Controlling Acquiescence Bias with Multidimensional IRT Modeling

Ricardo Primi, Nelson Hauck-Filho, Felipe Valentini, Daniel Santos and Carl F. Falk

Abstract Acquiescence is a commonly observed response style that may distort respondent scores. One approach to control for acquiescence involves creating a balanced scale and computing sum scores. Other model-based approaches may explicitly include an acquiescence factor as part of a factor analysis or multidimensional item response model. Under certain assumptions, both approaches may result in acquiescence-controlled scores for each respondent. However, the validity of the resulting scores is one issue that is sometimes ignored. In this paper, we present an application of these approaches under both balanced and unbalanced scales, and we report changes in criterion validity and respondent scores.

Keywords Acquiescence bias · Item response modeling

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). Correspondence should be addressed to: Ricardo Primi, Universidade São Francisco, Rua Waldemar César da Silveira, 105, Campinas, São Paulo, CEP 13045-510; e-mail: rprimi@mac.com.

R. Primi (✉) · D. Santos
Ayrton Senna Institute, Universidade São Francisco, and EduLab21, Campinas, Brazil
e-mail: rprimi@mac.com

D. Santos
e-mail: daniel.ddsantos@gmail.com

N. Hauck-Filho · F. Valentini
Universidade São Francisco, Campinas, Brazil
e-mail: hauck.nf@gmail.com

F. Valentini
e-mail: valentini.felipe@gmail.com

C. F. Falk
McGill University, Montreal, Canada
e-mail: carl.falk@MCGILL.CA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_4

1 Introduction

1.1 Large-Scale Assessment of Socioemotional Skills and the Self-report Method

Evidence has consistently indicated that socioemotional skills can predict many life outcomes (Ozer & Benet-Martínez, 2006), including job-related variables (Heckman, Stixrud, & Urzua, 2006), quality of life (Huang et al., 2017), psychopathology (Samuel & Widiger, 2008), and physical health (Allen, Walter, & McDermott, 2017). Among students, such skills have been associated with academic performance even when partialling out intelligence (Poropat, 2014), perhaps because these skills foster multiple learning strategies and positive self-beliefs (Zhang & Ziegler, 2018). Considering that such individual differences not only change over time (Soto, John, Gosling, & Potter, 2011), but can also be enhanced via school-based interventions (Lipnevich, Preckel, & Roberts, 2016), they represent key variables to modern national education policies.

Although many strategies exist for the assessment of socioemotional skills among students, the self-report method is recommended because it is simple, easy, and has a low cost compared to alternative techniques (Kyllonen, Lipnevich, Burrus, & Roberts, 2014). One recently published self-report inventory designed for the assessment of non-cognitive skills among students is SENNA (Primi, Santos, John, & De Fruyt, 2016). It contains 18 self-report scales using 5-point Likert-type items and provides researchers and public agencies with information on five broad dimensions of socioemotional skills: Open-mindedness (O), Conscientious Self-Management (C), Engaging with others (E), Amity (A), and Negative-Emotion Regulation (N) (John, Naumann, & Soto, 2008).

1.2 Self-report Method and Response Styles

Although the self-report method has many merits, it does not result in error-free information about respondents. Scores calculated on self-report data may be contaminated by random error or by systematic components unrelated to the trait of interest. Systematic biases include "response styles" (Paulhus, 1991) or "method variance" (McCrae, 2018). Response styles (RS) represent individual differences in the usage of response scales. For instance, some respondents will tend to manifest their agreement or disagreement with the content of an item by choosing the extremes of the Likert scale, while others will systematically avoid extremes. RS represent relatively stable individual differences (Weijters, Geuens, & Schillewaert, 2010; Wetzel, Lüdtke, Zettler, & Böhnke, 2015) and may account for up to 40% of item variance (McCrae, 2018). When separating trait and state components in repeated measures designs, response styles seem to be responsible for up to 59% of

systematic state variance (Wetzel et al., 2015). By adding nuisance variance to the data, RS can impair the validity and reliability of test scores (Ziegler, 2015).

1.3 Acquiescence and the Assessment of Socioemotional Skills

Acquiescence (ACQ) is one response style that deserves closer attention in the self-report assessment of socioemotional skills among youths. ACQ refers to a tendency to agree with items at the expense of their content (Paulhus, 1991). For instance, a student might indicate that he or she agrees (e.g., "4" on a 5-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree) with two items such as "I am an introvert" and "I am an extravert." Of course, such a response pattern is semantically contradictory, and it indicates agreement in detriment to consistency. In some cases, ACQ may reflect cognitive simplicity (Knowles & Nathan, 1997) as it occurs more often among under-educated people (Meisenberg & Williams, 2008), older adults (Weijters et al., 2010) and younger children and adolescents (Soto, John, Gosling, & Potter, 2008).

With respect to self-reports of socioemotional skills, ACQ has the potential to diminish correlations between semantically opposite items, creating method factors among negatively worded items (Kam & Meyer, 2015). ACQ can also increase correlations among items capturing unrelated traits (Soto et al., 2008). Accordingly, factor structure distortions are very likely to occur in the presence of ACQ. In a simulation study, ACQ caused classical parallel analysis and Hull methods to overestimate the number of factors to retain, and MAP and permutation parallel analysis to underestimate it (Valentini, 2017).

Moreover, ACQ can attenuate external validity (Mirowsky & Ross, 1991). ACQ tends to inflate scores of scales composed of mostly positively worded items. Thus, ACQ might impact the validity of a scale in a manner proportional to the amount of positively- and negatively-keyed items. At the same time, ACQ is often negatively related to achievement, suggesting that high ACQ can be explained in part by low language skills. Therefore, the criterion validity of socio-emotional skills may be suppressed by ACQ. In real data and using a classical scoring approach, Primi, De Fruyt, Santos, Antonoplis, and John (2018) found that partialling out ACQ resulted in disattenuated associations of socioemotional skills with achievement tests of language (from .13 to .21) and math (from .11 to .17).

1.4 Controlling for Acquiescence

One traditional way of controlling for ACQ is to create a balanced scale in which each positively worded item is paired with an antonym (a negatively worded item),

such as: *I am often talkative /I am often quiet*. On balanced scales, it is expected that subjects will give mirrored responses to antonym pairs (e.g., 5-1, 4-2, 3-3, 2-4 and 1-5 on a 5-point Likert-type item). If the response pattern of subject *j* is semantically consistent, then the average of subject *j*'s item responses before reverse coding will be the midpoint of the response options (in this case, 3; Soto et al., 2008). The person's average of the item responses before reverse coding negative items is the classical index of ACQ (acq_j).

Under certain assumptions (e.g., positively and negatively worded items are on average equally vulnerable to ACQ; Savalei & Falk, 2014a, b), classical scoring procedures will result in unbiased estimates of the respondents' scores. In essence, the effect of ACQ on positive and negatively worded items "cancels out" when computing a total score. For example, Primi et al. (2019) shows that scr_j, the classical average score of subject *j* on a balanced scale (with a 5-point Likert type item scored from 1 to 5), can be written as:

$$scr_j = 3 + \frac{1}{2}\left(\frac{\sum_{i=1}^{k^{(p)}} x_{ij}^{(p)}}{k^{(p)}} - \frac{\sum_{i=1}^{k^{(n)}} x_{ij}^{(n)}}{k^{(n)}}\right)$$

where $k^{(p)}$ equals the number of positive items, $k^{(n)}$ is the number of negative items, $x_{ij}^{(p)}$ and $x_{ij}^{(n)}$ are subject j's original responses (before reverse coding) on positive item i, and negative item i, respectively. Inside parentheses, the classical score is a function of the difference between the average agreement with positive versus negative items. The more inconsistent the responses to antonym items are, the more the term in parentheses will tend towards zero. Semantically consistent responses, however, will tend to result in either larger or smaller scr_j, depending on the subject's standing on the trait.

In unbalanced scales (i.e., $k^{(p)} \neq k^{(n)}$), classical scores may not be fully corrected and ACQ will not fully cancel out. In such a case, a form of within-person centering (or ipsatization) is sometimes recommended to control for ACQ (e.g., Soto et al., 2008). In the first step, an ACQ index (acq_j) is calculated as the average of only antonym pairs of items. Next, ACQ is removed from the raw item scores ($x_{ij} - acq_j$). Raw scores for the reverse-keyed items are then multiplied by -1, and scale scores are obtained by averaging these items with those of the positively worded items.

1.5 Item Response Theory with Questionnaires and Acquiescence

Item response theory (IRT) models are routinely used when scaling constructs derived from questionnaires in large-scale educational assessments. While much is known about the effect of ACQ in balanced and unbalanced classical scores (Ten Berge, 1999; Primi et al., 2019), less is understood about the effect of ACQ on latent trait scores estimated via IRT. Since it is known that ACQ, even with a balanced scale,

may contaminate the covariance structure when performing linear factor analysis (e.g., Savalei & Falk, 2014a, b), we conjecture that IRT-based models may also be vulnerable to the effects of ACQ. For instance, the graded response model and generalized partial credit models are commonly used IRT models for the analysis of ordered polytomous responses (De Ayala, 2009), but may not automatically correct for ACQ. There are, however, a number of model-based approaches that could be used to control for ACQ, such as those based on the random intercept model (Billiet & McClendon, 2000; Cai, 2010; Maydeu-Olivares & Coffman, 2006; Maydeu-Olivares, & Steenkamp, 2018).

Although we provide some details on these models later, some key questions emerge regarding the consequences of ACQ regardless of the method used. Simulations and analytical proofs are useful for studying whether a modeling approach can recover population parameters or results in bias, as well as the consequences of fitting a misspecified model. In practice, however, we never know the true model and whether a more complex modeling approach fits the data better because it is a better approximation to reality or because it is fitting noise. Supposing that we are interested in using self-management scores to predict an objective real-world outcome, we may wonder about the consequences of ignoring ACQ or using a specialized approach to control for it. For example, how does the use of one model versus another affect the validity of IRT scores? Are there differences if questionnaires are balanced or unbalanced? Are there any differences in scoring bias when comparing classical and IRT-based approaches? We therefore present an empirical study comparing the criterion validity of classical scores against four IRT approaches.

2 Method

Our main goal was to explore the criterion validity of self-management scores estimated via IRT. Previous research with classical scores suggests that ACQ suppresses criterion validity, and that ACQ-controlled scores show relatively higher validity (Primi, Santos, De Fruyt, & John, 2018). In the present study, we calculated scores via IRT, and then explored their criterion validity. We wanted to examine if classical scores are similar to ACQ-controlled trait estimates. We also compared these findings on a balanced versus an unbalanced item set.

2.1 Data

We reanalyzed data from Primi et al. (2018). Data comprised of 12,987 adolescents (52.7% female) from grades 7, 9, and 10, who ranged in age from 12 to 20 years ($M = 16$, $SD = 1.85$). Participants were regular students attending 425 public schools located in 216 cities of the state of Sao Paulo. Students completed SENNA as part of a

reading literacy program developed by the Ayrton Senna Institute and in partnership with the state secretariat.

2.2 Instruments

We focused on the 45-item Conscientious Self-Management Scale (C) from the SENNA inventory (Primi et al., 2018). The scale contains 30 antonym pairs, 15 positively-keyed and 15 negatively-keyed items, with an additional 15 positively-keyed items. The scale is therefore unbalanced. In what follows, we performed the analyses twice: Once on the 30 antonym pairs (the balanced item set), and a second time on the complete 45-item set (the unbalanced item set). Students responded using a 5-point scale. We also had two measures of students' academic achievement: standardized assessments for language and math (SARESP—Assessment of Educational Achievement at São Paulo State, in Portuguese—see http://saresp.fde.sp.gov.br). These scores were used as criterion measures.

2.3 Data Analysis and Multidimensional IRT Modeling

In synthesis, the study design crossed two features: (a) two types of item sets: Balanced versus unbalanced; and (b) five psychometric models to calculate scores: Classical, unidimensional IRT via a graded response model (GRM), a unidimensional partial credit model (PCM; e.g., see De Ayala, 2009; Embretson, & Reise, 2000), and two multidimensional IRT models that were an adaptation of the random intercept model but based on either the GRM or PCM. Our main focus was the correlation between self-management and standardized achievement in language and math.

When calculating classical scores, we obtained original scores (*Raw ave*) that are simply the average of item responses after reverse coding negative items (equivalent to computation of scr_j). We also calculated classical ACQ-controlled scores (*ACQ cntr*) using the procedure advocated by Soto et al. (2008) for unbalanced items as mentioned earlier in our manuscript, along with an acquiescence index (*ACQ*) via average endorsement of the 15 antonym pairs before reverse coding. Note that in the case of a balanced scale, *Raw ave* and *ACQ cntr* are equivalent; these indices differ only for unbalanced scales.

To understand the two random intercept models, consider boundary discrimination functions for the GRM as follows

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

where P_{ri} is short-hand for the probability of endorsing category r or higher for item i. For each 5-point Likert-type item there will be four of these equations modeling the

transitions 1 versus 2345, 12 versus 345, 123 versus 34, and 1234 versus 5. a_{1i} is the discrimination for item i on the substantive trait, θ_j, and a_{2i} is a set of fixed weights for item i associated with item wording and designed to capture ACQ. Values of a_{2i} were fixed to 1 if the item was positively worded, and -1 if the item was negatively worded. With this fixed set of weights, ζ_j represents ACQ. Finally, c_{ir} represents an intercept term. This model is similar to what Maydeu-Olivares and Steenkamp (2018) named the compensatory random-intercept model (see also Cai, 2010).

To estimate the model, we freed item discriminations (a_{1i}), and constrained the trait variance to 1. Since we fixed all a_{2i} parameters, we freed the variance of the acquiescence factor, ζ_j, and fixed the covariance between trait and acquiescence to zero for identification. We also estimated a second model with all specifications similar to the GRM but using a multidimensional PCM. This model fixed item discriminations to 1, and estimated substantive trait variance. After calibrating item parameters, we estimated subject factor scores using the *Expected* a Posteriori (EAP) algorithm. Trait and acquiescence scores were named *GRM f1* and *GRM f2*, respectively, for the GRM and *PCM f1* and *PCM f2*, respectively, for the PCM. (Chalmers, 2012)

3 Results

Table 1 shows descriptive statistics and criterion validity of the distinct types of scores investigated here. Whereas the upper half of the table shows scores calculated with a set of items balanced with respect to item wording, the lower half displays the same set of scores but calculated using the unbalanced set of items. The last two columns show zero-order correlations of various scores with standardized achievement in language and math.

Some key points are worth noticing in Table 1. First, considering classical scores in the balanced condition, we found that self-management was positively associated with achievement in magnitudes consistent with previous literature (see Poropat, 2009), while acquiescence tended to be negatively associated with achievement (Mirowsky & Ross, 1991). Second, *Raw ave* and *ACQ cntr* had the same association with achievement ($r = .22$ and .18 for language and math, respectively). Considering the unbalanced item set, *Raw ave* showed smaller correlations with achievement ($r = .16$ and .14) than did acquiescence-controlled scores, *ACQ cntr* ($r = .20$ and .16). This result is likely a consequence of the suppression effect of ACQ discussed earlier (see Primi et al., 2018). The negative correlation of *ACQ* with standardized achievement in language ($r = -.12$) was slightly stronger than its correlation with math ($r = -.08$), corroborating the idea that ACQ is associated with poor language skills.

When we consider IRT estimated scores from the balanced set of items, we also found a positive correlation between trait and achievement, but a negative correlation between acquiescence and achievement. It is interesting to note that only the PCM had validity coefficients that were of a similar magnitude as classical scores. On the one hand, this is not surprising as sum scores are a sufficient statistic for estimating

Table 1 Descriptive statistics and criterion validity of various scores based on classical, partial credit, graded response model and random intercept multidimensional IRT models

Variables	M	SD	Min	Max	Correlation	
					Lang.	Math
Balanced scale						
Classical scores						
Raw ave	0.57	3.55	1.13	5.00	0.22	0.18
ACQ cntr	0.57	0.55	−1.87	2.00	0.22	0.18
ACQ	0.35	2.95	1.00	5.00	−0.12	−0.08
Unidimensional IRT						
GRM[a]	0.96	0.00	−4.40	3.28	0.17	0.14
PCM[b]	0.54	0.00	−2.39	2.05	0.22	0.18
Random intercept MIRT						
GRM f1[c]	0.96	−0.01	−4.40	3.29	0.18	0.15
GRM f2	0.59	0.00	−3.58	3.67	−0.11	−0.07
PCM f1[4]	0.60	0.00	−2.67	2.28	0.21	0.17
PCM f2	0.30	0.00	−1.80	1.78	−0.10	- 0.07
Unbalanced scale						
Classical scores						
Raw ave	0.58	3.53	1.07	5.00	0.16	0.14
ACQ cntr	0.58	0.57	−2.00	2.11	0.20	0.16
Unidimensional IRT						
GRM[e]	0.98	0.00	−4.80	3.58	0.12	0.10
PCM[f]	0.61	0.01	−2.93	2.57	0.19	0.15
Random intercept MIRT						
GRM f1[g]	0.97	−0.01	−4.72	3.59	0.16	0.13
GRM f2	0.63	0.00	−3.99	4.08	−0.12	−0.08
PCM f1[h]	0.66	0.00	−3.09	2.71	0.20	0.16
PCM f2	0.35	0.00	−2.23	2.19	−0.10	−0.07

Note: *Raw ave* classical scores calculated via average item endorsing after reversing negatively phrased items; *ACQ cntr* classical scores controlled for acquiescence using the procedure by Soto et al. (2008); *ACQ* classical acquiescence index calculated via average endorsement of 15 antonym pairs of items before reversing negatively phrased items; *PCM* IRT estimated scores based on the partial credit model; *GRM* IRT estimated scores based on the graded response model; *GRM f1* and *GRM f2* trait and acquiescence scores estimated from the random intercept graded response model; *PCM f1* and *PCM f2* trait and acquiescence scores estimated from the random intercept partial credit model. Fit indices were: [a]*CFI* = .68, *RMSEA* = .08, *AIC* = 1,062,104, *BIC* = 1,063,008; [b]*CFI* = .73, *RMSEA* = .07, *AIC* = 1,043,494, *BIC* = 1,044,615; [c]*CFI* = .82, *RMSEA* = .06, *AIC* = 1,022,051, *BIC* = 1,023,179; [d]*CFI* = .77, *RMSEA* = .07, *AIC* = 1,042,179, *BIC* = 1,043,091; [e]*CFI* = .73, *RMSEA* = .08, *AIC* = 1,513,148, *BIC* = 1,514,829; [f]*CFI* = .78, *RMSEA* = .07, *AIC* = 1,554,601, *BIC* = 1,555,954; [g]*CFI* = .84, *RMSEA* = .06, *AIC* = 1,485,955, *BIC* = 1,487,643; [h]*CFI* = .82, *RMSEA* = .06, *AIC* = 1,5182,58, *BIC* = 1,519,618

the PCM (De Ayala, 2009), but surprising given that the GRM is often described as a more realistic model for data. Validity coefficients from the other three models (GRM and both random intercept models) were similar in magnitude, but were slightly lower than the validity of classical scores.

Although balanced scales have an equal number of items for each pole (15 items), we still found a difference in item discrimination under GRM across positively (1.27 on average) versus negatively worded (1.03 on average) items, which might yield an imbalance in the contribution of these items to EAP scores. Since the positive trait pole was favored, the correction process also becomes unbalanced, and is no longer similar to what occurred with classical scores. For instance, *Raw ave* correlated negatively with *ACQ* ($r = -.05$) in the balanced condition while *GRM* correlated positively with *ACQ* and *GRM f2* ($r = .09$ and .07 respectively) This indicates that the estimation of IRT scores may be slightly biased by ACQ even in balanced scales due to differences in the discrimination between positively- and negatively-keyed items. For instance, when discrimination is constrained equal across items (i.e., for the random intercept PCM), then the correlation between ACQ and the estimated trait becomes $r = -.04$.

For score estimates from the unbalanced item set, we found some noticeable differences. Since there were more positively worded items and they had higher discrimination (1.52 on average) than negatively worded items (0.79 on average) under GRM, this might lead to an even stronger positive association of ACQ with trait scores. In fact, classical acquiescence scores (*ACQ*) were positively correlated with *Raw ave*, $r = .18$, *PCM*, $r = .10$, and *GRM*, $r = .28$. By contrast, correlations between *ACQ* and self-management were lower when compared to EAP scores from the random intercept models, *GRM f1*, $r = .03$, and *PCM f1*, $r = .01$.

Score inflation due to ACQ tended to suppress the correlation between self-management and achievement. We see that the uncontrolled score *Raw ave* ($r = .16$ and .14 for language and math, respectively) had lower validity coefficients than *ACQ cntr* ($r = .20$ and .16). The random intercept GRM, *GRM f1*, had better coefficients ($r = .16$ and .13) than the *GRM* ($r = .12$ and .10). Rasch models tended to have better validities, as the unidimensional *PCM* ($r = .19$ and .15) and random intercept PCM, *PCM f1* ($r = .20$, .16), had the best validity coefficients of any IRT model. Overall, it is possible that this result indicates that random intercept models are producing scores that may be better controlling for ACQ.

Figures 1 and 2 show the effect of ACQ correction on scores. The upper part of Fig. 1 shows the relationship between *ACQ* (x-axis) and *Raw ave* (y-axis) for the balanced item set. When the ACQ index was near 3, scores had the full amplitude of variation from 1 to 5. As subjects tended to respond inconsistently, that is, tended to have $ACQ > 3$ or $ACQ < 3$, score variation was reduced. When agreeing was not completely consistent, scores were corrected towards the scale's center.

The lower part of Fig. 1 shows what happens in the unbalanced item set. We see the relationship between *ACQ* (x-axis) and *Raw ave* (y-axis) on the left, and between *ACQ* and *ACQ cntr* on the right (y-axis). In all graphs, we see a diamond shape characterizing the ACQ correction, with an important difference. Original scores (*Raw ave*) were correlated positively with *ACQ* ($r = .18$), but ACQ controlled

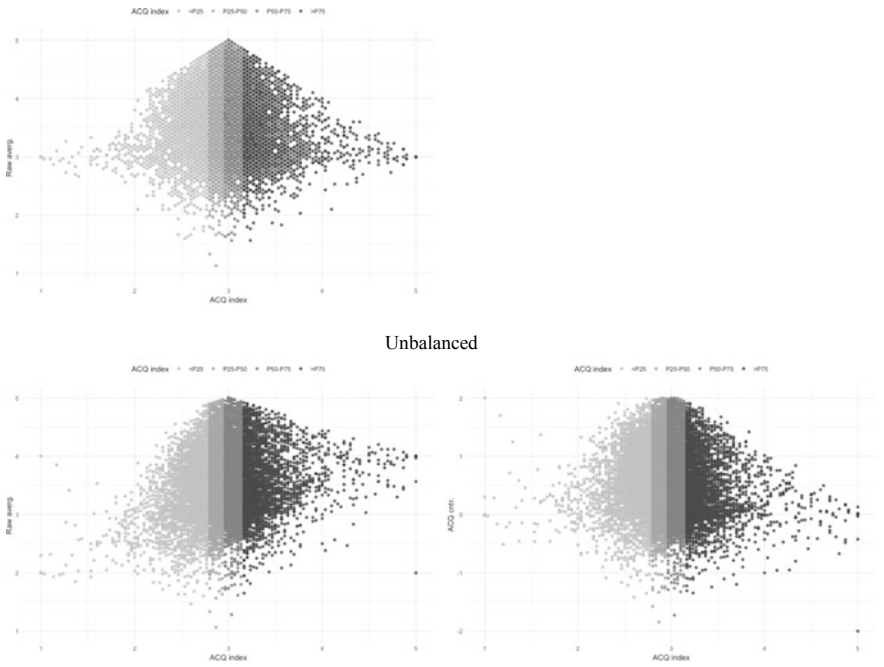

Fig. 1 Effects of acquiescence correction in classical scores of balanced scales (upper panel) versus unbalanced scales (lower panel). Scores on *y*-axis and ACQ indexes in *x*-axis

scores (*ACQ cntr*) were slightly negatively correlated with *ACQ* ($r = -.08$). Because the scale had more positively than negatively worded items, the correction process tended to produce the opposite effect, lowering high scores and increasing low scores if subjects exhibited ACQ or disacquiescence, respectively. This impacts validity coefficients because, in theory, ACQ is partialled out of *ACQ cntr*.

Figure 2 shows what happens with IRT scores, with *ACQ* always on the x-axis. The left columns show plots for the balanced item set, and the right columns for the unbalanced item set. On the y-axis, the upper panels represent *PCM*, the middle panels *GRM* and the lower panels *GRM f1*. We see patterns similar to what is shown in Fig. 1. In that scores may be corrected for ACQ. Nevertheless, we observe some variability among methods in the amount that scores are confounded with ACQ. There is no confounding for the *PCM* under the balanced item set ($r = -.04$), but some confounding under the unbalanced item set ($r = .10$). The *GRM* was slightly confounded in the balanced item set ($r = .09$), but much more confounded in the unbalanced item set ($r = .28$). Finally, random intercept models were less confounded. For example, for both balanced and unbalanced item sets, *GRM f1* correlated with *ACQ* near zero, $r = .03$.

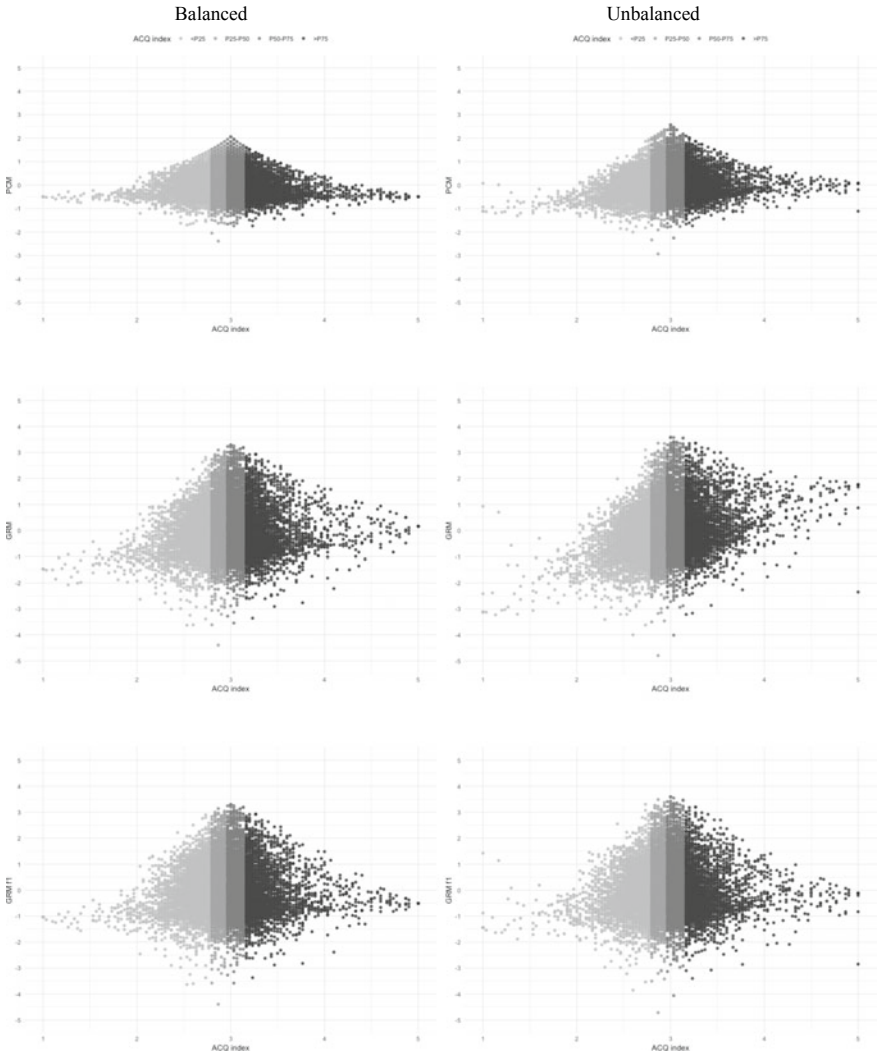

Fig. 2 Effects of acquiescence correction in IRT scores (partial credit—*PCM*, graded response—*GRM* and random intercept *GRM f1*) of balanced scales (left column) versus unbalanced scales (right column). Scores on *y*-axis and ACQ indexes in *x*-axis

4 Discussion

Acquiescence can negatively affect the criterion validity of self-report instruments. Balanced scales or acquiescence-controlled scores for unbalanced scales are ways to improve score validity (Mirowsky, & Ross, 1991; Primi et al., 2018; Soto & John, 2019). But less is known about ACQ corrections and the validity of IRT scores when scales are composed of both positively and negatively worded items. We tested five approaches spanning classical scoring, traditional IRT models and multidimensional IRT models based on the random intercept model (Billiet & McClendon, 2000; Cai, 2010; Maydeu-Olivares & Coffman, 2006; Maydeu-Olivares, & Steenkamp, 2018; Primi et al., 2018). The two modified versions of the random intercept models added an extra factor to explicitly model ACQ, and these were based on the GRM and PCM. These models produced ACQ-controlled IRT trait scores and also IRT ACQ index scores. The best of these models was the random intercept PCM.

We found that ignoring the possibility of ACQ is the worst-case scenario in terms of criterion validity. In balanced scales, the unidimensional PCM performed better than the GRM. With unbalanced scales, unidimensional GRM scores had the worst criterion validity. We suspect that either different item loadings for the GRM are picking up on some misspecification (lack of modeling ACQ) or that unique item content is important for criterion validity and is more equally considered under the PCM.

References

Allen, M. S., Walter, E. E., & McDermott, M. S. (2017). Personality and sedentary behavior: A systematic review and meta-analysis. *Health Psychology, 36*(3), 255–263. https://doi.org/10.1037/hea0000429.

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

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. Retrieved from http://www.jstatsoft.org/v48/i06/.

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

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

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482. https://doi.org/10.1086/504455.

Huang, I.-C., Lee, J. L., Ketheeswaran, P., Jones, C. M., Revicki, D. A., & Wu, A. W. (2017). Does personality affect health-related quality of life? A systematic review. *PLOS ONE, 12*(3), e0173806. https://doi.org/10.1371/journal.pone.0173806.

John, O., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York: Guilford Press.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality. *Organizational Research Methods, 18*(3), 512–541. https://doi.org/10.1177/1094428115571894.

Knowles, E. S., & Nathan, K. T. (1997). Acquiescent responding in self-reports: cognitive style or social concern? *Journal of Research in Personality, 31*(2), 293–301. https://doi.org/10.1006/jrpe.1997.2180.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: A prospectus for assessment and development. *ETS Research Report Series, 2014*(1), 1–48. https://doi.org/10.1002/ets2.12004.

Lipnevich, A. A., Preckel, F., & Roberts, R. D. (2016). *Psychosocial skills and school systems in the 21th century*. New York: Springer.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344.

Maydeu-Olivares, A., & Steenkamp, J. E. M. (2018). *An integrated procedure to control for common method variance in survey data using random intercept factor analysis models*. 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*(9), 1160–1173. https://doi.org/10.1037/pas0000566.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*(7), 1539–1550. https://doi.org/10.1016/j.paid.2008.01.010.

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*(2), 127–145. https://doi.org/10.2307/2786931.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*(1), 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrighsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996.

Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. *British Journal of Educational Psychology, 84*(2), 239–252. https://doi.org/10.1111/bjep.12019.

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S. & John, O. P. (2018). True or False? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. Submitted paper under review.

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology.*

Primi, R., Santos, D., John, O. P., & De Fruyt, F. D. (2016). Development of an inventory assessing social and emotional skills in Brazilian youth. *European Journal of Psychological Assessment, 32*(1), 5–16. https://doi.org/10.1027/1015-5759/a000343.

Savalei, V., & Falk, C. F. (2014a). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49*(5), 407–424. https://doi.org/10.1080/00273171.2014.931800.

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. *Clinical Psychology Review, 28*(8), 1326–1342. https://doi.org/10.1016/j.cpr.2008.07.002.

Savalei, V., & Falk, C. F. (2014b). 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., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*(4), 718–737. https://doi.org/10.1037/0022-3514.94.4.718.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*(2), 330–348. https://doi.org/10.1037/a0021717.

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.

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.

Valentini, F. (2017). Editorial: Influência e controle da aquiescência na análise fatorial [Editorial: Acquiescence and factor analysis]. *Avaliação Psicológica, 16,* 120–121. https://doi.org/10.15689/ap.2017.1602.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96–110. https://doi.org/10.1037/a0018721.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of extreme response style and acquiescence over 8 years. *Assessment*. https://doi.org/10.1177/1073191115583714.

Zhang, J., & Ziegler, M. (2018). Why do personality traits predict scholastic performance? A three-wave longitudinal study. *Journal of Research in Personality, 74,* 182–193. https://doi.org/10.1016/j.jrp.2018.04.006.

Ziegler, M. (2015). "F*** You, I Won't Do What You Told Me!"—Response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*(3), 153–158. https://doi.org/10.1027/1015-5759/a000292.