The Effects of Vignette Scoring on Reliability and Validity of Self-Reports

Applied Psychological Measurement

2018, Vol. 42(4) 291–306

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617730389

journals.sagepub.com/home/apm

Matthias von Davier¹, Hyo-Jeong Shin²,
Lale Khorramdel², and Lazar Stankov³

Abstract

The research presented in this article combines mathematical derivations and empirical results to investigate effects of the nonparametric anchoring vignette approach proposed by King, Murray, Salomon, and Tandon on the reliability and validity of rating data. The anchoring vignette approach aims to correct rating data for response styles to improve comparability across individuals and groups. Vignettes are used to adjust self-assessment responses on the respondent level but entail significant assumptions: They are supposed to be invariant across respondents, and the responses to vignette prompts are supposed to be without error and strictly ordered. This article shows that these assumptions are not always met and that the anchoring vignette approach leads to higher Cronbach's alpha values and increased correlations among adjusted variables regardless of whether the assumptions of the approach are met or violated. Results suggest that the underlying assumptions and effects of the anchoring vignette approach should be carefully examined as the increased correlations and reliability estimates can be observed even for response variables that are independent random draws and uncorrelated with any other variable.

Keywords

response styles, anchoring vignettes, Cronbach's alpha, convergent-discriminant validity, Program for International Student Assessment (PISA)

Introduction

Rating scales, such as Likert-type scales (Likert, 1932), are widely used in the assessment of noncognitive constructs and are also used in large-scale surveys to measure a variety of personality traits or attitudes using self-reports. Most commonly, Likert-type scales consist of at least four ordered response categories, sometimes with a middle category that reflects a neutral or undecided response (e.g., 1 = *strongly disagree*, 2 = *somewhat disagree*, 3 = *neutral or neither of agree/disagree*, 4 = *somewhat agree*, 5 = *strongly agree*).

¹National Board of Medical Examiners, Philadelphia, PA, USA

²Educational Testing Service, Princeton, NJ, USA

³Australian Catholic University, North Sydney, New South Wales, Australia

Corresponding Author:

Matthias von Davier, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA.

Email: mvondavier@nbme.org

Even though survey instruments are designed to be as concrete, objective, and standardized as possible, the intended comparability is not always guaranteed. First of all, the assignment of equal distances to ordered response categories is arbitrary, and the interpretations of the distances are subject to cultural and individual characteristics of the respondents (von Davier, 2010). Respondents from different cultural groups may exhibit different tendencies in responding to survey questionnaires, as shown by Chen, Lee, and Stevenson (1995); Hui and Triandis (1989); and King, Murray, Salomon, and Tandon (2004). For example, students in East Asia were found to be more likely to use midpoint categories than students in North America. In addition to cultural differences, individuals may understand the same question or response categories in different ways, and measurement based on respondents' perceptions of the situations can be biased (Brady, 1985; Sen, 2002).

Response styles (RSs) are of particular concern in self-reports using rating scales. A RS is defined as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (that is, what the items were designed to measure)" (Paulhus, 1991, p. 17). RSs refer to individual differences in the preference to choose among response categories, for example, the tendency to choose extreme responses more (or less) frequently than other categories independent of what the question or statement in the questionnaire is intended to measure. It has been shown that RSs affect the reliability and validity of survey data (e.g., Baumgartner & Steenkamp, 2001; Dolnicar & Grün, 2009; Greenleaf, 1992), although the direction and magnitude of change vary across the survey instruments and the sample of respondents. Previous studies described methods to define certain types of RSs, and argued that RS can affect conclusions about the relationship between scales by either increasing or decreasing the correlation between respondents' scores on survey instruments. Changing the correlations consequently affects the factor structure and regression analyses, which are frequently used techniques in secondary analysis of large-scale survey data. Thus, to correct and control for RS, several statistical approaches have been developed. For example, Rost, Carstensen, and von Davier (1997) used a mixture distribution item response theory (IRT) approach to correct for RS, an approach that was recently extended to multidimensional mixture IRT (von Davier, Naemi, & Roberts, 2012). Other recent studies by Böckenholt (2012), Bolt, Lu, and Kim (2014), de Jong, Steenkamp, Fox, and Baumgartner (2008), Johnson and Bolt (2010), Khorramdel and von Davier (2014), and von Davier and Khorramdel (2013) used different types of IRT models—or more broadly, factor analytic approaches—to reduce the effects of RSs on latent trait estimates.

Here, "anchoring vignettes" are studied that are administered in addition to the questionnaire items as part of the self-assessment, and are then used to correct for RSs affecting the questionnaire items (e.g., Hopkins & King, 2010; King et al., 2004; King & Wand, 2007).¹ Vignettes describe fictitious individuals with known characteristics who are rated by respondents using the same rating or Likert-type scale that is used for the self-assessment items. The vignettes usually take the form of a short text that describes another person's behavior associated with a certain value or response option on the rating scale. For example, students may be asked to read three texts describing imaginary scenarios in which each of the fictitious agents shows low, medium, and highly motivated behavior, respectively. Depending on their responses to these vignettes, students' self-ratings are adjusted. Hence, different levels on the rating scale of interest are described by example (the vignette) with respect to a particular concept and held constant across respondents. Thus, vignettes aim at providing scale anchors that are assumed to enable interpersonal comparisons. While this may seem like a plausible approach, it should be noted that there is strong empirical evidence that ratings of actors and observers are far from being perfectly aligned. Attribution theory has ample examples of how the same action or the same outcomes are evaluated differently, depending on whether the agent is the respondent

himself or herself, or some other persons. A prominent example is the tendency of applying an actor–observer bias (Jones & Nisbett, 1971) which is one form of self-serving bias (e.g., Heider, 1958; Kelley, 1973; Larson, 1977).

Although the anchoring vignette approach does not claim to resolve all issues related to RS, it has been argued that this approach has the potential to reduce bias, increase efficiency, and make measurements more interpersonally comparable than existing methods. For example, Kyllonen and Bertling (2014) argued that anchoring vignettes resolved the achievement–attitude anomaly identified in the Program for International Student Assessment (PISA). They found that correlations between achievement and attitude (e.g., motivation) at the within-country level increased, while between-country correlations were less negative after adjusting the scales using anchoring vignettes. Note, however, that in the PISA 2012 background questionnaire (BQ), about 25% of students' responses on vignettes violated assumptions of the approach and were not strictly ordered (Organisation for Economic Co-Operation and Development [OECD], 2014). Although increased attitude–achievement correlations appear to improve the results, it is unknown whether this increase mirrors the underlying associations between constructs better, or reflects artifacts of the transformation process.

Therefore, this article aims to investigate the underlying assumptions and the effects of the “anchoring vignettes” approach proposed by King and colleagues (2004) on the reliability and the validity of rating data. The authors of the present study describe both methods for anchoring vignettes—the nonparametric and the parametric method—and show how *in both cases* the person-specific adjustment by vignette scores can be viewed as a multidimensional function of original response and vignette scores. The following research questions are addressed in this article:

Research Question 1: Why does the nonparametric method fail to provide a well-defined transformation if the required strict ordering of vignette ratings is violated by some respondents?

Research Question 2: What is the distribution (mean, variance, and covariance) of vignette-anchored variables?

Research Question 3: What can be learned about the increased reliability estimates and changes to validity coefficients using the PISA 2012 BQ example?

To answer these questions, the effects of anchoring vignette approach on the distributional properties, including the mean, the variance, and the covariance of anchoring vignette-transformed scores, are examined. How the approach increases correlations between transformed variables even if the original responses are random and unrelated to any trait is shown. The authors proceed to illustrate the effects of anchoring vignette scoring on reliability and validity coefficients using a simulation study (results provided in the online appendix),² and in an empirical example based on data from the PISA 2012 BQ scales. Stankov, Lee, and von Davier (2017), also using the PISA 2012 scales, provide a somewhat different perspective by looking at the effects of vignette scoring on factor structure as it relates to construct validity across scales.

Correcting RSs Using Vignettes

Anchoring Vignette Scoring as a Multidimensional Response Transformation

Formally, a transformation of response vector using anchoring vignettes can be viewed as a function that maps the combined set of $i = 1, \dots, K$ self-assessment questions and $v = 1, \dots, J$ vignette questions ($X_1, \dots, X_K, Z_1, \dots, Z_J$) onto a new vector of vignette-transformed variables

$(Y_1, \dots, Y_K)$. The resulting variables Y_i are functions of the vector-valued input variables, and the statistical distribution of the resulting variables can be determined analytically if the distribution of the input variables is known. The authors refer to the mapping from the $K + J$ dimensional space of self-assessment responses plus vignette responses to the i th transformed (vignette-anchored) responses as $Y_i = f_i(X_1, \dots, X_K, Z_1, \dots, Z_J)$. A mapping of this type is called monotonic increasing in X_j if

$$f_i(X_1, \dots, X_j^*, \dots, X_K, Z_1, \dots, Z_J) > f_i(X_1, \dots, X_j', \dots, X_K, Z_1, \dots, Z_J)$$

for all $X_j^* > X_j'$, and $f()$ is called monotonic decreasing if $-f()$ is monotonic increasing. The same definitions apply to any of the Z components of the function. Note that real-valued functions with vector-valued arguments can be monotonic increasing in some components of the argument and monotonic decreasing in some, and nonmonotonic in others. The Equation $Y = X_1 - X_2 + X_3^2$ is a simple example.

For vignette-anchored response variables, the adjusted scores Y_i are not functions of all input variables. They are functions of a single original response X_i and monotonic in X_i , and depend on the set of vignette responses $(Z_1, \dots, Z_J)$. It can be written as

$$Y_i = Y_i(X_i | \mathbf{Z}) = f(X_i, Z_1, \dots, Z_J).$$

The specific function $f()$ that is applied to vignette scoring will be introduced and discussed in the next section. This functional dependency of Y_i on $X_i, Z_1, \dots, Z_J$ implies that the covariances $COV(Y_i, Y_k)$ depend not only on $COV(X_i, X_k)$ but also on the extent to which the Z_v determine the vignette-based transformation.

It is important to note that neither all monotonic nor all linear functions will necessarily induce correlations or change them in the transformed variables. Orthogonal transformations such as the ones used in factor rotation are a prominent example. Also, if all respondents chose the same vignette scores, then Y_i is a function of X_i only, and the Y_i remains uncorrelated if the X_i were uncorrelated.

In summary, transformed variables can be correlated or uncorrelated; correlations can change or stay the same between input and output variables as discussed above. The result depends on the exact nature of the transformation and is not a consequence of the transformation being monotonic or not. In the following, the effects of vignette scoring are reviewed by first describing the nonparametric and parametric approaches, and then exact results are derived for the expected value, the variance, and covariance of vignette-anchored transformed variables.

Nonparametric Method

An example of mapping responses using vignettes is illustrated below. Assume that there are three vignettes (Z_1, Z_2, Z_3) and five self-assessment questions (X_1, X_2, X_3, X_4, X_5). For both, assume respondents provide answers on a 5-point Likert-type scale—(1) *strongly disagree*, (2) *somewhat disagree*, (3) *neither disagree nor agree*, (4) *somewhat agree*, (5) *strongly agree*.

Using the *italicized* words in the text as a guide, the levels of each fictitious person are designed to be recognized as low-to-medium hard working for Peter, medium for Paul, and medium-to-high for Mary. Depending on the responses on vignette questions (Z_j), each number-correct score on the self-assessments, $X_i \in \{1, 2, 3, 4, 5\}$, is transformed into a vignette-anchored variable with seven levels, $Y_i \in \{1, 2, 3, 4, 5, 6, 7\}$, following the rules given below (Hopkins & King, 2010; King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007).

How much do you agree with the following **bolded** statements?

Item 1. Peter is a person who *sometimes* works hard on some assignments, and returns his work before the deadline and receives excellent grades. *Often*, he does provide work that is reasonable, but not quite on time, and *sometimes*, he is quite late and only provides what was minimally required to pass. **Peter is a hard worker.** (Vignette 1; Z_1)

Item 2. Paul is a person who *frequently* works hard on his assignments, and returns his work before the deadline and receives excellent grades. *Sometimes*, he does provide work that is reasonable, but not quite on time, and *rarely*, he is quite late and only provides what was minimally required to pass. **Paul is a hard worker.** (Vignette 2; Z_2)

Item 3. Mary is a person who *almost always* works hard on her assignments, and returns her work before the deadline and receives excellent grades. *Rarely*, she does provide work that is reasonable, but not quite on time, and *almost never*, she is quite late and only provides what was minimally required to pass. **Mary is a hard worker.** (Vignette 3; Z_3)

Item 4. **I work hard on my assignments.** (Student Motivation 1; X_1)

Item 5. **I do what it takes to get things done.** (Student Motivation 2; X_2)

Item 6. **I take responsibility for working on my goals.** (Student Motivation 3; X_3)

Item 7. **I complete my schoolwork regularly.** (Student Motivation 4; X_4)

Item 8. **I am good at staying focused on my goals.** (Student Motivation 5; X_5)

$$Y_i = 1 \text{ if } X_i < Z_1$$

$$Y_i = 2 \text{ if } X_i = Z_1$$

$$Y_i = 3 \text{ if } Z_1 < X_i < Z_2$$

$$Y_i = 4 \text{ if } X_i = Z_2$$

$$Y_i = 5 \text{ if } Z_2 < X_i < Z_3$$

$$Y_i = 6 \text{ if } X_i = Z_3$$

$$Y_i = 7 \text{ if } X_i > Z_3$$

For example, if Peter, Paul, and Mary are rated as (2), (3), and (4), respectively ($Z_1 = 2$, $Z_2 = 3$, $Z_3 = 4$), the responses on self-assessments (X_i) will be transformed as follows: ($x_i = 1$, $y_i = 1$), ($x_i = 2$, $y_i = 2$), ($x_i = 3$, $y_i = 4$), ($x_i = 4$, $y_i = 6$), and ($x_i = 5$, $y_i = 7$). Note that $y_i = 3$ and $y_i = 5$ cannot be observed in this example because there is no integer value between 2 (Z_1) and 3 (Z_2), or 3 (Z_2) and 4 (Z_3). More formally, let us assume that a categorical original variable X_i ranging from 1 to H is obtained on a number of self-assessment questions, $i = 1, \dots, K$. In addition, assume that there is a number of $j = 1, \dots, J$ responses on vignette questions Z_j on the same $1, \dots, H$ scale. The authors of the present study define new ordinal vignette-anchored variables, $Y_i \in \{1, \dots, 2J + 1\}$, transformed from the observed self-assessment scores depending on scores on vignettes. This transformed vignette-anchored variable y_{pi} of a respondent p will be determined based on the relative position of x_{pi} compared with the set of vignette responses z_{pj} .

The resulting variable is expected to be RS free, to have easily interpretable units, and to be able to be analyzed like any other ordinal variable. Note that when the nonparametric method for anchoring vignettes is used in practice, several assumptions are made: First, the set of Z_j must be strictly ordered, so that $Z_1 < Z_2 < Z_3$ and unordered values (i.e., ranking inconsistencies, ties) are not allowed. Second, a set of vignette responses Z_j are reflections of personal perception of “true” scale values (θ_j) underlying the vignettes, which means that vignette questions are assumed to be objective and invariant across respondents. Third, respondents evaluate vignette questions (Z_j) based on the same latent trait as the self-assessment questions (X_i). And fourth, because the RS are applied at the respondent level, for the given latent trait,

Table 1. Transformed Score Examples When Vignette Responses Are Strictly Ordered.

x_i	y_i depending on (z_1, z_2, z_3)					
	(1,2,3)	(3,4,5)	(1,3,5)	(2,3,4)	(1,4,5)	(1,2,5)
1	2	1	2	1	2	2
2	4	1	3	2	3	4
3	6	2	4	4	3	5
4	7	4	5	6	4	5
5	7	6	6	7	6	6
<i>M</i>	5.20	2.80	4.00	4.00	3.60	4.40
<i>SD</i>	2.17	2.17	1.58	2.55	1.52	1.52

the same RSs affect all items in the same way, and the same transformation is conducted for all self-reports using rating scales.

Among these, the combination of the first two assumptions is called “vignette equivalence” (between individuals) in the literature (King et al., 2004; King & Wand, 2007). This assumption requires that the levels intended in each vignette are understood in the same way on average by all respondents. For example, if there are three vignettes designed with the intent to illustrate a low, medium, and high level of motivation, the levels of motivation designed in vignettes should be understood correctly by all respondents and applied accordingly, except for random error and personal differences in choosing rating categories. Next, the combination of the third and the fourth assumption is called a (within-individual) “response consistency” assumption (King et al., 2004; King & Wand, 2007). It requires that each respondent uses the response categories in approximately the same way when he or she provides answers to the self-assessments and to the vignette questions. In addition, the same underlying latent trait is tapped into when respondents answer across vignette questions and self-assessment questions.

In summary, individual differences in responding to vignette questions are solely attributed to the RS. If the effects of RS (estimated from vignette questions) from the self-assessments are controlled for, then more genuine, RS-free responses should be obtained. However, those assumptions would be violated if (a) vignette questions cannot be objectively judged by respondents, or do not behave as expected, for example if ties or reversals are observed as opposed to eliciting only strictly ordered responses, and if (b) vignettes invoke other types of biases in some respondents (e.g., preference for the names or language used in the vignettes or desire for a respondent to represent themselves in a certain way when judging vignettes). As examples, tied or reversed responses to vignette questions have been observed, either due to poorly written or translated vignettes, or due to respondents’ lack of attention or ability to understand the vignettes (e.g., Kyllonen & Bertling, 2014; Möttus, Allik, Realo, Pullman, et al., 2012; Primi, Zanon, Santos, De Fruyt, & John, 2016).

In this study, it is not necessarily assumed that vignette responses (Z_i) are strictly ordered, but why the nonparametric method is not uniquely defined when ties or ranking reversals are observed is illustrated. The following tables present examples of how vignette-anchored variables are transformed (a) when vignette responses are strictly ordered (Table 1), (b) when vignette responses include ties (Table 2), and (c) when vignette responses include reversals as opposed to the intended order (Table 3). In each table, the first column shows the original number-correct scores on self-assessments (X_i), and the remaining columns show the transformed categorical variables (Y_i) resulting from different vignette response sets. The last two rows show the mean and standard deviation of the transformed vignette-anchored variables Y_i

Table 2. Transformed Score Examples With Tied Vignette Responses.

x_i	y_i depending on (z_1, z_2, z_3)					
	(3,3,3)	(1,1,4)	(3,3,5)	(1,1,1)	(5,5,5)	(2,2,3)
1	1	2/4	1	2/4/6	1	1
2	1	5	1	7	1	2/4
3	2/4/6	5	2/4	7	1	6
4	7	6	5	7	1	7
5	7	7	6	7	2/4/6	7
M	3.60-4.40	5.00-5.40	3.00-3.40	6.00-6.80	1.20-2.00	4.60-5.00
SD	3.00-3.13	1.14-1.87	2.30-2.35	0.45-2.24	0.45-2.24	2.55-2.88

Table 3. Transformed Score Examples With Reversed Vignette Responses.

x_i	y_i depending on (z_1, z_2, z_3)					
	(5,3,1)	(3,2,1)	(1,4,2)	(4,5,4)	(1,5,1)	(2,5,3)
1	1/6	1/6	2	1	2/6	1
2	1/7	1/4/7	3/6	1	3/7	2
3	1/4/7	2/7	3/7	1	3/7	3/6
4	1/7	7	4/7	2/6	3/7	3/7
5	2/7	7	7	4/7	4/7	4/7
M	1.20-6.80	3.60-6.80	3.80-5.80	1.80-3.20	3.00-6.80	2.60-4.60
SD	0.45-1.30	0.45-3.13	1.92-2.17	1.30-3.03	0.45-0.71	1.14-2.88

under a uniform distribution of the untransformed variables X_i (i.e., $P(x = 1) = P(x = 2) = \dots = P(x = H)$).

When vignette responses are strictly ordered as intended (Table 1), original responses on self-assessments (x_i) are adjusted to be higher when vignette responses (z_j) tend to be low while they are adjusted to be lower when vignette responses tend to be high. If vignette responses include extreme score categories such as 1 or 5, the transformed vignette-anchored scores (y_i) cover more central scores (2,...,6) excluding extreme categories (1,7). If vignette responses are more centered but exclude extremes, vignette scores are adjusted to cover extreme categories. These are intended effects described by King and colleagues (2004, 2007).

While there is a well-defined mapping from X to Y when vignette responses are strictly ordered, that is no longer the case when strict ordering is violated, such as when partial or full ties or reversals are observed. As shown in Table 2, in extreme cases where all three vignettes are not successfully differentiated but receive the same score, the original five different values (x_i) are reduced to only a couple of values among the possible seven values. Moreover, according to the rule, multiple values can be possibly assigned for cases where $x_i = z_j$, instead of just a single value. This is true because the definition of the vignette scoring rule becomes ambiguous as soon as vignette scores are not strictly ordered. There is no longer a unique rule for transforming original variables (x_i) to new variables (y_i)—for example, when $(z_1 = 3, z_2 = 3, z_3 = 3)$, $x_i = 3$ could be transformed into 2 ($x_i = 3 = z_1$) or 4 ($x_i = 3 = z_2$) or 6 ($x_i = 3 = z_3$) according to the transformation rule. In turn, single values for mean and standard deviation cannot be computed, but the ranges are provided instead in the two bottom lines in Tables 2 and 3.

If vignette responses include partial or full reversals in rankings, the situation becomes more ambiguous. With ties, there is no unique one-to-one relationship when $x_i = z_j$. However, when reversals in ratings are observed, ambiguity occurs for most x_i values. For example, in an extreme case, if $z_1 = 5$, $z_2 = 3$, and $z_3 = 1$, then $y_i = 3$ or $y_i = 5$ can never be observed because there is no integer between the two consecutive vignette responses according to the rule ($5 < X_i < 3$ or $3 < X_i < 1$). Instead, $y_i = 1$ for $x_i = 1, 2, 3, 4$ is possible because it meets the rule $X_i < Z_1$. Having more ambiguous sets of transformations leads to a much wider range in means and standard deviations compared with the sets with ties.

Parametric Method

As a complement to the nonparametric method described above, a parametric statistical model was developed, called the *chopit* model (King et al., 2004; Rabe-Hesketh & Skrondal, 2002). The chopit model is a compound-hierarchical-ordered probit model, which is a generalization of the ordered probit model. Modeling RS is achieved via threshold variation, with the vignettes providing the information underlying this variability.

There are two components in the chopit model: a model for self-assessment questions and a model for vignettes. First, for the self-assessment questions, the response X_{pi} on item i for person p is modeled as an ordinal probit model with the underlying latent trait θ_p and item specific latent responses x_{pi}^* :

$$\theta_p = \mathbf{W}_p' \boldsymbol{\beta} + \varepsilon_p.$$

The latent response of person p to item i is modeled as

$$x_{pi}^* \sim N(\theta_p, 1).$$

The $\mathbf{W}_p$ are covariates, and $\boldsymbol{\beta}$ are the fixed effects, and ε_p is a residual error term that follows a normal distribution $N(0, \omega^2)$. For the observed response categories $h = 1, \dots, H$, the latent trait is discretized via a threshold model with person-specific thresholds τ_p^h , so that

$$x_{pi} = h \text{ if } \tau_p^{h-1} \leq x_{pi}^* < \tau_p^h,$$

with strictly ordered thresholds $-\infty = \tau_p^0 < \tau_p^1 < \dots < \tau_p^H = \infty$. These ordered thresholds are modeled as

$$\tau_p^1 = \mathbf{V}_p \lambda_1,$$

$$\text{and } \tau_p^h = \tau_p^{h-1} + \exp(\lambda_h \mathbf{V}_p), \quad h = 2, \dots, H,$$

where $\mathbf{V}_p$ are covariates and λ_h are the parameters. The underlying latent trait θ_p can be viewed as the true perceived construct of respondent p on a scale that is comparable across individual respondents. The observed responses result from different individuals having comparable latent responses x_{pi}^* but by applying different person-specific thresholds τ_p^h . Hence, the observed responses x_{pi} are therefore no longer comparable, unless the person-specific thresholds are known.

Second, to extract information about the thresholds for self-assessment questions using the vignettes, it must be assumed that there is a true latent construct η_j associated with the hypothetical person (not the respondents!) described in the j th vignette, $j = 1, \dots, J$. Again, the perception of the survey respondents differs from η_j only by a random error term:

$$z_{pj}^* = \eta_j + u_{pj}, \quad u_{pj} \sim N(0, \sigma^2).$$

It is further assumed that the observed vignette responses are generated by applying the same person-specific ordered thresholds τ_p^h as for the self-assessment questions; that is, $z_{pj} = h$ if $\tau_p^{h-1} \leq z_{pj}^* < \tau_p^h$, and that person-specific thresholds for both self-assessments and vignettes are the same and predicted by the same variables V_p .

An assumption is made that the latent response to the hypothetical persons described in the vignettes is based only on the “true” trait levels η_j , and that only the ordered thresholds are person specific. Indeed, this shows that the parametric vignette scoring model assumes $J + 1$ person parameters, one person trait level θ_p measured by the K self-assessment items, and J person-specific thresholds τ_p^h measured by the J vignette items. Hence, the probability of an observed response x depends on $J + 1$ person parameters, that is

$$P(X_i = x|p) = f\left(x|\theta_p, \tau_p^1, \dots, \tau_p^H\right).$$

This shows that the parametric vignette scoring model is a multidimensional item response model, so that the structure of dependencies between the observed variables can be the result of differences in trait parameters, or, person-specific vignette thresholds, or both. Vonkova and Hulleger (2011) showed that adjustments using the parametric vignette scoring model do not always lead to the intended effects and often depend on the vignettes used.

In this study, the authors do not make use of the parametric version model but derive analytical results on the nonparametric method, which was also used in the BQ analysis of PISA 2012 and included in the public use data files.

Distribution of Vignette-Anchored Variables

Results on the distribution of the transformed vignette-anchored variables (Y_i) have been shown regarding (a) the expected value, (b) the variance, and (c) the covariance of two vignette-anchored variables. Recall that anchoring vignette is based on a mapping from original raw responses $\mathbf{X} = X_1, \dots, X_K \in \{1, \dots, H\}$, conditional on vignette responses $\mathbf{Z} = Z_1, \dots, Z_J \in \{1, \dots, H\}$ onto a new set of vignette-anchored variables $\mathbf{Y} = Y_1, \dots, Y_K \in \{1, \dots, 2J + 1\}$. Note that the transformation is applied in the same way to each original response variable $X_1, \dots, X_K$ based on the full set of vignette responses $Z_1, \dots, Z_J$. Therefore, the vignette-anchored variable Y_i is denoted as $Y_i(X_i|\mathbf{Z}) = f(X_i; Z_1, \dots, Z_J) = f(X_i; \mathbf{Z})$.

The Expected Value of a Vignette-Anchored Variable

The expected value of Y_i can be calculated using

$$E(Y_i) = E(E(f(X_i; \mathbf{Z})|\mathbf{Z})) = \sum_{\mathbf{z} = (z_1, \dots, z_J)} P(\mathbf{z})E(Y_i|\mathbf{z}),$$

which requires summation over the conditional distribution of all possible (or at least all permissible) vignette response vectors $z_1, \dots, z_J$. The conditional expectation is given by

$$E(Y_i|\mathbf{z}) = \sum_{x=1}^H P_i(x|\mathbf{z})Y_i(x|\mathbf{z}),$$

with $P_i(x|\mathbf{z}) = P(X_i = x|z_1, \dots, z_J)$ and $Y_i(x|\mathbf{z}) = f(x; z_1, \dots, z_J)$. This provides an analytic expression of the expected value of a vignette-anchored variable Y_i , given the observed variable $X_{i,s}$ distribution in levels of $\mathbf{Z} = \mathbf{z}$.

The Variance of a Vignette-Anchored Variable

The law of total variance enables us to perform a decomposition of the total variance of a vignette-anchored variable into two parts. This yields,

$$V(Y_i) = \sum_z P(z) V(Y_i|z) + V(E(Y_i|z)),$$

with the conditional variance, given z in the first part

$$V(Y_i|z) = \sum_{x=1}^H P_i(x|z) [Y_i(x|z)]^2 - \left[\sum_{x=1}^H P_i(x|z) Y_i(x|z) \right]^2,$$

and the variance of conditional expectations in the second part is as follows:

$$V(E(Y_i|z)) = \sum_z P(z) [E(Y_i|z)]^2 - \left[\sum_z P(z) E(Y_i|z) \right]^2.$$

The Covariance of Two Vignette-Anchored Variables

The covariance of two vignette-anchored variables, Y_i and Y_k , can be calculated as follows: First using the standard decomposition,

$$\text{COV}(Y_i, Y_k) = E_z(\text{cov}(Y_i, Y_k|z)) + \text{COV}_z(E(Y_i|z), E(Y_k|z)),$$

with

$$E_z(\text{COV}(Y_i, Y_k|z)) = \sum_z p(z) E(Y_i Y_k|z) - \sum_z p(z) E(Y_i|z) E(Y_k|z),$$

and per definition, it is given as

$$\text{COV}_z(E(Y_i|z), E(Y_k|z)) = \sum_z p(z) E(Y_i|z) E(Y_k|z) - \sum_z p(z) E(Y_i|z) \sum_z p(z) E(Y_k|z).$$

For the special case that original response variables X_i , X_k are independent and identically distributed (i.i.d.) random variables ($X_i \perp X_k$), it follows that

$$Y_i \perp Y_k | z \text{ as } Y_i(X_i|z) \perp Y_k(X_k|z)(x)|z.$$

Therefore, for any given vignette response set z , the following is obtained:

$$\text{COV}(X_i, X_k) = \text{COV}(Y_i, Y_k|z) = 0,$$

which makes the first part $E_z(\text{COV}(Y_i, Y_k|z)) = 0$. However, for the second part, the covariance of conditional expected values of vignette-anchored variables, the following is found:

$$\text{COV}_z(E(Y_i|z), E(Y_k|z)) = \text{COV}(E(Y_i|z), E(Y_i|z)) = V(E(Y_i|z)) \geq 0,$$

because $Y_i|z$ and $Y_k|z$ are i.i.d. for any given z , and $E(Y_k|z) = E(Y_i|z)$ is found for all levels of z . In turn, the following is obtained:

$$\text{COV}(Y_i, Y_k) = 0 + V(E(Y_i|z)) \geq 0.$$

This is a central result, which is not implied by the fact that the vignette scoring approach is monotonic in X_i , but follows from the specific form of the vignette transformation: If the original response variables (X_i, X_k) are independent and identically distributed random variables, then their covariance is 0, $\text{COV}(X_i, X_k) = 0$, but after transformation the covariance of the vignette-anchored variables $\text{COV}(Y_i, Y_k) > 0$ if the expected value of the vignette-anchored variables vary across respondents. As illustrated in Table 1, the vignette-anchored variables range from 1 to 7, as should be expected if vignettes are supposed to adjust for the original response variables. However, because the transformation applies at the respondent level and in the same way across all the original responses, vignette scoring introduces positive covariance even if the original response variables are completely random.

Effects on Reliability and Validity

The effects of the nonparametric anchoring vignette approach on estimates of reliability are presented via the well-known Cronbach's α (Cronbach, 1951) and the validity presented via convergent-discriminant validity (Campbell & Fiske, 1959). In the present study's context, with original response variables, $X_1, \dots, X_K$ and the sum score, $S = \sum_{i=1}^K X_i$, Cronbach's α is defined as

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{X_i}^2}{\sigma_S^2} \right).$$

Note that σ_S^2 denotes the variance of the observed total score on the self-assessment, and $\sigma_{X_i}^2$ denotes the variance of responses specific to item i . If the original response variables are uncorrelated, then $\alpha = 0$ because $\sum_{i=1}^K \sigma_{X_i}^2 = \sigma_S^2$. If original response variables are correlated, then the following is obtained:

$$\sigma_S^2 = \sum_{i=1}^K \sigma_{X_i}^2 + 2 \sum_{i < l} \text{cov}(X_i, X_l) > \sum_{i=1}^K \sigma_{X_i}^2,$$

and hence $\alpha > 0$. It was shown above that the transformation of anchoring vignettes introduces an increase in the covariance of vignette-anchored variables, contrary to the originally uncorrelated variables. Consequently, the reliability estimate obtained using Cronbach's α will also be increased, as all covariance terms of the vignette-anchored variables become positive.

The authors of the present study looked at the empirical results obtained using the PISA 2012 BQ data in which both original and vignette scored scales are available. In addition, a simulation study is provided in the online supplemental material. The simulation showed, as expected based on the mathematical derivations given above, that Cronbach's alpha and correlations increased in all simulation conditions after transformation. For the empirical example, results are presented in the next section that compare reliability and validity based on both original and vignette-anchored transformed variables.

Empirical Example

PISA 2012 used the anchoring vignette approach in the BQ to address issues of an apparent lack of cross-cultural comparability of responses in the educational survey context (OECD, 2014). To correct for RS found in previous assessment cycles (e.g., Buckley, 2009), two sets of vignettes were included in the PISA 2012 cycle using a 4-point Likert-type scale (*strongly agree*,

agree, disagree, strongly disagree) prior to self-assessment questions in the BQ using the same response format. Each vignette described behaviors of a fictitious mathematics teacher that were indicative of low, medium, and high levels of teacher support and classroom management. The first scale, "Teacher Support," consisted of four self-assessment questions, and the corresponding three vignettes described three teachers in terms of the frequency of setting and returning homework. The second scale, "Classroom Management," also consisted of four self-assessment questions, and the corresponding three vignettes described three teachers of different levels of punctuality for lessons and student behavior in class. Depending on students' rating standards and their interpretation of the four levels of the response format (*strongly agree, agree, disagree, strongly disagree*), students could place their perceptions on three vignette questions. Based on the "vignette equivalence" assumption, the actual levels for the fictitious teachers described in the vignettes were viewed as objective and invariant across respondents, and the reason for differences in vignette responses across respondents was attributed to individual RS. With the three vignette responses (Z) per scale, number-correct scores on the 4-point Likert-type scale (X) were adjusted to vignette-anchored scores on a 7-point scale (Y). The relationship between the original score and the vignette-anchored score follows the rule exemplified in the section on correcting RSs using vignettes.

Contrary to the requirements of the vignette approach, ties and reversals in vignette responses were observed quite frequently in the PISA 2012 BQ data. Across all participating countries, more than 25% of the vignette responses did not follow the intended strict ordering—for some countries even up to 50%.³ For cases with ties or reversals, the vignette responses (Z) were rescaled (refer to OECD, 2014, for details), and original variables (X) were transformed into vignette-anchored variables (Y) based on the rescaled vignette responses. More specifically, for ties, a "lower bound scoring" approach was chosen, where scores were adjusted based on the lower level. For example, if a student evaluated the fictitious teachers' classroom management with low and medium levels identical (e.g., $Z = \{3,3,5\}$), scores were adjusted based on the lower level (e.g., for $X = \{1,2,3,4,5\}$, the transform was $Y = \{1,1,2 \text{ (instead of } 4),5,6\}$). If ordering violations were present, they were reclassified into ties (Kyllonen & Bertling, 2014). That is, if a student rated the highest vignette lower than the medium vignette, responses for this student would be rescaled in a way that the ratings for the medium and high vignettes were tied (e.g., if $Z = \{1,4,2\}$, then use $Z' = \{1,4,4\}$ instead). For this case, ties were created at the highest response category chosen by the student. Interestingly, for the PISA 2012 BQ data, two sets of vignettes were used to adjust not only the associated scales but also all 12 scales in the BQ. This use of the vignettes method is based on an additional assumption that individual RSs were invariant across different contexts and scale contents whenever the same response format was used (Kyllonen & Bertling, 2014).

On the basis of Cronbach's alpha and convergent-discriminant validity, the correlation matrix of the 12 scales used in the PISA 2012 BQ data is computed. Figure 1 illustrates how the correlations of sum scores for each scale change after applying the anchoring vignette approach. As shown in Figure 1, it is apparent that all correlations increased after the transformation. Among the correlations of original responses (X), the highest correlations before transformation were $r = .64$, and were observed between "Instrumental Motivation for Math (V6)" and "Math Interest (V7)" and between "Math Self-Concept (V10)" and "Math Interest (V7)." These correlations increased to .77 and .79, respectively, after the transformation. The lowest correlation of original responses (X) was $r = .16$ and was observed between "Math Teacher's Classroom Management (V4)" and "Math Self-Concept (V10)," which became .61 after the transformation. However, in terms of vignette-anchored variables (Y), .61 was not the lowest correlation any more. After transformation, the lowest correlation was $r = .55$, observed between "Math Interest (V7)" and "Attitude towards School: Learning Activities (V2)." Regarding this change, one should note that V4 is one of the scales that was used for designing the vignettes.

Figure 1. Correlation matrix for PISA 2012 BQ scales, left for original variables, right for vignette-anchored variables.

Note. PISA = Program for International Student Assessment; BQ = background questionnaire.

The anchoring vignette approach systematically increased all correlations, and this may result in a changed interpretation of the relationships between the scales. The derivations above and the simulation study results in the online appendix imply that these increases are not the result of an intended effect, but rather an unintended one that moves all correlations in the same direction independent of the initial level. An alternative interpretation would be that the vignette scoring brings out the “true correlations” between items and scales, and provides the expected effect. However, the results obtained from the mathematical derivations and simulations indicate that increased correlations are seen in all cases, even in those where respondents provide completely random responses (e.g., they would either choose a response category randomly or roll dice to select a response) to both self-assessment and vignette items.

Summary and Conclusion

This article examined the underlying assumptions and the statistical effects of the nonparametric “anchoring vignettes” approach proposed by King and colleagues (2004) on the reliability and the validity of rating data. A mathematical derivation of the mean, variance, and covariance of vignette-anchored scores was combined with a simulation study and an illustration based on analyses of the BQ data from PISA 2012. Statistically, the nonparametric adjustment can be understood as respondent-dependent rescaling of the original scale ranging from 1 to H into a scale that ranges from 1 to $2J + 1$. These adjustments can either stretch or compress, or translocate the original responses in certain ways. In this process, it is assumed that RSs are effectively removed from the original responses on the self-assessments (X) based on the vignette scores (Z). A major assumption made is that the intended location and strict ordering of vignettes are understood by every respondent in the same way.

Several empirical studies revealed that this is not always true: On average, 25% of respondents in the PISA (Kyllonen & Bertling, 2014), 18% to 32% of respondents on the Big Five (Primi et al., 2016), and 8% to 35% of respondents in a cross-national study of conscientiousness (Möttus et al., 2012) did not provide strictly ordered vignette scores. Moreover, the adjustment is conducted at the respondent level and applied to all original response variables in the same way, which induces effects on reliability and validity estimates by increasing correlations among transformed variables. Among them are the following:

1. For uncorrelated and randomly chosen original responses (X), the resulting vignette-anchored variables (Y) have positive intercorrelations which results in a remarkable increase of the reliability estimate via Cronbach's α and may increase correlations in terms of indicators of convergent validity, for example, if the same vignettes are applied across scales.
2. For positively correlated original responses (X), which are expected from self-assessment items measuring the same latent trait, the simulation and empirical example showed that the correlation among vignette-anchored variables (Y) is always increased as is Cronbach's alpha, even if vignette scores are randomly chosen, or have ties or reversals.

These effects can lead to a more positive evaluation of the scale quality than warranted. The reported increase appears to be exacerbated if the ordering assumption of vignette scores is not met; for example, if ties occur or if reversals of the ordering of vignettes are observed due to higher levels of random or systematic response errors.

As King et al. (2004) noted, it can be as difficult to develop high-quality vignette questions as it is to develop high-quality self-assessment questions. In summary, no matter whether vignette responses are strictly ordered (as they are supposed to be) or not, correlations between scales always increase. In addition to the increase in reliability and changes to validity indicators, due to the homogenization of the covariance structure, secondary analyses that use vignette-anchored variables would obtain different factor structures (Stankov et al., 2017) or regression results from using the original variables.

One could argue that “corrected” or debiased validity and reliability coefficients should be reported for transformed scales. One way to do that would be to report validity coefficients that are weighted sums of coefficients obtained in groups of test takers who gave the same vignette responses (Z). While this is a viable way to obtain an approximate correction, it does not suffice. Any secondary analysis, for example, regression analyses or other forms of prediction models would need to consider that the “corrected” vignette-anchored scores are best compared (to avoid increased coefficients) in only homogeneous groups of test takers who gave the same vignette responses. This would call into question the utility of the anchoring vignette approach in the first place, as the correction of validity and reliability coefficients would essentially require rolling back the anchoring vignette approach and looking only at homogeneous groups, and then averaging over groups who were subjected to the same score transformation. Thus, researchers who intend to utilize the anchoring vignette approach or vignette-anchored variables should be aware of these effects, and should interpret the reliability and validity results of vignette score scales carefully.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

Notes

1. In their studies, they usually label Response styles (RSs) as differential item functioning (DIF).
2. Thanks to reviews received, the authors were able to move the simulation results to the online appendix. The mathematical study of vignette scoring in this article provides sufficient evidence to support the authors' interpretation of the empirical results obtained from examining the effects of Program for International Student Assessment (PISA) 2012 vignette scoring.
3. These high levels of response inconsistency were considered in the simulation design to mirror conditions found in real data.

References

- Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143-156.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17, 665-678.
- Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, 19, 528-541.
- Brady, H. E. (1985). The perils of survey research: Inter-personally incomparable responses. *Political Methodology*, 11, 269-291.
- Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. In *NCES Conference on the Program for International Student Assessment: What we can learn from PISA*. Retrieved from https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 6, 170-175.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- de Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45, 104-115.
- Dolnicar, S., & Grün, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education*. Retrieved from <http://jmd.sagepub.com/content/early/2009/04/22/0273475309335267.abstract>
- Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, 56, 328-351.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: John Wiley.
- Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, 74, 201-222.
- Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20, 296-309.
- Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, 35, 92-114.
- Jones, E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behaviors*. New York, NY: General Learning Press.
- Kelley, H. H. (1973). The process of causal attribution. *American Psychologist*, 28, 107-128. doi: 10.1037/h0034225
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49, 161-177.
- 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. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In Rutkowski L., von Davier M., Rutkowski D. (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277-286). Boca Raton, FL: CRC Press.
- Larson, J. R. (1977). Evidence for a self-serving bias in the attribution of causality. *Journal of Personality*, 45, 430-441. doi:10.1111/j.1467-6494.1977.tb00162.x
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*. Retrieved from <http://psycnet.apa.org/psycinfo/1933-01885-001>
- Möttus, 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.
- Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Technical Report*. Retrieved from <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>
- Paulhus, D. L. (1991). *Measurement and control of response bias*. Retrieved from <http://doi.apa.org/psycinfo/1991-97206-001>
- Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes. *European Journal of Psychological Assessment*, 32, 39-51.
- Rabe-Hesketh, S., & Skrondal, A. (2002). *Estimating chopit models in gllamm: Political efficacy example from King et al. (2002)*. Retrieved from <http://www.gllamm.org/chopit.pdf>
- Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Munster, Germany: Waxmann.
- Sen, A. (2002). Health: Perception versus observation: Self reported morbidity has severe limitations and can be extremely misleading. *British Medical Journal*, 324, 860-861.
- Stankov, L., Lee, J., & von Davier, M. (2017). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment*. Retrieved from <https://doi.org/10.1177/0734282917702270>
- von Davier, M. (2010). Why sum scores may not tell us all about test takers. *Newborn and Infant Nursing Reviews*, 10, 27-36. doi:10.1053/j.nainr.2009.12.011
- von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th Annual Psychometric Society Meeting* (pp. 463-488). New York, NY: Springer.
- von Davier, M., Naemi, B., & Roberts, R. D. (2012). Factorial versus typological models: A comparison of methods for personality data. *Measurement: Interdisciplinary Research and Perspectives*, 10, 185-208. doi:10.1080/15366367.2012.732798
- Vonkova, H., & Hulleger, P. (2011). Is the anchoring vignettes method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society: Series A*, 174, 597-620. doi:10.1111/j.1467-985X.2011.00704.x