Journal of Cross-Cultural Psychology
2017, Vol. 48(3) 319–334
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0022022116687395
journals.sagepub.com/home/jcc
SAGE

Effects of Anchoring Vignettes on Comparability and Predictive Validity of Student Self-Reports in 64 Cultures

Jia He[1,2], Janine Buchholz[2], and Eckhard Klieme[2]

Abstract

Anchoring vignettes are item batteries especially designed for correcting responses that might be affected by incomparability. This article investigates the effects of anchoring vignettes on the validity of student self-report data in 64 cultures. Using secondary data analysis from the 2012 Programme for International Student Assessment (PISA), we checked the validity of ratings on vignette questions, and investigated how rescaled item responses of two student scales, *Teacher Support* and *Classroom Management*, enhanced comparability and predictive validity. The main findings include that (a) responses to vignette questions represent valid individual and cultural differences; in particular, violations in these responses (i.e., misorderings) are related to low socioeconomic status and low cognitive sophistication; (b) the rescaled responses tend to show higher levels of comparability; and (c) the associations of rescaled Teacher Support and Classroom Management with math achievement, Student-Oriented Instruction, and Teacher-Directed Instruction are slightly different from raw scores of the two target constructs, and the associations with rescaled scores seem to be more in line with the literature. Namely, the associations among all self-report Likert-type scales are weaker with rescaled scores, presumably reducing common method variance, and both rescaled scale scores are more positively related to math achievement. The country ranking also changes substantially; in particular, Asian cultures top the ranking on Teacher Support after rescaling. However, anchoring vignettes are not a cure-all in solving measurement bias in cross-cultural surveys; we discuss the technicality and directions for further research on this technique.

Keywords

anchoring vignettes, measurement invariance, predictive validity, PISA

Validity, the degree to which evidence and theory support the inferences drawn from assessment data, is the most fundamental consideration in educational and psychological assessments (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In large-scale international assessments such as the Programme for International Student Assessment (PISA), concerns rising from cross-cultural

[1] Tilburg University, The Netherlands
[2] German Institute for International Educational Research, Frankfurt, Germany

Corresponding Author:
Jia He, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.
Email: j.he2@tilburguniversity.nl

comparability and predictive validity in the cognitive assessment (e.g., Kankaraš & Moors, 2014) and self-report data (e.g., van de Gaer, Grisay, Schulz, & Gebhardt, 2012) may have prevented full-fledged explorations of these data for basic research and evidence-based policy making (Goldstein, 2004; Gorur, 2014). Innovative designs of item formats and sophisticated psychometric methods promise to ameliorate such concerns in self-reports (Kyllonen & Bertling, 2014). This article investigates the effects of one particular design element, namely, anchoring vignettes, with student self-report data from 64 cultures in PISA. Specifically, we check (a) the validity of responses on vignette questions, (b) effects of anchoring vignettes on the cross-cultural comparability of two student scales: Teacher Support (TS) and Classroom Management (CM), and (c) their effects on the predictive validity of these two scales in relation to teaching strategies and student achievement. In the following, we first introduce how anchoring vignettes are designed to enhance validity in cross-cultural assessments, and then we review the anchoring vignettes applied in PISA (Organisation for Economic Co-Operation and Development [OECD], 2013).

The Technique of Anchoring Vignettes

The technique of anchoring vignettes was introduced to correct for measurement bias—in particular, the so-called reference-group effect—in Likert-type responses (King, Murray, Salomon, & Tandon, 2004). The reference-group effect stems from different standards (i.e., a reference group) that respondents use to evaluate themselves and their own behaviors, and may be related to various response styles, which refer to the tendency to use certain response options on some basis other than the target construct (Paulhus, 1991). These response biases, triggered by personality, idiosyncratic interpretation, and judgment of questions, or variations in survey contexts shift the alignment between the reported level of a trait and the true underlying trait. Thus, it is important to gauge their effects and better estimate the actual trait level of target constructs. Anchoring vignettes provide a common reference point for respondents using different response styles. It asks each respondent several additional vignette questions. Vignettes are descriptions of hypothetical persons with different levels of the target trait, and respondents rate the trait level of these hypothetical persons on the same response format as the self-assessment. The systematic differences in responses to the same vignette questions are supposed to reflect mainly differences in response styles, whereas responses on self-assessment are a combination of response style distortion and the true trait level. Subsequently, the measurement bias due to response styles from the self-assessment can be subtracted to yield a response-style-free estimate of the actual level of the target trait. There are two working assumptions of this approach: response consistency (i.e., participants use the same mechanisms to give responses to self-assessment questions and the vignette questions) and vignette equivalence (i.e., the vignettes are understood by all respondents in the same way).

Both nonparametric and parametric estimation strategies can facilitate data analysis with anchoring vignettes. The nonparametric approach basically rescales self-assessment responses (denoted as y) on the basis of responses of a total number of J ordered vignette questions (denoted as z_1 to z_j) to a single variable C (Equation 1; King & Wand, 2007). It is likely to encounter tied or inconsistently ordered vignette responses (e.g., $z_1 = z_2 = y$, or $z_2 < y = z_1$). In these cases, the self-assessment responses can take a vector of possible values instead of one scalar value. For instance, if the comparisons of self-assessment y with two vignettes z_1 (lower trait level) and z_2 (higher trait level) shows a pattern of $z_2 < y = z_1$, C may take any of the values from 2 to 5.

$$C = \begin{cases} 1 & if\ y < z_1 \\ 2 & if\ y = z_1 \\ 3 & if\ z_1 < y < z_2\,. \\ \vdots & \\ 2J+1 & if\ y > z_j \end{cases} \qquad (1)$$

Table 1. Anchoring Vignettes in the PISA 2012 Student Background Questionnaire (OECD, 2013).

Vignettes based on teacher support behaviors: How much do you agree with the statement "Mr./Ms (name) is concerned about students' learning"	
High level (ST82Q01)	Ms. (name) sets mathematics homework every other day. She always gets the answers back to students before examinations.
Medium level (ST82Q02)	Mr. (name) sets mathematics homework once a week. He always gets the answers back to students before examinations.
Low level (ST82Q03)	Ms. (name) sets mathematics homework once a week. She never gets the answers back to students before examinations.
Vignettes based on classroom management behaviors: How much do you agree with the statement "Mr./Ms. (name) is in control of his or her classroom"	
High level (ST84Q02)	The students in Ms. (name's) class are calm and orderly. She always arrives on time to class.
Medium level (ST84Q01)	The students in Ms. (name's) class frequently interrupt her lessons. She always arrives 5 min early to class.
Low level (ST84Q03)	The students in Mr. (name's) class frequently interrupt his lessons. As a result, he often arrives 5 min late to class.

Note. PISA = Programme for International Student Assessment; OECD = Organisation for Economic Co-Operation and Development.

The parametric approach is more sophisticated, especially in dealing with variable *C* as a range instead of a scalar value. It uses a generalization of the ordered probit model to distribute vector-valued responses according to the proportion of "similar" respondents who choose the categories spanned by the vector. As this approach is not central in the current study, further explanation can be found in King and colleagues (Hopkins & King, 2010; King et al., 2004; King & Wand, 2007).

Anchoring Vignettes in PISA

In PISA, using anchoring vignettes to rescale students' self-report scales of motivation and evaluation of teaching is suggested to have enhanced predictive validity of self-report scales (OECD, 2013). The most compelling evidence comes from how anchoring vignettes solve the paradoxical correlation of contextual factors and math achievement based on the 2012 field trial data (200 to 1,000 students conveniently sampled in each culture; Kyllonen & Bertling, 2014). According to the authors, when raw scale scores were used, contextual factors such as student self-efficacy and math motivation showed weak, positive correlations with math achievement at individual level (as theory predicts); yet, aggregated at culture level, the correlations were negative. Once scale scores were adjusted based on anchoring vignettes, the individual-level correlations became stronger, and the culture-level correlations were reversed to be positive as well.

In the PISA 2012 main study, two sets of vignettes questions were asked, targeting two student self-report scales: TS and CM. Each set had three vignette questions on high, medium, and low levels of traits, respectively (Table 1; OECD, 2013). However, before reaffirming the conclusion that anchoring vignettes improve the validity and interpretability of results, it is necessary to demonstrate whether responses to vignette questions represent valid individual and cultural differences, whether rescaling of self-assessment enhances levels of measurement equivalence in all cultures involved, and whether rescaled scale scores are more reasonably related to validity measures other than only math achievement. Expectations on these three aspects are detailed below.

Validity of Responses to Vignettes Questions in PISA

Respondents are expected to rate the vignettes logically according to the trait level described. Tied ratings and misorderings on vignettes intended for different trait levels are challenges for the validity of anchoring vignettes. In PISA, the vignettes were designed to ensure unidimensionality and considerable discriminative power in trait levels (OECD, 2013). Still, there may be caveats that these vignettes are not perfect; so, misorderings may be attributed to poor design of vignettes. Tied ratings could also be due to extremely high or low standards in respondents' perception so that the differences in vignettes cannot be detected. For example, a respondent considers running 2 km a day is already a strong indication of being active in sports, and if the vignettes describe persons running 5, 10, and 15 km, respectively. The low standard on this issue held by this respondent would result in a rating of *strongly agree* with being active in sports on all three vignettes. Therefore, tied ratings may be acceptable. Inconsistent ratings (e.g., rating a person who runs 5 km a day as more active than a person who runs 10 km a day) make it more difficult to interpret the scores, and they may be more due to personal and cultural characteristics that jeopardize the validity. Krosnick and colleagues (Krosnick, 1991; Narayan & Krosnick, 1996) found that respondents with limited socioeconomic resources and low in cognitive sophistication (e.g., literacy or math achievement) tend to satisfice rather than optimize the responses in questions requiring greater cognitive efforts. Inconsistent ratings might therefore result from the application of a satisficing strategy. It is hence expected in respondents low in socioeconomic status and math achievement (Hypothesis 1a).

Cross-cultural variations in responses to vignettes questions are mainly due to measurement bias stemming from cultural characteristics (King et al., 2004). In educational contexts, ample evidence shows that high expectations of achievement have an impact on achievement (e.g., OECD, 2015; Scheerens, 2016). We expect that high standards (i.e., expectation) in teaching and learning and high levels of student achievements reinforce each other, and perceptions of uniform standards share more consensus in high performing cultures than low performing cultures. The cultural strength of standards in TS and CM can be approximated as the mean ratings on vignettes in a culture, where a higher standard is indicated by generally stronger agreement on vignettes of the target construct. It is expected that higher standards in TS and CM are associated with higher student achievement, and the less agreement there is on standards, the lower student achievement at culture level (Hypothesis 1b). Meanwhile, the standard deviations of ratings within cultures are also an indicator of preference of endorsing end points of the scale, which has been consistently found to be related to higher uncertainty avoidance (e.g., He, van de Vijver, Dominguez-Espinosa, & Mui, 2014). The hypothesis that the within-culture variability in rating on vignette questions is positively related to uncertainty avoidance is tested (Hypothesis 1c).

Measurement Invariance Test of PISA Scales

Without both conceptual and statistical demonstration of comparability, comparisons of cross-cultural data are at best ambiguous and at worst erroneous (e.g., Chen, 2008). In PISA, meticulous design and implementation have lent much confidence in the comparability of the scales; yet, formal statistical testing of scale comparability has not been formally reported in previous cycles (OECD, 2013). Some attempts to establish measurement invariance in PISA reported limited comparability (e.g., Täht & Must, 2013). Three main levels of comparability (also called invariance) in scales can be distinguished and statistically tested: (a) Configural invariance means across cultures, the construct is understood as the same. In statistical terms, this level of invariance signals that items in a measure exhibit the same configuration of salient and nonsalient factor loadings across cultures. (b) Metric invariance indicates that items on the construct have the same factor loadings across cultures. With metric invariance, scale score comparisons can be made within cultures (e.g., TS can be compared between males and females within each culture), and the association of variables can be compared across cultures (e.g., correlations between TS and student-oriented teaching can be compared across cultures, if both scales reach metric invariance). (c) Scalar invariance implies that items have the same intercepts (i.e., point of origin) across cultures. Only with scalar invariance can mean scores of scales be validly compared across cultures (van de Vijver & Leung, 1997). The level of invariance (i.e., configural, metric, and scalar invariance) can be tested in hierarchical models using multigroup confirmatory factor analysis. The level of comparability can be inferred from the fit indexes in each model and the comparisons of fit indexes from different models (detailed in the "Results" section).

In large-scale assessment data involving dozens of cultures, scalar invariance is particularly difficult to satisfy. Given cross-cultural variations in response style preferences, measurement invariance tests should take response styles into consideration (Welkenhuysen-Gybels, Billiet, & Cambré, 2003). Rescaling with anchoring vignettes is expected to remove or reduce individual and cultural nuisance in response styles, and increase levels of data comparability (Hypothesis 2). We also explore how country ranking changes as a function of rescaling.

Predictive Validity of TS and CM

Efficient CM and TS are basic dimensions of teaching quality that contribute to better student outcomes (Capella, Aber, & Kim, 2016). The PISA 2012 Questionnaire Framework (Klieme et al., 2013) refers to these constructs. In addition, specific teaching practices were addressed, and two scales covering different instructional approaches were identified, namely, Teacher-Directed (clearly structured) versus Student-Oriented (participatory) Instruction (OECD, 2013). The former one should generally be associated with high levels of perceived teaching quality, while CM should be more difficult, and providing support should be easier in student-oriented settings. The relationship of TS and CM with math achievement, Teacher-Directed Instruction, and Student-Oriented Instructions is tested with the raw scores and the rescaled scores of the two target scales, respectively.

Method

Participants

The PISA student survey in 2012 assessed competencies of 15-year-olds in reading, mathematics, and science (with a focus on mathematics) in 64 cultures. International experts from participating countries built the assessment frameworks and the questionnaire framework, created and adapted items, and carried out extensive pretests to ensure the validity and reliability of measures (OECD, 2013). Students were recruited through a stratified sampling procedure to represent the schools and the 15-year-old student population of each country. Each student took a subset of the cognitive test that lasted 2 hr as well as a context questionnaire afterward. We based our analysis on complete responses on the vignette questions and the two target scales in the main study of PISA 2012.[1] A total of 296,415 students in 64 cultures were included. Sample size in each culture ranged from 176 (Liechtenstein) to 21,627 (Mexico).

Measures

Measures of vignettes. The vignettes questions for TS and CM were asked immediately prior to the self-assessment questions of the two scales. Content of the questions on high, medium, and low levels of the traits are presented in Table 1. Students responded on a 4-point Likert-type scale from 1 (*strongly agree*) to 4 (*strongly disagree*). Note that high scores refer to *less* support or *worse* CM assigned to the teacher described in the respective vignette.

Target scales. TS was measured with four items on the same scale as the vignettes (e.g., "The teacher helps students with their learning"). Values of Cronbach's alpha for this scale ranged from .628 (Liechtenstein) to .883 (Chinese Taipei) with a mean of .775 in the 64 cultures. CM was measured with four items on the same Likert-type scale (e.g., "My teacher gets students to listen to him or her"), and values of Cronbach's alpha ranged from .305 (Japan) to .792 (France) with a mean of .676.

Validity measures. Students' economic, social, and cultural status (ESCS) was a composite index consisting of three subcomponents: highest occupational status of parents, highest educational level of parents (in years of education), and the index of home possessions.

Students' self-report Teacher-Directed Instruction consisted of five items answered on a 4-point scale from 1 (*every lesson*) to 4 (*never or hardly ever*; for example, "The teacher asks questions to check whether we have understood what was taught"). Values of Cronbach's alpha ranged from .588 (Vietnam) to .809 (Jordan) with a mean of .718. Student-Oriented Instruction had four items on the same response scale, and values of Cronbach's alpha ranged from .505 (Czech Republic) to .799 (Jordan) with a mean of .685 (e.g., "The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster").

Students' *math achievement* was measured with different subsets of the cognitive test. In PISA, each student was administered only a subtest of the overall cognitive test to minimize test burden. By systematically varying items across student groups and using item response theory, these cognitive data were then scaled in a Rasch model and student ability was estimated as plausible values. Plausible values are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured. Five plausible values of math achievement for each student were produced and standard analyses with math achievement should be performed on each of the plausible values (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). As usual, high scores represent high levels of math achievement.

Culture-level *Uncertainty Avoidance* was extracted from Hofstede (2009), and the index was available for 54 cultures in PISA. Uncertainty avoidance is defined as a society's tolerance for ambiguity, where people embrace or avert unexpected or unknown events, or away from the status quo (Hofstede, 2001). Despite that these data were collected over a dozen years ago, cultural values are relatively stable in time, and these data should still be relevant.

Results

We describe the results in three parts: the validation of responses to vignette questions, the empirical test on the effects of anchoring vignettes on measurement invariance of target scales, and the predictive validity in multigroup path models.

Validation of Responses to Vignette Questions

The means of ESCS (*z* score standardized) and the five plausible values of *math achievement* were compared between students who had any inconsistent responses (i.e., misordering) and those who did not in *t* tests (Table 2). In both sets of vignettes, students who rated the vignettes inconsistently had lower ESCS and lower math achievement level. Hypothesis 1a was supported. Furthermore, the proportion of students with misorderings in TS vignettes ranged from 11% (Russian Federation) to 63% (Albania) across cultures, and that in CM vignettes ranged from 7% (Shanghai-China) to 35% (Romania).

Table 2. Results of *t* Test and Descriptive Statistics for Responses to Vignette Questions With and Without Violations.

	Groups							
	M	*SD*	*n*	*M*	*SD*	*n*		
Dependent variable	No violation in TS			Violation in TS			*t*	*df*
Standardized ESCS	0.016	1.005	219,733	−0.047	0.983	72,739	−14.781**	126722
Math PV1	481.100	101.531	221,725	454.456	98.552	74,690	−63.414**	131961
	No violation in CM			Violation in CM				
ESCS	0.192	1.001	242,575	−0.093	0.989	49,897	−23.051**	72510
Math PV1	480.777	100.79	245,737	443.31	98.899	50,678	−77.216**	74011

Note. The ESCS indicator is standardized to z scores across all cultures to enhance the comparability. Only mean comparisons of the first PV1 was reported here, the other four PVs showed very similar results. TS = Teacher Support (Anchoring Vignette); ESCS = Economic, Social, and Cultural Status; PV1 = Plausible Value; CM = Classroom Management (Anchoring Vignette).
$^{**}p < .01$.

The culture-level means and standard deviations of responses of the six vignette questions were correlated with culture-level math achievement and uncertainty avoidance (Table 3). As expected, higher standards in TS and CM—that is, vignettes intended to describe low TS/CM teachers being rated with high scores, indicating low levels on the target dimension—were related to higher achievement, whereas the variability in both sets was negatively associated with math achievement and positively with uncertainty avoidance. Hypothesis 1b and 1c were supported.

Enhanced Comparability in Measurement Invariance Tests

Rescaling of data. The rescaling based on vignette questions were carried out in the anchors package in R (Wand & King, 2007). Given the number of ties and misorderings, the discriminatory power of these two sets of vignettes (each with three questions) was moderate: 63% and 72% of students had neither ties nor misorderings in their responses to the TS and CM scale, respectively. When only two vignette questions (the high and the low trait levels) were evaluated, the percentages increased to 74% and 82%, respectively. In cases of any violation,[2] the rescaled responses had a range of possible values, and in the anchors package, the lowest (Cs) and the highest (Ce) possible rating could be produced. As there was no empirical evidence as to which approximate value should be used in such cases, rescaled scores with the lowest and the highest rating were checked, respectively. Moreover, to assess the sensitivity of numbers of vignettes used in the rescaling, the raw responses on the target scales were rescaled based on both three (high, medium, and low trait levels) and two vignette questions (only high and low trait levels), which resulted in two new 7-point scales and two new 5-point scales.

Analytical strategies and evaluation of model fit. A series of multigroup confirmatory factor analysis was performed in Mplus (Muthen & Muthen, 1998-2012) with the raw responses and the four sets of rescaled responses (i.e., the new 7-point with highest rating as proxy in cases of violations, 7-point with lowest rating, the new 5-point with highest rating, and 5-point with lowest rating), respectively. To ensure that each culture contributed equally in the model, student weights were rescaled to have a population of 1,000 in each culture, and these senate weights were used in the

Table 3. Culture-Level Correlations of Responses to Vignette Questions With Math Achievement and Uncertainty Avoidance.

	Mean correlation with math PVs	Uncertainty avoidance
Standards (*M*)		
TS-High	.080	−.069
TS-Medium	.122	.045
TS-Low	.652**	−.351**
CM-High	.035	.021
CM-Median	.460**	.045
CM-Low	.717**	−.426**
Variability (*SD*)		
TS-High	−.257*	.122
TS-Medium	−.381**	.373**
TS-Low	−.615**	.337*
CM-High	−.137	.282*
CM-Median	−.507**	.428**
CM-Low	−.736**	.549**

Note. PV = Plausible Value; TS = Teacher Support; CM = Classroom Management.
*p < .05. **p < .01.

models. Statistically, treating Likert-type scale responses as continuous is too ideal a situation, as normal distribution of these data is an assumption easily violated; therefore, it is expected that with both raw scores and rescaled scores, modeling data as ordered categorical would result in higher levels of comparability than as treating data as continuous (Desa, 2014; Rutkowski & Svetina, 2014). In this study, data were first treated as continuous and then as ordered categorical.

Model fit was evaluated by the Tucker–Lewis index (TLI; acceptable above .90), comparative fit index (CFI; acceptable above .90), and root mean square error of approximation (RMSEA; acceptable below .08; Cheung & Rensvold, 2002). The acceptance of a more restrictive model is based on the change of CFI and RMSEA. In the contexts of large-scale assessment with dozens of cultures, Rutkowski and Svetina (2014) proposed to set the cut point of change of CFI to .02 and that of RMSEA to .03 from configural to metric model, and from metric to scalar model, the changes of both CFI and RMSEA should be within .01. It should be noted that these criteria were based on simulation studies treating data as continuous, whereas the proper cut points for categorical models still await future research.

Results of measurement invariance testing. Table 4 presents the results of the measurement invariance tests. When raw responses were tested, neither TS nor CM demonstrated a good fit in scalar invariance model, with the exception of "TS continuous." This is understandable, given diverse response styles in different cultures and the restrictiveness in equality constraints in all 64 cultures. In the continuous models, the rescaled responses in both scales showed acceptable metric invariance, but no scalar invariance could be achieved. In the categorical models, the rescaled TS scale showed a better fit as scalar invariant compared with the scale of raw scores. In all four sets of rescaled responses, the changes of CFI from a less to a more restrictive model were below .012, although the changes of RMSEA were slightly larger. In the rescaled CM responses, the model fit was better compared with raw responses. In the rescaled 7-point scales and the 5-point scale taking lowest rating in cases of violations, only metric invariance was achieved, whereas in the 5-point scale taking the highest rating in cases of violations, scalar invariance was marginally acceptable (change of RMSEA .017).

Table 4. Results of Measurement Invariance Tests in Target Scales.

	Continuous model							Categorical model						
	χ^2	*df*	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA	χ^2	*df*	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA
Teacher Support														
Raw scores														
Configural	386.303**	128	.999	.996	.021			698.370**	128	1	.999	.031		
Metric	3,959.88**	317	.979	.975	.050	−.020	.029	21,543.760**	317	.984	.981	.120	−.016	.089
Scalar	22,543.700**	506	.873	.903	.097	−.106	.047	3,395.230**	758	.975	.987	.097	−.009	−.023
V7E														
Configural	504.190**	128	.998	.995	.025			2,417.176**	128	1	.999	.062		
Metric	2,755.010**	317	.988	.986	.041	−.010	.016	12,692.570**	317	.998	.998	.092	−.002	.030
Scalar	14,880.700**	506	.932	.948	.078	−.056	.037	81,126.210**	1514	.988	.997	.107	−.010	.015
V7S														
Configural	338.854**	128	.998	.995	.019			1,247.552**	128	1	.999	.043		
Metric	2,118.050**	317	.986	.983	.035	−.012	.016	13,193.140**	317	.996	.995	.094	−.004	.051
Scalar	12,285.600**	506	.906	.929	.071	−.080	.036	48,041.600**	1514	.985	.996	.081	−.011	−.013
V5E														
Configural	390.103**	128	.999	.996	.021			1,721.161**	128	1	.999	.052		
Metric	2,243.140**	317	.991	.989	.036	−.008	.015	1,1982.910**	317	.998	.998	.089	−.002	.037
Scalar	14,006.400**	506	.937	.952	.076	−.054	.040	43,364.690**	1010	.993	.997	.095	−.005	.006
V5S														
Configural	314.895**	128	.999	.996	.018			118.434**	128	1	.999	.042		
Metric	1,859.170**	317	.989	.987	.032	−.010	.014	12,894.120**	317	.996	.995	.093	−.004	.051
Scalar	11,922.200**	506	.918	.938	.07	−.071	.038	3,223.420**	1010	.990	.996	.082	−.006	−.011
Classroom Management														
Raw scores														
Configural	844.199**	128	.995	.986	.035			1,677.430**	128	.999	.996	.051		
Metric	5,604.400**	317	.964	.957	.060	−.031	.025	1,740.300**	317	.984	.981	.108	−.015	.057
Scalar	17,654.500**	506	.885	.912	.086	−.079	.026	33,538.440**	758	.969	.985	.097	−.015	−.011

(continued)

Table 4. (continued)

	Continuous model							Categorical model						
	χ^2	*df*	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA	χ^2	*df*	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA
V7E														
Configural	535.255**	128	.998	.993	.026			#11,975.490**	130	.997	.991	.140		
Metric	2,830.880**	317	.986	.983	.041	−.010	.015	7,548.990**	317	.998	.998	.070	.001	.070
Scalar	11,886.00 **	506	.937	.952	.07	−.049	.029	76,657.330**	1514	.980	.995	.104	−.018	.034
V7S														
Configural	405.013**	128	.997	.991	.022			#5,265.242**	130	.997	.992	.092		
Metric	2,207.790**	317	.98	.976	.036	−.015	.014	11,464.760**	317	.994	.992	.087	−.003	−.005
Scalar	10,695.500**	506	.893	.918	.066	−.058	.030	57,819.070**	1514	.969	.992	.090	−.025	.003
V5E														
Configural	479.515**	128	.998	.994	.024			#9,788.125**	130	.997	.991	.127		
Metric	2,048.160**	317	.99	.988	.034	−.008	.010	7,759.441**	317	.998	.997	.071	.001	−.056
Scalar	10,938.300**	506	.938	.953	.067	−.052	.033	37,091.470**	1010	.988	.995	.088	−.010	.017
V5S														
Configural	405.357**	128	.998	.993	.022			#5,416.041**	130	.997	.991	.094		
Metric	2,274.010**	317	.983	.98	.037	−.015	.015	11,186.950**	317	.994	.992	.086	−.003	−.008
Scalar	11,399.40**	506	.908	.93	.068	−.075	.031	39,082.320**	1010	.977	.991	.090	−.017	.004

Note. In models with #, the variance of the latent factor in Italy and Mexico was set to 1 to avoid nonconvergence. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; V7E = rescaled scores based on three vignette questions, and the highest possible rating taken in cases of violation; V7S = rescaled scores based on three vignette questions, and the lowest possible rating taken in cases of violation; V5E = rescaled scores based on two vignette questions, and the highest possible rating taken in cases of violation; V5S = rescaled scores based on two vignette questions, and the lowest possible rating taken in cases of violation.

**$p < .01$.

Table 5. Top and Bottom Five Cultures Based on Raw and Rescaled Factor Scores.

Construct	Raw	V5E	V7E
TS Lowest	Austria	Serbia	Serbia
	The Netherlands	Romania	Romania
	Luxembourg	Montenegro	Montenegro
	Germany	Peru	Peru
	Liechtenstein	Austria	France
TS Highest	Albania	Shanghai-China	Shanghai-China
	Kazakhstan	Hong Kong-China	Macao-China
	Jordan	Macao-China	Singapore
	Singapore	Singapore	Hong Kong-China
	Indonesia	Chinese Taipei	Kazakhstan
CM Lowest	Korea	Qatar	Indonesia
	The Netherlands	Argentina	Qatar
	Greece	Greece	Argentina
	Poland	Romania	Romania
	Finland	Slovenia	Thailand
CM Highest	Kazakhstan	Kazakhstan	Shanghai-China
	Albania	Shanghai-China	Costa Rica
	Jordan	Hong Kong-China	The United States of America
	Costa Rica	Singapore	The United Kingdom
	Lithuania	Costa Rica	Russian Federation

Note. V5E = rescaled scores based on two vignette questions, and the highest possible rating taken in cases of violation; V7E = rescaled scores based on three vignette questions, and the highest possible rating taken in cases of violation; TS = Teacher Support; CM = Classroom Management.

To sum up, in measurement invariance tests, strong evidence was found that anchoring vignettes improved the levels of comparability (Hypothesis 2). It also seems that using the highest possible rating in cases of violation produces a better model fit, compared with using the lowest possible rating. Thus, the responses taking the highest possible rating were used in the remaining analyses. Factor scores of the two scales were generated in a categorical model with the pooled sample, and the correlations of the raw responses with the 5-point and 7-point for TS were .492, and .520, respectively, that of CM were .565 and .560, respectively. The factor scores were reverse coded; thus, a larger value presented a higher trait level. The country/economy mean scores on TS and CM before and after rescaling are provided in the online appendix. Table 5 presents the top five and bottom five countries/economies on each construct when the raw factor scores, factor scores using rescaled scores based on two and three vignettes, are used, respectively. For both scales, changes are substantial: Overlap between rankings using raw and rescaled scores is no more than 40%, whereas there is some similarity between the two sets of rescaled scores. Another salient pattern is that Asian cultures rank top on TS after rescaling, which suggests that anchoring vignettes are effective in correction for the Asian modesty response bias.

Predictive Validity in Multigroup Path Models

To study the associations between the two classroom-level antecedents (TS and CM), teacher instructional behaviors (*Student-Oriented Instruction, Teacher-Directed Instruction*) and math achievement, multiple group path models using the raw versus rescaled TS and CM factor scores were carried out.[3] In the multigroup path model, the structural regression weights were constrained to be equal across 64 cultures.[4] The model with raw TS and CM scores fit well, χ^2(64,

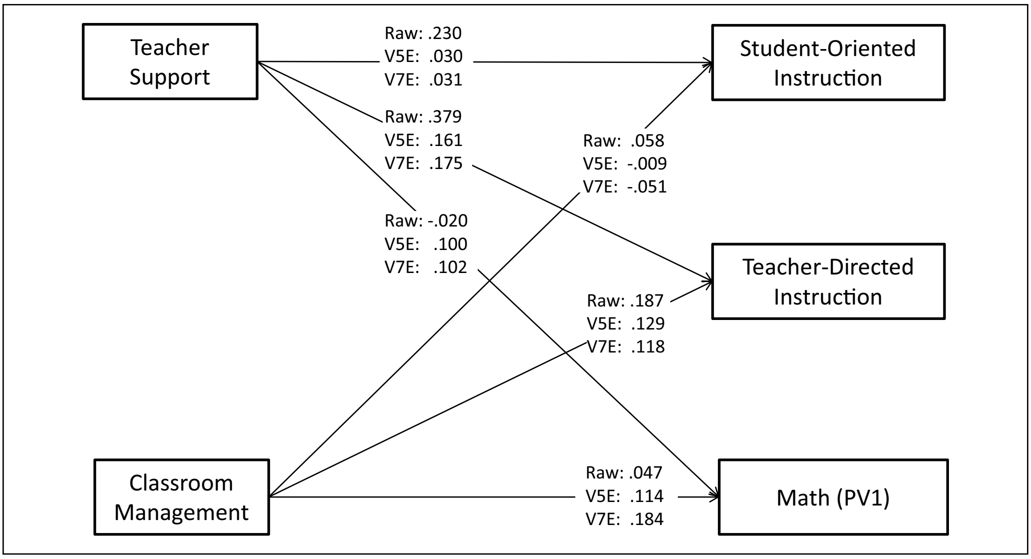

Figure 1. Standardized regression weights in the multigroup path model.
Note. The error terms between Student-Oriented Instruction and Teacher-Directed Instruction, and these between Student-Oriented Instruction and Math were correlated. The scales of Student-Oriented and Teacher-Directed Instruction were rated on a different response scale than the two sets of vignettes; thus, it is not appropriate to rescale these two scales based on the vignettes. In all models, the raw scale scores of these two scales were used. V5E = rescaled scores based on two vignette questions, and the highest possible rating taken in cases of violation; V7E = rescaled scores based on three vignette questions, and the highest possible rating taken in cases of violation; PV = Plausible Value.
All regression weights are significant at $p < .01$.

N = 295,347) = 2,578.242, $p < .01$, CFI = .974, TLI = .963, and RMSEA = .008. The model with the 5-point (Ce) TS and CM showed a good fit, $\chi^2(64, N = 295{,}347) = 4{,}019.459$, $p < .01$, CFI = .943, TLI = .918, and RMSEA = .008. Similarly, the model with the 7-point (Ce) TS and CM fit well, $\chi^2(64, N = 295{,}347) = 4{,}793.105$, $p < .01$, CFI = .940, TLI = .913, and RMSEA = .009. The standardized regression weights in the three models are illustrated in Figure 1.

In comparisons of regression weights in the three models, the negative association between raw TS and math achievement became positive after the rescaling of TS, and the positive relationship between CM and *Student-Oriented Instruction* become negative after rescaling, which confirmed our expectation. Furthermore, the relationships among the four self-report scales were attenuated from raw scores to rescaled scores, suggesting that the rescaling based on anchoring vignettes removed some common method variance (i.e., variance that is attributable to the measurement method rather than to the constructs the measures represent) in self-report Likert-type scales.

Discussion

Anchoring vignettes have been suggested as a technique that remedies various measurement biases in research areas such as health and political opinions, although not in personality (e.g., Mõttus et al., 2012; Rice, Robone, & Smith, 2010). This study systematically investigated the validity of responses to vignette questions, the improvement in measurement invariance with vignette rescaled responses, and the improved predictive validity of TS and CM in a large-scale international assessment involving students in 64 cultures. The main findings include that (a) responses to vignette questions represent valid individual and cultural differences, and, in particular, validity threats from violations in these responses are related to low ESCS and low cognitive sophistication; (b) the rescaled responses tend to show higher levels of comparability; scalar invariance is achieved when data are modeled as ordered categorical in many cases; and (c) the associations of rescaled TS and CM with math achievement, Student-Oriented Instruction, and Teacher-Directed Instruction are slightly different from raw scores, and the former seems to be more in line with the literature. Namely, the associations among all self-report Likert-type scales are weaker with rescaled scores, presumably reducing common method variance, and both rescaled scale scores are more positively related to math achievement. The benefit of anchoring vignettes in enhancing comparability, predictive validity, and interpretability seems rather promising. However, this approach is not an elixir. We focus our discussion on the technicality of using anchoring vignettes in large-scale international comparisons.

In the design of a set of vignettes, unidimensionality and sufficient heterogeneity in representation of trait levels are prerequisite to the success of the approach. It prevents confusions and reduces the likelihood of violations in vignette ratings (Hopkins & King, 2010). In this study, the vignettes for TS and CM worked relatively well, as evidenced in the low percentage of misorderings across all cultures. Nevertheless, the wording of the some vignettes is less than optimal. The vignettes for TS mainly speak about homework, which is not really addressed in the TS scale. The "low level" vignette for CM includes a strange causal statement ("as a result, . . .") which the "medium" and "high level" vignettes do not have. Moreover, each vignette in both constructs includes two stimuli; for TS vignettes, the second stimulus (i.e., always gets back in time) are identical; there is no difference between levels if Stimulus 1 does not apply to teacher (e.g., the concept of homework does not exist in a school or country). We would recommend investing into vignette quality and alignment with questionnaire scales in future studies.

Even a good design does not guarantee that each respondent would rate vignettes without ties or misorderings. Dealing with inconsistent ratings with proxies without further information on the cognitive process of respondents unavoidably adds measurement error to the rescaled scores. Still, we find similar results using multiple proxies in cases of violation (i.e., highest vs. lowest possible ranking) and using multiple numbers of vignettes (i.e., 3 vs. 2). The convergence of results speaks to the robustness of using proxies. It seems that using the highest possible ranking as a proxy always results in the best measurement invariance performance; this might be helpful to researchers who encounter violations in rescaling. In using different methods to test measurement invariance, ordered categorical models performed better compared with continuous models, and thus it is recommended.

Other debatable questions in the application of anchoring vignettes involve the number of sets of vignettes to be used and the estimation strategies. It is a luxury in this study to have two sets of vignettes that can be used to rescale responses of each of the two target scales. This is ideal to maximize the validity of rescaling based on vignettes, but in reality, such a design is extremely difficult to achieve, because adding vignette questions for each construct will at least triple the number of questions. To increase the economy of anchoring vignettes, some researchers proposed to use one set of vignettes to anchor various constructs of the same response format (e.g., Kyllonen & Bertling, 2014). We believe caution is needed to do so. It is reasonable to assume that respondents exhibit similar response styles in their responses to various questions (He & van de Vijver, 2015), yet the perception of standards depends on specific target constructs. For example, a higher standard in TS (i.e., lower rating on vignettes of TS) may not correspond to a higher standard in CM, or to a higher standard in math motivation.

If different sets of vignettes (vignettes target the trait or not) would be effective in rescaling one particular scale, the factor scores of this scale from different sets of vignettes should show strongly positive correlations. We empirically tested the interchangeability of the two sets of vignettes in this study. Specifically, we rescaled all item responses to TS based on the set of vignettes for CM, and all item responses to CM based on the set of vignettes for TS. We rescaled the items using three vignettes and two vignettes, respectively; thus, we later obtained scale scores based on 7-point and 5-point scores using either the highest or lowest possible rating in cases of ties and misordering (i.e., 7E, 7S, 5E, and 5S). It turns out that using rescaled scores in the categorical multigroup confirmatory factor analysis, both scales reached scalar invariance. The good fit of the scalar invariance model suggests similar response styles being controlled for in rescaled item responses. However, the individual-level correlations of rescaled scale scores of one particular scale based on target set of vignettes and nontarget set of vignettes ranged from .440 to .482, and the country-level correlations ranged from .509 to .721. The correlations among different sets of vignettes are not strong enough to conclude that one set of vignettes would work for the rescaling of different target constructs, which speaks against any one-size-fits-all application of vignettes (e.g., Primi, Zanon, Santos, De Fruyt, & John, 2016).

Similarly, dependent on the research questions and implementation of vignettes, various estimation strategies (i.e., nonparametric and parametric) can be used. We adopted the nonparametric approach, due to the necessity in measurement invariance tests which require a specific rescaled item response for each item. Moreover, this approach makes fewer assumptions. A parametric approach requires more statistical assumptions to be met, some of which may not hold. The interchanged use of vignettes above is a case in point.

Limitations and Future Directions

This study is not without limitations. First, the two assumptions of anchoring vignettes, namely, vignettes equivalence and response consistency, were not empirically demonstrated. Existing literature has shown mixed results on the soundness of the assumptions among different populations and topics of interest (e.g., Jürges & Winter, 2013; Kapteyn, Smith, Van Soest, & Vonkova, 2011). Future studies should test the tenability of these assumptions in the PISA context. Second, only the nonparametric approach was used in the study. Future efforts can make use of the parametric approach to study the predictive validity of measures, and apply it to complex survey designs, as is always the case of large-scale international surveys. Third, anchoring vignettes do not work well in some domains but not others (cf. Mõttus et al., 2012); the domain-specificity should be further researched. All in all, this study has provided evidence for the better measurement invariance and better predictive validity of anchoring vignettes. More progress on anchoring vignettes is contingent on careful design of vignettes, further research on the tenability of assumptions, and the development of more appropriate rescaling approaches.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is conducted with sponsorship provided by the Alexander von Humboldt Foundation as well as to the donor, the Federal Ministry for Education and Research to the first author.

Notes

[1] There are three forms of student background questionnaires, and the target scales and vignettes item were only asked in Form B and Form C (N = 318,229). As the effects of anchoring vignettes can only be demonstrated with cases with responses on both vignette items and the two target scales, cases with missing values on these items were deleted. The percentage of missing in the two forms was 6.85%. The small percentage of missing and the large sample size that remains are believed to guarantee adequate power to detect meaningful effects.

2. The term *violation* is used to indicate ties and misorderings, and it does not refer to “right” or “wrong” responses to vignettes.
3. Before testing the path model, the metric invariance of Student-Oriented Instruction and Teacher-Directed Instruction was tested in categorical multigroup confirmatory factor analysis. Across the 64 cultures, Student-Oriented Instruction reached metric invariance, whereas for Teacher-Directed Instruction metric invariance was not acceptable. Caution is needed in comparing association of this scale with other variables across cultures.
4. It is not necessary to assume that the structural weights are the same across all cultures. For ease of presentation, the average effects of Teacher Support (TS) and Classroom Management (CM) on the outcome variables were presented instead of 64 sets of regression weights. The culture-specific structural weights in the unconstrained model were checked as well, and they followed the same patterning of change. The analyses were done with each of the five math achievement plausible values, and results were extremely similar; thus, only results based on the first plausible value of math achievement was reported.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Capella, E., Aber, J. L., & Kim, H. Y. (2016). Teaching beyond achievement tests: Perspectives from developmental and education science. In D. Gitomer & C. Bell (Eds.), *Handbook of research on teaching* (pp. 249-348). Washington, DC: American Educational Research Association.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005-1018. doi:10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. doi:10.1207/s15328007sem0902_5

Desa, D. (2014). *Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses*. Paris, France: Organisation for Economic Co-Operation and Development.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education Principles Policy and Practice*, *11*, 319-330. doi:10.1080/0969594042000304618

Gorur, R. (2014). Towards a sociology of measurement in education policy. *European Educational Research Journal*, *13*, 58-72. doi:10.2304/eerj.2014.13.1.58

He, J., & van de Vijver, F. J. R. (2015). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences*, *31*, 129-134. doi:10.1016/j.paid.2014.09.009

He, J., van de Vijver, F. J. R., Dominguez-Espinosa, A. D., & Mui, P. H. C. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross-Cultural Management*, *14*, 306-322. doi:10.1177/1470595814541424

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage.

Hofstede, G. (2009). *Dimension data matrix: Dimension data matrix*. Retrieved from http://www.geerthofstede.eu/dimension-data-matrix

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*, 201-222. doi:10.1093/poq/nfq011

Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics*, *22*, 1-13. doi:10.1002/hec.1806

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*, 381-399. doi:10.1177/0022022113511297

Kapteyn, A., Smith, J. P., Van Soest, A., & Vonkova, H. (2011, February). *Anchoring vignettes and response consistency* (Working Paper WR-840). Santa Monica, CA: RAND Corporation.

King, G., Murray, C. J. L., 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. doi:10.1017/S000305540400108X

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*, 46-66. doi:10.1093/pan/mpl011

Klieme, E., Backhoff, E., Blum, W., Buckley, J., Hong, Y., Kaplan, D., . . . Vieluf, S. (2013). PISA 2012 context questionnaires framework. In Organisation for Economic Co-Operation and Development (Ed.), *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy* (pp. 167-258). Paris, France: Organisation for Economic Co-Operation and Development.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236. doi:10.1002/acp.2350050305

Kyllonen, P. C., & Bertling, J. J. (2014). 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, FL: CRC Press.

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

Muthen, L. K., & Muthen, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58-88. doi:10.1086/297739

Organisation for Economic Co-Operation and Development. (2013). *PISA 2012 technical report*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2015). *Education at a glance 2015*. Paris, France: Author.

Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. *1*, pp. 17-59). San Diego, CA: Academic Press.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment*, *32*, 39-51. doi:10.1027/1015-5759/a000336

Rice, N., Robone, S., & Smith, P. C. (2010). International comparison of public sector performance: The use of anchoring vignettes to adjust self-reported data. *Evaluation*, *16*, 81-101. doi:10.1177/1356389009350127

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142-151. doi:10.3102/0013189x10363170

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31-57. doi:10.1177/0013164413498257

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Dordrecht, The Netherlands: Springer.

Täht, K., & Must, O. (2013). Comparability of educational achievement and learning attitudes across nations. *Educational Research and Evaluation*, *19*, 19-38. doi:10.1080/13803611.2012.750443

van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, *43*, 1205-1228. doi:10.1177/0022022111428083

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Wand, J., & King, G. (2007). *Anchoring vignettes in R: A (different kind of) vignette*. Retrieved from https://cran.r-project.org/web/packages/anchors/vignettes/anchors.pdf

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology*, *34*, 702-722. doi:10.1177/0022022103257070