

True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement

Ricardo Primi, Filip De Fruyt, Daniel Santos, Stephen Antonoplis & Oliver P. John

To cite this article: Ricardo Primi, Filip De Fruyt, Daniel Santos, Stephen Antonoplis & Oliver P. John (2019): True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement, International Journal of Testing, DOI: [10.1080/15305058.2019.1673398](https://doi.org/10.1080/15305058.2019.1673398)

To link to this article: <https://doi.org/10.1080/15305058.2019.1673398>



Published online: 22 Oct 2019.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement

Ricardo Primi 

*Psychology, Universidade São Francisco, Brazil;
EduLab21, Ayrton Senna Institute, Brazil*

Filip De Fruyt 


*Department of Developmental, Personality and Social Psychology,
Ghent University, Belgium;
EduLab21, Ayrton Senna Institute, Brazil*

Daniel Santos 

*Faculty of Economics, Administration and Accounting of Ribeirão Preto,
Universidade de São Paulo, Brazil;
EduLab21, Ayrton Senna Institute, Brazil*

Stephen Antonoplis 

Department of Psychology, University of California, USA

Oliver P. John 

*Department of Psychology and Institute of Personality and Social Research,
University of California, USA;
EduLab21, Ayrton Senna Institute, Brazil*

What type of items, keyed positively or negatively, makes social-emotional skill or personality scales more valid? The present study examines the different

criterion validities of true- and false-keyed items, before and after correction for acquiescence. The sample included 12,987 children and adolescents from 425 schools of the State of São Paulo Brazil (ages 11–18 attending grades 6–12). They answered a computerized 162-item questionnaire measuring 18 facets grouped into five broad domains of social-emotional skills, i.e.: Open-mindedness (O), Conscientious Self-Management (C), Engaging with others (E), Amity (A), and Negative-Emotion Regulation (N). All facet scales were fully balanced (3 true-keyed and 3 false-keyed items per facet). Criterion validity coefficients of scales composed of only true-keyed items versus only false-keyed items were compared. The criterion measure was a standardized achievement test of language and math ability. We found that coefficients were almost as twice as big for false-keyed items' scales than for true-keyed items' scales. After correcting for acquiescence coefficients became more similar. Acquiescence suppresses the criterion validity of unbalanced scales composed of true-keyed items. We conclude that balanced scales with pairs of true and false keyed items make a better scale in terms of internal structural and predictive validity.

Keywords: 21st century skills measurement, acquiescence, criterion validity, big-five

INTRODUCTION

Most of the self-report measures of socio-emotional skills that we use in education rely on some form of Likert rating scales (Abrahams et al., 2019; John, Caspi, Robins, Moffitt, & Stouthamer-Loeber 1994; Primi, Santos, John, & De Fruyt, 2016; Santos & Primi 2014). When constructing measures, researchers need to decide whether to include items that measure not only the high pole of the construct but also the low pole, that is, items that are false-or reverse-keyed. For example, a scale designed to measure Conscientious Self-Management might include a true-keyed item like “Once I’ve started working on a task, nothing can distract me” but also a false-keyed item like “I get distracted very easily.”¹ Should researchers bother with such false-keyed items?

¹False-keyed items are commonly created in two ways: negation form or polar-opposite form of content reversal (Bentler, Jackson, & Messick, 1971). The first strategy uses negations, so that the true-keyed item *I fulfill tasks I commit to* could be negated to create a false-keyed version like *I do not fulfill tasks I commit to*. An alternative strategy uses antonyms, such as *I have difficulties in fulfilling what I promised* (Bertling & Alegre, 2018). The advantage of the antonym strategy is that it doesn’t require negations (which may be harder for respondents to understand), instead phrasing the item content in the form of an affirmative statement. When we advocate the use of false-keyed items here, we mean items created employing the antonym strategy without the use of negations.

Psychometricians seem to disagree about the value of false-keyed items. In internal analyses, such as alpha reliability or factor analyses, items that are false-keyed often seem to perform poorly, especially in younger samples (e.g. ages 8–14). In scales that include both true and false-keyed items, the false-keyed items tend to have lower item-total correlations. In factor analyses, false-keyed items sometimes define a separate factor, even in scales designed to be one-dimensional; for example, Marsh (1996) found that the false-keyed items on the Rosenberg Self-esteem scale formed an additional factor, separate from the factor defined by the true-keyed items. Worse, in multi-scale analyses, false-keyed items have all loaded together on a single factor even though they were designed to measure different constructs (e.g. Benson & Hocevar, 1985; Maydeu-Olivares & Coffman, 2006; Rammstedt, Goldberg, & Borg, 2010). Some authors argue that false-keyed items are linguistically more difficult to understand and may not measure the same construct as the true-keyed items (Suárez-Álvarez et al., 2018). Some guidelines (e.g. Gehlbach & Brinkworth, 2011; Gehlbach & Artino, 2017) even recommend avoiding negatively phrased items altogether. Van Sonderen, Sanderman, and Coyne (2013) concluded “An instrument with all items formulated in the same direction and referring to the intended concept (i.e. fatigue or fitness, depression or happiness) is to be preferred” (p. 6). Considering the apparent problems that false-keyed items cause with internal analyses, it would hardly seem worth the trouble to include false-keyed items in our measures.

However, there is an alternative perspective, first articulated by Cronbach (1946), that focuses on validity and the role of acquiescence:

[S]ince response tendencies affect an answer only when the student is to some degree uncertain about the content of the item, acquiescence tends to make false items more valid, and true items less valid. The poor student, guessing, tends to be right on the true item because of acquiescence, but tends to be wrong on the false item. False items alone are often as reliable and valid as the entire test of double the length (p. 480).

In this article, we tested Cronbach’s (1946) hypothesis that false-keyed items may in fact be more valid than true-keyed items, and that acquiescence plays a key role in this difference. Specifically, we systematically compared how well true and false-keyed items predict standardized scores on language and math achievement tests. That is, we focused on *criterion validity*, which is studied much less frequently in the socio-emotional literature than internal-structure aspects of validity. As Messick (1995) emphasized in his unified view of validity, both *score meaning* (how well intercorrelations among items are consistent to what is known about the construct domain) and *test use* (the utility of tests scores in the applied setting) are critical for the evaluation of a

measure. In addition, because we expected acquiescence to differentially affect responses to true and false-keyed items, we examined how acquiescence influenced (a) the internal consistency (alpha) of socio-emotional scales consisting solely of true-keyed or of false-keyed items and (b) their factor structure.

We were optimistic about the value of false-keyed items, for several reasons. First, including them in a scale guarantees that the researcher has a full range of construct representation at both low and high levels of the latent trait continuum, something that item-response theory emphasizes.

Second, and perhaps most important, if Cronbach is correct in suggesting that true-keyed items are more contaminated with acquiescence bias than false-keyed items, then correcting true-keyed items for acquiescence is required to realize their true validity. Acquiescence correction can be done only when false-keyed, antonymous items are included in a scale. Including false-keyed items is absolutely necessary if one wants to assess and control acquiescence bias.

Several studies of adults have shown that including false-keyed items and controlling (or modeling) acquiescence effects helps recover the theoretically expected factor structure – or the *true* factor structure when the study used simulation (e.g. Garrido, Golino, Nieto, Peña, & Molina, 2018; Lorenzo-Seva & Ferrando, 2009; Maydeu-Olivares & Coffman, 2006; Rammstedt et al., 2010; Soto & John, 2017; Savalei & Falk, 2014; Ten Berge, 1999). Although few studies have examined criterion validity, there is some recent evidence that fully balanced scales (i.e. that include equal numbers of true and false keyed items) tend to perform better than non-fully balanced scales, even in terms of external validity (Soto & John, 2019; Primi, Hauck-Filho, Valentini, Santos, & Falk, 2019a). In contrast, most studies reporting problems with false-keyed items examined internal structure *without* taking into account the effects of acquiescence. Taken together, these findings suggest that the apparent advantage of true-keyed items in analyses of internal structure may well be an artifact of acquiescence bias (Ferrando & Lorenzo-Seva, 2010).

ACQUIESCENCE BIAS

Rating scales are particularly susceptible to a response bias called acquiescence (e.g. Cronbach, 1946). Acquiescent responding is an individual's general tendency to consistently agree (yea-saying) or disagree (nay-saying) with questionnaire items, regardless of their content. Whereas the importance of acquiescence has long been recognized in adults (e.g. Jackson & Messick, 1958), Soto, John, Gosling, and Potter (2008) studied acquiescence effects in children. They expected more pronounced individual differences in acquiescence for

children compared to adults which, in turn, could be responsible for the less-clear factor structures often found in younger respondents. Indeed, individual-difference variance in acquiescence was much greater in children than in adults; it was highest at age 10 (the youngest age group studied) and then decreased substantially (by about half) to age 20. Even more important, acquiescence variance seriously distorted the factor structures of socio-emotional self-ratings in the youngest children. At age 10, the standard Big Five personality structure (John, Naumann, & Soto, 2008) could not be recovered in the raw data but did emerge in children's self-reports even at this young age when the substantial individual differences in acquiescence were controlled. These findings have led to a renewed interest in understanding the causes and consequences of acquiescence and in identifying ways to control for acquiescence effects (e.g. Rammstedt, Danner, & Bosnjak, 2017).

Jackson and Messick (1958) presented the now widely held view that respondents need to clearly understand the meaning of the items if we want to limit bias and error in their item responses. Consistent with this view, several recent studies of adults have shown that acquiescence is related (with r s of about $-.15$) to lower levels of cognitive functioning, such as IQ, verbal skill, and education (e.g. Rammstedt et al., 2017; Rammstedt et al., 2010). Respondents that find the items on a questionnaire difficult to understand, unclear, or confusing will rely more on their habitual response style than on the meaning of the items, thus increasing acquiescence-related error in the data and lowering reliability and validity. Much of the past research has focused on adults. The present research was designed to achieve two goals: extend this work to late childhood and adolescence and to examine differences in acquiescence effects between true- and false-keyed items.

CONTROLLING ACQUIESCENCE WITH BALANCED SCALES

Why include negatively phrased items? One good reason is that they are needed to make possible the measurement and control of acquiescence in the individual respondent. When left uncontrolled, individual differences in acquiescence tend to bias the correlations between items on the same scale: Acquiescence increases the (positive) correlations between two items keyed in the same direction but decreases the (negative) correlations between items keyed in the opposite direction (one true- and one false-keyed; see also Maydeu-Olivares & Steenkamp, 2018). Many psychometricians recommend controlling the effects of acquiescence by including equal numbers of true- and false-keyed items to create balanced scales (Jackson & Messick, 1958; Lorenzo-Seva & Ferrando, 2009; Primi, Santos, De Fruyt & John, 2019c; Savalei & Falk, 2014; Soto & John, 2017; Ten Berge, 1999).

Soto et al. (2008; see also Soto & John, 2017) developed a method to correct for acquiescence in personality ratings that is not confounded with substantive personality variance. Specifically, they proposed using pairs of items with opposite meanings to estimate an acquiescence score for each subject that indicates how much that subject's mean response is shifted up or down from the actual mid-point of the rating scale. This mean response can then be used to re-center each subject's item responses evenly on the rating scale. The logic of this process is as follows: If two items are antonyms (or opposites) of each other, then *agreement* with one item ought to be coupled with *disagreement* with the other item. For instance, when answering the Conscientious Self-Management items, individuals who respond "*completely like me*" (=5) to the true-keyed item "Once I start working on a task, nothing can distract me" should also respond "*not at all like me*" (=1) to the false-keyed item "I get distracted very easily." Similar response patterns should be seen for other pairs of opposite items, such as "I like artistic activities" (true-keyed) vs. "I find art boring and useless" (false-keyed) from the Open-mindedness domain.

On a scale ranging from 1 to 5, with a mid-point of 3, individuals who use the rating scale in an evenly balanced way ought to have an overall response mean of 3. If the overall response mean is *larger* than 3, it indicates acquiescence or "yea-saying", a tendency to favor the high end of the rating scale and *agree more* with items regardless of their content. If the overall response mean is *smaller* than 3, it indicates dissent or "nay-saying", a tendency to favor the low end of the rating scale and *disagree more* regardless of item content. The subject's overall response mean before reversing false-keyed items forms the acquiescence index. To correct for acquiescence in individual items, this acquiescence index is subtracted from each item score, yielding individually re-centered item scores that are no longer correlated with the individual's acquiescence bias. As a consequence, variance due to acquiescence is removed and its biasing effects are minimized (see Primi et al., 2019c, for a detailed discussion of the psychometric implications of this method of recoding item responses).

As this discussion shows, the measurement of acquiescence requires that tests and measures include not only true-keyed items but false-keyed items as well.

ACQUIESCENCE EFFECTS ON VALIDITY

Could acquiescence also bias the correlations of items and scales with external variables? For any two variables X and Y , we suggest, acquiescence may have either a suppressing or inflating effect on the correlation between X and Y (r_{XY}). Let's consider an example to understand how acquiescence may affect

criterion validity. Suppose we are studying the criterion validity of *Conscientious Self-management* (X) with cognitive achievement (Y). Suppose they have a true positive relationship, $r_{XY} > 0$. Imagine that we have a scale composed of only true-keyed items. Individuals with high acquiescence tend to agree indiscriminately; consequently, they will tend to have high scores on C. But these individuals will also tend to have lower scores on cognitive achievement (Rammstedt et al., 2010). They will have a high-C and low-achievement profile. That is, with respect to C, they will look like high-achievers, even though they are not. Their C scores no longer mean what they should, for they now partially reflect a construct irrelevant to C (i.e. acquiescence). Now imagine we compute the criterion validity from two samples A and B. Sample A contains individuals with evenly balanced responses, that is, neither high nor low acquiescence; and sample B contains a high proportion of individuals with high acquiescence. In this scenario the correlation between C and achievement computed in sample B will tend to be lower than the one computed in sample A regardless of what is the true value of this correlation. This will occur because in sample B contains a high proportion of high-C and low-achievement profiles that is incongruent with positive relationship between these variables.

Now imagine we have a scale composed of only false-keyed items. Individuals with high acquiescence will tend to agree. When computing scale scores, their item responses will be reversed, and their scores on C calculated by the sum of item responses will tend to be low. Now these individuals will have a Low-C and low achievement profile, therefore, a profile congruent a positive relationship between these variables. In this case the bias could increase the size of correlations between C and Achievement.

Psychometric and simulation studies provide support for these conceptual predictions. For instance, Ferrando, Lorenzo-Seva, and Chico (2003) showed that acquiescence can have a suppressor effect on the relationship between a content factor and an external variable. Namely if a content factor is contaminated by acquiescence and acquiescence tends to have zero correlations with external variables related to the content factor, the contaminated content factor contains more noise and would, therefore, tend to be less correlated with the external variable than if it was not contaminated. Mirowsky and Ross (1991) elaborated in some detail the statistical reasons why acquiescence could suppress correlations of true-keyed (i.e. unbalanced) scales with external criteria and lead to inflated reliability estimates.

In sum, acquiescence (*Acq*) can potentially (a) suppress the relationship between two variables *Y* and *X* when it has a positive association with one variable and a negative one with the other simultaneously, and (b) inflate the relationship when it has a positive relationship with both variables or a

negative relationship with both variables simultaneously. Incidentally, distorted factor structures and decreased reliabilities of scales that use false keyed items can be explained by these two effects occurring simultaneously.

For instance, *Acq* may inflate the correlation between two true-keyed items *X* and *Y* because it has a positive effect on both. *Acq* may inflate the correlation between two false-keyed items *X* and *Y* because it has a negative effect on both. *Acq* may suppress the correlation between a true-keyed item *Y* and a false-keyed item *X* because it has a positive effect on *Y* and a negative effect on *X*. Therefore, acquiescence may affect factor structures in such a way that true-keyed items will load on one factor and false-keyed items on another even though they measure the same dimension.

Note that this problem is not inherently caused by negatively phrased items, as the conventional view has been. When researchers factor analyze items without correcting for acquiescence, they are likely to encounter distorted structures and then decide to remove false-keyed items. Then, scales composed of only true-keyed scales may appear to have better internal structure and reliability. Based on this evidence they conclude that negatively phrased items have problems with structural validity. But note that acquiescence bias may still be there, alive and well, but disguised: Part of the positive covariance among true-keyed items is still due to acquiescence bias. But the problem now is that it is confounded with the substantive trait - construct-irrelevant variance is confounded with true trait variance. Therefore, increases in reliability are due in part to acquiescence variance and may not indicate true variance. Indeed, McCrae (2015) estimated that about 34% of systematic variance in self-reports is due to method variance (which includes acquiescence and other forms of response bias). Ferrando and Lorenzo-Seva (2010) explained analytically, with a simulation and with an empirical study, how this bias operates and why it remains unidentifiable in unbalanced scales.

THE PRESENT RESEARCH

To clarify the role of acquiescence in the external validity of item responses, we compared the criterion validities of fully unbalanced scales, that is, scales including either only true-keyed items or only false-keyed items. Content and scale length of these two kinds of scales was held constant: For example, we compared two Open-mindedness scales, one based on nine true-keyed items and the other based on nine false-keyed items. In other words, the scales were equivalent, except for the keying direction of the items being used. As external validity criteria, we used theoretically and practically important school outcomes, namely objective, standardized-test measures of learning progression in math and in language.

We tested three main hypotheses about the differential effects of acquiescence on responses to true-keyed and false-keyed items in terms of (a) reliability, (b) factor structure, and (c) criterion validity. Studies in the US and Europe have shown that acquiescence can substantially affect both internal consistency reliability and factor structure but these effects were based mostly on college students and adults. Here we sought to replicate these effects with elementary and secondary school students and extend them to the Brazilian context. In particular, because the raw, uncorrected response data include the biasing effects of acquiescence, we expected alpha reliability coefficients to be *higher* than in acquiescence-corrected data, which eliminate this shared bias variance (*Hypothesis 1*).

We further expected the acquiescence-corrected data to show a *clearer* factor structure, conforming more closely to the expected five-factor solution, than the raw, uncorrected responses (*Hypothesis 2a*). One of the reasons for the clarified factor structure is, we propose, the greater bipolarity of true and false keyed items when acquiescence is controlled; we therefore tested whether the correlations between the true and false keyed scales of the same construct increase when acquiescence is controlled (*Hypothesis 2b*).

Our third hypothesis extended this analysis of acquiescence effects to understanding how keying direction and acquiescence would jointly affect external criterion validity. We split this hypothesis into four possible patterns of results, each one representing different views in the literature.

Hypothesis 3a is the “null” hypothesis, stating that true and false-keyed items would show equal criterion validity and that acquiescence would not affect scores. In principle, both the true and false-keyed items had been developed to measure the same underlying construct; assuming successful scale development, true- and false-keyed items should show equivalent criterion validity, and construct-irrelevant variance should not be systematic.

Hypothesis 3b captures the “conventional wisdom” that false-keyed items are problematic, especially for younger respondents. If the conventional wisdom about false-keyed items is correct, false-keyed items should have lower criterion validity than do true-keyed items. According to this perspective, acquiescence does not play a role, so correcting for it should not change criterion validity.

In contrast, *Hypotheses 3c* and *3d* capture Cronbach’s (1946) proposal that acquiescence has *differential* effects on responses to true and false-keyed items and thus their criterion validity.

For true-keyed items (*Hypothesis 3c*), we expected that their criterion validity would be *suppressed* because acquiescence in self-reports is positively correlated with scales consisting of true-keyed items (i.e. agreeing to true-keyed items leads to higher scores). In contrast, there is no reason why acquiescence

bias in self-reports should be positively correlated with objectively scored learning achievement outcomes; in fact, as discussed above, acquiescence bias in self-reports tends to show a small negative correlation with measures of cognitive functioning (see Rammstedt et al., 2017). Hence, controlling for acquiescence should undo this suppression effect and thus *increase* the criterion validity of true-keyed items.

For false-keyed items (*Hypothesis 3d*), we expected acquiescence would serve to *inflate* (or enhance) their criterion validity. *Agreeing* with false-keyed items lowers one's scale score, so after reverse-scoring these items, acquiescence should be correlated *negatively* with scales consisting solely of false-keyed items. Combined with acquiescence's small negative correlation with learning achievement, these two negative correlations would induce a confounding effect, inflating the correlations between false-keyed items and learning achievement. Hence, controlling for acquiescence should slightly *decrease* the criterion validity of false-keyed items.

Putting Hypotheses 3c and 3d together, the “suppressor” and “inflation” hypotheses predict that false-keyed items should have higher criterion validity than true-keyed items *when using uncorrected responses*, and that the criterion validities should become more equal *when using acquiescence-corrected responses*.

METHOD

Participants

The sample included 12,987 adolescents (52.7% female) from 425 public schools located in 216 cities in the State of São Paulo in Brazil. They attended grades 7 ($N = 840$), 9 ($N = 6,474$) or 10 ($N = 5,673$) and ranged in age from 12 to 20 years ($M = 16$, $SD = 1.85$). All data were collected in the course of social-emotional skill assessments conducted by researchers from Edulab21 at the Ayrton Senna Institute in São Paulo, Brazil.

Criterion Variables: Academic Achievement Test Scores

We measured the students' academic achievement with a standardized test (SARESP) for language ($M = 257.3$, $SD = 49.9$) and for math ($M = 273.4$, $SD = 48.2$) in 2015. The scores for each student were provided by the Secretariat of Education of the State of São Paulo, which administers these measures of school performance as part of their regular assessment cycle. As expected, the language and math scores were positively correlated ($r = .67$). Details can be found here: http://file.fde.sp.gov.br/saresp/saresp2015/Arquivos/SE_2015_online.pdf.

Self-Reports of Socio-Emotional Skills

Students completed SENNA v2.0 (Primi, Santos, De Fruyt & John, 2019b; Primi et al., 2016), which is a computerized 162-item questionnaire developed specifically in Brazil to measure five broad social-emotional skill domains in children and adolescents: Open-mindedness (O), Conscientious Self-Management (C), Engaging with others (E), Amity (A), and Negative-Emotion Regulation (N). These five Brazilian dimensions are conceptually akin to those of the Big Five model of personality (John et al., 2008). Each of these five domains is carefully defined in terms of several more specific facets; for example, the E dimension (Engaging with others) is measured in terms of the three facets of Social Initiative, Assertiveness, and Enthusiasm. Each domain (and its facets) is measured by an even amount of true-keyed and false-keyed items. Therefore, the scales are completely balanced with respect to true- and false-keyed items. Students responded using a 5-point 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

The fully balanced structure of the SENNA v2.0 questionnaire allowed us to compute two scale scores for each of the 5 domains: one using only the true-keyed items and the other using only the false-keyed items. In total, we had ten scale scores, five domains each split into two keying directions. We reversed the false-keyed item scores before calculating domain scores so that high scores always meant high socio-emotional skills, even for the scales based on false-keyed items. For the N domain (Negative-emotion regulation), both scales were keyed in the desirable, emotionally stable direction.

In addition, to study the effects of acquiescence, we generated two sets of scores for each of these ten scales. The first set simply used the raw response scores, whereas the second corrected the scores for acquiescence as described by Soto et al. (2008).

Computing the Acquiescence Index and Correcting the Raw Data in a Content-Balanced Way

The item pairs for the SENNA v.2 questionnaire were selected to be both conceptually and empirically opposite, drawing from a large item pool of more than 500 items in three successive empirical studies. These procedures ensured an equal number of true- and false-keyed items in each domain scale (Primi et al., 2019a, 2019b). To compute the acquiescence index, we used 54 pairs of

²The SENNA questionnaire also includes self-efficacy questions for each of the 5 domains and its facets; however, these items are all true-keyed. Thus, they are not relevant to the present research comparing true and false keyed items and will not be considered here.

opposite items that represented all five broad domains measured in the SENNA v.2 questionnaire, ensuring the index was content-balanced. We calculated the acquiescence index (*mean endorsement*) by averaging an individual's self-ratings on the 108 individual items included in the 54 opposite pairs *before any reverse-scoring of the false-keyed items*. Thus, because 54 items in this index were positively keyed and the other 54 matching items were negatively keyed, the resulting mean score does not reflect any construct-related variance but only individual differences in the use of the rating scale (high vs. low acquiescence) plus random measurement error.

In our sample of public school students, the mean of this acquiescence index was 2.92, quite close to the expected normative value of 3 (i.e. the mid-point of our 1-5 rating scale). However, the standard deviation of 0.35 indicates important individual differences in scale usage: The average adolescent in this study did not use the rating scale in the normative way centered around 3; instead, their mean response was shifted .35 scale points, either down from the observed mean to 2.57 or shifted up to 3.25.

To correct for acquiescence, we subtracted each individual's raw item responses from the individual's acquiescence score: $Y_{cij} = Y_{oij} - Mean_j$ where Y_{cij} is the corrected item score of individual j on item i , Y_{oij} is the observed (raw) score of individual j on item i , and $Mean_j$ is the index of acquiescence of subject j . This formula transforms the original metric 1 to 5 to -2 to $+2$. After acquiescence correction, false-keyed items were reverse-scored by multiplying by -1 before summing. In addition to the 10 raw-score scales, we computed another set of 10 scores based on these corrected scores. We call them the acquiescence-corrected scores, five true-keyed and five false-keyed.

Summary of Design

In total, this yielded 20 scores defined by a fully crossed design ($5 \times 2 \times 2$): five socio-emotional domains (O, C, E, A and N) times 2 keying directions (true-keyed scale, false-keyed scale) times 2 methods of scoring (raw score, acquiescence-controlled). False-keyed items were always reverse-scored prior to computing scales scores. Thus, *high* scores on all 20 scales always represent the high end of the dimensions, including for N which was always keyed in the emotionally stable direction.

RESULTS

Acquiescence and Internal Consistency

Acquiescence inflates internal consistency by adding common variance to all items that are keyed in the same direction. Thus, correcting for acquiescence

TABLE 1
Reliability and Descriptive Statistics of 10 Indexes Formed by Combining Five
Socio-Emotional Constructs with Two Keying Directions

	Alpha		Mean		SD	
	Raw	Acqu-Crct	Raw	Acqu-Crct	Raw	Acqu-Crct
O: Open-Mindedness						
True-Keyed	.79	.72	3.33	0.41	0.72	0.62
False-Keyed	.75	.64	3.76	0.68	0.69	0.60
C: Conscientious Self-Management						
True-Keyed	.87	.84	3.50	0.59	0.65	0.60
False-Keyed	.86	.79	3.60	0.51	0.68	0.58
E: Engaging with Others						
True-Keyed	.72	.60	3.47	0.55	0.64	0.55
False-Keyed	.68	.56	3.32	0.24	0.68	0.59
A: Amity						
True-Keyed	.69	.58	3.37	0.45	0.57	0.50
False-Keyed	.78	.65	3.68	0.60	0.64	0.52
N: Negative-Emotion Regulation						
True-Keyed	.70	.56	3.09	0.18	0.65	0.56
False-Keyed	.78	.68	3.28	0.19	0.78	0.66
Mean						
True-Keyed	.75	.66	3.35	0.44	0.65	0.57
False-Keyed	.77	.66	3.53	0.44	0.69	0.59

Note. "Acqu-Crct" refers to acquiescence-corrected scores. False-keyed items were reverse-scored prior to analyses.

should remove this common variance from all items, deflating internal consistency of item composites but better reflecting the true consistency of the item set.

Consistent with this prediction, we found that the internal consistencies of all true-keyed and all false-keyed scales decreased after correcting for acquiescence. Table 1 displays this effect for the composites of true- and false-keyed items for each of the Big Five domains (in the columns under "Alpha"). Averaging across all Big Five domains, true-keyed item composites had a Cronbach's *alpha* of .75; false-keyed item composites, .77. After correcting for acquiescence, these alpha estimates both decreased to .66. This pattern held across all true- and false-keyed scales for all five content domains.

Correcting for Acquiescence and Improving Factor Structure and Bipolarity

A common argument against the inclusion of false-keyed items in scales has been that they load poorly on expected common factors but this problem may

TABLE 2
Congruence Coefficients of Five Socio-Emotional Skills Using Raw and Acquiescence-Corrected Responses

Empirical Factors	Theoretical Factors				
	O	C	E	A	N
<i>Raw Responses</i>					
(O)	.21	.13	.05	.54	-.14
C	.30	.85	-.05	.25	.06
E	.36	-.05	.79	.04	.11
(A)	.09	-.06	-.10	.64	.13
N	.03	.10	.10	.20	.83
<i>Acquiescence-Corrected Responses</i>					
O	.84	.09	.10	.36	.02
C	.06	.90	.00	.22	.05
E	.07	-.02	.84	.05	.10
A	.01	.00	-.08	.81	.00
N	.05	.02	.05	.20	.88

Note. Fit indexes for raw responses was $\chi^2 = 18,019.87$, $df = 460$, $RMSEA = .05$ and $SRMR = .03$; and acquiescence corrected responses $\chi^2 = 43,192.1$, $df = 460$, $RMSEA = .08$ and $SRMR = .03$.

Bold values are $> .70$. Values $> .84$ indicate fair similarity (Lorenzo-Seva & ten Berge, 2006).

be due to the biasing effects of acquiescence that reduce the correlations of false-keyed items with true-keyed items. Table 2 replicates recent work in the U.S. and Europe in our Brazilian sample, using factor congruence coefficients to index the clarity of the observed factor structure. We fitted a five-factor model with target rotation using Exploratory Structural Equation Modeling (ESEM; Muthén & Muthén, 1998–2017). Indicators for the five factors were 36 facet scores calculated from 3 items each of: 18 facet scores \times 2 keying directions combination. Specifically, loadings from factor analysis were correlated with a matrix with theoretically perfect loadings (i.e. each facet loads entirely and exclusively on its expected factor). Each cell in Table 2 represents the correlation of the observed facet loadings with the theoretically perfect vector column. The upper half shows the congruence coefficients when using the uncorrected responses. Only three of the five Big Five traits (C, E, and N) emerged in this analysis and show reasonable levels of congruence with the ideal loadings. The theoretically defined dimension of Openness did not show a clear congruence with any of the observed factors. Amity was poorly defined. The lower half of Table 2 shows the same analysis after correcting for acquiescence. Each empirically observed factor now shows substantial and clear congruence with its expected theoretical factor, and the congruence coefficients on the diagonal are all higher than for the raw, uncorrected scores in the upper half of Table 2.

TABLE 3
Correlations of True-Keyed and False-Keyed Scales with Acquiescence and Between True- and False-Keyed Scales With and Without Acquiescence Correction

	Correlation with Acquiescence		Correlation between True and False		
	True	False	Acquiescence		
			Raw	Corrected	Partialled
O	.50	-.50	.23	.65	.65
C	.41	-.53	.43	.84	.85
E	.51	-.48	.24	.64	.64
A	.50	-.58	.19	.69	.68
N	.51	-.53	.23	.69	.69
<i>Mean</i>	.49	-.52	.26	.70	.70

Note. “Corrected” correlations are correlations between acquiescence-corrected true- and false-keyed items. “Partial” correlations are correlations between true- and false-keyed items with acquiescence partialled out.

Table 3 shows why this is the case. Before correcting for acquiescence, the true-keyed and false-keyed scales for the same Big Five domain were not highly correlated, seemingly questioning the bipolarity of the underlying constructs. In fact, averaging across the five domains, the high pole (true-keyed items) and low pole (false-keyed items) correlated only .26. After correcting for acquiescence, the average correlation between true- and false-keyed scales increased to .70, a substantial jump. Table 3 also shows that this correlation obtained via acquiescence correction was identical to that obtained when partialing out acquiescence from true- and false-keyed item composites. Thus, the correction technique used here yielded the same result as the traditional partialing technique for removing acquiescence bias. The advantage of the present technique over the traditional technique of removing common variance from acquiescence is seen in the following section on criterion validity. Namely, the traditional technique would require that researchers use a multiple regression framework to correct for acquiescence, even though this can introduce complex interactions and may remove meaningful variance from the outcome variable. Our technique avoids these problems by directly correcting (re-centering) the scores of only the targeted predictor variables.

Effects on Criterion Validity

We considered four hypotheses about the ways that acquiescence could affect the criterion validity of our true and false-keyed scales. Given that true- and

false-keyed items are designed to assess the same construct, they ought not correlate with other constructs differently. In an ideal world, measures of the same thing do not function differently. Thus, “improved criterion validity” should mean “more equal criterion validity”. This conclusion entails that the criterion validities for true- and false-keyed items using raw scores be more different from each other than the criterion validities using acquiescence-corrected scores.

In the case of our criteria (language and math scores), the results presented so far suggest that acquiescence correction should raise external validity for true-keyed items and lower validity for false-keyed items. In raw form, true-keyed items should correlate with the criteria in the opposite direction of acquiescence, creating a “suppressor” situation (*Hyp 3c*) that depresses the observed correlation relative to the true correlation. False-keyed items should show the opposite pattern: when the construct is measured with reverse-scored false-keyed items, acquiescence correlates negatively with these scale scores, as shown in [Table 3](#). Combined with the small negative correlations of acquiescence with the validity criteria, this negative correlation should create a confounding effect (*Hyp 3d*) that inflates the correlation of the false-keyed scales with the criteria in the raw, uncorrected data.

[Table 4](#) shows these effects. Prior to acquiescence-correction, false-keyed items predicted the criteria much better than true-keyed items across the Big Five (*language*: $r_{\text{avg}} = .21$ vs. $r_{\text{avg}} = .08$; *math*: $r_{\text{avg}} = .17$ vs. $r_{\text{avg}} = .08$). After acquiescence-correction, the average criterion validity for false- and true-keyed scales became equal (*language*: $r_{\text{avg}} = .17$ for both; *math*: $r_{\text{avg}} = .15$ for both). Hence, correcting for acquiescence improved the criterion validity of true- and false-keyed items by moving them closer to equal. This equalizing means that for the true-keyed scales criterion validity increased and for the false-keyed scales it decreased, changes consistent with the suppressor hypothesis (*Hyp 3c*) and the confounding hypothesis (*Hyp 3d*), respectively.

Another way to examine these effects is via multiple regression. We ran two regressions predicting language and math. In each regression, we first included only true-keyed items, and then added false-keyed items. For language with only true-keyed items $R^2 = .046$ (95% CI [.04,.05]), but adding false-keyed items increased $R^2 = .129$ (95% CI [.12,.14]). Math scores followed a similar pattern: $R^2 = .024$ (95% CI [.02,.03]) and $R^2 = .074$ (95% CI [.07,.08]), respectively. All increments were significant. [Table 5](#) shows the coefficients for the final model. These results reinforce the idea that false-keyed items had incremental value for predicting the criterion. The predicted cognitive achievement variance more than doubled when false-keyed items were included in the model.

TABLE 4
 Criterion Validity of the 5 True-Keyed and the 5 False-Keyed Scales, Before and After
 Correction for Acquiescence

	Language		Math		Mean Lang and Math	
	Raw	Acqu-Crct	Raw	Acqu-Crct	Raw	Acqu-Crct
O: Open-Mindedness						
True-Keyed	.12	.21	.10	.17	.11	.19
False-Keyed	.30	.27	.21	.20	.26	.24
C: Conscientious Self-Management						
True-Keyed	.13	.21	.11	.17	.12	.19
False-Keyed	.24	.21	.19	.18	.22	.20
E: Engaging with Others						
True-Keyed	.04	.12	.04	.10	.04	.11
False-Keyed	.19	.14	.15	.12	.17	.13
A: Amity						
True-Keyed	.15	.26	.13	.21	.14	.24
False-Keyed	.23	.20	.17	.15	.20	.18
N: Negative-Emotion Regulation						
True-Keyed	-.02	.05	.04	.10	.01	.08
False-Keyed	.09	.04	.12	.09	.11	.07
<i>Mean</i>						
True-Keyed	.08	.17	.08	.15	.08	.16
False-Keyed	.21	.17	.17	.15	.19	.16

Note. “Acqu-Crct” refers to acquiescence-corrected scores. False-keyed items were reverse-scored prior to analyses.

Figure 1 demonstrates these suppressor and confounding effects. It shows scatterplots for the correlations between true- and false-keyed Conscientiousness scores with Math scores. Ellipses and dashed lines in each graph show the general relationship between Conscientiousness scores with Math scores without conditioning by acquiescence. Superimposed to these “main effects” are the scatterplots “pulled apart” across the 4 quartiles of acquiescence, from very low acquiescence on the left to very high on the right.

For the *true-keyed* Conscientiousness scale (in the upper half of the figure), the overall scatterplot is pulled rightward and downward, fully suppressing the correlation between Conscientiousness and Math scores. This movement (right and down) is gradually reduced moving leftward through relatively lower acquiescence scores, and the correlation between Conscientiousness and Math scores reemerges with this countermovement.

For the *false-keyed* Conscientiousness scale (in the lower half of the figure), this pattern did not occur. Instead, the correlation between Conscientiousness

TABLE 5
Regression Results Using Language and Math as the Criterion and True, False-Keyed
Items as Predictors

Predictor	<i>b</i>	<i>b</i> 95% CI [LL, UL]	β	<i>r</i>
Criterion Variable: Language				
(Intercept)	130.97**	[123.31, 138.62]		
True-Keyed				
O: Open-Mindedness	4.20**	[2.62, 5.78]	0.06	.12**
C: Conscientious Self-Management	-0.06	[-2.11, 1.99]	-0.00	.13**
E: Engaging with Others	-7.14**	[-8.99, -5.30]	-0.09	.04**
A: Amity	14.56**	[12.61, 16.51]	0.17	.15**
N: Negative-Emotion Regulation	-2.87**	[-4.66, -1.08]	-0.04	-.02*
False-Keyed				
O: Open-Mindedness	12.45**	[10.79, 14.11]	0.17	.30**
C: Conscientious Self-Management	7.09**	[5.00, 9.18]	0.10	.24**
E: Engaging with Others	10.86**	[9.26, 12.45]	0.15	.19**
A: Amity	0.89	[-1.07, 2.85]	0.01	.23**
N: Negative-Emotion Regulation	-4.43**	[-5.89, -2.98]	-0.07	.09**
Criterion Variable: Math				
(Intercept)	169.36**	[161.73, 176.99]		
True-Keyed				
O: Open-Mindedness	3.17**	[1.60, 4.74]	0.05	.10**
C: Conscientious Self-Management	-0.31	[-2.36, 1.74]	-0.00	.11**
E: Engaging with Others	-6.18**	[-8.02, -4.34]	-0.08	.04**
A: Amity	12.02**	[10.08, 13.97]	0.14	.13**
N: Negative-Emotion Regulation	0.71	[-1.07, 2.50]	0.01	.04**
False-Keyed				
O: Open-Mindedness	7.43**	[5.77, 9.09]	0.11	.21**
C: Conscientious Self-Management	5.89**	[3.80, 7.97]	0.08	.19**
E: Engaging with Others	8.42**	[6.83, 10.01]	0.12	.15**
A: Amity	-1.18	[-3.13, 0.78]	-0.02	.17**
N: Negative-Emotion Regulation	0.18	[-1.28, 1.63]	0.00	.12**

Note. *b* represents unstandardized regression weights. β indicates the standardized regression weights. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

*Indicates $p < .05$.

**Indicates $p < .01$.

and Math scores remained stable across the 4 levels of acquiescence. If anything, the correlation strengthened at higher levels of acquiescence, as people lower on Conscientiousness and Math are pushed farther down and leftward. This movement is consistent with our confounding hypothesis that sharpens the criterion validity of false-keyed items.

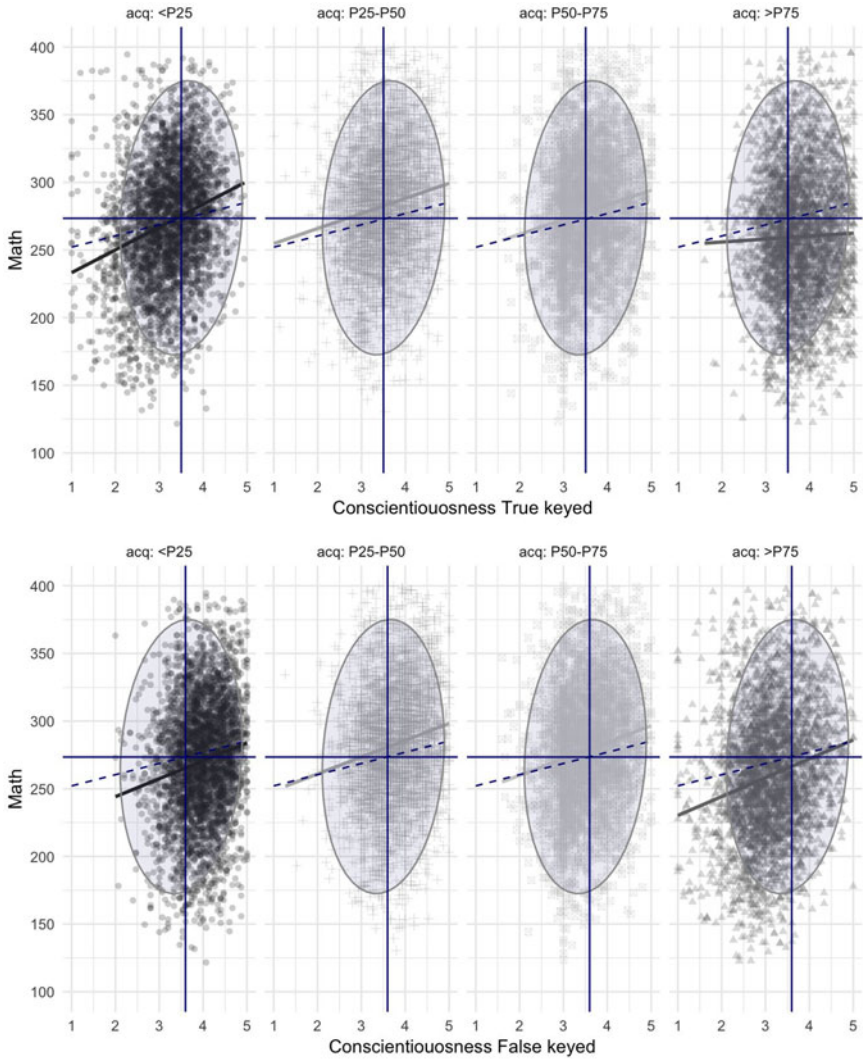


FIGURE 1

Scatterplots of conscientiousness scales (x-axis) versus math achievement (y-axis). Upper half shows true keyed scales and lower half false keyed scales. Figures are split horizontally as a function of the quartiles of acquiescence index from low (left) to high (right).

Horizontal and vertical lines are placed at the means of corresponding y-axis and x-axis variables. Ellipses with dashed lines shows general trend of the relationship without conditioning on acquiescence.

DISCUSSION

What Have We Learned About True-Keyed Items, False-Keyed Items, and Acquiescence?

How should we measure a construct in a self-report questionnaire? Conventional wisdom suggests that only true-keyed items—those that describe the high end of a construct—are necessary. False-keyed items have been found to load poorly on theorized unidimensional factors, and they seem more difficult for children and adolescents to answer. Our results suggest this wisdom is shortsighted. The apparent factor structure problems seem to result largely from acquiescence bias—agreeing or disagreeing with items irrespective of content—present in both *true*- and *false*-keyed items, a result demonstrated in the present research and in past and recent work (e.g. Benson & Hocevar, 1985; Ferrando & Lorenzo-Seva, 2010; Garrido et al., 2018; Lorenzo-Seva & Ferrando, 2009; Soto et al., 2008; Rammstedt et al., 2010; Ten Berge, 1999). Critically, *this bias is detectable only with the inclusion of content-balanced false-keyed items*, for their presence permits constructing a general acquiescence index or factor that captures bias variance that can then be removed from each individual's item responses. Doing so in our sample of Brazilian public school students improved unidimensionality dramatically—increasing correlations between the true- and false-keyed scales (representing the high and low pole of the construct, respectively) from below .30 all the way to .70.

Moreover, correcting for acquiescence provided a more accurate representation of scale internal consistency by removing a common source of construct-irrelevant variance from all items, a result found both in the present and in past research (e.g. Soto et al., 2008).

Improving factor structure, however, merely shows that false-keyed items do indeed measure their intended construct. This is hardly sufficient evidence to warrant their inclusion in scales. We would also want to know that false-keyed items tell us something additional about people, via the prediction of valued outcomes, relative to true-keyed items. The main contribution of the present research was demonstrating that, prior to correcting for acquiescence, false-keyed scales possessed *greater* criterion validity than true-keyed scales. Correcting for acquiescence—an action possible only when false-keyed items were included—nearly doubled the criterion validity of true-keyed items (see explanation of this suppression in Ferrando et al., 2003; Mirowsky & Ross, 1991). Together, these results suggest that false-keyed items may be generally better at predicting outcomes than true-keyed items and that false-keyed items are necessary to achieve the expected criterion validity of true-keyed items. Ironically, then, if researchers wanted to quickly measure constructs that validly predict valued outcomes, they may be better off to stay away from the

conventional true-keyed items entirely and use only false-keyed items. While unexpected, these results were anticipated as early as 1946 by Lee Cronbach.

Assuming researchers were to include as many true- as false-keyed items in a scale, our research suggests that the false-keyed items would, indeed, be nearly as reliable and valid as all the true- and false-keyed items combined. These results suggest that a reevaluation of the utility of false-keyed items is greatly needed in psychometrics. They appear to be a boon, rather than a bane.

Limitations and Future Directions

One important point for future research is the development of construct validation studies of the acquiescence index itself. Prior studies suggest it may have multiple causes: low language skills characteristic of younger and less educated populations and potentially careless responding by inattentive or unmotivated participants, to name just two (Garrido et al., 2018; Huang, Liu, & Bowling, 2015; Meijer, Egberink, Emons, & Sijtsma, 2008; Niessen, Meijer & Tendeiro, 2016). Studies of response processes can help us understand why some people provide inconsistent responses when asked about the same content using true- and false-keyed items. Such studies may expand our understanding of what the acquiescence index measures.

Additional studies should check that the patterns observed here replicate using other measures. Though the SENNA v2.0 questionnaire can be thought of as a Big Five measure in Portuguese and tailored specifically to children and adolescents, testing whether inventories intended for older or different populations, and even inventories for different constructs (e.g. intergroup tolerance), are needed to test the scope of the effects of acquiescence on criterion validity.

In addition, progress in learning math and language are undoubtedly important targets of education, but other skills and outcomes are also valued and should be included in future research, such as happiness and satisfaction with life, belonging and close relationships, income, or even successful emotion regulation. Correcting for acquiescence may help researchers in these fields of investigation better estimate how various constructs impact or predict these valued outcomes. Obtaining more accurate estimates is critical for program and personnel evaluation, for avoiding both Type I and Type II errors. Researchers are unlikely to want to dismiss useful programs because of underestimates of effect sizes and ought not want to oversell, or devote excessive time to, overestimated programs. Correcting for acquiescence may help researchers better decide where to aim.






Conclusion

To answer our initial question of whether we should use true or false items: We conclude that a better scale, at least in terms of criterion validity, will require both true- and false-keyed items. A careful selection of opposite pairs of items to compose a balanced scale will make an instrument more valid by permitting assessment and removal of acquiescence bias. Although false-keyed items are often perceived as more difficult to answer, well-constructed false-keyed items seem to have better criterion validity than true-keyed items, and their inclusion permits correcting for acquiescence and improving the criterion validity of true-keyed item. Hence, we recommend using *both* true- and false-keyed items when measuring social-emotional skill and personality constructs.

FUNDING

This article is part of a research program on socio-emotional skills sponsored by the Ayrton Senna Foundation. The first author also receives a scholarship from the National Council on Scientific and Technological Development (CNPq) and São Paulo Research Foundation (FAPESP).

ORCID

Ricardo Primi  <http://orcid.org/0000-0003-4227-6745>
 Filip De Fruyt  <http://orcid.org/0000-0002-5552-0754>
 Daniel Santos  <http://orcid.org/0000-0002-2605-2736>
 Stephen Antonoplis  <http://orcid.org/0000-0002-2789-5516>
 Oliver P. John  <http://orcid.org/0000-0003-0171-0971>

REFERENCES

- Abrahams, L., Pancorbo, 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*(4), 460–599. doi:10.1037/pas0000591
- Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231–240. Retrieved from <http://www.jstor.org/stable/1435036>. doi:10.1111/j.1745-3984.1985.tb01061.x
- Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186–204. doi:10.1037/h0031474
- Bertling, J., & Alegre, J. (2018). *PISA 2021 context questionnaire framework* (p. 80). Princeton: Educational Testing Service (ETS).

- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6(4), 475–494. doi:10.1177/001316444600600405
- Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: a theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, 63(2), 427–448. doi: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(3), 353–374. doi: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, NY.
- Gehlbach, H., & Artino, A. R. (2017). The survey checklist (Manifesto). *Academic Medicine*, 93(3), 1. doi:10.1097/ACM.0000000000002083
- Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, 15(4), 380–387. doi:10.1037/a0025704
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845. doi:10.1037/a0038510
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55(4), 243–252. doi:10.1037/h0045996
- John, O. P., Caspi, A., Robins, R. W., Moffitt, T. E., & Stouthamer-Loeber, M. (1994). The “little five”: exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development*, 65(1), 160–178. doi:10.1111/j.1467-8624.1994.tb00742.x
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of personality: Theory and research* (3rd ed. pp. 114–158). New York, NY, US: Guilford Press.
- Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology*, 62(2), 319–326. doi:10.1348/000711007X265164
- Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64. doi:10.1027/1614-2241.2.2.57
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, 70(4), 810–819. doi:10.1037/0022-3514.70.4.810
- Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11(4), 344–362. doi: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. 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. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, 19(2), 97–112. doi:10.1177/1088868314541857
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, 90(3), 227–238. doi:10.1080/00223890701884921

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741
- 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. doi:10.2307/2786931
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. doi:10.1016/j.jrp.2016.04.010
- Primi, R., Hauck-Filho, N., Valentini, F., Santos, D., & Falk, C. F. (2019a). Controlling acquiescence bias with multidimensional IRT modeling. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology. IMPS 2017. Springer proceedings in mathematics & statistics* (vol. 265). Cham: Springer. doi:10.1007/978-3-030-01310-3_4.
- Primi, R., Santos, D., John, O. P., & De Fruyt, F. (2016). Development of an inventory assessing social and emotional skills in Brazilian youth. *European Journal of Psychological Assessment*, *32*(1), 5–16. doi:10.1027/1015-5759/a000343
- Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019b). *SENNA V2.0 technical manual*. São Paulo: Instituto Ayrton Senna (IAS).
- Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019c). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*. Advance Online publication. doi:10.1111/bmsp.12168
- Rammstedt, B., Danner, D., & Bosnjak, M. (2017). Acquiescence response styles: A multilevel model explaining individual-level and country-level differences. *Personality and Individual Differences*, *107*, 190–194. doi:10.1016/j.paid.2016.11.038
- 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. doi:10.1016/j.jrp.2009.10.005
- Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*(5), 407–424. doi:10.1080/00273171.2014.931800
- Santos, D., & Primi, R. (2014). *Social and emotional development and school learning: A measurement proposal in support of public policy*. São Paulo: Ayrton Senna Institute. Retrieved from <http://educacaosec21.org.br/wp-content/uploads/2013/07/Social-and-emotional-developmente-and-school-learning1.pdf>
- 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. doi: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*(4), 444–590. doi:10.1037/pas0000586
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*(4), 718–737. doi:10.1037/0022-3514.94.4.718
- Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M. B., García-Cueto, E., Cuesta, M., & Muñoz, J. G. F. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. <https://doi.org/10.7334/psicothema2018.33>.

- 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. doi:[10.1207/s15327906mbr3401_4](https://doi.org/10.1207/s15327906mbr3401_4)
- Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *Plos One*, *8*(7), e68967. doi:[10.1371/journal.pone.0068967](https://doi.org/10.1371/journal.pone.0068967)