Original Article

Anchoring Vignettes

Can They Make Adolescent Self-Reports of Social-Emotional Skills More Reliable, Discriminant, and Criterion-Valid?

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

[1]Graduate Program in Psychological Assessment, Universidade São Francisco, Itatiba, Brazil, [2]Economics Department, Universidade de São Paulo, Ribeirão Preto, Brazil, [3]Department of Developmental, Personality, and Social Psychology, Ghent University, Belgium, [4]Department of Psychology, University of California, Berkeley, CA, USA, [5]EduLab21, Ayrton Senna Institute, São Paulo, Brazil

Abstract. Individuals differ in the way they use rating scales to describe themselves, and these differences are particularly pronounced in children and early adolescents. One promising remedy is to correct (or "anchor") an individual's responses according to the way they use the scale when they rate an anchoring vignette (a set of hypothetical targets differing on the attribute of interest). Studying adolescents' self-reports of their socio-emotional attributes, we compared traditional self-report scores with vignette-corrected scores in terms of reliability (internal consistency), discriminant validity (scale intercorrelations), and criterion validity (predicting achievement test scores in language and math). A large and representative sample of 12th grade Brazilian students (N = 8,582, 62% female, mean age 18.2) were administered a Portuguese-language self-report inventory assessing social-emotional skills related to the Big Five personality dimensions. Correcting scores according to vignette ratings led to increases in the reliability of scales measuring Conscientiousness and Openness, but discriminant validity and criterion validity increased only when each scale was corrected using its own corresponding vignette set. Moreover, accuracy in rating the vignettes was correlated with language achievement test scores, suggesting that verbal factors play a role in providing both normative vignette ratings of others and self-reports that are reliable and valid.

Keywords: response styles, Big Five, person differential functioning, large-scale educational assessment, 21st century skills, socio-emotional learning

Response styles (RSs) are a recognized phenomenon referring to the tendency to answer Likert-type scales in a systematic direction, regardless of their descriptive content. Likert scales are frequently used in social science research to examine preferences and attitudes, but also for the assessment of personality traits and social-emotional skills. RS can take multiple forms, including a propensity to agree with items irrespective of their content (acquiescent response style), to use the extreme or end points of the scale (extreme response style), or to endorse the middle-response options (middle-response style; see Bolt & Johnson, 2009; Hamilton, 1968; He, Bartram, Inceoglu, & van de Vijver, 2014; Lentz, 1938; Soto & John, 2009). RS can be influenced by culture (Harzing, 2006; Marin, Gamba, & Marin, 1992; Van Herk, Poortinga & Verhallen, 2004), personality characteristics, such as dominance and competitiveness (He et al., 2014), but also by gender and age (Austin, Deary, & Egan, 2006; Hamilton, 1968). Soto, John, Gosling, and Potter (2008), for example, demonstrated that individuals differ in acquiescent responding and that these differences can bias self-reports of personality traits, even in adults; however, the effects were most pronounced in preadolescents (ages 10–12 years) and then decreased in magnitude throughout adolescence.

RS can threaten the validity and reliability of scores, adding construct-irrelevant variance, with the potential to introduce a new latent dimension degrading the construct validity of single scales or distorting the internal structure of a multi-construct inventory. Past research has indicated that RS can be understood as person differential functioning (PDIF), where individuals with the same level of a latent trait (Johanson & Osborn, 2004) respond differentially to Likert scales, causing variance in total scores that is unrelated to the intended construct.

Social scientists have proposed different ways to handle RS (e.g., Paulhus, 1991; Soto & John, 2009). Here we focus on a psychometric approach that corrects raw scores using *anchoring vignettes* (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007; Mõttus, R., Allik, J., Realo, A., Rossier et al. 2012), which holds the promise to separate true score variance and RS. However, this approach has been mainly advocated in studies assessing

DOI: 10.1027/1015-5759/a000336

1. Low	Aline leaves her belongings in a mess, hates cleaning the house, and usually doesn't complete her homework. How organized do you think is Aline? 1 = Not at all | 2 = A little | 3 = Moderately | 4 = Very much | 5=Completely
2. Average	Manuela has a good sense for order but sometimes she leaves her room messy for a couple of days. She tends to complete school assignments just in time for the due date. How organized do you think is Manuela? 1 = Not at all | 2 = A little | 3 = Moderately | 4 = Very much | 5=Completely
3. High	Juliana is very careful and dedicated. She regularly cleans her room, is careful in her homework, and always finishes it well before the deadline. How organized do you think is Aline? 1 = Not at all | 2 = A little | 3 = Moderately | 4 = Very much | 5=Completely

Figure 1. Anchoring vignette set for Conscientiousness.

political attitudes. The current paper examines the usefulness of the anchoring vignette approach to correct personality and social-emotional skill assessments using Likert scales.

Using Anchoring Vignettes to Assess Response Style and Correct for Its Effects

Anchoring vignettes have been used frequently in political science research to improve comparability among assessments of attitudes and preferences in self-report questionnaires (King & Wand, 2007). The method tries to assess RS via the presentation of short descriptions of hypothetical persons (vignettes) that vary systematically in the latent traits represented in the inventory. Respondents are requested to rate the persons described in the vignettes on an item similar to those used for the respondents' self-descriptions, adopting the same response format and rating scale. An example of a *vignette set* we used to measure Conscientiousness is presented in Figure 1, which uses three vignettes to capture different levels (low, medium, and high) of the latent trait of Conscientiousness. The first vignette describes an individual (Aline) low in Conscientiousness; the second (Manuela) is average; and the third (Juliana) is high.

Relying on respondents' ratings of the persons in the vignettes, one can infer how respondents translate anchor levels into the 5-point Likert rating scale. Since every subject rates the same set of vignettes, these can be used as fixed anchor points to equate responses across respondents and thus correct their personality descriptions. Using Item Response Theory (IRT) terminology, the potential variability on vignette ratings can be conceived as thresholds varying within subjects, translating latent trait levels into ordered categorical responses (here from 1 to 5). Since the persons in the anchoring vignettes are constant across subjects, the variance in vignette ratings can be interpreted as an indicator of individual differences in response styles, which can be used to estimate PDIF. To account for PDIF, respondents' self-ratings are recoded in comparison with their ratings on the vignettes using a nonparametric recoding algorithm (King et al., 2004; Tandon, Murray, Salomon, & King, 2003; see also further in the Method section below). The key benefit of using anchoring vignettes is that an external source of data (i.e., distinct from the substantive descriptive information enclosed in self-ratings) is used for estimating RS (Hamamura, Heine, & Paulhus, 2008). Correcting for RS is presumed to improve the psychometric qualities of the scale.

If the RS assessed with anchoring vignettes reflects only a single, general individual difference (e.g., acquiescent responding across all attributes rated), then one would expect that these styles will affect individuals' responses similarly across different constructs. In other words, a vignette written to correct RS for Conscientiousness items should work equally well to correct Openness items, and vice versa. Some preliminary evidence for this claim has been provided by Kyllonen and Bertling (2014b) who found that an anchoring vignette written for a teacher support scale also worked to correct scale scores assessing student-teacher relations and student interest in mathematics. The present study provides a first opportunity to examine whether such corrective effects are also manifested when correcting a specific socio-emotional skill characteristic with an anchoring vignette designed for a *different* skill.

RSs can affect scale variance at the level of the individual, but they can also hamper the interpretation of mean (or group-level) scores, for example, when comparing means obtained after aggregating across children within schools or across people from different cultural groups (e.g., East Asian vs. Western countries). In these instances, RSs may cause systematic deviations between the scale midpoint and the obtained means, independent from the latent means. Children of two different schools may have identical latent

trait means, but the observed aggregated means may vary due to differential endorsement of extreme items. Comparisons relying on such confounded means therefore include a risk to lead to anomalous conclusions. Findings from the Program for International Student Assessment (PISA), for example, showed that mathematical self-concept was positively related to mathematics achievement (.40) at the within-country level, but aggregated self-concept scores were associated negatively (−.20, $N = 37$ countries) with country-mean scores when correlations were computed across countries (Kyllonen & Bertling, 2014a, 2014b).

Similar conceptual reversals for within and between-culture level comparisons have been reported for Conscientiousness and indices of achievement. Conscientiousness correlates positively with indicators of school (John, Caspi, Robins, Moffitt, & Stouthamer-Loeber, 1994) and work achievement (Wille, De Fruyt, & De Clercq, 2013) within various cultures, but Conscientiousness scores at the culture level were unrelated to Gross Domestic Product or the Human Development Index (McCrae et al., 2005).

Although associations between variables observed at one level cannot necessarily be assumed to translate to another level (the ecological fallacy), Kyllonen and Bertling (2014a) demonstrated that, using anchoring vignettes in PISA 2012, the association between teacher support and achievement rose from .03 to .13 when the analyses were run *within* countries, suggesting that correcting for anchor ratings increased predictive validity at the individual level. Surprisingly, at the culture level this association was −.45; however, after correcting the teacher support scores using the vignettes, the association with achievement became positive and rose to .29, suggesting parallel findings at both levels of observations. In a similar vein, Mõttus and colleagues (2012) demonstrated that country rankings were affected when conscientiousness' self-ratings were controlled for variance in anchoring vignettes' ratings. In sum, there is some first evidence that correcting for RS using anchoring vignettes may (a) lead to improved scale properties, (b) translate into increased predictive validities within samples, and (c) lead to very different conclusions about the associations between variables when looking at aggregated scores across groups (schools or cultures).

Large-Scale Study of Socio-Emotional Skills in Adolescents in Brazil: The Big Five

Social, emotional, and personal skills (e.g., goal-setting, perseverance, optimism, emotional control, gratitude, social intelligence, or curiosity), also referred to as 21st century skills, constitute a group of competencies considered crucial for individuals' development in current and future societies (Ananiadou & Claro, 2009; Trilling & Fadel, 2009). Broadly, social-emotional skills can be defined as individual characteristics that (a) originate in the reciprocal interaction between biological predispositions and environmental factors, (b) are manifested in consistent patterns of thoughts, feelings, and behaviors, (c) continue to develop through formal and informal learning experiences, and (d) influence important socioeconomic outcomes throughout the individual's life (De Fruyt, Wille, & John, 2015).

In Brazil, interest in the assessment and development of these skills has been growing over the past decade, as a means to improve general welfare and prepare youth for upcoming challenges via education and intervention programs in schools. The Institute Ayrton Senna (IAS) is a foundation that has played a key role in Brazil in raising awareness for this challenge and has initiated and supported a range of education projects, including the development of a reliable and valid instrument to assess these skills in the school context (Santos & Primi, 2014; see also Primi et al., 2016).

Socio-emotional skills have been mapped onto the Big Five taxonomy (McCrae & John, 1992) as a theoretical framework to guide assessment, research, and intervention programs (Kyllonen, Lipnevich, Burrus, & Roberts, 2008; McCrae & John, 1992). The Big Five taxonomy refers to five broad factor-analytically derived personality dimensions considered to be the largest common denominator underlying the variety of personality characteristics represented both in the natural language and in structured personality inventories (for a review, see John, Naumann, & Soto, 2008). The dimensions are commonly referred to as Extraversion, Agreeableness, Conscientiousness, Emotional stability (vs. Neuroticism), and Openness to experience. Extraversion and Agreeableness refer to individual differences in, respectively, the frequency and quality of social interactions, whereas Emotional stability represents individual variation in emotional strength, vulnerability, and regulation. Conscientiousness denotes differences in task engagement, performance, concentration, and achievement orientation, and Openness to experience describes individual variability in creativity, originality, and fantasy. There is compelling support for the cross-cultural validity of the five dimensions (McCrae, & Terracciano, 2005), and they have been also found valid to describe the personalities of children and adolescents (De Fruyt et al., 2006; John et al., 1994).

Over the past two decades, an international research effort has demonstrated that these Big Five dimensions predict a variety of significant life outcomes. John et al. (1994) showed that two of the Big Five, Conscientiousness and Openness to experience, are particularly important for predicting educational outcomes. Summarizing 15 years of subsequent research studies, Poropat (2009) conducted a meta-analysis of associations between the Big Five and academic performance, and found significant corrected correlations of .22 for Conscientiousness and .12 for Openness. Poropat's meta-analytic summary clearly points to Conscientiousness and Openness as key individual-difference dimensions to understand school achievement, and we will focus on these two dimensions in the present study.

The Present Research

Socio-emotional skills like self-discipline, persistence, intellectual curiosity, and passion for learning can be conceived as contextualized manifestations of these

underlying traits. They are chiefly assessed via self-reports, because participants find it easy to describe themselves, and self-reports are inexpensive to collect in large assessments for low-stake purposes (Kyllonen et al., 2008). Despite these advantages, traditional self-rating methods assume implicitly that (a) participants interpret and use response categories in the same way and (b) RSs do not meaningfully affect item responses. PDIF, however, is a validity threat to this assumption since it points to the unequal use of the rating scale by participants with identical levels on the latent trait (Cronbach, 1946). Attempts to account for and control RS using anchoring vignettes are hence an important objective in the search for optimally assessing personal, socio-emotional, and 21st century skills to achieve comparable measurements across individuals and schools (King & Wand, 2007).

The present study reports initial results of using anchoring vignettes to correct item responses in socio-emotional skill ratings in a large sample of adolescents, recruited from more than 200 schools in the federal state of Rio de Janeiro, Brazil. We compare Conscientiousness and Openness scale scores computed from raw-score (or uncorrected) item responses with scale scores computed from corrected item responses in terms of reliability (internal consistency), discriminant validity (correlations between the scales), and criterion validity (predicting school-based standardized achievement tests). Based on the literature reviewed above, we tested the following hypotheses: (a) Correcting item responses with the appropriate anchoring vignette for that skill domain (e.g., correcting Conscientiousness items using responses to the Conscientiousness vignette set) will improve the psychometric characteristics of that scale, so that corrected scales will have higher alpha reliabilities (Hypothesis 1a); (b) Correcting for an anchoring vignette will improve the discriminant validity (i.e., lower the intercorrelation) between constructs because it removes shared response-style variance from specific skill scores (Hypothesis 1b); and (c) Correcting for an anchoring vignette representing a specific skill domain will increase the validity for that domain to predict school achievement (Hypothesis 1c).

A second set of hypotheses tested Kyllonen and Bertling's (2014a) suggestion that response styles may be fairly general individual-difference factors; if so, it may be sufficient to correct with a single vignette (regardless of attribute domain), rather than correcting with multiple vignettes, each unique for the attribute domain being assessed. Specifically, correcting for a single anchoring vignette (even if mismatched in attribute content) would have the same beneficial effects on the reliability (Hypothesis 2a), discriminant validity (Hypothesis 2b), and criterion validity (Hypothesis 2c) for multiple scales.

Method

Participants

A representative sample of 12th grade students was recruited from 216 schools (median number of students per school was 50, min = 4, max = 150) in the federal state of Rio de Janeiro, Brazil (N = 8,582, 62% female); these adolescents participated in a school-based assessment and completed paper-and-pencil questionnaires. Mean age was 18.2 years (SD = 1.1). However, reflecting the large number of students in the public school system dropping out and then returning, the age range extended from 16 all the way to 24 years; in fact, more than a quarter of the sample (N = 2,293) was between 19 and 24 years old, an age range that in many Western countries would seem to fall well outside the expected age for secondary education. Such numbers are common, however, in the Brazilian educational system because students commonly drop out of school to help their families or work for a year or more and later reenroll to continue their education. As potential control variables, we also collected background information. Mothers' level of education was coded from 1 (= *never received formal schooling or did not complete first grade*) to 5 (= *completed college degree or higher*). Our index of economic standing of the adolescents' family was based on eight basic indicators, such as the availability of running water and electricity in the building, number of interior bathrooms available, number of cars, and so on.

Measures

Self-Reports of Socio-Emotional Characteristics Using SENNA 1.0

The SENNA 1.0 questionnaire is a Brazilian self-report inventory specifically designed for children and adolescents (see Primi, Santos, John, & De Fruyt, 2016; Santos & Primi, 2014), assessing core qualities of social-emotional skills with strong links to the Big Five personality taxonomy (John et al., 2008). For 12th graders, the measure includes 92 items and assesses youth versions of the Big Five personality dimensions plus External Locus of Control. Given their importance for understanding educational achievement (Poropat, 2009), we focus here on the Conscientiousness scale (18 items) and the Openness scale (14 items). Adolescents provided self-ratings on a 5-point Likert scale, using the rating scale in Figure 1.

Anchoring Vignettes

The vignette set we used for Conscientiousness is shown in Figure 1. Similar sets of vignettes were written for the other SENNA domains, each referring to a person low, average, or high on that attribute. The students rated each hypothetical person's standing on the corresponding trait using the same 5-point Likert scale as for the SENNA self-report items. Because many of the adolescents participating in this research come from poor families and some have limited attention and reading skills, testing time needed to be constrained to a single class period (i.e., no more than 50 min). Therefore, each student was asked to answer only three vignette sets (i.e., for a total of nine items): One set (three hypothetical target persons to be rated) for Conscientious-

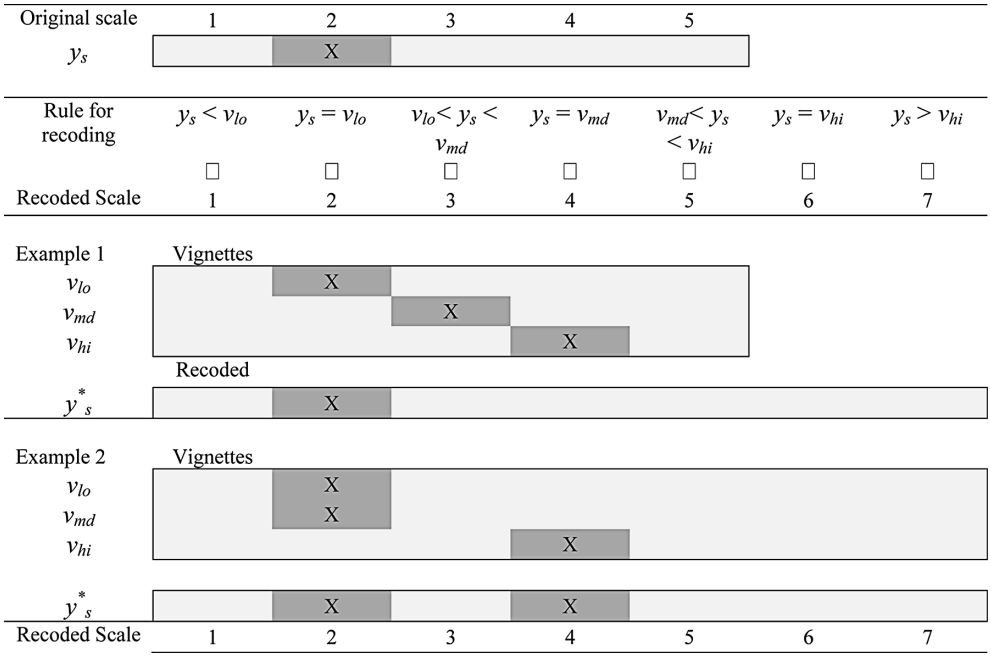

Figure 2. Two examples demonstrating the nonparametric recoding of three vignettes.

ness (C), one set for Neuroticism (N) and, depending on the vignette booklet, a third set for either Openness (O), Extraversion (E), or Agreeableness (A). So, there were three booklets – one with vignettes for C, N, and O, one for C, N, and E, and one for C, N, and A – which were randomly assigned within each school, in order to ensure random samples of students answering each booklet from each school. For the present study, ratings were available for $N = 8{,}458$ on the C-vignette set, and for $N = 2{,}816$ on the O-vignette set. In other words, data for the C-vignette set was available for the entire sample, and data or the O-vignette for a random 1/3 subsample; analyses involving corrected data using the O-vignette were thus based on the subsample of $N = 2{,}816$. In analyses using only C-vignette-corrected scores, or only original C and O scores, we were able to use the entire sample ($N = 8{,}458$). The items used to rate the vignettes (and later the self) asked the students to rate each person in terms of how *organized* (for the three C-vignettes) or how *creative* (for the three O-vignettes) they thought that person is.

Standardized Achievement Measures

As part of their regular school activities, students completed the standardized school test (SAERJINHO) for assessing achievement in language (i.e., Portuguese) and in mathematics. We obtained their scores for the academic semester directly from their school records.

Procedure

Data were collected in October 2013. Students rated the three vignette sets and the 92 SENNA self-report items, and answered a short demographic background questionnaire, including their gender, age, race, home context (including the economic standing indicators), and parental behaviors. Data collection was conducted during regular classroom activities in a block of 50 min, using traditional paper-and-pencil materials. A professional testing company trained the test administrators (i.e., the tests were not given by the schoolteachers) and processed the raw data.

Data Analyses: Vignette Violations and Two Vignette-Based Corrections

Data analysis focused on comparing the psychometric properties (reliability and validity) of raw versus vignette-corrected self-report item scores.

Addressing Violations in Ratings of Vignettes

The first step was to evaluate how normatively consistent the ratings of vignettes were and then perform the recoding. We used the nonparametric recoding procedure provided by the *Anchors Package* developed by Wand, King, and Lau (2011) for the R program (R Development Core Team, 2014). Specifically, we followed the procedure outlined by Kyllonen and Bertling (2014b) used in part to correct scales in PISA 2012. The procedure was as follows.

Let y_s be subject s response to a self-rating item, and v_{lo}, v_{md}, and v_{hi} responses to the three vignettes anchoring, respectively, low, medium, and high levels on the trait and let y^*_s be the recoded responses. Figure 2 presents two examples of recoding along with the recoding rules. Most common responses to vignettes are ordered correctly,

as defined by the anchors and illustrated in Example 1. In this case, the self-rating was $y_s = 2$. By applying the set of rules we discover that s response is equal to the one given to the low vignette ($y_s = 2 = v_{lo}$) resulting in the recoded score of $y^*_s = 2$. If it had been $y_s = 4$ it would have resulted in $y^*_s = 6$ ($y_s = 4 = v_{hi}$). These rules compare the response to see if it is lower than, equal to, or higher than the responses given to the vignettes, transforming the original scale to a $2k + 1$ point scale where k is the number of vignettes in the set (in our case, it transforms a 5-point scale to a $2 \times 3 + 1 = 7$-point scale).

As in previous research, we found that a subset of responses was *not* ordered in the way intended by the three anchors. This is illustrated in Example 2 in Figure 2. In this example, the low and medium vignette descriptions are tied ($v_{lo} = v_{md} = 2$). This causes indeterminacy in the way the responses should be recoded, resulting in more than one possible solution. Two rules are true: $y_s = v_{lo} = 2$ resulting in a $y^*_s = 2$ or $y_s = v_{md} = 2$ resulting in a $y^*_s = 4$. Even worse cases exist when responses to vignettes violate the preestablished order by reversals, for instance, when $v_{md} < v_{lo}$. Ties and order violations require some type of special treatment, which we will not discuss here in detail due to space limitations.

Among various treatments of anomalies, we followed the one described by Kyllonen and Bertling (2014b). We constructed a matrix of all possible response patterns to the vignettes (for three vignettes rated on 5-point scales, there are $5^3 = 125$ response patterns). Using this matrix, we tied all the violations to the orders and then recoded all the responses to the vignettes in such a way that all the violations became tied. We then applied the recoding scheme provided by the anchoring function of the *Anchors Package*[1] (see Wand et al., 2011, p. 3). In case of ties (and violations) the software produces the minimum and maximum values among all the rules outlined in Figure 2 that hold true for the responses. Kyllonen and Bertling (2014b) found that choosing the lowest interval value for recoding produced higher reliabilities. The same procedure was used here.

Two Kinds of Corrections: Using Either the Domain-Corresponding or the Other Vignette

We compared two kinds of corrections: Scale responses were corrected once for the vignette set corresponding to the characteristic being assessed (e.g., the Conscientiousness vignette was used to correct the Conscientiousness item scores), and once using the vignette set unrelated to the characteristic being assessed (e.g., the Conscientiousness vignette was used for correcting the Openness item score). This allowed us to examine our second set of hypotheses about the generalizability of corrections using a single vignette set for more than one dimension. In other words, we analyzed item scores recoded by their own, thus corresponding, vignette (e.g., items of the C-scale corrected by the C-vignette set), as well as item scores recoded using the vignette set from the other domain (e.g., C-items recoded with the O-vignette). All analyses were performed in R, mainly with functions from the Psych package (Revelle, 2014).

Statistical Analyses: Predicting Language and Math Test Scores

In the Results section, we compare the raw and the two kinds of vignette-corrected variables in terms of their criterion validity, predicting school-based achievement test scores in language and math. The procedures we used to predict the achievement scores were derived from economists' approaches in analyzing these types of school-based data (e.g., Hausman & Taylor, 1981). In particular, we used a linear regression estimation of the equation:

$$Y_{ij} = \beta_0 + \beta_C X_{Cij} + \beta_O X_{Oij} + \gamma_1 Z_{1ij} + \ldots + \gamma_p Z_{pij} + \sum_{k=2}^{j} \mu_j D_k + e_{ij} \quad (1)$$

where β_0 is the intercept, Y_{ij} represents a test score (either Portuguese or Math) of student i in school j; X_{Cij} and X_{Oij} represent SENNA scores on Conscientiousness and Openness, respectively, $Z_{1ij} \ldots Z_{pij}$ refers to a set of p variables socioeconomic controls at the individual level; and $D_2 \ldots D_j$ refers to dummy codes (one for each school minus 1) such as $D_k = 1$ if student i is in school j and $K = 0$ otherwise so each μ_j is a school j fixed-effect. We call this as a Fixed-Effect (FE) model since we estimate a large set of individual effects for each school. Our goal was to test whether β_C or β_O are statistically different from 0, after assuming $(e_{ij}|X_C, X_O, Z_p, D_k) = 0$, that is, errors are uncorrelated with predictors of the model. Under these assumptions, it would be possible to identify the regression function and we could estimate β_C and β_O with no bias by Ordinary Least Squares (eventually controlling for potential intra-school correlations in e by using a robust variance estimation).

The inclusion of school fixed-effects in the model aims to prevent potentially confounding effects that will prevent us from finding consistent estimates of β_C or β_O, which is essentially an identification problem. Imagine, for instance, that the true relationship between Conscientiousness and language achievement is zero (we know that this is likely an unrealistic example but just pretend that this is true); imagine that we have a subset of highly selective schools that use cognitive variables to select students, so their test scores will be above average $\bar{Y}_j > \beta_0$. Imagine also that these schools care a lot about organization and dependability and then they select students high in Conscientiousness (so $\bar{X}_{cj} > \bar{X}_c$). This would likely generate a positive relationship between Language test scores and

[1] The Anchors Package recodes one variable at a time with the anchors' function. We modified this function to recode an entire data frame doing the recoding all at once.

Table 1. Descriptive statistics: Responses to vignette sets for Conscientiousness (C) and Openness (O)

	C		O	
Response characteristics	Number	%	Number	%
Total number of responses	8,458		2,816	
Consistent responses	6,944	.821	1,934	.687
Ties				
One tie	484	.057	420	.149
Two ties	50	.006	44	.016
Order violations				
One violation	662	.078	284	.101
Two violations	233	.028	99	.035
Three violations	85	.010	35	.012
Correlation between violations in C and O				.260

Notes. C = Conscientiousness; O = Openness. Consistent responses are those that follow the intended ordering of the three vignettes (high, medium, low).

Conscientiousness caused by school selection practices and not by a direct relation between these two variables. If we don't take schools into account in the estimation of β_C we would have inconsistently found that $\beta_C > 0$ when in fact it is $\beta_C = 0$.

The fixed-effects model is not, however, the only option to account for school heterogeneity. A popular alternative to the FE model is the called Multilevel Model, Random Effects (RE, Hox, 2010) or Random Intercept model. It can be written as:

Level 1 equation:

$$Y_{ij} = \beta_{0j} + \beta_C X_{Cij} + \beta_O X_{Oij} + \gamma_1 Z_{1ij} + \ldots + \gamma_p Z_{pij} + e_{ij} \quad (2)$$

Level 2 equation:

$$\beta_{0j} = \beta_{00} + \mu_{0j} \quad (3)$$

And substituting the second equation into the first results in the general model:

$$Y_{ij} = \beta_{00} + \beta_C X_{Cij} + \beta_O X_{Oij} + \gamma_1 Z_{1ij} + \ldots + \gamma_p Z_{pij} + \mu_{0j} + e_{ij} \quad (4)$$

Notice that in this case, μ_{0j} is treated as a random variable, instead of a fixed-effect. Because of that it is now just one variable, the variance is estimated, $\sigma^2{}_{\mu_0}$ and is not a series of fixed quantities, as in the FE model. Maximum Likelihood or Generalized Least Squares can do the estimation of this model. The RE model is more efficient than the FE model but it comes with a price. Identification of this model requires not only that $E(e_{ij}|X_C, X_O, Z_p, \mu_{0j}) = 0$ but also that $E(\mu_{0j}|X_C, X_O, Z_p) = 0$, that is, we have to assume not only that errors are independent of predictors at Level 1 but also that random effects μ_{0j} are independent of e_{ij}. In other words, the multilevel model requires that the school heterogeneity is (mean) independent of both the constructs and the socioeconomic variables of the students. Suppose that μ_{0j} captures mainly the quality of the school. By this assumption we would rule out the possibility that schools in poor neighborhoods are at the same time of low quality (which could be the case if these schools have more difficulty to attract the best teachers), and have low SES students. In any case, multilevel models typically require stronger assumptions to be unbiased.

In the econometric literature there are detailed discussions comparing FE and RE models (e.g., Hausman & Taylor, 1981). A practical solution is to compare both models and check if there are differences in the coefficients. When differences are found, it is advisable to retain the fixed-effect coefficients (Antonakis, Bendahan, Jacquart, & Lalive, 2010). Our strategy here was to present results from both models.

All variables (outcome and predictors) were *z*-standardized before running the analyses. Therefore, the effects we report are similar to standardized coefficients in multiple regressions.

Results

In a preliminary set of analyses, we examined how consistently the adolescents responded to the vignettes by examining descriptive statistics for the response patterns ordered according to the normative model (i.e., the predicted ordering of the three hypothetical persons), including ties (two adjacent vignettes receiving the same rating) and violations (inversion of the predetermined anchored order of the vignettes). The upper part of Table 1 presents this information for the vignettes for C and O. The proportion of responses that was consistent with the intended anchor order was 0.82 for C, and somewhat lower at 0.68 for O. Overall, then, the vignettes worked as intended for more than two-thirds of the sample, but vignette difficulty (or quality) also varied somewhat, with lower levels of model consistency for the O-vignette.

Despite that mean difference, the number of order violations an adolescent made when responding to the C-vignette set (out of three questions) correlated *positively* and significantly ($r = .26$) with the number of order

Table 2. Internal consistency and discriminant validity of the scales scored from original and corrected item responses

	Internal consistency for each scale	
	Conscientiousness (C)	Openness (O)
Uncorrected scores	.87	.83
Scores recoded		
... with C-vignette set	*.95*	.92
... with O-vignette set	.93	*.91*
... with ***own*** vignette set	.95	.91
	Discriminant validity	
	Discriminant correlations between C and O	
Uncorrected scores		
Scores recoded	0.355	
... with C-vignette set	0.742	
... with O-vignette set	0.674	
... with ***own*** vignette set	0.182	

Note. C = Conscientiousness; O = Openness.

violations the adolescent made for the O-vignette set. In other words, adolescents who made no mistakes with one vignette set were also more likely to make no mistakes with the other vignette. The effect size of .26 is not large but similar to what one would expect for two single items measuring the same construct. This finding suggests that the vignettes may, in part, be tapping into a broader construct assessing whether students use personality rating scales in a normatively consistent way.

The internal consistency (alpha) coefficients for the C and O scales are reported in the upper part of Table 2, before and after correcting for the construct-corresponding vignette set (e.g., C-items corrected with the C-vignette) or for the construct-unrelated vignette set (e.g., C-items corrected with the O-vignette). The results were clear: internal consistency coefficients always increased after correcting, regardless of the vignette set that was used for correction.

How different were the corrected scores from the raw, uncorrected scores? When the number of violations students made in the vignette ratings was not taken into account, the correlations between raw and corrected scores (using the domain-corresponding vignette) in the entire sample were .60 for C and .67 for O. In other words, the vignette-based corrections substantially changed the ordering of the individuals on the constructs of interest. When analyses were restricted to the 82% of the sample who had no ties and no violations on the C-vignette set, the correlations between raw and corrected scores increased, as expected, for C (from .60 to .81) but not for O (from .67 to only .69). A similar pattern was observed when restricting analyses to the 68% of students who had no ties and no violations on the O-vignette: correlations increased for O (from .67 to .78) but not for C (from .60 to .61).

The discriminant validity correlations are shown in the lower part of Table 2. The uncorrected C and O scales correlated moderately (.36). However, when both C and O item responses were corrected using the *same single* vignette, the intercorrelation of .36 (r^2 = .13) more than tripled in variance terms: either to .67 (r^2 = .45) when correcting both C and O items with the O-vignette set, or to .74 (r^2 = .55) when correcting both C and O items with the C-vignette set. These effects represent a substantial loss in discriminant validity; using the same vignette set for correction introduced substantial collinearity (see this effect further elaborated in the regression models shown below).

In contrast, when responses were corrected for their own corresponding vignette, the correlation between C and O did not increase but dropped from .36 to .18, hence providing a meaningful decrement of collinearity, that is, an improvement of discriminant validity.

Table 3 reports the criterion validity correlations with the scholastic achievement test scores for Portuguese language and mathematics. Here we compare the uncorrected and corrected C and O scale scores when using their own corresponding vignette only. These zero-order external validity correlations (uncorrected for attenuation due to unreliability of either predictors or criteria) were consistently positive though small in size for both C and O. They held for both language and math, with the correlations slightly stronger for language than for math achievement, and slightly stronger for O than for C.

Correcting O for its corresponding vignette led to a small increase in validity when predicting language achievement, from .17 to .20. The other three zero-order validity coefficients did not increase.

Table 3 also shows the correlations between the number of vignette order violations (ranging from 0 to 3) and

Table 3. Predicting language and math achievement from raw and corrected personality scale scores and from number of vignette order violations: zero-order correlations

	Correlations with language test			Correlations with math test		
Personality domain	Raw scale scores	Corrected scale[a]	Number of violations[b]	Raw scale scores	Corrected scale[a]	Number of violations[b]
C	.11	.12	−.10	.07	.07	−.07
O	.17	*.20*	−.16	.12	.12	−.12

Notes. C = Conscientiousness; O = Openness. All coefficients are significant at $p < .05$. [a]Scale scores corrected with their own corresponding vignettes. [b]Number of order violations is the sum of normatively inconsistent ratings on that vignette set and does not include ties.

Table 4. Fixed-effects and multilevel models predicting math and Portuguese-language achievement test scores from either raw (uncorrected) or corrected conscientiousness (c) and openness (o) scores and number of vignette violations, controlling for adolescents' gender, family economic (econ.) standing, mother's education (edu.), sex, and school attended (Coded as dummy variables – fixed-effect models or random effects – multilevel models)

Models and predictors	Coefficients			
Model 0 (Null model)	*Language* (R^2 = .22)		Math (R^2 = .32)	
Intercept	−.03	–	−.03	–
$\sigma^2_{\mu_0}$		.22		*.32*
σ^2_e		.79		.68
	Language (R^2 = .26)		Math (R^2 = .32)	
Model 1 (No C and O Scores Included)	β^1	β^2	β^1	β^2
Intercept	−.01	–	−.02	–
Econ.Standing	.10	.10	.06	.07
Mother's Edu.	.06	.06	.04	.04
Sex	.09	.09	−.06	.05
	Language (R^2 = .26)		Math (R^2 = .34)	
Model 2 (C and O Scores *Uncorrected*)	β^1	β^2	β^1	β^2
Intercept	−.01		−.02	
Econ.Standing	.09	.09	.06	.06
Mother's Edu.	.06	.06	.04	.04
Sex	.09	.09	−.06	−.06
*Un*corrected *C*	.06	.07	.07	.07
*Un*corrected *O*	.12	.12	.06	.05
	Language (R^2 = .25)		Math (R^2 = .33)	
Model 3 (Both C and O Scores *Corrected with Same C-Vignette Only*)	β^1	β^2	β^1	β^2
Intercept	−.01		−.01	
Econ.Standing	.09	.08	.06	.06
Mother's Edu.	.06	.06	.04	.04
Sex	.09	.09	−.06	−.06
C corrected by C-vignette	*.00	*.00	.04	.04
O corrected by C-vignette	.13	.13	.05	.05
Violations in C	*−.01	*−.01	* .01	*.02
	Language (R^2 = .32)		Math (R^2 = .38)	
Model 4 (Both C and O Scores *Corrected with Same O-Vignette Only*)	β^1	β^2	β^1	β^2
Intercept	.02		−.01	
Econ.Standing	.10	.09	.06	.05
Mother's Edu.	.05	.05	*.03	*.03
Sex	.07	.07	−.11	−.12
C corrected by O-vignette	*.01	*.02	*.03	*.04
O corrected by O-vignette	.13	.13	*.02	*.10
Violations in O	−.07	−.06	−.07	−.06
	Language (R^2 = .33)		Math (R^2 = .38)	
Model 5 (C and O Scores Each *Corrected with Own Corresponding Vignette*)	β^1	β^2	β^1	β^2
Intercept	.02		−.01	
Econ.Standing	.10	.09	.06	
Mother's Edu.	.05	.05	*.03	*.03
Sex	.07	.06	−.11	−.12
C corrected by C-vignette	.09	.09	.10	.10
O corrected by O-vignette	.12	.13	.03	.04
Violations in C	−.05	.04	*.00	*.01
Violations in O	−.06	−.06	−.08	−.07

Notes. R^2: percentage of total variance explained by the fixed-effect models; *p > .05; C = Conscientiousness; O = Openness; Econ. Standing: Family Economic Standing, Mother's Edu. = Mother's Education. Sex was keyed so that positive regression weights mean girls' high scores then boys and vice versa. [1]Standardized coefficients referent to the fixed-effect multiple regression analysis models; [2]Standardized coefficients referent to the random intercept multilevel models.

objectively measured school achievement. The number of vignette violations correlated negatively with both achievement tests. Specifically, language achievement correlated −.10 with violations on the C-vignette and −.16 with violations on the (more difficult) O-vignette. In other words, students showing poorer language performance also made more mistakes in the personality vignette ratings.

We also conducted a series of control analyses because in the recoding scheme we used here, students with more violations tend to receive lower corrected scale scores on both C and O. This raises the question whether corrected C and O scores might be confounded with the number of violations, which might in turn complicate the interpretation of the observed correlations of C and O with school achievement. That is, corrected scale scores may be lower because these students have truly lower scores on the latent constructs scores, but they might also reflect comprehension difficulties indexed by vignette violations (see Table 3).

To examine these alternative explanations, we conducted five multiple regression models, regressing achievement on either the uncorrected personality scales or the different kinds of corrected personality scale scores, including also relevant control variables: gender, mother's education, economic standing, school attended, and finally number of vignette violations (computed separately for the C- and the O-vignette sets).

Because students were recruited through the school they attended and are thus statistically "nested" inside schools, we also conducted multilevel analyses with individual students as Level 1 and the 216 schools as Level 2 variables; this type of analysis allowed us to estimate the effects of schools as random effects. We first estimated the null model with only school as a Level 2 predictor. As shown in Table 4, this model accounted for 22% of the variance in language achievement scores and 32% in math scores. Briefly put, these variance percentages indicate that, as expected, schools in Brazil differ substantially from each other in the achievement of their students.

In Model 1, we added the student's sex, their mother's education, and their family's economic standing as Level 1 variables to the model, resulting in 26% and 32% of explained variance, respectively. Models 2–5 then serve to test the effect of including Conscientiousness and Openness as predictors, comparing uncorrected scales (in Model 2) and the variously corrected scales (in Models 3–5).

Specifically, in Model 2 we added the uncorrected Conscientiousness and Openness scores as fixed predictors. The explained was 26% and 34%, respectively. The next two models included corrected scale scores when both C and O were corrected by the same vignette set (i.e., using the C-vignette set to correct both in Model 3 and the O-vignette set to correct both in Model 4). This kind of correction increased the correlation between C and O (see Table 2) and should thus create collinearity (overlap) between the two predictors in these analyses; we also included the number of violations for that vignette set as a control. Indeed, Models 3 and 4 did not increase the explained variance beyond Model 2 and C isn't statistically significant.

Finally, Model 5 included the Conscientiousness and Openness scores where each scale was corrected by its own, corresponding vignette set. These final models accounted for 33% and 35% of variance in achievement. These analyses produced similar results for C and O, even when we used the C and O scores both corrected with the same single vignette set potentially generating collinearity among predictors. In all, the criterion validity findings appear to be due to the unique contribution of each trait predictor, rather than the reading difficulty of the vignettes. But we only see a small improvement in criterion validity when we compare scores of Models 3–5 with uncorrected scores of Model 2.

General Discussion

This paper reports an initial test of three major hypotheses about the effects of using anchoring vignettes to correct for RS in large-scale self-report assessments of C and O in Brazilian youth attending the 12th grade. We examined the effects of correcting for vignette ratings on internal consistency, discriminant validity, and predictive validity.

The two vignettes written for C and O generally performed well in the present study. The proportion of normative responses (82% for C and 68% for O) was in line with percentages reported in previous work (ranging from 63% to 91% in Kyllonen & Bertling, 2014b, and from 65% to 92% in Mõttus, R., Allik, J., Realo, A., Rossier et al. 2012) although there were more ties and violations for the Openness than for the Conscientiousness vignette set. These differences could be due to the particular vignette sets under consideration, but we suspect they reflect substantive differences between these two attribute domains: whereas Conscientiousness refers to concrete and easily observable behaviors (e.g., punctual arrival; neat and organized work space; completing tasks), Openness refers to internal and experiential characteristics of the individual (e.g., curiosity; interest in learning and understanding; aesthetic experiences) that are difficult to observe, evaluate, and rate even for adults (John & Robins, 1993; see also Vazire, 2011, and Soto et al., 2008).

Our finding that the number of vignette violations was correlated across domains shows these measures have a common core and suggests further hypotheses for future research; it is possible, for example, that vignette anchor ratings provide a test of the individual's skills in constructing and verbally communicating social perceptions that reflect the norms and expectations of a culture or community.

Increases in the reliability of the C and O scales were observed regardless of which vignette was used for correction, confirming our Hypotheses 1a and 2a. This finding could indicate that correction reduces error variance in item responses, increasing the ratio of true versus total score variance, and advancing within-domain coherence of self-ratings (Soto et al., 2008). However, an unexpected finding was that using the other, noncorresponding vignette set for correcting responses (i.e., correcting C-items with the

O-vignette set, and vice versa) led to a dramatic increase of the interdomain correlations, threatening the discriminant validity of the scales and rejecting our Hypothesis 2b. Correction using only the corresponding vignette set, in contrast, served to reduce the intercorrelation between C and O, thus improving discriminant validity and confirming Hypothesis 1b.

The surprising findings on the poor discriminant validity may suggest an alternative explanation for the observed increases in internal consistency for corrected scores. To the extent that there are no individual differences in the vignette ratings, for example when most children are giving a rating of three to the high vignette or when there is a high proportion of ties, the correction method could impute within-subject dependency among responses, because a set of values would be reduced to a single value within subjects (value 4 or 5 reduced to 7 for instance). This would explain the increase in internal consistency accompanied by an increase in the correlation between C and O, because responses in both domains are restricted relative to the same set of anchor points provided by the vignettes. This does not happen when using different vignette sets for each domain; of course, even though the dependency cannot occur between domains, it could still be occurring *within* a domain.

The final set of hypotheses was related to criterion validity. The correlations of C and O with school achievement observed here were generally consistent with what has been reported in the personality literature over the past 20 years (John et al., 1994; Poropat, 2009; Von Stumm & Ackerman, 2012). Interestingly and new was that the number of vignette violations showed correlations with school achievement as well.

An interesting finding was that openness tended relatively to be more associated to language achievement and conscientiousness more to math achievement when analyzing the unique contribution of each domain in the multiple regressions (Model 5). This suggests that imagination and aesthetic sensitivity could be more related to engaging in regular reading habits that in turn could develop language skills, whereas order and discipline could be more important for the persistence and repeated practice necessary for developing the complex acquaintance underlying math knowledge. These findings suggest the intriguing hypothesis that C and O have unique moderation effects in the process of investment of potential abilities into learning activities that turn potential into crystalized knowledge (Von Stumm & Ackerman, 2012).

Original and corrected scores for C and O showed generally similar magnitudes of criterion validity, providing no support for our Hypotheses 1c and 2c. Results suggest that correcting responses using corresponding domain vignettes helped to more clearly identify the unique contribution of each domain with achievement via multiple regression (openness with language achievement vs. conscientiousness with math). To increase criterion validity for C and O, it seems more promising to recode each domain with its respective vignette set.

Like all empirical research, this study has important limitations. We studied a large sample of high school students within the public school context, and made great efforts to adapt our measurement to that particular context (Santos & Primi, 2014). However, the school context and the students in Brazil are considerably more diverse, both within classrooms and across schools, than in many more developed countries, such as Finland or Japan. Thus, our findings await replication in other cultures and languages, and in developed and developing countries.

Future research should examine the effects of age; we suspect that the benefit of using vignette ratings may be even more pronounced for younger age groups than the high school age group studied here. Moreover, we need information about the developmental course of normative performance on vignette ratings; like Soto et al.'s (2008) work on acquiescence response set, one might test the hypothesis that vignette performance is worse in childhood and preadolescence and then improves throughout the middle or even high school years and when it reaches asymptote.

Finally, our findings are consistent with the idea that the vignettes may well tap into particular skills in social perception and communication. If so, we wonder whether those skills can be taught. What might be the effect of continued practice on personality vignettes, especially with normative feedback on performance? One could even envision students sharing and discussing their social perceptions, initially using the vignettes but then extending to real individuals they know, and even to themselves. Ultimately, this kind of work might give rise to advances in our understanding of the many ways social and self-perception processes operate in the real world.

Acknowledgments

This article is part of a research project financed by the Ayrton Senna Foundation (http://www.institutoayrtonsenna.org.br, http://educacaosec21.org.br). First author also receives a scholarship from National Council of Technological Research (CNPq).

References

Ananiadou, K., & Claro, M. (2009). *21st century skills and competences for New Millennium Learners in OECD countries*. Paris, France: Centre for Educational Research and Innovation (CERI) – New Millennium Learners.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*, 1086–1120.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235–1245.

Bolt, M. D., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335–352. doi: 10.1177/0146621608329891

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475–494.

De Fruyt, F., Bartels, M., Van Leeuwen, K. G., De Clercq, B., Decuyper, M., & Mervielde, I. (2006). Five types of

personality continuity in childhood and adolescence. *Journal of Personality and Social Psychology, 91*, 538–552. http://dx.doi.org/10.1037/0022-3514.91.3.538

De Fruyt, F., Wille, B., & John, O. P. (2015). Employability in the 21st century: Complex (interactive) problem solving and other essential skills. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 276–281.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*, 932–942.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69*, 192–203.

Harzing, A.-W. (2006). Response styles in cross-national survey research. *International Journal of Cross-Cultural Management, 6*, 243–266.

Hausman, J., & Taylor, W. (1981). Panel data and unobservable individual effects. *Econometrica, 49*, 1377–1398.

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

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge.

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

John, O. P., Caspi, A., Robins, R., 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*, 160–178.

John, O. P., Naumann, L., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: Discovery, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement: The Big Five, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521–551.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*, 191–207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*, 46–66.

Kyllonen, P. C., & Bertling, J. P. (2014a). Innovative questionnaire assessment methods to increase cross-country comparability. In L Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 277–286). Boca Raton: Taylor & Francis.

Kyllonen, P. C., & Bertling, J. P. (2014b). *Draft report: anchoring vignettes reduce bias in noncognitive rating scale responses*. Princeton, NJ: ETS/OECD.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2008). Personality, motivation, and college readiness: a prospectus for assessment and development,. Non printed technical report. Educational Testing Service, Princeton. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/ets2.12004/epdf

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. *Psychological Bulletin, 35*, 659.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology, 23*, 498–509.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407–425. doi: 10.1037/0022-3514.89.3.407

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., … Tseung, C. N. (2012). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality, 26*, 303–317. doi: ez1.periodicos.capes.gov.br/10.1002/per.840

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., … Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*, 1423–1436. doi: 10.1177/0146167212451275

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. 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*, 322–338.

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

R Development Core Team. (2014). R: *A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Revelle, W. (2014). *psych: Procedures for personality and psychological research*. Evanston, IL, USA: Northwestern University.

Santos, D., & Primi, R. (2014). *Social and emotional development and school learning: a measurement proposal in support of public policy* Technical report for Organization for Economic Cooperation and Development (OCDE) Rio de Janeiro State Education Department (SEEDUC) and Ayrton Senna Institute. São Paulo, Brazil: Ayrton Senna Institute.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality, 43*, 84–90.

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

Tandon, A., Murray, C. J., Salomon, J. A., & King, G. (2003). Statistical models for enhancing cross-population comparability. In *Health systems performance assessment: Debates, methods and empiricism* (pp. 727–746). Geneva: World Health Organization. Retrieved from http://www.who.int/healthinfo/paper42.pdf

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco, CA: Jossey-Bass.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346–360. doi: 10.1177/0022022104264126

Von Stumm, S., & Ackerman, P. L. (2012). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139*, 841–869.

Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software*. Forthcoming, Retrieved from http://www.jstatsoft.org

Wille, B., De Fruyt, F., & De Clercq, B. (2013). Expanding and reconceptualizing aberrant personality at work: Validity of Five-Factor Model aberrant personality tendencies to predict career outcomes. *Personnel Psychology, 66*, 173–223.

Received: October 28, 2014
Revision received: April 23, 2015
Date of acceptance: September 5, 2015
Published online: April 22, 2016

Ricardo Primi

Universidade São Francisco
Laboratory of Psychological and Educational Assessment
Rua Alexandre Rodrigues Barbosa, 45
CEP 13251-900
Itatiba, São Paulo
Brazil
Tel. +55 11 45348118
Fax +55 11 45348118
E-mail rprimi@mac.com

