS* | *RAS* |
|---|---|---|---|---|---|---|
| | **F1** | **F2** | **F1** | **F2** | **F** | **F** |
| 1 | 0.40 (p) | 0.16 | 0.39 (p) | 0.17 | 0.77 (p) | -0.45 (n) |
| 2 | -0.40 (n) | 0.33 | -0.41 (n) | 0.32 | -0.68 (n) | 0.82 (p) |
| 3 | 0.34 (p) | 0.06 | 0.34 (p) | 0.07 | 0.71 (p) | -0.55 (n) |
| 4 | 0.28 (p) | 0.05 | 0.27 (p) | 0.06 | 0.82 (p) | -0.41 (n) |
| 5 | 0.31 (p) | 0.04 | 0.31 (p) | 0.05 | -0.75 (n) | 0.79 (p) |
| 6 | -0.55 (n) | 0.12 | -0.54 (n) | 0.11 | -0.69 (n) | 0.81 (p) |
| 7 | -0.33 (n) | 0.34 | -0.33 (n) | 0.33 | 0.89 (p) | 0.40 (p) |
| 8 | -0.55 (n) | 0.21 | -0.55 (n) | 0.20 | 0.73 (p) | -0.50 (n) |
| 9 | -0.65 (n) | -0.22 | -0.65 (n) | -0.23 | -0.61 (n) | 0.73 (p) |
| 10 | 0.70 (p) | 0.10 | 0.70 (p) | 0.11 | -0.67 (n) | 0.59 (p) |
| 11 | 0.63 (p) | 0.40 | 0.62 (p) | 0.41 | | -0.30 (n) |
| 12 | 0.67 (p) | 0.14 | 0.66 (p) | 0.15 | | -0.79 (n) |
| 13 | -0.29 (n) | 0.30 | -0.30 (n) | 0.29 | | -0.52 (n) |
| 14 | -0.52 (n) | 0.24 | -0.53 (n) | 0.23 | | 0.40 (p) |

*Note. p means positive loadings. n means negative loadings.*

As discussed above, in the three cases the structure of the 'content' factor has the perfect bipolarity that theory suggests. However, more information can be obtained from the table. First, note that the LOC solution is clearly weaker than the other two. The average absolute values of the loadings are 0.47 (LOC), 0.73 (SSS-V-TAS) and 0.58 (RAS). In particular, the SSS-V-TAS solution is quite strong for a personality test. Second, the weak condition of balance was approximately met for the LOC and the SSS-V-TAS, but less so for the RAS. The sums of the loadings are 0.04 (LOC), 0.52 (SSS-V-TAS) and 1.03 (RAS).

The second factor in the LOC solution has mainly positive loadings. This is the expected direction if it was the acquiescence factor, understood here as the tendency to choose the first statement (as discussed above). The magnitude of the loadings in this factor reflects the strength of this tendency, and here 4 of these loadings are above 0.30. Indeed, we first checked whether this cluster might arise because of some artefact caused by

shared specificities among the items. However, the four items did not appear to have common content or redundancies in the wording. While this evidence reinforces the interpretation that this factor is the acquiescence factor, in our opinion, it is still not sufficient for a clear interpretation. Nevertheless, if we tentatively accept this explanation, the immediate question is why this factor appears in the LOC scale but not in the other two. The following gives two plausible explanations. First, the LOC has a 'content' factor that is clearly weaker than the other two, and AR seems to be more likely to appear in weak (poorly discriminating) items (Condon, Ferrando & Demestre, 2006). Second, even though the LOC is a personality measure, the content of the items refers more to attitudes and beliefs than to behaviours and feelings. And it is widely acknowledged that AR is more problematic in attitude measurement than in personality measurement (Paulhus, 1981). More specifically, the LOC item stems contain the kind of broad generalizations that are thought to be the most susceptible to AR (Schuman & Presser, 1981). In this regard, Mukherjee (1969) conjectured that the use of specific and personal-reference statements in FC items minimised the manifestation of AR, and the LOC items do quite the opposite. Indeed, these post-hoc explanations can only be taken as plausible conjectures, and further research on this issue is clearly needed. Finally, we note that the centroid and the WLS solutions are virtually the same, as expected.

## DISCUSSION AND CONCLUSIONS

This paper aims to make a double contribution to the response-biases literature: methodological and substantive. At the methodological level, the present results suggest that the model-based procedure we proposed works well with FC personality scales. So, we believe that applied researchers in personality now have at their disposal a useful tool for assessing the impact of AR in empirical studies based on FC measures.

At the substantive level, we applied the proposed procedure to three large-sample datasets in which well-known and widely used normative FC measures had been administered. Given these conditions, the empirical study cannot be considered as a mere illustration of the procedure. However, it has limitations, does not reach totally clear conclusions, and has to be considered as a first step within a potentially important future body of research. In our opinion, further empirical research should, on the whole, take two directions. First, more controlled studies should be undertaken in which the same measures are administered in conventional and FC formats (this would require longitudinal and/or multiple group

designs). Second, additional information based on external variables (i.e. validity assessment) should be collected to enhance the interpretation of results.

In our opinion, the results obtained so far warrant the further research outlined above. Except for the LOC case, the strong bipolar structure of the first canonical factor, and the weak structure of the second factor (too weak to deserve interpretation) suggested that the items measured mainly 'content', and perhaps, secondarily, random responding. So, in general, our results suggest that FC scales that use clear, short, specific and personal-reference (i.e. behaviours and/or feelings) statements are practically free from the impact of AR. If this result could be generalized it would be of considerable practical interest. It would mean that a well designed FC questionnaire would need only to control for AR and not to balance the items.

In conclusion, we should also point out that the model proposed for FC responding could also serve as a basis for a model-based procedure intended to assess the impact of faking and social desirability. Even though this type of response bias has received far more attention in the normative FC measurement literature, no model-based assessment study seems to exist at present. Extending our procedure in this direction is also an objective for future research.

## RESUMEN

**Impacto de la aquiescencia en ítems de elección forzosa: un análisis basado en un modelo de respuesta.** El objetivo general del presente estudio es evaluar la potencial utilidad que tiene el formato de elección forzosa para reducir el impacto de la respuesta aquiescente. Las contribuciones con vistas a conseguir este objetivo son de dos tipos: metodológicas y substantivas. Metodológicamente, se propone un procedimiento basado en un modelo de análisis que, a su vez, se desarrolla a partir de un mecanismo de respuesta. Dicho procedimiento permite una evaluación rigurosa del impacto de la aquiescencia. Substantivamente, el procedimiento propuesto se aplica a tres cuestionarios con formato de elección forzosa ampliamente utilizados en personalidad: La escala de Locus de Control de Rotter, la escala de búsqueda de sensaciones de Zuckerman y la escala de aumentación-reducción de Vando. Los resultados sugieren que el impacto de la aquiescencia es mínimo excepto en el caso de la escala de Rotter. Estos resultados diferenciales podrían explicarse por ciertas características de la escala que se discuten en el artículo.

# REFERENCES


Barrick, M.R., & Ryan, A.M. (2003). *Personality and work: Reconsidering the role of personality in organizations*. San Francisco: John Wiley & Sons, Inc.

Berkowitz, N.H., & Wolkon, G.H. (1964). A Forced Choice forms of the F Scale-Free of Acquiescent Response Set. *Sociometry, 27,* 54-65.

Berndt, D.J. (1978). Construct validation of the personal and sociopolitical dimensions of Rotter's internal-external locus of control scale. *Psychological Reports, 42*, 1259-1263.

Bock, R.D., & Lieberman, M. (1970) Fitting a response model for n dichotomously scored ítems. *Psychometrika, 35*, 179-197.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.) *Testing structural equation models* (pp 136-162). Newbury Park: Sage.

Browne, M.W., & Du Toit, S.H.C. (1992). Automated fitting in nostandard models. *Multivariate Behavioral Research, 27,* 269-300.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

Choulakian, V. (2003). The optimality of the centroid method. *Psychometrika, 68*, 473-475.

Christiansen, N.D., Burns, G.N., & Montgomery, G.E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18,* 267-307.

Clapper, R.L. (1990). Adult and adolescent arousal preferences: The revised reducer-augmenter scale. *Personality and Individual Differences***,** *11,* 1115–1122.

Condon, L., Ferrando, P.J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality and Individual Differences, 40,* 403-407.

Coombs, C.H. (1948). A rationale for the measurement of traits in individuals. *Psychometrika, 13,* 59-68.

Edwards, A.L. (1970). *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart and Winston.

Edwards, A.L., & Abbott, R.D. (1973). Measurement of Personality Traits: Theory and Technique. *Annual Review of Psychology, 24*, 241-278.

Eysenck, S.B.G., & Haapasalo, J. (1989). Cross-cultural comparison of personality: Findland and England. *Personality and Individual Differences, 10*, 121-125.

Feldman, M.J., & Corah, N.L. (1960). Social desirability and the forced choice method. *Journal of Counsulting Psychology, 24,* 480-482.

Ferrando, P.J., Vigil-Colet, A., Tous-Pallares, J., & Lorenzo-Seva, U. (1993). Spanish adaptation of the reducer−augmenter scale: relations with EPI-A scales. *Personality and Individual Differences, 14*, 513-518.

Ferrando, P.J., & Chico, E. (2001). The construct of sensation seeking as measured by Zuckerman's SSS-V and Arnett's AISS: a structural equation model. *Personality and Individual Differences, 31,* 1121-1133.

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

Ferrando, J.P. (2006). Two item response theory models for analysing normative forced-choice personality items. *British Journal of Mathematical and statistical Psychology, 59,* 379-395.

Ferrando, P. J., & Lorenzo-Seva, U. (2009). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, (in press, pre-print available online).

Ford, L.H. (1964). A forced-choice, acquiescence-free, social desirability (defensiveness) scale. *Journal of Consulting Psychology, 28,* 475.

Gordon, L.V. (1951). Validities of the forced-choice questionnaire methods of personality measurement. *Journal of Applied Psychology, 35*, 407-412.

Guilford, J.P. (1954). *Psychometric Methods.* New York: McGraw-Hill

Harvey, R.J., & Murry, W.D. (1994). Scoring the Myers-Briggs Type Indicator: empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment, 62*, 116-129.

Hicks, L.E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167-184.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Jackson, D.N., Wroblewski, V.R., & Ashton, M.C. (2000). The Impact of Faking on Employment Tests; Does Forced Choice Offer a Solution?. *Human Performance, 13,* 371-388

Johnson, C.E., Wood, R., & Blinkhorn, S.F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61,* 153-162.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide.* Chicago: Scientific Software.

Kohn, P.M., Hunt, R.W., Cowles, M.P., & Davis, C.A. (1986). Factor structure and construct validity of the Vando Reducer–Augmenter Scale. *Personality and Individual Differences, 7*, 57-64.

Lawley, D.N. (1955). A statistical examination of the centroid method. *Proceedings of the royal society of Edinburgh. A, 54*, 175-189.

McDonald, R.P. (1985). *Factor analysis and related methods*. Hillsdale: LEA.

Miller, T.R., & Cleary, T.A. (1993). Direction of wording effects in balanced scales. *Educational and Psychological Measurement, 53,* 51-60.

Morgeson, F.P., Campion, M.A., Dipboye, E.L., Hollenbeck, J.R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683-729.

Mukherjee, B.N. (1969). Absence of acquiescence in a forced-choice test of achievement value. *Psychological Reports, 25*, 70.

Myers, I.B. (1962). *The Myers-Briggs Type Indicator manual.* Princeton: Educational Testing Service.

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

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

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

Ray, J.J. (1989). Acquiescence and problems with forced-choice Scales. *The Journal of Social Psychology, 130,* 397-399.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (1, whole Nº. 609).

Saltz, E., Reece, M., & Ager, J. (1962). Studies of forced-choice methodology: individual differences in social desirability. *Educational and Psychological Measurement, 22*, 365-370.

Schuman, H., & Presser, S (1981). *Questions and answers in attitude suveys: Experiments in question forms, wording and context*. London: Academic Press.

Smyth, J.D., Dillman, D.A., Christian, L.M., & Sterns, M.J. (2006). Comparing check-all and forced.choice question formats in web surveys. *Public Opinion Quarterly, 70,* 66-77.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P.E. Shrout and S.T. Fiske (Eds.) *Personality Research, Methods, and Theory* (pp. 161-181). Hillsdale: LEA.

Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.

Vando, A. (1974). The development of the R–A scale: a paper-and-pencil measure of pain tolerance. *Personality and Social Psychology Bulletin, 1*, 28–29.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*, 117-124.

Zuckerman, M. (1996). Item revisions in the Sensation Seeking Scale form V (SSS-V). *Personality and Individual Differences, 20,* 515.

(Manuscript received: 15 December 2009; accepted: 16 February 2010)