JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Spring 2018, Vol. 55, No. 1, pp. 3–31

Cross-Country Heterogeneity in Students' Reporting Behavior: The Use of the Anchoring Vignette Method

Hana Vonkova
Charles University
Gema Zamarro
University of Arkansas
Collin Hitt
Southern Illinois University School of Medicine

Self-reports are an indispensable source of information in education research but they are often affected by heterogeneity in reporting behavior. Failing to correct for this heterogeneity can lead to invalid comparisons across groups. The researchers use the parametric anchoring vignette method to correct for cross-country incomparability of students' reports on teacher's classroom management. Their analysis is based on the data from the Programme for International Student Assessment 2012. The results show significant variation in implicit standards across countries. Correlations between countries' average teacher classroom management levels and external variables like students test scores and public expenditure per pupil change substantially after vignette adjustment. The researchers conclude that the anchoring vignettes method shows potential to enhance the comparability of self-reported measures in education.

Student surveys are an indispensable source of information for education research and policy. Unfortunately, self-reports can also be a flawed source of information. In this article, we focus on heterogeneity in students' reporting behavior across countries, which may occur when different students use different frames of reference (i.e., implicit standards or, in other words, differences in which respondents use response categories) to answer the same question. For example, when students are asked to rate the competencies of their teachers, the individual's own standard for teacher quality impacts the rating that the student assigns the teacher. A student's notion of what it means for a teacher to keep his/her class in order is most probably affected by cultural and country characteristics. So, two students from different countries whose teachers may actually possess the same classroom skills may rate their teachers' classroom management skills differently.

Students within the same country may also exhibit different implicit standards when evaluating their teachers due to, for example, differential socioeconomic background and classroom context (see, e.g., Vonkova, Bendl, & Papajoanu, 2017). Similarly, research has shown the possibility of students having different frames of reference for male and female teachers in the context of teacher evaluations in higher education (see, e.g., MacNell, Driscoll, & Hunt, 2015). Throughout this article we examine international differences in student reports of their teachers' skills, but the

problem we are concerned with extends to virtually any case where student responses are affected by frame of reference. Evidence of this bias in fact has been found when comparing responses of students attending different schools even in the same city (see, e.g., West et al., 2016). Students in Boston attending schools with different academic expectations showed very different standards in self-reports of their own "grit." This is consistently problematic when comparing students attending schools in different countries (Heine, Lehman, Peng, & Greenholtz, 2002). For instance, a consistent finding in international education research is that student attitudes toward learning are, oddly, negatively correlated with achievement test scores across countries—yet, more intuitively, the correlation is positive within countries (Kyllonen & Bertling, 2013). The most plausible explanation for this paradox is that countrywide norms exist that impact how students express their attitudes about learning, and those norms themselves are related to learning outcomes. This creates major problems for between-country comparisons.

A potential solution exists. The anchoring vignettes method was introduced in the social sciences by King, Murray, Salomon, and Tandon (2004) to adjust for such differences in reporting behavior. Anchoring vignettes are hypothetical scenarios representing different levels of a specific concept we desire to measure. Survey respondents are asked to rate a situation described in a vignette, allowing the researcher to gather information on the implicit standards used by the respondent to evaluate their own specific situation.

Self-reports of the concept of interest can then be adjusted based on responses to the vignettes to correct for heterogeneity in implicit standards. Since its introduction, the anchoring vignettes method has been used in social science research in areas such as health, work disability, life satisfaction, job satisfaction, and satisfaction with contacts, but often among adults and not in the context of education (Angelini, Cavapozzi, Corrazini, & Paccagnela, 2012; Bago d'Uva, van Doorslaer, Lindeboom, & O'Donnell, 2008; Bonsang & van Soest, 2012; Grol-Prokopczyk, Freese, & Hauser, 2011; Kapteyn, Smith, & van Soest, 2007; Kristensen & Johansson, 2008; Peracchi & Rossetti, 2012).

Recognizing the potential of anchoring vignettes, the administrators of the renowned Programme for International Student Assessment (PISA) study included vignettes in the student surveys that are administered alongside their tests of academic content knowledge. In this article, we use vignettes data from PISA 2012 and the parametric model of the anchoring vignettes method to adjust student responses to certain key survey items related to teachers' performance in the classroom. Our use of the parametric anchoring vignettes method has potential to improve cross-country comparisons of students' reports on an important dimension of teacher quality: teacher's classroom management skills.

Specifically, we study (a) the heterogeneity in student's assessments of teachers' classroom management skills across countries and (b) the use of the anchoring vignettes method for improving comparability of measures of teacher's classroom management levels across countries. As far as we are aware, this is the first study of the use of the parametric model of the anchoring vignettes method to adjust students' perceptions on a dimension of teacher quality and its potential consequences for comparisons across countries.

Short History of the Anchoring Vignette Method in Education Research

Education research trails other fields in the use of the anchoring vignettes method, although this is starting to change. Buckley and Schneider (2007) and Buckley (2008) used anchoring vignettes for the comparison of parents' satisfaction measures in charter and public schools. Vonkova and Hrabak (2015) studied the use of anchoring vignettes for improving comparability of self-assessments of knowledge and skills of information and communication technologies (ICT) among upper secondary school ICT and non-ICT students. Vonkova et al. (2017) studied heterogeneity in reporting behavior and its impact on the analysis of self-reports of dishonest behavior in schools across secondary school students of different socioeconomic backgrounds.

Additionally, Kyllonen and Bertling (2013) used data from the PISA study in 2012 to showcase the use of nonparametric vignettes methods. In particular, the authors studied the use of nonparametric vignettes methods to correct student reports related to the degree of support received by their teachers and compared the nonparametric vignettes methods to other alternative methods such as forced choice in questions and signal detection correction. Finally, He and van de Vijver (2016) also used data from the PISA study 2012 and nonparametric vignettes methods to explain the paradoxical result that at the country-level student's reports on motivation appear negatively correlated with average student performance, whereas within countries the correlation appears positive. The authors conclude that part of this result is due to differences in implicit standards used by students in different countries.

Our article contributes to this growing literature by illustrating the heterogeneity across countries in the use of implicit standards in students' evaluations of teachers' classroom management skills. We also showcase the use of parametric models of the anchoring vignette method to correct for such heterogeneity in reporting and investigate possible sources of different implicit standards.

Data

This article uses data from the PISA 2012 assessments, which include standardized tests and survey data collected from over 485,000 students, representative of the population of all 15-year-olds in each participating country, enrolled in randomly selected public and private schools in 68 country-regions. These data include measures of student aptitudes in the subject areas of reading, mathematics, and science. Additionally, measures of student attitudes, learning experiences, demographics, and school organization and environment were collected, based on the survey responses of both students and principals.

In the 2012 student questionnaire, PISA included two sets of anchoring vignettes, which were written to describe varying levels of a hypothetical teacher's classroom management and support.[1] A randomly chosen group of students within each school was asked to respond to the anchoring vignettes questions, through the introduction of a rotation design for the student questionnaire. Our analysis is based on this subsample of students asked to complete the anchoring vignettes evaluations, which includes observations of more than 310,000 students in 68 country-regions (a list of

Table 1
Classroom Management Questions: Summary Statistics

Question	1 Strongly Disagree	2 Disagree	3 Agree	4 Strongly Agree	Mean	*SD*
My teacher gets students to listen to him or her.	2.5%	10.6%	52.6%	34.4%	3.19	.72
My teacher keeps the class orderly.	2.8%	14.0%	52.0%	31.2%	3.11	.74
My teacher starts lessons on time.	2.5%	13.7%	47.3%	36.5%	3.18	.76
The teacher has to wait a long time for students to quiet down.	20.7%	42.3%	26.8%	10.3%	2.73	.90

Note. All tabulations and statistics calculated using final student weights.

country-region names and abbreviations used in the 2012 PISA study can be found in Table A1 in Appendix A).

In particular, students in our sample were asked to use the following four-point scale: 1, *strongly disagree*; 2, *disagree*; 3, *agree*; 4, *strongly agree*[2] to report the extent they agree with the following statements:

- Question 1: My teacher gets students to listen to him or her.
- Question 2: My teacher keeps the class orderly.
- Question 3: My teacher starts lessons on time.
- Question 4: The teacher has to wait a long time for students to quiet down.[3]

To combine all questions on teachers' classroom management skills in our empirical model, described below, and to help interpretation of results, we reverse-coded responses to Question 4, and so higher value estimates will correspond to higher levels of teacher's classroom management. Table 1 shows responses for each item of the classroom management scale, across the entire sample. For Questions 1 through 3, answers are skewed heavily toward positive ratings of teacher classroom management skills, with more than 83% of students agreeing or strongly agreeing with positive statements about their teachers' classroom management skills. For Question 4, a negatively worded statement was put to students about teacher classroom management skills, and only 63% disagreed or strongly disagreed with the statement. The average scores across the first three items are practically identical, whereas teachers are on average scored lower on the fourth item.[4]

While these questions aim to measure students' perceptions of teacher classroom management, comparing these raw percentages across countries (or other settings) can be troublesome. This is because any differences in these ratings may be due to differences in students' implicit standards, that is, in how students interpret the reporting scale, rather than actual differences in teachers' classroom management skills. For instance, one would conclude that teachers' classroom management skills

are much better in the United States than in certain high-performing European countries such as The Netherlands. For example, 30% of students in the United States strongly agreed with the statement "My teacher keeps the class orderly," while that proportion was only 17% in the Netherlands. While it is possible that this result reflects actual differences in teacher's classroom management skills across these two countries, it is also very plausible that they are influenced by differences in students' implicit standards.

In order to be able to correct for discrepancies on implicit standards, we can use information from student answers to vignettes on teacher quality. Students were asked to rate the following three scenarios about hypothetical teachers, using the same 4-point scale they used for the evaluations of their actual teacher:

- Vignette 1 (High level): The students in Ms. <name's> class are calm and orderly. She always arrives on time to class. **Ms. <name> is in control of her classroom.**
- Vignette 2 (Medium level): The students in Ms. <name's> class frequently interrupt her lessons. She always arrives five minutes early to class. **Ms. <name> is in control of her classroom**.
- Vignette 3 (Low level): The students in Mr. <name's> class frequently interrupt his lessons. As a result, he often arrives five minutes late to class. **Mr. <name> is in control of his classroom.**

Table 2 shows the average answers to each vignette, across the entire PISA sample. The responses to the vignettes differ substantially. It is interesting and reassuring to see that the average ratings and item scores for the vignettes follow the order hypothesized. The teacher in Vignette 3 has the weakest classroom management skills, the teacher in Vignette 1 has the strongest, and the skills of the teacher in Vignette 2 are somewhere in between. In Table B2 in Appendix B we show descriptive average responses across all countries for the high vignette question presented above.[5] This table suggests that there is a large amount of heterogeneity of ratings not only across countries but also within a country. A decomposition of the variance of ratings to the vignettes into between-and within-country variation suggested that most of the observed overall variance was due to within-country differences in vignette ratings, especially for the high vignette. However, a considerable amount of variation was still observed across countries (29% of the variation in the low vignette, 25% of the

Table 2
Vignettes Questions: Summary Statistics

Vignette	1 Strongly Disagree	2 Disagree	3 Agree	4 Strongly Agree	Mean	*SD*
1 (High)	1.4%	5.9%	39.4%	53.2%	3.44	.67
2 (Medium)	12.3%	41.8%	31.7%	14.2%	2.48	.88
3 (Low)	42.6%	37.4%	14.1%	5.9%	1.83	.88

Note. All tabulations and statistics calculated using final student weights.

variation in the medium vignette and 14% of the variation in the high vignette were explained by across-country differences in reporting). As data sets like PISA are often used for cross-country comparisons, we decided to focus on the effect of different implicit standards across countries in this article.

Coming back to our previous illustrative example comparing the United States and the Netherlands, we observe that Dutch students tend to be more exacting when evaluating higher teacher management skills. For instance, 53% of U.S. students strongly agreed with the statement that this vignette represented a teacher in control of her/his class while only 36% of students in the Netherlands strongly agreed with this statement. Similar differences in reporting behavior across these two countries were also observed by Kapteyn et al. (2007) when studying self-reports of work disability in the adult population. In the next section, we describe how we make use of the information from the PISA vignettes to correct students' perceptions of their teachers' performance in the classroom.

The Parametric Model of the Anchoring Vignettes Method

This article extends the parametric model of the anchoring vignettes method, introduced by King et al. (2004) as the compound hierarchical ordered probit (CHOPIT) model,[6] to the case of having four student assessments related to their actual teacher's classroom management skills along with three vignette assessments, which are all in a 4-point Likert-like scale, as it is the case in our data.[7]

The model consists of two components: a student's assessments of his or her teacher's classroom management skills, and the student's assessment of hypothetical teachers described in the vignettes. For the student's assessment component, let us denote perceived teacher's classroom management skill level by student $i = 1, 2, \ldots, N$ in a given question $q = 1, 2, 3, 4$ by a latent continuous variable Y_{qi}^* and assume that the latent variable is a linear function of observed variables X_i and a normally distributed error term ε_{qi}:

$$Y_{qi}^* = X_i'\beta + \varepsilon_{qi}, \tag{1}$$

$$\varepsilon_{qi} \sim N\left(0, \sigma_q^2\right). \tag{2}$$

For our analysis purposes, the observed variables X_i include country-area dummies for each country–area included in PISA 2012. As it is the case in the traditional ordered probit model, we do not observe Y_{qi}^* directly. What we observe are the answers of student i to the four questions, described above, about the extent of agreement with statements about his/her teacher's classroom management skills on the 4-point ordinal scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, and 4 = *strongly agree*).

Then, the student reported teacher's classroom management level in a given dimension q, as perceived by student i, Y_{qi}, is equal to $j = 1, 2, 3, 4$ if the latent variable Y_{qi}^* is between thresholds τ_i^j and τ_i^{j-1}:

$$Y_{qi} = j \;\leftrightarrow\; \tau_i^{j-1} < Y_{qi}^* \leq \tau_i^j \qquad j = 1, 2, 3, 4. \tag{3}$$

It should be stressed that our model resembles the standard ordered probit model with the key difference that the thresholds are allowed to be student-specific (τ_i^j), to capture differences in implicit standards. In particular, in our model they are allowed to vary with each student's characteristics X_i in the following way:

$$\tau_i^1 = X_i^{'}\gamma^1, \tag{4}$$

$$\tau_i^j = \tau_i^{j-1} + \exp X_i'\gamma^j \qquad j = 2, 3 \tag{5}$$

$$\tau_i^0 = -\infty, \qquad \tau_i^4 = \infty, \tag{6}$$

where γ^j are vectors of unknown parameters. In our case X_i denotes a given country–area for student i. As this is the only explanatory variable allowed to affect the thresholds in our model, thresholds' estimates should be interpreted as the average implicit standards used in a given country to evaluate teacher's classroom management skills. By allowing the thresholds to vary across students, our model captures potential country differences in the usage of reporting scales.

If the only information available are students' assessments of their teachers' performance in the classroom, one would not be able to separately identify the parameters β and γ^j, above, as one would not be able to separate the objective teacher performance level from a different usage of scale. Therefore, more information is needed to separately identify these parameters. This is the information that is provided by the vignettes.

For the vignettes component of the CHOPIT model let us denote with the latent continuous variable Z_{vi}^* the teacher's classroom management skills described in vignette $v = 1, 2, 3$ as it is perceived by student i, and assume

$$Z_{vi}^* = \phi_v + \varsigma_{vi}, \tag{7}$$

$$\varsigma_{vi} \sim N\left(0, \sigma_v^2\right), \tag{8}$$

where the parameter ϕ_v captures the actual level of classroom management skills described in Vignette v and ς_{vi} is an error term independent of ε_{qi}. As it was the case for the students' assessments of their teacher's classroom management skills, what is observed are the actual ordered vignettes evaluations Z_{vi} on a 4-point scale:

$$Z_{vi} = j \;\leftrightarrow\; \tau_i^{j-1} < Z_{vi}^* \leq \tau_i^j \qquad j = 1, 2, 3, 4. \tag{9}$$

Note that the thresholds (τ_j) are assumed to be the same as in the student's assessment component of the CHOPIT model described in Equation 3. This key assumption is known as response consistency. Note also that the teacher classroom management skills described in the vignette are assumed to be the same independent of the student's country of residence. This assumption is called vignette equivalence. These assumptions allow us to identify the true country differences and threshold parameters. As it is also the case in the traditional ordered probit model, parameters in the CHOPIT model described above are not identified unless we make some additional parametrization assumptions. In our case, we take the United States as our reference country and set its coefficient β to zero. In addition, the variance of the

error terms for the four questions are assumed to be equal ($\sigma_q^2 = 1, \quad q = 1, 2, 3, 4$).[8] Note that the rest of parameters are left unrestricted and are estimated by maximum likelihood.

Results

Heterogeneity in the Use of Reporting Scales

The heterogeneity in implicit standards is captured by the estimated thresholds' parameters using the CHOPIT model described in the previous section. We find significant differences in the estimated thresholds across countries (see Table B3 in Appendix B for the estimated values of all thresholds for each country). This suggests a considerable amount of heterogeneity in implicit standards across countries. In Figure 1, we illustrate the estimated values of threshold between *strongly agree* and *agree* (Threshold 3) in a world map.

In particular, based on the estimated thresholds we can classify countries in the following way:

Low Threshold 1. Estimated Threshold 1 values, representing the cutoff point for responding *disagree* versus *strongly disagree* vary significantly across countries. Countries with low Threshold 1 tend to label a given level of teachers' classroom management with the scale option *disagree* rather than *strongly disagree*. One could say that students in these countries have lower standards concerning the choice between these two categories. Or, put differently, they are more optimistic about their teachers' classroom management skills. Some provinces in China (e.g., Macao-China), other Asian countries like Indonesia, Malaysia, Thailand, and

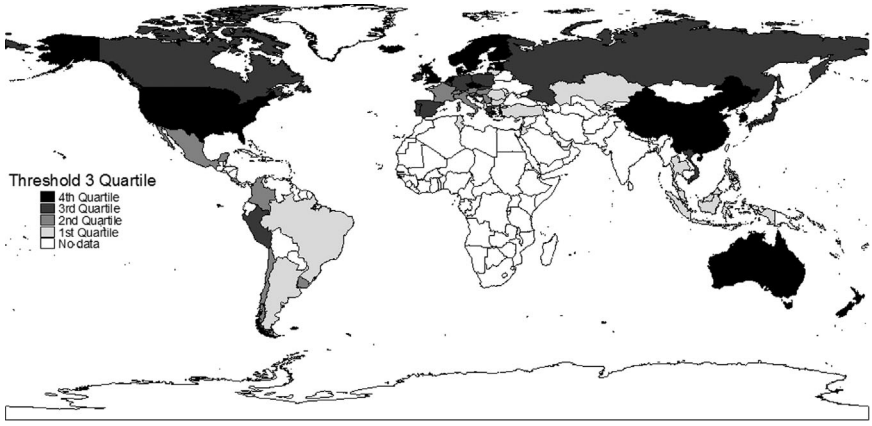

Figure 1. Geographic distribution of estimated threshold between the categories *strongly agree* and *agree* using CHOPIT model.
Note. We divided countries into four categories based on where country's threshold between the categories *strongly agree* and *agree* stands in the overall distribution of values. Light grey indicates lower threshold values. Countries with lower threshold tend to label a given level of teachers' classroom management with the scale option *strongly agree* rather than *agree*. For China, only results for Shanghai are presented.

Vietnam, countries located in the Middle East (e.g., Jordan), Eastern European countries (e.g., Romania, Bulgaria, and Hungary) and South American countries like Brazil and Colombia display the lowest estimated Threshold 1 values.

High Threshold 1. The countries with the highest Threshold 1 values include the United States, Shanghai-China, the United Kingdom, Ireland, Israel, and countries in Continental Europe (e.g., Austria, Germany, France, and Luxembourg).

Low Threshold 3. Countries with low Threshold 3 tend to label a given level of teacher's classroom management with the highest end-point of the scale (*strongly agree*) more often than students in other countries. As shown in Figure 1, the lowest Threshold 3 values are seen in countries located in the Middle East, Western Asia, the Balkans, and South America. This is the case for countries such as Jordan, Indonesia, Qatar, Albania, Romania, Malaysia, United Arab Emirates, Tunisia, Bulgaria, Turkey, Thailand, Lithuania, Brazil, and Argentina, among others.

High Threshold 3. Countries with higher estimated Threshold 3 values are located in Asia (e.g., Shanghai-China and Korea), North America (i.e., the United States), Northern and continental Europe (e.g., the Netherlands, Norway, and Denmark), and the continent of Australia (e.g., New Zealand and Australia).

The relationship between Thresholds 1 and 3 is not strong (correlation of the order of .46), suggesting the existence both of countries using more end-points (high Threshold 1 and low Threshold 3) and countries using more mid-points (low Threshold 1 and high Threshold 3). Figure 2 presents a comparison of the estimated Thresholds 1 and 3.

High Threshold 1 and low Threshold 3. Countries in this group include, for example, Lithuania, Luxembourg, Turkey, Tunisia, Chile, Iceland, Switzerland, and Costa Rica.

Low Threshold 1 and high Threshold 3. Countries or country–areas in this group include Korea, the province of Macao in China, Chinese Taipei, Finland, Russia, Hong Kong, Slovak Republic, Peru, and to some extend also Portugal and Vietnam.

In addition, it is also of interest to study whether there are countries whose estimated thresholds are all high and so whose reporting scale is shifted to the right. Students in these countries would tend to have higher standards or be more pessimistic on average when evaluating their teachers' classroom management behavior. On the other side, there would be countries whose estimated thresholds are all low, indicating their reporting scale is shifted to the left; that is, students in these countries would have lower implicit standards or be generally more optimistic when evaluating their teachers.

All thresholds are high (scale is shifted to the right). A clear example of a country-region in this group is the case of Shanghai in China, followed by the United States, the U.S. states of Florida, Massachusetts, and Connecticut, the United Kingdom, Iceland, and Austria.

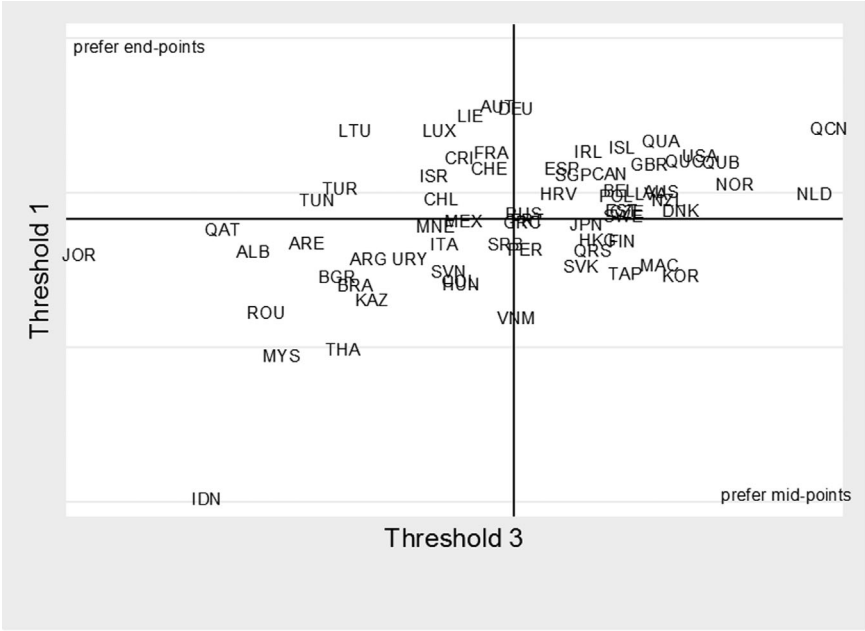

Figure 2. Relationship between estimated Threshold 1 and Threshold 3 using CHOPIT model.
Note. High Threshold 1 (Threshold 3) indicates the tendency to use end point *strongly disagree* (*agree*) rather than *disagree (strongly disagree)*, when evaluating a given level of teachers' management skills. Solid lines represent the mean of the distribution of Threshold 1 and Threshold 3 values.

All thresholds are low (scale is shifted to the left). The extreme case in this group is Indonesia, followed by Malaysia, Thailand, Jordan, Romania, Albania, Kazakhstan, Qatar, and Bulgaria, among others.

Finally, it should be pointed out that there are also country–areas that do not appear in any of the classifications above and so their students do not make use of the reporting scale in any extreme way. Countries in this group would be, for example, Japan, Singapore, Croatia, Peru, and Mexico, among others.

Adjusted Versus Unadjusted Levels of Teachers' Classroom Management Skills

Given the high level of heterogeneity on how scales are used in different countries we expected that our adjustments using anchoring vignettes would make a difference on how countries compare in terms of their teachers' classroom management skills—and this is exactly what we found. Table 3 presents the estimated country coefficients from both the CHOPIT model (representing adjusted country average teachers' classroom management skills) and traditional ordered probit models (representing unadjusted country average teachers' classroom management skills) and a ranking based on these coefficients, such as the lowest numbers represent top positions. Results from the ordered probit models are presented as comparison to adjustments due to different implicit standards across countries performed through the CHOPIT model.

Table 3
Estimated Country Effects and Ranking of Countries Using CHOPIT and Ordered Probit Models

Country	Ranking		Effects		Country	Ranking		Effects	
	CH	OP	CH	OP		CH	OP	CH	OP
Shanghai–China	1	19	.209***	.028***	Montenegro	35	20	–.321***	.027**
			(.0212)	(.0096)				(.0204)	(.0098)
Connecticut (USA)	2	16	.064**	.039***	Macao–China	36	43	–.324***	–.148***
			(.0293)	(.0139)				(.0200)	(.0099)
Massachusetts (USA)	3	15	.023	.040***	Portugal	37	39	–.346***	–.128***
			(.0294)	(.0136)				(.0199)	(.0092)
United States of America	4	27	.000	.000	Spain	38	56	–.352***	–.220***
			Reference	Reference				(.0161)	(.0074)
Costa Rica	5	3	–.022	.256***	The Netherlands	39	68	–.354***	–.390***
			(.0213)	(.0098)				(.0208)	(.0104)
Russian Federation	6	5	–.026	.196***	Croatia	40	53	–.361***	–.197***
			(.0200)	(.0099)				(.0203)	(.0093)
Florida (USA)	7	26	–.031	.000	Luxembourg	41	50	–.394***	–.176***
			(.0273)	(.0125)				(.0207)	(.0090)
Perm (Russian Federation)	8	7	–.059**	.105***	Poland	42	64	–.395***	–.285***
			(.0281)	(.0146)				(.0209)	(.0096)
United Kingdom	9	32	–.064***	–.018***	Sweden	43	61	–.396***	–.259***
			(.0176)	(.0080)				(.0210)	(.0101)
Latvia	10	29	–.068***	–.007	Turkey	44	21	–.400***	.024***
			(.0221)	(.0105)				(.0203)	(.0098)
Iceland	11	31	–.071***	–.016*	Slovak Republic	45	55	–.423***	–.204***
			(.0235)	(.0106)				(.0207)	(.0102)

(Continued)

Table 3
Continued

Country	Ranking		Effects		Country	Ranking		Effects	
	CH	OP	CH	OP		CH	OP	CH	OP
Kazakhstan	12	1	–.083***	.404***	Hong Kong–China	46	51	–.429***	–.184***
			(.0193)	(.0097)				(.0207)	(.0099)
Liechtenstein	13	9	–.114**	.081***	Serbia	47	46	–.431***	–.157***
			(.0714)	(.0272)				(.0204)	(.0096)
Canada	14	30	–.120***	–.014**	Chinese Taipei	48	59	–.436***	–.237***
			(.0164)	(.0075)				(.0188)	(.0092)
Estonia	15	33	–.124***	–.032***	Finland	49	65	–.436***	–.299***
			(.0217)	(.0097)				(.0188)	(.0087)
Singapore	16	14	–.125***	.046***	Tunisia	50	25	–.437***	.009
			(.0201)	(.0095)				(.0209)	(.0099)
Ireland	17	36	–.140***	–.059***	France	51	62	–.443***	–.266***
			(.0213)	(.0091)				(.0220)	(.0091)
Lithuania	18	4	–.141***	.236***	Chile	52	45	–.447***	–.157***
			(.0214)	(.0095)				(.0192)	(.0088)
Japan	19	28	–.189***	–.004	United Arab Emirates	53	13	–.469***	.046***
			(.0200)	(.0103)				(.0174)	(.0082)
Belgium	20	38	–.204***	–.103***	Bulgaria	54	22	–.487***	.017***
			(.0188)	(.0085)				(.0198)	(.0094)
Albania	21	2	–.220***	.375***	Slovenia	55	44	–.507***	–.149**
			(.0221)	(.0109)				(.0197)	(.0094)
Norway	22	57	–.222***	–.231***	Uruguay	56	40	–.517***	–.131***
			(.0215)	(.0104)				(.0203)	(.0099)
Australia	23	48	–.227***	–.166***	Hungary	57	52	–.525***	–.193***
			(.0172)	(.0079)				(.0204)	(.0095)

(Continued)

Table 3
Continued

Country	Ranking		Effects		Country	Ranking		Effects	
	CH	OP	CH	OP		CH	OP	CH	OP
Mexico	24	8	–.228***	.087***	Greece	58	66	–.526***	–.304***
			(.0157)	(.0073)				(.0200)	(.0093)
Peru	25	23	–.245***	.013*	Italy	59	54	–.546***	–.199***
			(.0198)	(.0099)				(.0158)	(.0073)
Austria	26	41	–.250***	–.136***	Romania	60	11	–.552***	.055***
			(.0222)	(.0093)				(.0199)	(.0096)
Czech Republic	27	47	–.256***	–.158***	Korea	61	67	–.557***	–.347***
			(.0209)	(.0093)				(.0203)	(.0106)
Switzerland	28	34	–.261***	–.050***	Brazil	62	37	–.576***	–.097***
			(.0179)	(.0080)				(.0164)	(.0077)
Israel	29	24	–.267***	.013*	Malaysia	63	17	–.613***	.035***
			(.0206)	(.0094)				(.0208)	(.0103)
Germany	30	49	–.272***	–.172***	Jordan	64	6	–.635***	.148***
			(.0232)	(.0096)				(.0191)	(.0090)
Colombia	31	10	–.288***	.075***	Thailand	65	35	–.657***	–.058***
			(.0180)	(.0088)				(.0195)	(.0100)
New Zealand	32	58	–.296***	–.232***	Qatar	66	42	–.717***	–.139***
			(.0218)	90.0102)				(.0177)	(.0084)
Denmark	33	60	–.300***	–.239***	Argentina	67	63	–.723***	–.283***
			(.0196)	(.0092)				(.0196)	(.0096)
Vietnam	34	12	–.307***	.051***	Indonesia	68	18	–.830***	.035***
			(.0213)	(.0108)				(.0211)	(.0107)

Note. ***, **, and * indicate statistical significance at 1%, 5%, and 10%, respectively. CH is the abbreviation for the CHOPIT model, OP is the abbreviation for the ordered probit model. Country effects label estimated beta parameters from both CHOPIT and ordered probit models.

In the ordered probit models, thresholds are assumed to be constant across countries and therefore the model does not capture across-country heterogeneity in implicit standards. The estimated thresholds in this case are "average thresholds" across all countries in the sample.

In particular, our results show that some countries significantly improve their position in the ranking when based on adjusted average teachers' classroom management skills. This is the case for several countries, including Norway, the Netherlands, Denmark, New Zealand, Australia, the United States, and the United Kingdom. For example, Norway moved up from 57th position to 22nd after the adjustments. In contrast, other countries' position worsened after the adjustments. This is the case, for example, for Jordan, Indonesia, Romania, Malaysia, and the United Arab Emirates, among others. For instance, in the extreme case, Jordan moved from 6th position to 64th position after adjustments.

Relationship Between Teachers' Classroom Management Performance and External Variables

In this section we try to shed light on whether our adjustments move us closer to a truer measure of teachers' classroom management skills. We cannot observe teacher quality directly in these data, of course, but we can observe factors that we would expect to correlate with true teacher quality. Specifically, we studied the correlations of country average levels of teachers' classroom management skills, before and after adjustments for heterogeneity in implicit standards, and country-level variables including average math and reading test scores, public expenditure per pupil, GDP per capita, percentage of private schools, and whether the country had a curriculum-based external high school exit exam. Our unadjusted measures, in this case, are based on estimates of an ordered probit type model where we model together responses to Questions 1 and 4 but restrict the thresholds to be constant across countries. Adjusted measures are based on country effects obtained through the CHOPIT model, as described above.

Table 4 shows simple correlations among adjusted and unadjusted average teachers' classroom management skills and average math and reading test scores at the country level. It is very interesting to observe that unadjusted measures of teacher classroom management skills, at the country level, are negatively correlated with average math and reading scores. That is, it seems that, across countries, those countries with lower reported levels of teacher classroom management skills are

Table 4
Correlation Among Estimated Adjusted and Unadjusted Averaged Teachers' Classroom Management Levels at the Country Level and Averaged Test Scores

	Ordered Probit (No Adjustment)	CHOPIT (With Vignette Adjustment)
Average math scores	–.325	.452
Average reading scores	–.373	.477

Note. Number of observations: 68 country–regions.

Table 5
Regression Estimates of Ordered Probit and CHOPIT Estimated Country-Level Effects on Country-Level Characteristics

	Ordered Probit (No Adjustment)	CHOPIT (With Vignette Adjustment)
Public expenditure per pupil	–.0051*	.0043
	(.0026)	(.0044)
GDP per capita: 1,000$	–.0017	.0038**
	(.0011)	(.0018)
% Private schools	–.0013	–.0009
	(.0090)	(.0009)
Math exit exam	.0294	–.0758
	(.0403)	(.0510)
Constant	.0807	–.4999***
	(.0774)	(.1161)

Note. Number of observations: 42 countries; robust standard errors in parenthesis; *** represents significance at the 1% level,** at the 5% level, and * at the 10% level.

those that have higher average performance on reading and math. These results are totally reversed when using the vignette-adjusted teacher classroom management measures. In this case, as one would expect, we observe that countries with higher levels of teacher classroom management skills are those with higher average math and reading scores.

We also obtained correlations of adjusted and unadjusted teacher classroom management skills and other external country-level variables using linear regression models. These results are presented in Table 5. An interesting pattern is observed in this table. Average teacher classroom management skills from an unadjusted ordered probit model are negatively correlated with public expenditure in education per student. That is, countries that spend more on education from their public funds have lower levels of teachers' classroom management. Once we adjust for heterogeneity in the use of the reporting scales, by means of the CHOPIT model, we see that the relationship between country's performance in teachers' classroom management skills and public expenditure in education changes signs and becomes not significant. A similar effect is observed for GDP per capita. A negative but insignificant effect is observed without adjustments while a positive and significant relationship between GDP per capita and teachers' classroom management skills is observed after correcting for heterogeneity in reporting behavior. Finally, we do not find a significant correlation between the proportion of private schools and the existence of a curriculum-based external high school exit exam (CBEEE),[9] with either unadjusted or adjusted teachers' classroom management skills measures.

Conclusion

Self-reports are a central source of information for education research. Students are often surveyed on topics such as teacher performance in the classroom or the safety environment of their schools. These types of student self-reports are

increasingly being used to shape policy and personnel decisions. Students are also often surveyed about their own behavior: self-reports are the basis for the measurement of student's socioemotional skills. However, comparisons of self-reported measures, across individuals in different countries or groups within a country, can be biased if respondents differ on their use and interpretation of the different scales in the provided questions. This issue, which has been referred in the education literature as reference group bias, is an important problem that has been mostly ignored in most education research.

We explore a potential solution to this problem, using information from anchoring vignettes to correct self-reports. Although successfully used in other areas of social sciences among adults, this approach is relatively new among students in the context of education and more research is needed to study its validity.

In this article, we use data from PISA 2012 to study the heterogeneity in students' assessments of teacher performance. In particular we examine differences in implicit standards across countries. We use a parametric anchoring vignettes method as a way to correct for this heterogeneity in reporting behavior. This approach is not common in education research. A unique set of new questions in the PISA 2012 student surveys made this analysis possible. PISA 2012 asked students not only to assess their own teacher's classroom management; the survey also asked students to rate fictional teachers described in vignettes.

Our results show significant differences between the adjusted and unadjusted distributions of teachers' classroom management skills across countries. We also show that these differences in scale usage might be geographically related—for example, students in some Northern European countries tend to have higher standards whereas some Southeast Asia countries tend to present lower standards when evaluating their teacher's classroom management levels.

Put plainly, we find that countries' relative rankings in student-reported teacher quality is sensitive to adjustments for differential use of reporting scales. In turn, the apparent associations between student-reported teacher quality and certain policy variables is sensitive to adjustments for students' implicit standards. We show that correlations between countries' student-reported teacher classroom management levels and external variables, like average test scores and public expenditure in education per student, go from negative to positive after adjusting for the heterogeneity in reporting behavior, moving to more intuitive results. Like He and van de Vijver (2016), our results show that these paradoxical correlations found across countries are in large part due to differential reporting styles across countries.

We must offer one caveat. Within the PISA study we cannot know for certain what the true levels of classroom management skills actually are. Without additional, more objective measures of teacher quality, we are not able to determine to what extent our adjustments lead to values of the estimated correlations that are closer to the real situation. However, the fact that the estimated correlations between teacher quality measures and test scores or policy variables change signs toward more intuitive values leads us to think that these adjustments bring us closer to truer measures of teacher quality. Our findings suggest strongly that, when making international comparisons, any student surveys of teacher quality should include anchoring vignettes. We conclude that the parametric anchoring vignettes

method has potential to enhance the validity and international comparability of self-reported measures in education—and for this reason we strongly recommend that vignettes be developed for other topics on which students are surveyed. Students are increasingly being asked to rate themselves on topics such as grit and self-efficacy, to borrow examples from PISA and the U.S. National Assessment of Educational Progress (NAEP) tests. If cultural differences influence how students rate their teachers, those differences almost certainly influence how students view themselves.

Acknowledgments

This study was supported by a grant by the Czech Science Foundation through the project "The relationships between skills, schooling and labor market outcomes: a longitudinal study" (P402/12/G130). We thank Vera DeBerg for research assistance in early versions of this article. We would also like to thank Cara Jackson and conference and seminar participants at the 2015 AEFP 40th Annual Conference, CIES 2015 conference, ECER 2015 conference, and the Department of Education Reform at the University of Arkansas for valuable feedback on a previous version of this article.

Appendix A

Table A1
Country Names and Abbreviations

Abbreviation	Country Name	Abbreviation	Country Name
USA	United States of America	KOR	Korea
ALB	Albania	LIE	Liechtenstein
ARE	United Arab Emirates	LTU	Lithuania
ARG	Argentina	LUX	Luxembourg
AUS	Australia	LVA	Latvia
AUT	Austria	MAC	Macao–China
BEL	Belgium	MEX	Mexico
BGR	Bulgaria	MNE	Montenegro
BRA	Brazil	MYS	Malaysia
CAN	Canada	NLD	The Netherlands
CHE	Switzerland	NOR	Norway
CHL	Chile	NZL	New Zealand
COL	Colombia	PER	Peru
CRI	Costa Rica	POL	Poland
CZE	Czech Republic	PRT	Portugal
DEU	Germany	QAT	Qatar
DNK	Denmark	QCN	Shanghai–China
ESP	Spain	QRS	Perm (Russian Federation)
EST	Estonia	QUA	Florida (USA)
FIN	Finland	QUB	Connecticut (USA)
FRA	France	QUC	Massachusetts (USA)
GBR	United Kingdom	ROU	Romania
GRC	Greece	RUS	Russian Federation

(Continued)

Table A1
Continued

Abbreviation	Country Name	Abbreviation	Country Name
HKG	Hong Kong–China	SGP	Singapore
HRV	Croatia	SRB	Serbia
HUN	Hungary	SVK	Slovak Republic
IDN	Indonesia	SVN	Slovenia
IRL	Ireland	SWE	Sweden
ISL	Iceland	TAP	Chinese Taipei
ISR	Israel	THA	Thailand
ITA	Italy	TUN	Tunisia
JOR	Jordan	TUR	Turkey
JPN	Japan	URY	Uruguay
KAZ	Kazakhstan	VNM	Vietnam

Appendix B

Table B1
Descriptive Responses to Question 2: My Teacher Keeps the Class Orderly

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
North America					
Canada	13,898	3.40	14.48	51.66	30.46
Connecticut (USA)	1,079	1.58	12.14	54.25	32.04
Florida (USA)	1,241	2.63	13.86	52.97	30.54
Massachusetts (USA)	1,120	1.93	11.37	55.54	31.15
USA	3,226	2.56	12.50	54.82	30.13
Central America					
Costa Rica	2,861	1.93	10.43	43.78	43.87
Mexico	22,134	1.96	11.79	48.52	37.73
South America					
Argentina	3,689	5.22	20.65	49.61	24.52
Brazil	11,964	2.95	18.82	48.20	30.02
Chile	4,490	3.67	21.37	48.76	26.20
Colombia	5,443	1.96	11.84	50.99	35.21
Peru	3,636	1.19	11.90	53.45	33.46
Uruguay	3,299	2.82	17.20	50.66	29.31
Northern Europe					
Denmark	4,740	2.45	17.03	60.60	19.93
Estonia	3,162	2.21	14.69	52.59	30.50
Finland	5,652	4.32	23.26	54.46	17.96

(Continued)

Table B1
Continued

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
Iceland	2,240	3.03	15.78	49.00	32.19
Ireland	3,313	4.10	16.95	47.71	31.24
Latvia	2,816	2.08	12.78	54.44	30.70
Lithuania	3,071	3.25	11.28	37.34	48.14
Norway	2,959	2.74	19.16	59.44	18.66
Sweden	2,985	4.15	22.38	54.03	19.44
United Kingdom	8,240	3.05	16.08	52.38	28.49
Southern Europe					
Albania	2,671	.82	4.28	37.77	57.13
Croatia	3,312	4.88	19.07	49.61	26.43
Greece	3,366	6.68	26.73	45.94	20.65
Italy	20,424	5.54	17.89	50.16	26.40
Montenegro	3,054	3.64	12.84	49.18	34.34
Portugal	3,711	3.98	18.50	49.49	28.04
Serbia	3,029	3.80	17.43	51.04	27.73
Slovenia	3,775	2.36	18.50	49.71	29.42
Spain	16,518	4.64	19.99	50.00	25.37
Eastern Europe					
Bulgaria	3,335	3.28	14.43	46.06	36.23
Czech Republic	3,431	3.70	16.44	53.12	26.75
Hungary	3,157	4.80	21.16	47.13	26.90
Perm (Russian Fed.)	1,157	.93	10.40	52.84	35.84
Poland	3,028	4.73	21.08	51.47	22.72
Romania	3,340	2.47	10.20	44.46	42.88
Russian Federation	3,451	1.29	8.93	47.00	42.78
Slovak Republic	3,025	2.52	14.81	59.05	23.63
Western Europe					
Austria	3,087	5.06	19.84	42.86	32.23
Belgium	5,405	3.56	14.32	54.75	27.38
France	2,978	8.15	22.32	45.65	23.88
Germany	2,737	5.36	22.45	43.43	28.76
Liechtenstein	188	2.77	14.63	43.84	38.77
Luxembourg	3,385	7.24	18.94	42.88	30.94
The Netherlands	2,799	5.80	22.24	55.14	16.82
Switzerland	7,342	4.00	17.63	47.54	30.83
Middle East					
Israel	3,149	2.90	12.07	44.49	40.54
Jordan	4,495	3.14	8.27	38.00	50.59
Qatar	6,575	5.70	13.76	44.43	36.12
Tunisia	2,743	5.33	11.99	42.19	40.49
Turkey	3,175	4.05	11.23	45.13	39.59
United Arab Emirates	7,293	2.93	11.59	44.54	40.94

(Continued)

Table B1
Continued

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
Central Asia					
Kazakhstan	3,830	.30	2.40	42.95	54.35
Eastern Asia					
Chinese Taipei	4,007	3.46	18.94	56.15	21.46
Hong Kong–China	3,016	3.58	16.19	59.88	20.34
Japan	4,131	4.64	24.84	48.82	21.70
Korea	3,356	3.45	19.92	62.46	14.17
Macao–China	3,527	2.89	17.94	60.91	18.26
Shanghai–China	3,456	1.88	13.80	54.76	29.56
South-Eastern Asia					
Indonesia	3,663	.55	4.31	57.51	37.63
Malaysia	3,358	1.19	6.67	53.00	39.14
Singapore	3,660	1.28	9.16	53.53	36.04
Thailand	4,378	.56	7.04	59.88	32.52
Vietnam	3,299	.93	9.15	63.38	26.53
Oceania					
Australia	9,293	3.43	19.46	52.33	24.78
New Zealand	2,753	3.59	21.70	52.89	21.83

Note. All tabulations and statistics calculated using final student weights.

Table B2
Descriptive Responses to High Vignette "The Students in Ms. <name's> Class Are Calm and Orderly. She Always Arrives on Time to Class. Ms. <name> Is in Control of Her Classroom"

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
North America					
Canada	13,872	.80	3.32	35.39	60.50
Connecticut (USA)	1,070	.70	4.35	42.98	51.98
Florida (USA)	1,240	.65	4.12	41.12	54.11
Massachusetts (USA)	1,120	.68	3.13	38.43	57.75
USA	3,216	.75	.43	41.56	53.36
Central America					
Costa Rica	2,865	1.17	3.24	32.72	62.87
Mexico	22,130	1.59	5.56	32.18	60.67
South America					
Argentina	3,637	3.16	8.79	39.69	48.36
Brazil	11,965	1.94	9.20	41.37	47.48

(Continued)

Table B2
Continued

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
Chile	4,476	1.33	4.85	29.55	64.27
Colombia	5,432	1.37	5.97	40.10	52.55
Peru	3,633	1.23	5.24	41.16	52.36
Uruguay	3,256	1.83	4.84	36.60	56.73
Northern Europe					
Denmark	4,746	1.41	5.86	31.76	60.98
Estonia	3,154	.65	2.92	27.05	69.37
Finland	5,662	1.04	3.75	30.32	64.88
Iceland	2,245	1.99	5.27	28.96	63.78
Ireland	3,314	.52	1.96	31.03	66.49
Latvia	2,815	.81	5.21	31.09	62.89
Lithuania	3,049	2.53	7.35	22.76	67.36
Norway	2,968	4.26	14.05	35.44	46.26
Sweden	2,975	3.41	11.27	40.22	45.10
United Kingdom	8,225	.71	3.84	36.41	59.03
Southern Europe					
Albania	2,663	1.44	4.08	26.33	68.15
Croatia	3,292	.91	4.30	34.69	60.11
Greece	3,364	2.11	8.24	38.90	50.75
Italy	20,383	1.70	5.10	34.75	58.44
Montenegro	2,985	3.32	7.06	37.86	51.75
Portugal	3,705	.82	2.82	33.23	63.14
Serbia	2,998	3.01	8.40	39.69	48.90
Slovenia	3,758	1.54	6.85	34.14	57.47
Spain	16,519	1.83	5.07	28.22	64.89
Eastern Europe					
Bulgaria	3,319	2.39	9.02	40.55	48.04
Czech Republic	3,429	1.56	4.89	36.15	57.40
Hungary	3,151	1.33	5.42	36.05	57.20
Perm (Russian Fed.)	1,155	.75	8.70	40.09	50.46
Poland	3,028	2.36	7.07	35.22	55.35
Romania	3,342	2.77	9.36	38.03	49.84
Russian Federation	3,446	1.45	8.52	35.58	54.45
Slovak Republic	3,010	2.15	9.22	43.00	45.63
Western Europe					
Austria	3,079	4.42	8.13	17.99	69.46
Belgium	5,370	1.63	4.91	36.99	56.47
France	2,953	2.00	4.38	27.31	66.32
Germany	2,719	3.06	8.34	19.72	68.89
Liechtenstein	190	4.29	7.21	22.29	66.21
Luxembourg	3,369	4.59	8.85	24.53	62.03

(Continued)

Table B2
Continued

	N	% 1 Strongly Disagree	% 2 Disagree	% 3 Agree	% 4 Strongly Agree
The Netherlands	2,805	1.07	6.40	56.60	35.93
Switzerland	7,323	2.78	8.01	24.98	64.22
Middle East					
Israel	3,147	2.37	5.06	28.73	63.85
Jordan	4,463	2.36	7.68	29.22	60.74
Qatar	6,549	3.80	12.72	36.01	47.47
Tunisia	2,714	4.01	9.46	33.04	53.48
Turkey	3,164	2.09	7.27	31.37	59.27
United Arab Emirates	7,262	2.00	6.63	32.66	58.72
Central Asia					
Kazakhstan	3,823	1.00	4.65	35.91	58.44
Eastern Asia					
Chinese Taipei	4,007	1.76	6.30	49.66	42.28
Hong Kong–China	3,014	1.05	8.20	56.63	34.12
Japan	4,157	2.26	10.71	46.22	40.81
Korea	3,327	1.55	6.74	57.68	34.02
Macao–China	3,528	.96	5.92	47.32	45.79
Shanghai–China	3,456	.67	3.40	46.57	49.36
South-Eastern Asia					
Indonesia	3,675	.45	3.34	49.51	46.70
Malaysia	3,366	.90	6.27	38.03	54.80
Singapore	3,653	1.46	3.13	39.60	55.81
Thailand	4,378	.44	4.82	40.13	54.61
Vietnam	3,295	.44	3.33	40.73	55.50
Oceania					
Australia	9,246	.84	3.81	38.22	57.12
New Zealand	2,755	1.03	4.74	41.90	52.33

Note. All tabulations and statistics calculated using final student weights.

Table B3
Estimated Thresholds' Parameters from CHOPIT Model

Country	Threshold 1	Threshold 2	Threshold 3
Shanghai–China	.087***	.074***	.024***
	(.0206)	(.0160)	(.0099)
Connecticut (USA)	–.020	.033*	.014
	(.0305)	(.0247)	(.0142)
Massachusetts (USA)	–.016	.008	–.013
	(.0315)	(.0246)	(.0140)

(Continued)

Table B3
Continued

Country	Threshold 1	Threshold 2	Threshold 3
United States of America	–1.882***	.046***	.266***
(Reference)	(.0121)	(.0116)	(.0069)
Costa Rica	–.008	–.148***	–.183***
	(.0213)	(.0182)	(.0112)
Russian Federation	–.188***	.045***	–.106***
	(.0203)	(.0159)	(.0102)
Florida (USA)	.048**	–.051***	–.043***
	(.0273)	(.0216)	(.0136)
Perm (Russian Federation)	–.306***	.143***	–.015
	(.0302)	(.0224)	(.0142)
United Kingdom	–.026*	.023**	–.060***
	(.0176)	(.0136)	(.0083)
Latvia	–.123***	.087***	–.037***
	(.0216)	(.0168)	(.0106)
Iceland	.026	.001	–.120***
	(.0225)	(.0175)	(.0115)
Kazakhstan	–.465***	–.005	–.028***
	(.0206)	(.0168)	(.0102)
Liechtenstein	.129**	–.151***	–.298***
	(.0615)	(.0546)	(.0363)
Canada	–.057***	–.003	–.062***
	(.0166)	(.0128)	(.0078)
Estonia	–.176***	.128***	–.069***
	(.0211)	(.0157)	(.0102)
Singapore	–.061***	–.098***	–.028***
	(.0205)	(.0169)	(.0098)
Ireland	.015***	.018*	–.172***
	(.0211)	(.0158)	(.0104)
Lithuania	.080***	–.302***	–.300***
	(.0205)	(.0177)	(.0113)
Japan	–.221***	.032**	.009
	(.0192)	(.0153)	(.0094)
Belgium	–.110***	.051***	–.054***
	(.0185)	(.0140)	(.0089)
Albania	–.309***	–.133***	–.212***
	(.0221)	(.0186)	(.0121)
Norway	–.091***	.086***	.039***
	(.0207)	(.0152)	(.0098)
Australia	–.116***	.078***	–.021***
	(.0175)	(.0131)	(.0081)
Mexico	–.213***	–.038***	–.091***
	(.0159)	(.0123)	(.0075)
Peru	–.302***	.064***	–.029***
	(.0201)	(.0157)	(.0096)

(Continued)

Table B3
Continued

Country	Threshold 1	Threshold 2	Threshold 3
Austria	.159***	–.085***	–.349***
	(.0210)	(.0160)	(.0116)
Czech Republic	–.180***	.125***	–.056***
	(.0204)	(.0148)	(.0099)
Switzerland	–.039***	–.042***	–.206***
	(.0177)	(.0135)	(.0087)
Israel	–.066***	–.061***	–.243***
	(.0204)	(.0160)	(.0109)
Germany	.152***	–.080***	–.317***
	(.0221)	(.0165)	(.0119)
Colombia	–.404***	.056***	–.021***
	(.0187)	(.0145)	(.0088)
New Zealand	–.141***	.094***	–.009
	(.0222)	(.0161)	(.0103)
Denmark	–.178***	.116***	.015*
	(.0194)	(.0143)	(.0091)
Vietnam	–.523***	.053***	.131***
	(.0223)	(.0178)	(.0096)
Montenegro	–.229***	–.043***	–.109***
	(.0198)	(.0157)	(.0102)
Macao–China	–.352***	.078***	.145***
	(.0204)	(.0155)	(.0091)
Portugal	–.209***	.062***	–.099***
	(.0199)	(.0149)	(.0096)
Spain	–.039***	–.005	–.138***
	(.0162)	(.0123)	(.0077)
The Netherlands	–.120***	.107***	.128***
	(.0208)	(.0153)	(.0097)
Croatia	–.122***	.031**	–.102***
	(.0199)	(.0149)	(.0098)
Luxembourg	.079***	–.143***	–.303***
	(.0196)	(.0153)	(.0106)
Poland	–.127***	.065***	–.057***
	(.0204)	(.0149)	(.0100)
Sweden	–.194***	.073***	–.002
	(.0205)	(.0153)	(.0099)
Turkey	–.106***	–.203***	–.217***
	(.0197)	(.0162)	(.0106)
Slovak Republic	–.355***	.136***	.015*
	(.0210)	(.0152)	(.0100)
Hong Kong–China	–.273***	–.052***	.122***
	(.0215)	(.0163)	(.0091)
Serbia	–.286***	.065***	–.067***
	(.0199)	(.0148)	(.0098)

(Continued)

Table B3
Continued

Country	Threshold 1	Threshold 2	Threshold 3
Chinese Taipei	–.380***	.142***	.078***
	(.0189)	(.0140)	(.0089)
Finland	–.277***	.163***	–.022***
	(.0188)	(.0135)	(.0088)
Tunisia	–.143***	–.158***	–.254***
	(.0201)	(.0167)	(.0111)
France	.009***	–.054***	–.237***
	(.0216)	(.0153)	(.0107)
Chile	–.139***	–.035***	–.189***
	(.0191)	(.0147)	(.0096)
United Arab Emirates	–.283***	–.140***	–.154***
	(.0176)	(.0140)	(.0086)
Bulgaria	–.390***	–.047***	–.097***
	(.0202)	(.0156)	(.0096)
Slovenia	–.374***	.070***	–.070***
	(.0194)	(.0148)	(.0095)
Uruguay	–.332***	.030**	–.120***
	(.0202)	(.0155)	(.0101)
Hungary	–.414***	.154***	–.103***
	(.0205)	(.0145)	(.0100)
Greece	–.214***	.071***	–.111***
	(.0200)	(.0145)	(.0098)
Italy	–.285***	–.001	–.086***
	(.0160)	(.0122)	(.0075)
Romania	–.508***	–.010	–.123***
	(.0206)	(.0159)	(.0099)
Korea	–.386***	.087***	.182***
	(.0199)	(.0150)	(.0090)
Brazil	–.417***	.008	–.098***
	(.0168)	(.0128)	(.0079)
Malaysia	–.647***	–.008	.007
	(.0212)	(.0169)	(.0096)
Jordan	–.319***	–.316***	–.315***
	(.0187)	(.0160)	(.0101)
Thailand	–.627***	–.028**	.078***
	(.0207)	(.0162)	(.0089)
Qatar	–.240***	–.242***	–.233***
	(.0172)	(.0139)	(.0087)
Argentina	–.332***	–.034***	–.116***
	(.0193)	(.0147)	(.0095)
Indonesia	–1.107***	.100***	.167***
	(.0229)	(.0170)	(.0093)

Note. Standard errors in parentheses; countries in order by adjusted estimates of teacher's classroom management levels; ***, **, and *indicate statistical significance at 1%, 5%, and 10%, respectively.

Notes

[1]More details regarding the two sets of anchoring vignettes included in PISA can be found in the PISA 2012 Technical Report (OECD, 2014). As explained in this report, when designing the vignettes the PISA team studied the degree of clear interpretation that students gave to these sets of vignettes, in terms of the relative ordering that students gave to the different levels of classroom management and support reported in the vignettes. Results from field trials and the main survey showed that students gave a clearer interpretation to the vignettes capturing classroom management behaviors, which are the focus of this article. Other key assumptions to be considered when creating and developing vignettes are response consistency—that is, students use the same implicit standards to respond to self-reports and vignettes, and vignette equivalence—that is, vignettes are interpreted the same way across students. The literature has introduced statistical tests for these key assumptions (Bago d'Uva, Lindeboom, O'Donnell, & van Doorslaer, 2011; Kapteyn, Smith, van Soest, & Vonkova, 2011) and discussed how vignettes should be formulated and vignette data collected to minimize their violation (Grol-Prokopczyk, Verdes-Tennant, McEniry, & Ispany, 2015; Vonkova et al., 2017). We recommend, in line with He, Buchholz, and Klieme (2017), that PISA vignettes should be tested for these assumptions in the PISA context.

[2]Note that the original PISA 2012 data set assigned values 1 to *strongly agree*, 2 to *agree*, 3 to *disagree*, and 4 to *strongly disagree*. We changed the labels of the values to ease interpretation. This change does not affect our results.

[3]All but Question 4 attributed higher values of response to higher levels of teachers' classroom management skills.

[4]In Table B1 in Appendix B, we present raw percentages of responses for Question 2 presented above for each country and region participating in PISA. Descriptive statistics for Questions 1, 3, and 4 are available from the authors upon request.

[5]Descriptive statistics for the Low and Medium Vignettes are available from the authors upon request.

[6]A simple nonparametric approach, also introduced by King et al. (2004), could be used as an alternative to the parametric CHOPIT model. The idea under the nonparametric approach is to recode a respondent's self-assessment relative to the respondent's vignettes evaluations. Although the nonparametric approach can be used easily, it presents several disadvantages for our application. In particular, it has been shown to be problematic with models that include multiple covariates and when more than a few groups of countries are being compared, as in our case (see van Soest and Vonkova, 2014). Also, the nonparametric approach can be hampered by ties (respondents evaluate two naturally ordered vignettes the same way) or inconsistencies (respondents evaluate two naturally ordered vignettes with the reverse order). Finally, most of the applications in the literature using anchoring vignettes have adopted the parametric approach as we do in this article.

[7]It should be stressed this is not the usual case in the anchoring vignettes literature as it is often the case that only one assessment is available along with the vignettes questions.

[8]We also estimated models under an alternative assumption for identification where the variance of the error for the first question was set to be 1 and the variance of the errors for the rest of questions were allowed to be different and estimated by the model. This alternative assumption did not affect significantly the estimated results.

[9]We followed the CBEEE definition put forth by Bishop (1998) in order to develop our inclusion criteria. Using this definition, we further restricted our CBEEE indicator to countries where the exit exam is in mathematics and it is required for all students to graduate or receive a high school diploma. We primarily used information from TIMSS (Mullis et al., 2012) to identify countries that met our CBEEE criteria; however, for countries in our sample that were not participants in TIMSS 2011, we relied on country-level education system reports produced by UNESCO (en.unesco.org).

References

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2012). Age, health and life satisfaction among older Europeans. *Social Indicators Research*, *105*(2), 293–308. https://doi.org/10.1007/s11205-011-9882-x

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, *46*, 875–906. https://doi.org/10.3368/jhr.46.4.875

Bago d'Uva, T., van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, *17*, 351–375. https://doi.org/10.1002/hec.1269

Bishop, J. H. (1998). The effect of curriculum-based external exit exams systems on student achievement. *Journal of Economic Education*, *29*, 171–182. https://doi.org/10.1080/00220489809597951

Bonsang, E., & van Soest, A. (2012). Satisfaction with social contacts of older Europeans. *Social Indicators Research*, *105*(2), 273–292. https://doi.org/10.1007/s11205-011-9886-6

Buckley, J. (2008). *Survey context effects in anchoring vignettes*. New York University. Working paper. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.9907&rep=rep1&type=pdf

Buckley, J., & Schneider, M. (2007). *Charter schools: Hope or hype*? Princeton, NJ: Princeton University Press.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-related health. *Journal of Health and Social Behavior*, *52*, 246–261. https://doi.org/10.1177/0022146510396713

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispany, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, *52*, 1703–28. https://doi.org/10.1007/s13524-015-0422-1

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology*, *48*, 319–334. https://doi.org/10.1177/0022022116687395

He, J., & van de Vijver, F. J. R. (2016). The motivation-achievement paradox in international educational achievement tests: Toward a better understanding. In R. B. King & A. B. I. Bernardo (Eds.), *The psychology of Asian learners: A Festschrift in Honor of David Watkins* (pp. 253–268). Singapore: Springer. https://doi.org/10.1007/978-981-287-576-1_16

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*, 903–918. https://doi.org/10.1.1.333.4536

Kapteyn, A., Smith, J. P., & van Soest, A. (2007). Vignettes and self-reports of work disability in the US and the Netherlands. *American Economic Review*, *97*, 461–473. https://doi.org/10.1257/aer.97.1.461

Kapteyn, A., Smith, J. P., van Soest, A., & Vonkova, H. (2011). Anchoring vignettes and response consistency. RAND Working Paper WR-840. Retrieved from http://www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR840.pdf

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*, 567–583. https://doi.org/10.1017/S000305540400108X

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, *15*, 96–117. https://doi.org/10.1016/j.labeco.2006.11.001

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski von Davier & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis* (pp. 277–285). Boca Raton, FL: Taylor & Francis.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4

Mullis, I. V. S., Martin, M. O., Minnich, C. A., Stanco, G. M., Arora, A., Centurino, V. A. S., & Castle, C. E. (Eds.). (2012). *TIMSS 2011 encyclopedia: Education policy and curriculum in mathematics and science, Vols. 1 and 2*. Chestnut Hill, MA: IEA TIMSS & PIRLS International Study Center, Boston College. Retrived from http://timssandpirls.bc.edu/timss2011/encyclopedia-timss.html)

OECD. (2014). *PISA 2012 technical report*. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, *42*, 513–538. https://doi.org/10.1007/s00181-011-0530-8

Van Soest, A., & Vonkova, H. (2014). Testing the specification of parametric models by using anchoring vignettes. *Journal of the Royal Statistical Society Series A*, *177*, 115–133. https://doi.org/10.1111/j.1467-985X.2012.12000.x

Vonkova, H., Bendl, S., & Papajoanu, O. (2017). How students report dishonest behavior in school: Self-assessment and anchoring vignettes. *Journal of Experimental Education*, *85*(1), 36–53. https://doi.org/10.1080/00220973.2015.1094438

Vonkova, H., & Hrabak, J. (2015). The (in)comparability of ICT knowledge and skill self-assessments among upper secondary school students: The use of the anchoring vignette method. *Computers and Education*, *85*, 191–202. https://doi.org/10.1016/j.compedu.2015.03.003

West, M., Kraft, M., Finn, A., Martin, R., Duckworth, A., Gabrieli, C., & Gabrieli, J. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*, 148–170. https://doi.org/10.3102/0162373715597298

Authors

HANA VONKOVA is an Associate Professor at the Faculty of Education, Charles University, Myslikova 7, Prague, 110 00, Czech Republic; h.vonkova@gmail.com. Her primary research interests include measurement in education and health.

GEMA ZAMARRO is an Associate Professor at the Department of Education Reform, University of Arkansas, 219–B Graduate Education Building, Fayetteville, AR 72701; gzamarro@uark.edu. Her primary research interests include educational assessment and evaluation.

COLLIN HITT is an Assistant Professor at the Department of Medical Education, Southern Illinois University School of Medicine, 913 N. Rutledge St., Springfield, IL 62794; chitt47@siumed.edu. His primary research interests include medical education and survey methods.