British Journal of Mathematical and Statistical Psychology (2019), 72, 447–465

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Paper

Comparison of classical and modern methods for measuring and correcting for acquiescence

Ricardo Primi[1,2]*, Daniel Santos[2,3], Filip De Fruyt[2,4] and Oliver P. John[2,5]

[1]Postgraduate Program in Psychology, Universidade São Francisco, Campinas, São Paulo, Brazil
[2]EduLab21, Ayrton Senna Institute, São Paulo, Brazil
[3]Faculty of Economics, Administration and Accounting of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil
[4]Department of Developmental, Personality and Social Psychology, Ghent University, Belgium
[5]Department of Psychology and Institute of Personality and Social Psychology, University of California, Berkeley, California, USA

Likert-type self-report scales are frequently used in large-scale educational assessment of social-emotional skills. Self-report scales rely on the assumption that their items elicit information only about the trait they are supposed to measure. However, different response biases may threaten this assumption. Specifically, in children, the response style of acquiescence is an important source of systematic error. Balanced scales, including an equal number of positively and negatively keyed items, have been proposed as a solution to control for acquiescence, but the reasons why this design feature worked from the perspective of modern psychometric models have been underexplored. Three methods for controlling for acquiescence are compared: classical method by partialling out the mean; an item response theory method to measure differential person functioning (DPF); and multidimensional item response theory (MIRT) with random intercept. Comparative analyses are conducted on simulated ratings and on self-ratings provided by 40,649 students (aged 11–18) on a fully balanced 30-item scale assessing conscientious self-management. Acquiescence bias was explained as DPF and it was demonstrated that: the acquiescence index is highly related to DPF; balanced scales produce scores controlled for DPF; and MIRT factor scores are highly related to scores controlled for DPF and the random intercept is highly related to DPF.

1. Introduction

Likert-type self-report scales are increasingly used in large-scale assessments of social-emotional skills in primary and secondary education (Abrahams *et al.*, 2019). One long-standing issue with the use of self-report Likert-scales is that they are influenced by

*Correspondence should be addressed to Ricardo Primi, Postgraduate Program in Psychology, Universidade São Francisco, Rua Waldemar César da Silveira, 105, Campinas, São Paulo, CEP 13045-510, Brazil (email: rprimi@mac.com).

DOI:10.1111/bmsp.12168

response biases. Response bias refers to systematic factors that affect responses but are not related to the construct of interest. McCrae (2018) has suggested that the influence of response biases represents up to 40% of the variance in self-reports. This is a substantial amount of 'method variance' that limits the validity and utility of self-reports.

One important response bias is acquiescence, which refers to the tendency to choose responses stating agreement regardless of the content of the item. Conversely, disacquiescence refers to the tendency to choose responses stating disagreement regardless of the content of the item (He, Bartram, Inceoglu, & van de Vijver, 2014; Wetzel & Carstensen, 2015). Acquiescence has pronounced effects, especially in self-reports of children and adolescents. Soto, John, Gosling, and Potter (2008) have shown that acquiescence variance is twice as large for children as for adults. Specifically, acquiescence variance was highest at age 10 and then dropped year by year until age 18, at which point it reached stable adult levels. Individual differences in acquiescence influence the internal structure of assessment instruments as well as their criterion validity (Primi, De Fruyt, John, & Santos, 2018; Rammstedt, Goldberg, & Borg, 2010; Soto & John, 2017). Finally, individual differences in acquiescence are stable even across 8 years (Wetzel, Lüdtke, Zettler, & Böhnke, 2016). In summary, extensive research has shown that acquiescence is an important, stable, and influential response bias that must be addressed in student assessments and needs to be understood thoroughly, including how to measure it and how to use the resulting measures to correct or reduce the effects of acquiescence on children's self-report responses.

This paper presents a comparison of three methods of measuring and correcting for acquiescence: classical method using balanced scales; differential person functioning (DPF); and multidimensional item response theory (MIRT) with random intercept. Our goal is to show the empirical connections between the classical acquiescence index, DPF and MIRT estimated factor scores.

1.1. Classical measurement and correction for acquiescence using balanced scales

The psychometric formulation of this model was developed by Ten Berge (1999). One way to examine acquiescence is to include positively keyed (PK) and negatively keyed (NK) items, that is, markers of opposite poles of a trait. Imagine an item designed to measure negative emotional regulation, such as *i*+: 'I adapt easily to new situations without worrying too much', with students asked to respond on a scale with '1' (not at all like me), '2' (a little like me), '3' (moderately like me), '4' (a lot like me) and '5' (completely like me). Also, an antonym pair item is included such as *i*−: 'I have trouble controlling my anxiety in difficult situations'. Now consider a person with *no* acquiescence, who is perfectly calibrated on the five-point response scale with a mid-point of 3. If that person is high on emotional regulation they would give differentiated responses to these items, reflected across the actual mid-point of the rating scale, for example '4' to the true-keyed item *i*+ and correspondingly '2' to the false-keyed item *i*−. If that same person was very *high* on acquiescence, their responses would be systematically higher on both items, such as 5 and 3, respectively (an acquiescence effect of +1). In contrast, if that same person was *low* on acquiescence (i.e., showed disacquiescence), both item responses would be lower than the perfectly calibrated person, such as 3 and 1 (an acquiescence effect of −1).

Soto *et al.* (2008) computed acquiescence in the Big Five Inventory by calculating for each individual *j* a mean on 16 antonym item pairs (index of acquiescence, acq_j). Since these items measure the same trait but come from opposite ends of the trait continuum, agreement with positive items should co-occur with disagreement with negative items. Assuming that within each antonym pair, each PK and NK item has the same difficulty, the

expected score on this index will be $\text{acq}_j = 3$. If $\text{acq}_j > 3$ or $\text{acq}_j < 3$ it will indicate high acquiescence or disacquiescence, respectively. Then test items responses y_{ij} of individual j to the i items of the test can be transformed using the formula[1] : $x^r_{ij} = x_{ij} - \text{acq}_j$. This recentring of all item scores is done to partial out the effect of acquiescence on all the individual's responses.

If we have fully balanced scales where the number of pairs is equal to half the number of items, $n_i/2$, the vector of transformed item scores x_{ij} for an individual observation j will be in an ipsative form because their sum will be equal to zero for each individual. It might appear that by applying this formula we will remove all between-subject variance and be left with only within-subject variance (Chan & Bentler, 1998). But when calculating subjects' scores, we first reverse half of the x^r_{ij} recoded scores of NK items and then sum (or average) the item recoded scores to calculate total scores. This final total score does not sum to a constant across individuals: there is still between-subject variance left, which has been purged of acquiescence. Ipsatization that occurred before reversing NK items removes a parcel of x_{ij} variance related to the positive manifold across positive and negative items, that is, acquiescence variance that is captured by the acquiescence index acq_j. But after reversing NK items, the vector of transformed (and half reversed) item scores x^r_{ij} is no longer in an ipsative form.

Back in 1999, Ten Berge noted that an ipsative transformation of balanced personality scales containing positive and negative items is a special case with peculiar properties. What happens when scores are recoded to partial out the mean acq_j (acquiescence)? Consider a six-item scale composed of three pairs of antonym items scored on a five-point Likert scale. Let $i = 1, 2, 3$ be the positive and $i = a, b, c$ the corresponding negative items, and x_{ij} the original response of subject j on item i. Acquiescence controlled scores can be written as[2]

$$\text{scr.rec}_j = \frac{1}{6}\left[\sum_{i=1}^{3} x_{ij} - \sum_{i=a}^{c} x_{ij}\right].$$

The simplified formulation of recoded scores scr.rec_j can be used to understand how acquiescence is partialled out. Table 1 shows eight response patterns on the six items described above and their corresponding: average (scr_j), acquiescence (acq_j) and recoded (scr.rec_j) scores. Assume that each item pair ($i1/ia$, $i2/ib$, $i3/ic$) has the same level of difficulty and relates equally to acquiescence.

These eight patterns vary in terms of trait scores and acquiescence scores. Patterns P1 and P2 show the maximum and minimum scores that are possible when an individual is maximally consistent, that is, with no acquiescence. Recoded scores scr.rec_j vary from -2 to 2. Patterns P3 and P4 also show maximum and minimum possible scores, but now when a person is maximally inconsistent, that is, with acquiescence or disacquiescence maximized. Now recoded scores scr.rec_j are 0, the mid-point of the scale.

---

[1] There is an alternative method of recoding responses that divides $x_{ij} - \text{acq}_j$ by the within-person standard deviation of responses SD_j on item responses. Then recoded scores calculated by: $\text{scr.z.rec}_j = \frac{1}{6}\left[\sum_{i=1}^{3}[(x_{ij} - \text{acq}_j)/SD_j] - 1\sum_{i=a}^{c}[(x_{ij} - \text{acq}_j)/SD_j]\right]$. When doing so, the amplitude of the scores will be similar for acquiescence individuals and non-acquiescence individuals. Since SD_j is dependent on acq_j, this transformation may undo part of the DPF correction (see https://github.com/rprimi/acqu_mirt for more details).

[2] A detailed mathematical derivation of this equation and other issues of classical method can be found at: https://github.com/rprimi/acqu_mirt

Table 1. Eight response patterns on six items: three PK (*i*1, *i*2, *i*3) and three NK (*ia*, *ib*, *ic*) and their corresponding average (scr_j), acquiescence (acq_j) and recoded ($scr.rec_j$) scores

Response patterns	Trait level	Acquiescence/inconsistency	*i1*	*i2*	*i3*	*ia*	*ib*	*ic*	scr_j	acq_j	$scr.rec_j$
P1	Max	Min	5	5	5	1	1	1	5.0	3	2.0
P2	Min	Min	1	1	1	5	5	5	1.0	3	−2.0
P3	Max	Max	5	5	5	5	5	5	3.0	5	0.0
P4	Min	Max	1	1	1	1	1	1	3.0	1	0.0
P5	High	Low	5	4	4	1	2	2	4.3	3	1.3
P6	Low	Low	1	2	2	5	4	4	1.7	3	−1.3
P7	High	Moderate high	5	4	4	2	3	3	3.8	3.5	0.8
P8	Low	Moderate high	2	3	3	5	4	4	2.2	3.5	−0.8

Individuals with patterns P5 and P7 are high on the trait and have the same answers on PK items, but P5 has no acquiescence, and P7 high acquiescence. The expected response on negative items of an individual with responses '5', '4', '4' on the three positive items is '1', '2', '2'. P7 instead answered '2', '3, '3', that is, he/she did not answer in the same way when confronted with questions measuring the same trait with the same difficulty, but phrased negatively in an NK item. Therefore, for this individual, it is relatively easy to agree because he/she agrees to anything. As a consequence, scores on PK items from an acquiescent individual will be relatively biased toward the high than from a non-acquiescent individual. The correction then lowers P7's score. The recoded score is 0.8 for P7 as compared to 1.3 for P5. Examples P6 and P8 have the same logic but now at the low (false-keyed) end. The correction now makes scores less negatively extreme because it takes into account acquiescence.

These examples illustrate how acquiescence variance is partialled out. The amplitude of recoded scores was reduced from 2.6 (−1.3 to 1.3) to 1.6 (−0.8 to .8). Since part of the item endorsement is explained by acquiescence, the recoding makes the item scores less extreme. In this way, acquiescence variance is partialled out. In extreme cases with maximum acquiescence, scores will be in the middle with no variance (examples P3 and P4). It is worth noting that all these adjustments occur unnoticed when computing total scores for balanced scales (see Appendix A).

1.2. Acquiescence as differential person functioning (DPF method)

Johanson and Osborn (2004) proposed a new operational definition of acquiescence as DPF. Differential item functioning (DIF) occurs when subjects with the same level on the construct, but from different groups, have different probabilities of endorsing a particular answer, for example, agreeing with an item. Johanson and Osborn proposed that one can also look at groups of items instead of persons. For instance, let us assume that true-keyed and false-keyed items measure the same construct with the same level of difficulty. If responses to these items happen to be different we are seeing DPF, that is, a person reacts differently to two types of items that measure the same trait with equivalent difficulty but differ in another item feature.

While solving DIF is relatively straightforward, solving DPF is more challenging. DIF is solved by splitting the items by groups of interest and estimating different item parameters for each group, treating items as if they were different across groups. Since we are interested in person parameters, this makes sense because it disentangles group differences of non-target dimensions from person parameters.

DPF can be estimated in three steps (Andrich, 1978; de Ayala, 2009; Linacre, 2017): first, estimating item response theory (IRT) item parameters in one calibration, considering all items with reversed NK items; second, estimating person parameters twice, using PK items θ_{pj} and then using reversed NK items θ_{nj} while keeping items anchored at the item parameters from the main analysis; and third, calculating $DPF_j = \theta_{nj} - \theta_{pj}$. If persons function in the same way while answering items they should have similar trait estimates. The difference should be within the confidence intervals computed from the standard error of measurement. Therefore, the difference between these two scores indexes DPF.

Although we can estimate DPF with this method, it does not make sense to estimate two latent traits (one based on PK and another on NK items) when we have substantive evidence that this trait is unidimensional. Therefore, this method proposes a way to measure acquiescence but not to solve the problem. So how can DPF be solved? We argue

that Ten Berge's (1999) classical method is a way to control for DPF but it is operationalized under classical test theory. Conceptualizing response styles (acquiescence included) as DPF opens up the possibility of modelling both substantive traits and response styles using MIRT methods. This is coherent with the definition of bias as an additional dimension of individual differences pertaining to persons.

1.3. Random intercept modelling approach (MIRT method)

Savalei and Falk (2014) revised three statistical models to analyse balanced scales: a factor model for ipsatized variables of Chan and Bentler (1998); an extension of Ten Berge (1999) using exploratory factor analysis with the target rotation proposed by Ferrando, Lorenzo-Seva, and Chico (2003) (see also Ferrando & Lorenzo-Seva, 2010); and a confirmatory factor-analytic approach detailed by Maydeu-Olivares and Coffman (2006), called random-intercept factor analysis (Billiet & McClendon, 2000). Although the three approaches performed very well, the random-intercept approach performed best. This was also the case for a simulation study conducted by Garrido, Golino, Nieto, Peña, and Molina (2018).

In unidimensional scales the random-intercept model can be written as

$$x_{ij} = (\mu_i + \zeta_j) + \lambda_i f_j + e_{ij},$$

where x_{ij} is the response of subject j on item i, μ_i is the general item i intercept (reverse of item difficulty), ζ_j is the random intercept of individual j, λ_i is the factor loading of item i on latent factor f, f_j is the score of individual j on the latent factor and e_{ij} is the error of prediction of the response of individual j on item i. Note that item difficulty is split in two components: general item difficulty μ_i that varies between items but is the same for all individuals; and the random intercept ζ_j that varies between individuals but is the same for all the items that the individual answers. The random intercept is estimated by fixing a loading of 1 in all items before reversing the NK items. This model also assumes that the random intercept is uncorrelated with e_{ij} and f_j. Note that, in balanced scales, the random intercept ζ_j captures positive covariance between all items and therefore is an estimate of acquiescence similar to the simple acq_j score.

This model assumes x_{ij} is a continuous variable. For ordered categorical responses (e.g., five-point rating scale items) we can adapt the random-intercept model as a MIRT model (MIRT-RI; Chalmers, 2012; Kamata & Bauer, 2008; Takane & de Leeuw, 1987) based on Samejima's graded response model (GRM) as follows (de Ayala, 2009; Embretson & Reise, 2000; Samejima, 1968):

$$P_{x_{ij}}(\theta_j, \zeta_j) = \frac{1}{1 + \exp(-(a_{1i}\theta_j + a_{2i}\zeta_j + c_{ri}))},$$

where $P_{x_{ij}}$ is the probability of subject j obtaining x_{ij} or higher in item i versus lower categories scores ($x_{ij} \geq 2$ versus 1, $x_{ij} \geq 3$ versus 1 or 2 and so forth). a_{1i} is the discrimination for item i on the substantive trait, θ_j. a_{2i} is fixed at 1 for PK items and -1 for NK items (because we fix all a_{2i} we estimate variance of ζ_j); ζ_j represents acquiescence and c_{ir} is an intercept term (Cai, 2010; Maydeu-Olivares & Steenkamp, 2018). With this set-up, we predict that the random intercept will be highly correlated with the traditional acquiescence index, and the latent substantive trait will be similar to the scores corrected for acquiescence.

1.4. Research questions

Following this conceptual outline, the three methods to deal with acquiescence will now be applied and compared. We first present an empirical study to evaluate the links proposed here, namely that DPF calculated via an IRT method is highly correlated with the acquiescence index acq_j; that acq_j is highly correlated with the random intercept ζ_j; and that recoded scores $scr.rec_j$ are highly correlated with θ_j, that is, highly acquiescent individuals will have less extreme scores θ_j than non-acquiescent individuals given similar substantive-response patterns. We then present a simulation study checking whether we can replicate these connections, verify how well each method recovers true parameter values of trait and acquiescence, and further examine whether those relationships generalize to the case of unbalanced scales.

2. Empirical example

The general purpose of the empirical study was to demonstrate the primary relationships among classical indices of acquiescence and acquiescence-controlled scores with DPF and scores calculated from an MIRT version of a random-intercept modelling approach. Previous studies have noticed the relationship between acquiescence and DPF (Johanson & Osborn, 2004) and have modelled acquiescence with random intercept (Maydeu-Olivares & Coffman, 2006; Savalei & Falk, 2014), but no study has connected acquiescence with DPF in the framework of MIRT modelling (see Appendix B).

2.1. Method

2.1.1. Participants

The participants were 40,649 students (55.1% females) from 495 public schools located in 232 cities in the state of São Paulo in Brazil, aged from 11 to 18 (M = 14.7, SD = 1.99) and attending grades 6 (N = 6,411), 7 (N = 4,357), 8 (N = 5,558), 9 (N = 6,272), 10 (N = 6,534), 11 (N = 5,763) and 12 (N = 5,754). All participated in a social-emotional skill study conducted by Edulab21 at the Ayrton Senna Institute. The sample was relatively diverse, including many more students from lower socioeconomic status backgrounds than the more typical convenience samples of college students or web volunteers. Three indicators of socioeconomic status were available: mother's educational attainment (29.8% of the mothers did not complete elementary school, 23.6% completed elementary school, 34.3% completed high school, and 12.3% completed graduate school); families so poor that they qualified for the federal financial assistance programme (27%) versus not; and living in a neighbourhood where there was no sidewalk (15%) versus not.

2.1.2. Measures

Students provided self-ratings on SENNA v2.0. (Primi, Santos, De Fruyt, & John, 2019), but the analyses reported here are focused on one domain, conscientious self-management. This scale is composed of five facets measuring achievement, focus, order, perseverance, and sense of responsibility. Each facet is assessed by six items, three PK (e.g., 'I do my tasks the best way I can') and three NK (e.g., 'I put little effort into my tasks'), forming a fully balanced 30-item scale consisting of 15 pairs of antonym items. Items were answered on a five-point Likert scale: '1' (not at all like me), '2' (a little like me), '3' (moderately like me), '4' (a lot like me) and '5' (completely like me).

2.1.3. Design and data analysis

We calculated several indices for each of the three methods described in Section 1. For the classical method (Ten Berge, 1999) we calculated total average agreement scores (scr_j), indices of acquiescence (acq_j), and acquiescence-controlled scores ($scr.rec_j$). For the DPF method (Andrich, 1978; Linacre, 2017), we first calibrated Samejima's GRM parameters for all items. Then we estimated person parameters twice: once using PK items θ_{pj} and the second time using the NK items θ_{nj} while keeping items anchored at the item parameters from the main analysis. Then we calculated $DPF_j = \theta_{nj} - \theta_{pj}$. Finally, we employed the MIRT-RI random-intercept method using Samejima's GRM as described in Section 1. We calibrated item thresholds and discrimination parameters for the trait factor, whereas for the acquiescence factor we fixed item discrimination parameters as +1 for the PK and −1 for the NK items and estimated the factor's variance. In a second step, we estimated expected *a posteriori* factor scores for trait (θ_j) and acquiescence (ζ_j). Data and code are available at (https://github.com/rprimi/acqu_mirt). All analyses were run in the *mirt* package in R using an expectation–maximization (EM) algorithm for parameter estimation (Chalmers, 2012).

2.2. Results

Table 2 shows descriptive statistics and correlations among the main parameters estimated under the three models. As predicted, DPF, acquiescence acq_j and random intercept parameters ζ_j correlated very highly (r-values from .96 to .98). Substantive trait θ_j correlated very highly with scr_j and $scr.rec_j$ ($r = .97$).

The upper part of Figure 1 shows the relationship between the recoded scores $scr.rec_j$ on the y-axis and the acquiescence index acq_j on the x-axis. It also shows DPF measures as shades of grey. This figure clearly illustrates how DPF relates to recoded scores. If acquiescence acq_j is close to the expected score of 3 (DPF = 0), the recoded scores cover a full amplitude from −2 to 2. The more the acquiescence index goes to the extremes (towards 5, indicating high acquiescence, or towards 1, indicating disacquiescence), the less extreme the recoded scores become, reducing the amplitude of variation. The lower part of Figure 1 shows the corresponding parameters but now obtained in MIRT-RI. It shows a parallel pattern. These results are consistent with the idea that a MIRT-RI is close to classical scores corrected for acquiescence.

3. Simulation study

Here our goal was to replicate the relationships found in the empirical study using a simulation with unbalanced scales, which is the most common situation in practical applications.

3.1. Method

We started from a model based on MIRT-RI of two uncorrelated factors representing trait and acquiescence. Substantive-trait and acquiescence factors were simulated from normal distributions with $M = 0$ and $SD = 1$, and $M = 0$ and $SD = 0.40$, respectively. Samples of 10,000 subjects were generated for each replication. We generated three tests of 12 Likert scale items with five response categories according to Samejima's GRM with the following

Table 2. Descriptive statistics and correlations among estimated measures via three models: classical, DPF, and MIRT-RI

Variable	*M*	*SD*	1	2	3	4	5
Classical method							
1. Original: scr_j	3.58	0.58					
2. Recoded: $scr.rec_j$	0.58	0.58	1.00**				
3. Acquiescence: acq_j	2.97	0.35	−.04**	−.04**			
DPF							
4. DPF_j	0.03	0.96	−.08**	−.08**	.96**		
MIRT-RI							
5. F1 (Trait, θ_j)	−0.01	0.96	.97**	.97**	.05**	.00	
6. F2 (RI, ζ_j)	−0.00	0.56	−.08**	−.08**	.98**	.98**	.01

Notes. DPF = differential person functioning; *M* = mean; *SD* = standard deviation.
Scores from classical method: scr_j = classical scores calculated via average item endorsement after reversing negatively phrased items; $scr.rec_j$ = classical scores controlled for acquiescence, partialling out the within-subject mean; acq_j = classical acquiescence index calculated via average endorsement of antonym pairs of items before reversing negatively phrased items. Score from DPF method: DPF_j, difference between person parameters estimated with positively versus negatively phrased items. Scores from MIRT-RI method: F1, trait θ_j estimated substantive trait; F2 RI ζ_j, estimated random intercept ζ_j measuring acquiescence.
**$p < .01$.

features: a balanced test having six PK items and six NK items (BL); a test unbalanced towards the positive side having nine PK and three NK items (unbalanced positive, UP); and a test unbalanced towards the negative side having three PK and nine NK items (unbalanced negative, UN). In each of these conditions 50 replications of 10,000 subjects were generated.

Item discrimination parameters for the trait factor were generated from a lognormal distribution with $M = 1.26$ and $SD = 0.42$. Item discrimination parameters for the acquiescence factor were fixed at +1 for PK and −1 for NK items. Item thresholds were randomly selected from a set of 30 item calibrations (15 PK and 15 NK) from our previous empirical example. We selected PK and NK item thresholds from the subsets of 15 PK and 15 NK items, respectively. The selected values of parameters defined above (*SD* for acquiescence, *M* and *SD* for discrimination parameters and thresholds) were based on the values found in the empirical example so as to simulate a typical set-up of large-scale assessments of socio-emotional skills.

In total we had simulated 150 samples of 10,000 subjects (50 × 3 conditions representing the item–keying balance in the scale). We calculated the same set of scores for the three methods (classical, DPF and MIRT-RI) as described in the empirical study. We further estimated scores of a unidimensional GRM solution to compare what happens when acquiescence ('do-nothing approach', examined by Savalei & Falk, 2014) is not modelled. The primary variables of interest were the scores estimated with different methods (DPF trait scores and acquiescence scores) and their relationships. Thus, we computed correlations among all scores and with the true values of trait and acquiescence for each sample. We then averaged values across replications in the three balance conditions (BL, UP, and UN).

Analyses were again done in the *mirt* package in R using the EM algorithm (Chalmers, 2012); see https://github.com/rprimi/acqu_mirt.

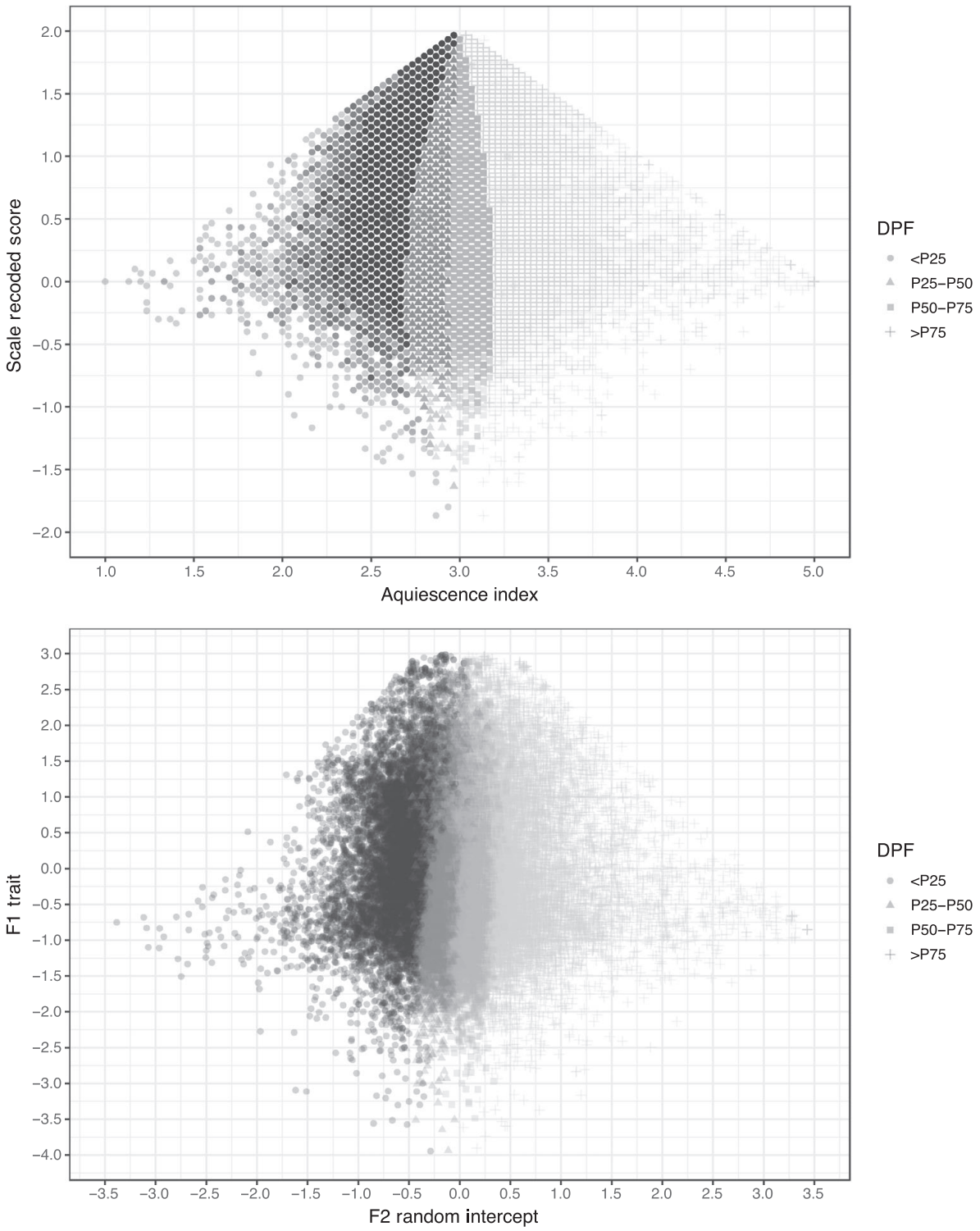

Figure 1. (Top) Relationship between recoded scores scr.rec$_j$ (y-axis) and acquiescence index acq$_j$ (x-axis). (Bottom) MIRT-RI substantive trait F1 (θ_j) and MIRT-RI F2 (ζ_j), Both figures are shaded by the level of differential person functioning.

3.2 Results

Table 3 presents the main results of the simulations: the average correlations among scores calculated from three different methods (classical, DPF and MIRT RI) in three conditions (BL, UP and UN). The first three variables are scores from the classical method

(average agreement, scr_j; recoded scores, $scr.rec_j$; and acquiescence index, acq_j). Variable 4 is the DPF index. Variables 5 and 6 are trait (θ_j) and acquiescence (ζ_j) scores estimated by the MIRT random intercept model. Variable 7 is a trait score estimated with a unidimensional model. Variables 8 and 9 are the true trait and acquiescence scores from where the data were simulated.

First, we consider the classical method. Original scores scr_j correlated perfectly with recoded scores $scr.rec_j$ ($r = 1$), that is, in balanced scales the total scores are already corrected for acquiescence. In the case of unbalanced scales scr_j and $scr.rec_j$, correlations

Table 3. Average descriptive statistics and correlations across 50 replications by conditions (balanced [BL], unbalanced positive [UP] and unbalanced negative [UN]) among estimated measures via three models: classical scores, differential person functioning (DPF), MIRT random intercept and true values

	M	*SD*	1	2	3	4	5	6	7	8
					BL condition					
Classical method										
1. scr_j	2.62	0.67								
2. $scr.rec_j$	0.62	0.67	1.00							
3. acq_j	1.98	0.35	−.09	−.09						
DPF										
4. DPF_j	0.0	0.76	−.02	−.02	.96					
MIRT RI										
5. F1 trait θ_j	0.0	0.92	.98	.98	−.05	.01				
6. F2 RI ζ_j	0.0	0.47	−.03	−.03	.97	.98	.01			
Unidimensional										
7. F1 1D (GRM)	0.0	0.91	.98	.98	−.04	.02	1.00	.02		
True person parameters										
8. F1 true trait θ_j	0.0	1.00	.91	.91	−.05	.01	.92	.00	.92	
9. F2 true acq. ζ_j	0.0	0.63	−.02	−.02	.73	.73	.01	.75	.01	.00
					UP condition					
Classical method										
1. scr_j	2.58	0.67								
2. $scr.rec_j$	0.45	0.53	.95							
3. acq_j	2.27	0.43	.72	.47						
DPF										
4. DPF_j	0.0	0.76	.45	.15	.92					
MIRT RI										
5. F1 trait θ_j	0.0	0.91	.98	.96	.62	.34				
6. F2 RI ζ_j	0.0	0.53	.29	−.03	.85	.95	.15			
Unidimensional model										
7. F1 1D (GRM)	0.0	0.92	.98	.89	.79	.57	.97	.40		
True person parameters										
8. F1 true trait θ_j	0.0	1.00	.89	.88	.57	.31	.91	.14	.88	
9. F2 true acqu ζ_j	0.0	0.63	.20	−.02	.61	.68	.10	.71	.28	.00
					UN condition					
Classical method										
1. scr_j	2.63	0.69								
2. $scr.rec_j$	0.46	0.50	.95							

Continued

Table 3. (*Continued*)

	M	*SD*	1	2	3	4	5	6	7	8
3. acq_j	1.66	0.49	−.81	−.59						
DPF										
4. DPF_j	0.00	0.78	−.47	−.19	.88					
MIRT RI										
5. F1 trait θ_j	0.0	0.91	.96	.97	−.67	−.31				
6. F2 RI ζ_j	0.0	0.52	−.33	−.03	.81	.95	−.14			
Unidimensional model										
7. F1 1D (GRM)	0.0	0.92	.98	.91	−.83	−.54	.97	−.38		
True person parameters										
8. F1 true trait θ_j	0.0	1.00	.88	.89	−.61	−.29	.91	−.13	.88	
9. F2 true acqu ζ_j	0.0	0.63	−.23	−.02	.58	.68	−.10	.72	−.27	.00

Note. Scores from the classical method: scr_j = classical scores calculated via average item endorsement after reversing negatively phrased items; $scr.rec_j$ = classical scores controlled for acquiescence, partialling out the within-subject mean; acq_j = classical acquiescence index calculated via average endorsement of antonym pairs of items before reversing negatively phrased items. Score from the DPF method: DPF_j, difference between person parameters estimated with positively versus negatively phrased items. Scores from MIRT-RI method: F1, trait θ_j estimated substantive trait; F2 RI ζ_j, estimated random intercept measuring acquiescence (the same notation is used for the correspondent true generation parameters). F1 1D (graded response model [GRM]), estimated graded response scores of a unidimensional model.

are slightly lower than 1. When items are balanced, the correlation of acquiescence acq_j with scr_j/$scr.rec_j$ was close to 0. When items are unbalanced, acq_j correlates positively with scr_j/$scr.rec_j$ in UP and negatively in UN (r = .72 and r = −.81, respectively). Considering the correlations of $scr.rec_j$ and acq_j with true values of trait θ_j and acquiescence ζ_j, we see that in BL it achieves r = .91/.73, in unbalanced UP r = .88/.61 and in unbalanced UN r = .89/.58. In unbalanced conditions we see a small correlation of original scores with true acquiescence in the direction of unbalancedness: UP r = .20 and UN r = −.23, but recoded scores do not correlate with true acquiescence (both r = −.02). Also, in unbalanced conditions acquiescence scores acq_j correlate with true trait (UP r = .57, UN r = −.61). This suggests that the unbalanced condition biases acquiescence and trait scores. In unbalanced conditions recoded scores $scr.rec_j$ tend to be less dispersed (SD = 0.53/0.50 for UP/UN, SD = 0.67 for BL) due to partialling out the effect of acquiescence from the original scores.

Second, we consider the DPF index. As predicted, DPF is highly related to the acquiescence index (r = .96, .92, .88 for BL, UP, UN, respectively). It also shows the expected correlation with true acquiescence (r = .73, .68, .68, respectively). It is also slightly biased by trait information in unbalanced conditions.

Third, we consider scores from the MIRT-RI method. Again, relationships are consistent with predictions and the results of the empirical study. Random intercept parameters ζ_j are highly correlated with acquiescence acq_j (r = .97/.85/.81) and DPF (r = .98/.95/.95). Relationships of estimated trait and acquiescence with true values are particularly important since this method is the same as the true model behind simulations. Estimated trait θ_j parameters correlated r = .92/.91/.91 with true values in all conditions,

and random intercept ζ_j correlated r = .75/.71/.72 with true values of acquiescence.[3] Regardless of condition, the MIRT-RI method performed very similarly. Nevertheless, the correlation between estimated trait and random intercept seems to be slightly biased, positively in UP (r = .15) and negatively in UN (r = −.14).

Finally, we consider the unidimensional GRM. The unidimensional GRM solution was highly correlated with true trait (r = .92, .88 and .88 for BL, UP and UN, respectively). In UP and UN it also correlates with true acquiescence (r = .28, −.27). It performed poorly like the original scores in unbalanced conditions. GRM model fit indices were lower than the MIRT-RI version (the correct model). An average fit index across 150 simulated data sets for the GRM model was M2 = 204.3, p = .02, with root mean square error of approximation (RMSEA) .023, Tucker–Lewis index (TLI) .893 and comparative fit index (CFI) .93, as compared to M2 = 17.5, p = .47, RMSEA = .002, TLI = .99, CFI = .99 for the MIRT-RI model. Taken together, the simulation study replicated the conceptual connections between acquiescence and DPF found in the empirical study.

4. Discussion

Self-report questionnaires rely on the assumption that their items reflect information only about the traits they are supposed to measure; that is, persons will respond mainly to the content of the questions. The item response process, however, is complex and multidimensional, and involves other factors that may substantially affect subjects' responses (McCrae, 2018; Wetzel & Carstensen, 2015). Especially in children, the response style of acquiescence is an important source of systematic error (Soto *et al.*, 2008). Balanced scales have been proposed as a solution to control for acquiescence (Jackson, 1971), but the reason why they work has not been fully examined using modern psychometric models.

We compared three methods for measuring and controlling for acquiescence (classical, DPF, and MIRT-RI). Johanson and Osborn's (2004) proposal conceptualizes acquiescence as DPF. We elaborated DPF in the context of a multidimensional IRT model as an extra factor beyond substantive trait variance. Results from both empirical data and simulations show converging evidence that the acquiescence index is highly associated with DPF, that balanced scales automatically correct for the influence of acquiescence and DPF (see Ferrando & Lorenzo-Seva, 2010, for a similar conclusion), and that MIRT-RI models acquiescence and DPF and produces acquiescence-controlled trait scores close to balanced scales, even when scales are unbalanced.

To understand the functioning of balanced scales, two aspects of the response process are of crucial importance: responding to the rating scale (e.g., asking for levels of

[3] Since this is the correct model it is surprising that correlation of estimated acquiescence with true values does not achieve values close to 1. One feature to note is that the information we have to estimate the random intercept is much lower than the trait. The proportion of variance explained by the random intercept is generally around 7%, compared to 29% for the trait. Since this is related to the item's loading/discriminations, and discrimination is related to information (inverse of error of measurement), the error of measurement to estimate the random intercept is large compared to the error when estimating the trait. Therefore, the magnitude of correlations of estimated values with true values obtained from the MIRT random intercept can be considered a high benchmark possible given the error of measurement. It may be difficult to see the relative lower information of acquiescence factor since item discrimination parameters are all fixed to 1 or −1. Because it explains a lower amount of item response variance, to compensate for these high values of fixed item discrimination, the variance of the latent factor of acquiescence is reduced (see the discussion on IRT parameterizations in Baker, 1990). In a very similar study, Ferrando and Lorenzo-Seva (2010) derived theoretically expected correlations of true versus simulated data sets of a similar model. Expected correlations never reached 1 and were very similar to the ones reported here.

agreement versus disagreement) and responding to the item content. Mean ratings (or endorsement) on the rating scale *before* reversing any scores will capture how strongly the individual reacts to the different levels of the rating scale (i.e., choosing generally higher or lower numbers). Reactions to item content will produce consistent responses of agreement/disagreement because for PK items a high position on the latent trait is associated with agreement, whereas for NK items this position is associated with disagreement. Therefore, the process of responding to the substantive content of the item will be captured by the consistency of reflected responses on antonym pairs, that is, by the behaviour in the direction of the keying. The bias component in the response process (e.g., difficulty or ambiguity of understanding the item) is captured by the acquiescence score, that is, overall behaviour of agreement (or disagreement in the case of disacquiescence) with antonym pairs (see Primi *et al.*, 2018).

A key aspect of balanced scales is the design of items measuring the same construct but from opposite poles. This makes it possible to disentangle method from trait variance. These concepts of item design that try to purify measurement by manipulating item features are commonly used in cognitive measurement (De Boeck & Wilson, 2004; Embretson, 1983, 1994; Primi, 2014) but only rarely in personality assessment. A thoughtful example of item design was presented by Mirowsky and Ross (1991) who analysed internal and external locus of control with items that allowed them to isolate acquiescence and defensiveness. They used the same principle of balancing item features with difference scores between groups of items to identify substantive information separately from bias.

Conceptually, the MIRT random intercept is consistent with this approach of using item design features, as it defines item difficulty parameters that are split into two components (c_i and ζ_j). Note that ζ_j was only identified as acquiescence because of item design features that produced antonym pairs PK and NK and fixed item discrimination in order to capture a person's main effect of how easy it is to agree with all items regardless of their content. Since ζ_j is a random effect and varies over j persons, it is similar to a second latent factor. We showed that this parameter could be interpreted as a person-tailored adjustment of item difficulty based on their tendency to agree indiscriminately. A person A with the same trait level as person B on substantive trait θ_j, but manifesting high acquiescence $\zeta_A > \zeta_B$, will have an inflated trait score if not adjusted down proportionally to the difference in acquiescence. Therefore ζ_j indexes how globally easy it is to agree (or disagree) for a particular person j varying over all persons and adjusting test difficulty individually, partialling out this difference from substantive trait variance that becomes less influenced by acquiescence and DPF.

We showed that acquiescence and DPF correction happens automatically in balanced scales. But what happens when scales are unbalanced? Our simulations replicated findings of Ferrando and Lorenzo-Seva (2010) that one of the most important consequences of imbalance is a biased correlation between trait and acquiescence. Even if in the true model there is no correlation between the two factors, having more PK items produced a biased positive correlation and having more NK items produced a biased negative correlation between trait and acquiescence. This can potentially bias correlations with external criteria (see the detailed discussion in Ferrando *et al.*, 2003; Mirowsky & Ross, 1991; Primi *et al.*, 2018; Rammstedt *et al.*, 2010; Soto & John, 2019). We found limited evidence in our simulation study that the MIRT random intercept is relatively influenced by this imbalance. This finding, however, needs to be interpreted with some caution given that the data were simulated from MIRT-RI (where the parameter recovery was imperfect) and only one condition was considered in the simulation study.

Some limitations of the present paper should be noted. First, acquiescence measurement relies on antonym pairs and on the assumption that they measure the same trait in opposite poles having equivalent difficulty and sensitivity to acquiescence. But Ten Berge (1999) has argued that items may be differentially influenced by acquiescence requiring different item loadings and discrimination parameters. Indeed, MIRT-RI modelled item difficulty and discrimination differently. But keying balance is based on a strong assumption that antonym pairs are equal and cancel each other out. We have not explicitly tested this assumption and let item discrimination be freely estimated across antonym pairs. This may have biased the estimation of the acquiescence factor. More flexible models, as described by Falk and Cai (2016), could be used to test these assumptions. Finally, our simulation study may be limited in terms of generality because it was based on a single (albeit large) sample and mimicked the empirical features of that sample. In conclusion, future simulation studies are needed to test the assumption of equal discrimination within antonym pairs, to examine other features of the response process (e.g., the strength of the influence of acquiescence on item responses) and to explore different IRT models.

Acknowledgements

This article is part of a research project supported by the Ayrton Senna Foundation. The first author also received a scholarship from the National Council on Scientific and Technological Development (CNPq).

References

Abrahams, L., Pancorbo Valdivia, G., Primi, R., Santos, D., Kyllonen, P., John, O. P., & De Fruyt, F. (2019). Social-emotional skill assessment in children and adolescents: Advances and challenges in personality, clinical, and educational contexts. *Psychological Assessment*, *31*, 591–600. https://doi.org/10.1037/pas0000591

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. https://doi.org/10.1007/bf02293814

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement*, *14*(2), 139–150. https://doi.org/10.1177/014662169001400203

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. https://doi.org/10.1207/s15328007sem0704_5

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612. https://doi.org/10.1007/s11336-010-9178-0

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63*, 369–399. https://doi.org/10.1007/bf02294861

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), Cognitive assessment. A multidisciplinary perspective (pp. 107–135). New York: Plenum Press.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328–347. https://doi.org/10.1037/met0000059

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427–448. https://doi.org/10.1348/000711009x470740

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. https://doi.org/10.1207/s15327906mbr3803_04

Garrido, L. E., Golino, H., Nieto, M. D., Peña, K. G., & Molina, A. M. (2018). *A systematic evaluation of wording effects modeling under the ESEM framework*. Paper presented at the International Meeting of the Psychometric Society (IMPS), New York.

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology*, *45*, 1028–1045. https://doi.org/10.1177/0022022114534773

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, *78*, 229–248. https://doi.org/10.1037/h0030852

Johanson, G. A., & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education*, *29*, 535–548. https://doi.org/10.1080/02602930410001689126

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Linacre, J. M. (2017). *Winsteps Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com/winman/

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362. https://doi.org/10.1037/1082-989x.11.4.344

Maydeu-Olivares, A., & Steenkamp, J.-B. E. M. (2018). *An integrated procedure to control for common method variance in survey data using random intercept factor analysis models*. Retrieved from https://www.academia.edu/36641946/An_integrated_procedure_to_control_for_common_method_variance_in_survey_data_using_random_intercept_factor_analysis_models

McCrae, R. R. (2018). Method biases in single-source personality assessments. *Psychological Assessment*, 30, 1160–1173. https://doi.org/10.1037/pas0000566

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly*, *54*, 127–145. https://doi.org/10.2307/2786931

Primi, R. (2014). Developing a fluid intelligence scale through a combination of Rasch modeling and cognitive psychology. *Psychological Assessment*, *26*, 774. https://doi.org/10.1037/a0036712

Primi, R., De Fruyt, F., John, O. P., & Santos, D. (2018). Validities of true and false keyed social-emotional skill items before and after acquiescence correction to predict educational achievement. *Paper presented at the 11th International test Commission Conference*, Montreal, Canada.

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). *SENNA V2.0 technical manual*. São Paulo: Instituto Ayrton Senna.

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*(1), 53–61. https://doi.org/10.1016/j.jrp.2009.10.005

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, *1968*(1), 1–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424. https://doi.org/10.1080/00273171.2014.931800

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment*, *31*, 586–590. https://doi.org/10.1037/pas0000586

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

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. https://doi.org/10.1007/bf02294363

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*(1), 89–102. https://doi.org/10.1207/s15327906mbr3401_4

Wetzel, E., & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*, 352–364. https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279–291. https://doi.org/10.1177/1073191115583714

Received 16 February 2018; revised version received 7 February 2019

Appendix A: Acquiescence correction as a noise-cancelling mechanism

A good analogy for what is going on when we use acquiescence correction is to consider the way noise-cancelling headphones work. These headphones have a microphone inside their cups to 'hear' external noise and neutralize its distorting effects. A new inverted version of the noise sound wave (180 degrees out of phase) is created that is sent into a speaker. Through the speaker, you now hear signal, noise, and inverted noise waves. Since there are two noise waves, one positive and one negative, they cancel each other out and a clearer signal emerges. This is analogous to what happens with balanced personality scales: trait is signal, bias is noise. The only difference is that signal, not noise, is inverted (via PK and NK items) and inserted into the scale. Noise, that is, acquiescence is a positive signal that happens to occur with both PK and NK items. When scoring the test, NK items are reversed, and item scores are summed into a scale score. While doing this, NK items are inverted, aligning them to the positive pole, transforming the signal in the same 'phase' as positive items, finally ending up with equal amounts of positive and negative noise that cancel each other out. Thus, if we carefully develop scales with balanced items to isolate bias in the positive manifold, the remaining variance after ipsatizing will be a purified measure. The formula $\text{scr.rec}_j = 1/6\left[\sum_{i=1}^{3} x_{ij} - \sum_{i=a}^{c} x_{ij}\right]$ makes the analogy with noise-cancelling headphones more salient. In headphones you have an audio signal (+ noise) *plus* (− reflected noise), therefore noise cancels out. In balanced scales, positive items are composed of +trait signal *plus* (+acquiescence bias), and negative items are composed of −trait signal *plus* (+acquiescence bias). When we compute [+trait signal +

(+acquiescence bias)] *minus* [−trait signal + (+acquiescence bias)], the acquiescence cancels out and we end up only with the trait signal.

Appendix B: MIRT, DIF and DPF

In order to make the relationships of random intercept/acquiescence ($\zeta_j \simeq \text{acq}_j$) and the latent trait/DPF corrected measure ($\theta_j \simeq \text{scr.rec}_j$ more salient, let us consider two semantically opposite items, one PK item *ip* and another NK item *in*, such as, 'I do my tasks the best way I can' versus 'I put little effort into my tasks', where a participant should give a binary yes/no answer. The *yes* response is scored as $s = 1$. Let us reverse the NK item, computing $1 - s_{in}$ so that its score will be in the same direction as the PK item s_{ip}. Assume also that both questions measure the same latent trait, have the same difficulty $b_{ip} = b_{in} = b_i$, the same discrimination $a_{ip} = a_{in} = a_i$, and finally that the latent trait is not correlated with acquiescence ($r_{z\theta} = 0$; see justifications for this assumption in Ferrando *et al.*, 2003). The log-odds of scoring 1 on these two items are, for the positively worded item,

$$\ln\left(\frac{P_{s_{ip}}}{1 - P_{s_{ip}}}\right) = a_i\theta_j + (1\zeta_j) + b_i,$$

and for the negatively worded item,

$$\ln\left(\frac{P_{s_{in}}}{1 - P_{s_{in}}}\right) = a_i\theta_j + (-1\zeta_j) + b_i.$$

If there is no DPF, subjects will answer both items in the same way but reflected. Since the negative item was reversed, the scored answers are expected to be the same. For a subject with a given θ_j the probability of scoring 1 will be equal in both items, $P_{s_{ip}} = P_{s_{in}}$. So the item characteristic curves (ICCs) will be the same. If there is DPF, subjects no longer answer each item in the same way and the ICCs will differ. This is indeed a characteristic situation that defines DIF, that is, a given level on the primary trait having different probabilities of scoring 1 on the item. If the random intercept ζ_j is an index of DPF, it should capture these differences. If we subtract $P_{s_{ip}} - P_{s_{in}}$ we see that this relationship holds true:

$$\ln\left(\frac{P_{s_{ip}}}{1 - P_{s_{ip}}}\right) - \ln\left(\frac{P_{s_{in}}}{1 - P_{s_{in}}}\right) = a_i\theta_j + (+1\zeta_j) + b_i - [a_i\theta_j + (-1\zeta_j) + b_i] = 2\zeta_j.$$

That is, for a subject with $\zeta_j = 0$, indicating no acquiescence, the ICCs will be equal (the difference between probabilities will be equal to zero), indicating that they answered positive and negative items equally. If there is acquiescence, then $\zeta_j > 0 \Rightarrow P_{s_{ip}} > P_{s_{in}}$, that is, it will be relatively easier to score 1 on the positive than on the negative item. In the case of disacquiescence, $\zeta_j < 0 \Rightarrow P_{s_{ip}} < P_{s_{in}}$, that is, the opposite pattern will occur, and it will be relatively easier to score 1 on the negative item (remember that the negative item was already reversed). This relationship can be extended to Likert scales. The difference is that there will be one equation for each binary response with different item difficulties indexed by c_{it}.

When PK and NK items measure the same construct but have different parameters the relationship will be more complex, given by

$$\ln\left(\frac{P_{s_{ip}}}{1-P_{s_{ip}}}\right) - \ln\left(\frac{P_{s_{in}}}{1-P_{s_{in}}}\right) =$$
$$a_{ip}\theta_j + (+1\zeta_j) + b_{ip} - [a_{in}\theta_j + (-1\zeta_j) + b_{in}] = \theta_j(a_{ip} - a_{in}) + (b_{ip} - b_{in}) + 2\zeta_j.$$

Therefore, more generally, differences in likelihood of scoring 1 in semantically opposite items are related to the differences in the item–trait relationship, $a_{ip} - a_{in}$, differences in item difficulties, $b_{ip} - b_{in}$, and the presence of DPF, $2\zeta_j$. One important advantage of IRT modelling is that all item parameters are modelled simultaneously, making it possible to identify DPF unconfounded with eventual global differences in PK and NK item characteristics.