Education Reform Faculty and Graduate Students Publications

Education Reform

6-1-2015

Comparisons of Student Perceptions of Teacher's Performance in the Classroom: Using Parametric Anchoring Vignette Methods for Improving Comparability

Hana Vonkova
Charles University in Prague

Gema Zamarro
University of Arkansas, Fayetteville, gzamarro@uark.edu

Vera DeBerg

Collin Hitt
University of Arkansas, Fayetteville

Follow this and additional works at: https://scholarworks.uark.edu/edrepub

Part of the Educational Assessment, Evaluation, and Research Commons, Educational Leadership Commons, and the Other Educational Administration and Supervision Commons

Comparisons of Student Perceptions of Teacher's Performance in the Classroom: Using Parametric Anchoring Vignette Methods for Improving Comparability

Hanka Vonkova, Gema Zamarro, Vera Deberg, and Collin Hitt*

Abstract
Self-reports are an indispensable source of information in education research but might be affected by reference group bias if the frame of reference (i.e. implicit standards), used to answer the questions, differs across students. The anchoring vignettes method was introduced, in other areas of social science, precisely to correct for this source of bias. However, studies that make use of this approach in education are rare and more research is needed to study its potential. This paper uses data from PISA 2012 to investigate the use of the parametric model of the anchoring vignettes method to correct for differential implicit standards in cross-country comparisons of student's perceptions of an important dimension of teacher quality: teacher's classroom management. Our results show significant heterogeneity in implicit standards across countries. We also show how correlations between countries' average teacher classroom management levels and external variables can be improved substantially when heterogeneity in implicit standards is adjusted for. We conclude that the anchoring vignettes method shows a good potential to enhance the validity and comparability of self-reported measures in education.

Key Words: Self-reported measures, teacher quality, reference group bias, anchoring vignettes methods, PISA study.

_____

*Vonkova: Charles University in Prague, Faculty of Education, Institute for Research and Development in Education, Myslikova 7, Prague 1, 110 00, Czech Republic (email: h.vonkova@gmail.com, hana.vonkova@pedf.cuni.cz); **Corresponding author:** Zamarro: University of Arkansas, Department of Education Reform, 219-B Graduate Education building, Fayetteville, AR 72701 (email: gzamarro@uark.edu: phone: 479-575-7024; fax: 479-575-3196); DeBerg: University of Arkansas, Department of Education Reform, 219-B Graduate Education building, Fayetteville, AR 72701 (email: vdeberg@uark.edu); Hitt: University of Arkansas, Department of Education Reform, 219-B Graduate Education building, Fayetteville, AR 72701 (email:cehitt@email.uark.edu).

1. Introduction

Student surveys are an indispensable source of information for education research. But self-reports can also be a flawed source of information. Students are often surveyed on topics, such as teacher performance in the classroom or the safety environment of their schools, and this type of student self-reports are increasingly being used to shape policy and personnel decisions. For example, student reports of teacher quality are being included as part of new teacher evaluation systems in several states in the U.S[1]. Students are also often surveyed about their own behavior: self-reports are the basis for the measurement of character traits and non-cognitive skills. Unfortunately, most education research ignores an important problem of such self-reports called reference group bias – a problem that limits the usefulness of surveys in making valid comparisons across students from different backgrounds, classroom settings, cultures or countries.

Reference group bias occurs when individuals have different frames of reference (i.e. implicit standards), they use to answer the question they are being asked. For example, when students are asked to rate the competencies of their teachers, the individual standard for teacher quality impacts the rating that the student assigns the teacher. For instance, the notion of what it means for a teacher to keep his/her class in order is most probably affected by cultural and classroom context and so, two students who face the same performance from their teachers may rate their teachers' classroom management skills differently.

---

[1] Up to date seven states in the U.S require student perception surveys as part of their teacher's evaluation systems. This is the case in Alaska, Georgia, Hawaii, Iowa, Massachusetts, New York, and Utah. In addition, five more states allow for student surveys to be part of teacher's evaluations. This is the case of Colorado, Connecticut, Mississippi, Missouri, and New Mexico.

1

Evidence of reference group bias has been found when comparing responses of students in different school environments within a country (see, e.g., West et. al., 2014) but it can be especially problematic in cross-country comparisons (see, e.g. Heine et al., 2002; Peng, Nisbett & Wong, 1997; Chen, Lee, & Steverson, 1995). For instance, a consistent finding in international education research is that questions related to students' attitudes toward learning have been shown to be positively correlated, at the student level, with achievement within countries, but negatively correlated across countries (Kyllonen & Bertling, 2013). A potential solution exists. The anchoring vignettes method was introduced in the social sciences by King et al. (2004) precisely to adjust for such heterogeneity in reporting behavior and obtain comparable responses across groups.

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 reference points 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 reporting behavior. Since its introduction, the anchoring vignettes method has been largely used in social-science research in areas such as health, work disability, life satisfaction, job satisfaction and satisfaction with contacts (see, e.g. Peracchi & Rossetti, 2012; Vonkova & Hullegie, 2011; Grol-Prokopczyk, Freese, & Hauser, 2011; Bago d'Uva et al., 2008; Angelini, Cavapozzi, & Paccagnela, 2012 and 2014; Kapteyn, Smith, & van Soest, 2007 and 2010; Kristensen & Johansson, 2008; Bonsang & van Soest, 2012).

Education research trails other fields in the use of the anchoring vignettes method, although this is just starting to change. Buckley & Schneider (2007) and Buckley (2008) used

anchoring vignettes for the comparison of parents' satisfaction measures in charter and public schools. Vonkova & Hrabak (2015) studied the use of anchoring vignettes for improving comparability of self-assessments of ICT knowledge and skills among upper secondary school ICT and non-ICT students. Vonkova, Bendl, & Papajoanu (2015) 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 socio-economic backgrounds. Additionally, Kyllonen and Bertling (2013) used data from the Programme for International Student Assessment (PISA) study in 2012 to showcase 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. More research is still needed, however, to study the potential of the anchoring vignette method for improving measures in education research.

Recognizing the potential of anchoring vignettes, the administrators of the renowned PISA study included vignettes in the student surveys that are administered alongside their tests of academic content knowledge. In this paper, 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, b) the use of the anchoring vignettes method for improving comparability of measures of teacher's classroom management levels across countries, and c) possible sources of such heterogeneity in reporting behavior. As far as we are

aware, this is a 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.

Insofar as this new approach improves our understanding of international differences in student-reported teacher quality, it also has potential to impact how student survey data is used in the United States. As mentioned, researchers and policymakers are increasingly interested on alternative measures of teacher quality and one candidate for such measures is the use of student perception surveys rating teacher practices (see, e.g., Berk, 2005; Bill & Melinda Gates Foundation, 2013; Glazerman et al., 2011). Students are first hand witnesses of what occurs in the classrooms and so they are in a privileged position to provide feedback on teaching practices. Teacher's classroom management is one of the dimensions often measured in students' perceptions surveys, and it was found to be one of the most predictive components of student achievement gains (see Ferguson & Danielson, 2014).

Naturally, policymakers will be interested in the factors that are associated with student-reported teacher quality. For example, they may examine which schools or districts have the highest student ratings of teacher skills. Our results show that correcting for heterogeneity in reporting behavior can be an essential step in this process. As mentioned, previous literature shows the paradoxical finding that, across countries, student reports of teacher quality are negatively related to average test scores – we replicate this finding using unadjusted scores. But once we adjust for heterogeneity in reporting behavior, by using the parametric anchoring vignettes method, the correlation becomes positive. Similarly, we show that, when adjusting scores using vignettes, the correlations between country-level policy variables and student-

reported teacher quality can be reversed. Generally speaking, correlation is weak (0.4) between the adjusted and unadjusted country-level ratings of teacher skills.

The rest of the paper goes as follows. Section 2 describes the PISA 2012 study and the information on teachers' classroom management skills and vignettes that we used for the analysis. Section 3 describes the intuition underlying the use of the anchoring vignettes method and the details of the parametric model for anchoring vignettes method as it is implemented in the analysis. Section 4 presents the resulted distribution of teacher quality measures based on teachers' classroom management behavior before and after adjusting for heterogeneity in reporting behavior. In this section, we also present evidence on the prevalence of such heterogeneity in reporting and a descriptive analysis of factors related with it. In addition, we also study whether correlations between teachers' classroom management skills and external variables improve after adjustments using the parametric model for anchoring vignettes method. Finally, Section 5 presents our conclusions and further discussion on the implications of our findings.

2. Data

This paper uses data from the PISA 2012 study which is based on standardized-test and survey data collected from over 485,000 students enrolled in public and private schools in 68 country-regions[2]. These data include measures of student aptitudes in the subject areas of reading, mathematics, and science, as indicated by performance on a multiple-choice assessment. Additionally, measures of student attitudes, learning experiences, demographics, and school

---

[2] These include 60 unique countries and the following regions within countries that took the PISA study individually: Connecticut (USA), Massachusetts (USA), Florida (USA), Chinese Taipei, Hong-Kong (China), Macao (China), Shanghai (China), Perm (Russia Federation).

organization and environment were collected, based on the survey responses of both students and principals.

In each participating country, schools are randomly selected to participate in the PISA assessment and survey. Although participating schools are selected randomly, a sampling frame is defined by PISA in order to select a sample of students that is representative of the total population of all 15-year-olds in each participating country. This allows for an easier comparison of students across countries, as the organization of grade levels tends to differ across participating countries.

While the initial PISA studies only collected data on student achievement, since 2000 PISA has administered a student survey in addition to the assessment. This survey asks students to report on statements by selecting among the following categories on a four-point Likert-type scale: "strongly agree," "agree," "disagree," and "strongly disagree". This scale is used to measure a variety of constructs, including measures of student attitudes regarding learning, the school and classroom environment, and teacher behavior in the classroom.

Since the introduction of these self-report scale items to the student questionnaire, diverse studies using PISA data noted inconsistencies in the predictive validity of some constructs in relation to achievement outcomes when using variation within country, across students, as opposed to across countries. For instance, student's scores on mathematics self-concept scales, based on items such as "I learn mathematics quickly"-"strongly agree", "agree", "disagree", "strongly disagree", were found to present positive correlations with mathematics achievement within countries. On the other hand, when the correlation was studied between country-mean mathematics self-concept and average country math achievement it turned out negative (see Kyllonen and Bertling, 2013). Although it is possible that such inconsistencies in relationships

across students within country and across countries hold in reality, it is also possible that these inconsistencies can be attributed to cross-country differences in reporting behaviors related to the use of scales for student's responses in these questions.

The PISA 2012 study aimed to address these observed inconsistencies through the introduction of anchoring vignettes. 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. Although each student participating in the PISA 2012 study completed the student survey questionnaire and assessment, not all students were asked to respond to the anchoring vignettes questions. This is because vignettes were included in only two of the three versions of the student survey questionnaire. However, vignette questions were randomly assigned to participating students within each school through the introduction of a rotation design for the student questionnaire. Therefore, although only a sub-sample of students received questionnaires that included vignettes, the responses of this sub-sample follow a distribution similar to those seen at the aggregate country-level. Our analysis is based on this sub-sample, which includes observations of more than 310,000 students in 68 region-countries.[3]

Student's Reports on Teacher Quality Measures in PISA 2012: This paper focuses on the study of teacher quality measures based on student's assessments on teachers' classroom behaviors. In this respect, the 2012 PISA student survey included four items that measure the classroom management levels of students' mathematics teachers. In particular, students were

---

[3] See Appendix 1 for a list of country names and abbreviations used in the 2012 PISA study.

asked to use the following four-point scale: 1."strongly disagree", 2."disagree", 3."agree", 4."strongly agree"[4], 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.[5]

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 towards 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 negative 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 typically ranked worse (with a lower numerical score) on the fourth item.

We also conducted a confirmatory analysis of the classroom management scale items and found that all items loaded onto one factor, though the loading of Question 4 was relatively weak

---

[4] Note that the original PISA 2012 dataset 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.

[5] It should be pointed out that all but the fourth question attributed higher values of response to higher levels of teachers' classroom management skills.

(at 0.45) compared to the other items (0.74 to 0.83)[6]. The internal reliability of the Classroom Management Scale is acceptably high with an overall Chronbach's alpha of 0.69. The responses to Question 4 are relatively weakly correlated with answers to the remaining items, as was previously suggested by our factor analysis. If Question 4 is omitted, the internal reliability of the scale increases to alpha = 0.77. In our main analyses below, we include all four questions in the classroom management scale-scores. As a robustness check, we removed Question 4, which did not affect our results.

That said, the international average scores on the classroom management scale are of only limited interest to us. The point of this study is that there are considerable differences between countries on the classroom management scale, and that there are also considerable cross-country differences in how the reporting scales are viewed. Table 2 shows raw percentages of responses for Question 2 presented above[7], for each country and region participating in PISA. While these questions aim to measure students' perceptions of teacher classroom management, comparing these raw percentages can be troublesome. As previously mentioned, this is because any differences in these ratings may be due to differences in how students interpret the reporting scale, rather than actual differences in teachers' classroom management skills. For instance, looking at Table 2 one would conclude that teacher's classroom management skills are much better in the U.S. than in certain high performing European countries such as the Netherlands. For example, 30% of students in the U.S. strongly agreed with the statement "My teacher keeps the class orderly", while that proportion was only 17% in the Netherlands. While it is possible

---

[6] Results available from the authors upon request.

[7] Descriptive statistics for Question 1, 3 and 4 are available from the authors upon request.

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 reporting styles.

In order to be able to correct for discrepancies on reporting behavior, the PISA 2012 study included vignettes related to hypothetical teacher's classroom management behaviors. Students were then asked to rate the following three hypothetical scenarios using the same four-point scale (1."strongly disagree", 2."disagree", 3."agree", 4."strongly agree") 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 3 shows the average answers to each vignette, across the entire PISA sample. As discussed below, the responses to the vignettes differ substantially – which, again, is the primary impetus for our study. It is interesting and reassuring to see that the average ratings and item scores for the vignettes follows 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.

Table 4 shows descriptive average responses across countries for the high vignette question presented above.[8] As it can be seen in this table, there is considerably heterogeneity across countries in the responses to these hypothetical scenarios. Coming back to our previous illustrative example comparing the U.S. and The Netherlands, we observe that Dutch students tend to be more demanding when evaluating higher teacher management skills. For instance, if we look at the results for the high vignette in Table 4 we see that 53% of American 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, Smith and Van Soest (2007), when studying self-reports of work disability in the adult population. In the next section, we describe how we make use of this information, on differential reporting behavior contained in responses to the vignettes, to correct student's perceptions of their teacher's performance in the classroom.

3. **Methods**

   3.1 **Intuition Underlying the Use of the Anchoring Vignettes Method**

As explained above, we are concerned that comparisons of self-reported measures across countries might be biased due to differential use of reporting scales by students in different countries. The anchoring vignettes method proposes to reduce this bias using information of student's responses to hypothetical scenarios or "vignettes". Since all respondents are asked to evaluate the same scenarios represented in the vignettes, differences on evaluations to the vignette questions will be interpret as heterogeneity in the use of reporting scales. This section presents the intuition underlying the use of the anchoring vignettes method for correcting self-

---

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

reported student's perceptions of teacher's classroom management abilities. The basic idea underlying this method in the context of comparisons of teacher's classroom management skills can be illustrated studying the case of two hypothetical countries. For this description, we follow work by Kapteyn et al. (2007), Vonkova (2013), and others, who explained the basic idea of the anchoring vignettes method for the comparison of self-reported health in two hypothetical countries that differ in their use of reporting scales.

Suppose your aim is to compare teachers' classroom management levels in two hypothetical countries, country A and country B, using answers to the following question in a student's questionnaire: "To what extent do you agree that your teacher is in control of your classroom?" with response options in a four-point Likert-like scale (strongly disagree, disagree, agree, and strongly agree). If individuals in the two countries differ in their use of reporting scales to answer this question, we would be concerned that observed differences between the two countries would be in part reflective of this different uses of the scale. Figure 1 which presents objective and reported levels of classroom management in the two countries illustrates this problem. The curves presented in this figure represent the actual unobserved distribution of teacher's classroom management skills in each country. The distribution for country A is shifted to the left compared to the distribution for country B implying that the objective level of teacher's classroom management is better in country B. However, researchers do not observe the actual distribution of teacher's classroom management levels. Instead they observe the proportion of individuals reporting in each category, from strongly disagree to strongly agree. That is, what is observed is the area of the objective distribution that lies between the thresholds for each category. As Figure 1 shows, the "strongly agree" and "agree" categories are more frequent in country A, while categories "strongly disagree" and "disagree" are more frequent in

country B. In this sense, comparisons of the student's reported teacher's classroom management levels would lead to the conclusion that teacher's classroom management is better in country A. This is the opposite conclusion one would make based on the objective levels of teacher's classroom management. This wrong conclusion is due to the difference in students' scale usage across the two countries. Students in country A have lower standards (are less strict / more positive) when evaluating their teacher's behavior. In Figure 1, this can be observed by comparing student's reports in both countries for a given level of objective teacher's classroom management quality represented by the horizontal dashed line. While a student from country A will agree that his/her teacher is in control of his/her classroom, a student from country B will report he/she strongly disagrees.

The anchoring vignettes method can help us solve this problem by using and individual's response to several vignettes to adjust their final scale score. Students in both country A and country B are asked to evaluate not only their actual teacher's classroom management level, but also the quality of classroom management of the hypothetical teachers described in the vignettes. All students are given the same vignettes, so any observed differences in their evaluations may be interpreted as differences in the usage of reporting scales. For instance, students may be asked to evaluate a vignette describing a classroom management situation corresponding to the objective level depicted by the horizontal dashed line in Figure 1. While students from country A will agree that the hypothetical teacher is in control of his/her classroom, students from country B will strongly disagree. This will tell us that students in Country A have lower standards, as they are less strict to evaluate the same hypothetical scenario than students in Country B. We then use this information about heterogeneity in scale usage, identified by the evaluation of the vignette, to adjust students' evaluations of their teacher's classroom management. In this process,

the scale of one of the countries could be chosen as the benchmark scale and evaluations of the other country could then be expressed in terms of this benchmark scale. The comparison of the two countries would then be corrected and we would rightly conclude that teacher's classroom management levels are better in country B.

Generally, in the case where the actual teacher's classroom management levels in country B are better than in country A (B>A), the comparison of students' reported teachers' classroom management levels could lead us to the opposite conclusion than the comparison based on unobserved actual levels (b<a) or, could lead us to find no differences in the observed students' reported levels (a=b) or, could lead us to maintain the same conclusion than the one based on unobserved actual levels (b>a). However, even if the comparison of students' reported teacher's classroom management levels leads to the same conclusion, the same one that would be obtained based on unobserved actual teacher's classroom management (b>a), we still would like to know if our comparison based on students' reports leads to underestimated differences across countries (thresholds of country B are, for example, shifted to the right compared to thresholds of country A but the conclusion b>a still holds) or overestimated (thresholds of country B are, for example, slightly shifted to the left in comparison to the thresholds for country A). The next section presents, in better detail, the econometric model we use to correct for student's differential use of scales when evaluating their teachers.

3.2 The Parametric Model of the Anchoring Vignettes Method

This paper makes use of the parametric model of the anchoring vignettes method to correct students' reports of their teacher's classroom management skills. This method is equivalent to a Compound Hierarchical Ordered Probit (CHOPIT) model as introduced by King et al. (2004). In this section we present an econometric model specification for the case of having four student's

assessments related to their teacher's classroom management skills along with three vignettes evaluations, which are all in a four-point Likert-like scale, as it is the case in our data. 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. The model consists of two components: the students' classroom management assessments for their teachers and the vignettes component.

For the students' assessment component, let us denote perceived teacher's classroom management skill 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}:

$$Y_{qi}^* = X_i' \beta + \varepsilon_{qi}$$
$$\varepsilon_{qi} \sim N(0, \sigma_q^2)$$

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 behavior related to classroom management skills on the 4-point ordinal scale (1=strongly disagree, 2=disagree, 3=agree, and 4=strongly agree).

Then, the students' 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}: $Y_{qi} = j \leftrightarrow \tau_i^{j-1} < Y_{qi}^* \leq \tau_i^j \quad j = 1,2,3,4 \quad (1)$

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). In particular, in our model they are allowed to vary with student's characteristics X_i in the following way:

$$\tau_i^1 = X_i^{'}\gamma^1$$
$$\tau_i^j = \tau_i^{j-1} + exp(X_i^{'}\gamma^j); j = 2,3 \quad (2)$$
$$\tau_i^0 = -\infty, \tau_i^4 = \infty$$

Where γ^j are vectors of unknown parameters. In our case X_i denotes a given country-area for student i. 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^1, 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_{iv}^* the teacher's classroom management level described in vignette $v = 1,2,3$ as it is perceived by student i, and assume:

$$Z_{iv}^* = \phi_v + \varsigma_{iv}$$
$$\varsigma_{iv} \sim N(0,\sigma_v^2)$$

Where the parameter ϕ_v captures the actual level of classroom management described in vignette v and ς_{iv} 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 it is observed are the actual ordered vignettes evaluations Z_{iv} on a four-point scale:

$$Z_{iv} = j \leftrightarrow \tau_i^{j-1} < Z_{iv}^* \leq \tau_i^j \quad j = 1,2,3,4 \quad (3)$$

$$\tau_i^0 = -\infty, \tau_i^4 = \infty$$

Note that the thresholds (τ_j) are assumed to be the same as in the first component of the CHOPIT model described in (1). Then, equations (1), (2), and (3) represent the version of the CHOPIT model we estimate. However, 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 U.S. 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 ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$)[9]. Although it is uncommon to find in the literature a CHOPIT model, like ours, including multiple assessments, variance parametrizations are often found for the case of the CHOPIT model containing one assessment along with vignette information (see, e.g. Angelini et al., 2012; Bonsang & van Soest, 2012). Note that the rest of parameters are left unrestricted and are estimated by Maximum Likelihood.

As an alternative to the parametric anchoring vignettes method, described above, one could consider a simple nonparametric approach as described by King et al. (2004) and used by Kyllonen & Bertling (2013) and Vonkova & Hrabak (2015). This simple approach would consist on relating student's assessments relatively to his/her vignettes' evaluations. For example, in the case of having three vignette evaluations and only one assessment, we would create a new corrected student's assessment variable such as it takes value one if the student's assessment is less than the evaluation of the first vignette, two if the student's assessment equals the evaluation of the first vignette, three if the student's assessment is in between the evaluations of the first and

---

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

second vignettes, and so on until value seven corresponding to the student's assessment being greater than the evaluation of the third vignette. Although this approach has the advantage of being very simple and easy to compute, the parametric anchoring vignettes method has several advantages over this simple approach (see, van Soest and Vonkova, 2014).

Firstly, the simple nonparametric approach could lead to results that are difficult to interpret in the case of ties and inconsistencies, for instance, if vignettes evaluations do not preserve a natural order. If this is the case, one needs to make additional assumptions in this nonparametric approach to build the corrected assessments described above. Note that this is not a problem in the parametric anchoring vignettes model because the error terms would explain any violation of the natural ordering. Secondly, the parametric vignettes model allows us to directly estimate the effect of other relevant student's characteristics both in the levels of adjusted assessments and in the thresholds (i.e. country-area effects in our case). Thirdly, the parametric anchoring vignettes model, through the chosen parametrization, allows us to choose the benchmark scale of a reference group (student's in the U.S. in our case) and express adjusted assessments of other groups in reference to the scale of this reference group. In addition, the adjusted assessments are then expressed on the same scale as the unadjusted assessments (i.e., the four-point Likert scale). Finally, estimation under the nonparametric approach requires more data. This is so because both assessments and vignettes' evaluations must be collected for all students in order to construct the corrected student's assessment variable under the nonparametric approach. On the other hand, the nonparametric approach has the advantages that technics like the one developed by King and Wand (2007) could be used to better selecting vignettes based on how well they can extract information on differential use of scales. Therefore, if the aim of the study is to do a selection of vignettes which could provide the best information,

then the nonparametric approach may be preferred. However, given the advantages of the parametric approach described above, we believe this is the most suitable approach for getting a better understanding of the properties of cross-country comparisons on student's assessments of teacher's performance in the classroom and so, we decided to follow this approach in this paper.

4   **Results**

   4.1 Students' perceptions on teacher's classroom management performance across PISA countries

This section presents our estimated thresholds and country effects using Maximum Likelihood methods on the CHOPIT model as described in section 3.2 above.

   ### 4.1.1   Heterogeneity in the use of reporting scales

Our first analysis has to do with the amount of heterogeneity on the use of reporting scales by students in different country-areas, when evaluating the classroom management skills of their teachers. In this respect, Figure 2 shows geographic differences in estimated thresholds values where countries are divided into four categories, or quartiles, based on where each country's threshold values stand in the overall distribution of values for a given threshold[10]. Light grey indicates lower threshold values, while darker grey or black indicate higher threshold values. By looking at the results presented in this figure, we observe that there are significant differences in the estimated thresholds across countries. This suggests considerably amount of heterogeneity in the use of reporting scales across countries. In particular, looking at these figures we can classify countries according to the levels of their estimated thresholds in the following way:

   a) *Low Threshold 3*: One could say that students in these countries have lower standards or that they are more optimistic about their teachers' classroom management skills as they tend to label a given level of teacher's classroom management with the highest end-point

---

[10] See Appendix B, Table B.1 for the exact estimated values for each country.

of the scale ("Strongly Agree") more often than students in other countries. As shown in Figure 2, 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.

b) *High Threshold 3:* Students in countries with higher threshold 3 values can be considered to be more demanding, or to have higher standards, and be less optimistic when evaluating their teacher's classroom management skills. This is so because students in countries with higher threshold 3 values are more likely to use the scale option "Agree" rather than the option "Strongly Agree," when classifying a given teacher classroom management behavior. As illustrated in Figure 2, countries with lower estimated threshold 3 values are located in Asia (e.g. Shanghai-China and Korea), North America (i.e. U.S), Northern and continental Europe (e.g. The Netherlands, Norway and Denmark), and the continent of Australia (e.g. New Zealand and Australia).

c) *Low Threshold 2:* The second threshold distinguishes between the reporting categories "agree" and "disagree" therefore, differences in estimated values in this threshold would be of special interest if there were considerations for changing the response scale to be a two-point reporting scale. Estimated threshold 2 results show similar patterns that the ones discussed for threshold 3 above. Lowest threshold 2 values are observed in countries in the Middle East, Western Asia and the Balkans. These countries include among others Indonesia, Thailand, Malaysia, Jordan, Romania, Kazakhstan, Viet Nam, and Qatar.

d) *High Threshold 2:* Students in Shanghai- China, Continental Europe (e.g. Germany, Netherlands), U.S., Northern Europe (e.g. Norway and Denmark), United Kingdom,

Ireland, Iceland, Eastern Europe (e.g. Latvia, Czech Republic, Poland), South of Europe (e.g. Spain, France), and the continent of Australia (e.g. Australia) present the highest estimated threshold 2 values (i.e. tend to use more the reporting category "disagree" than "agree" for a given level of their teacher's classroom management skills).

e) *Low Threshold 1:* Estimated threshold 1 values, representing the cutoff point for responding "disagree" versus "strongly disagree", also vary significantly across countries. Some provinces in China (e.g. Macao-China), other Asian countries like Indonesia, Malaysia, Thailand or Viet Nam, countries located in the Middle East (e.g. Jordan), Eastern European countries (e.g. Romania, Bulgaria and Hungary) and South American countries like Brazil or Colombia present the lowest estimated threshold 1 values.

f) *High Threshold 1:* Countries with higher values of threshold 1 tend to use "strongly disagree" more than "disagree" and so they can be seen as having higher standards or being more pessimistic as they tend to use the lowest end-point of the scale more. The countries with the highest threshold 1 values include the U.S., Shanghai-China, United Kingdom, Ireland, Israel, and countries in Continental Europe (e.g. Austria, Germany, France, and Luxembourg).

After analyzing the distribution of each estimated threshold separately it is also relevant to study how the estimated thresholds are correlated to each other. For instance, by studying how thresholds 1 and 3 correlate we will be able to determine to what extend countries tend to use the end-points of the scale more, classifying a shorter range of teachers' classroom management levels using the mid-point categories of the scale (high threshold 1 and low threshold 3) or, if they prefer to use mid-points (low threshold 1 and high threshold 3), classifying a large range of

21

values of teachers' classroom management skills in the mid-point categories of the scale. Looking at the relationship between the estimated threshold 1 and 3 in our sample of countries we observe that the correlation is not very strong (correlation of the order of 0.46), suggesting the existence of both countries using more end-points and countries using more mid-points. Figure 3 presents a comparison of the estimated thresholds 1 and 3.

g) *High Threshold 1 and Low Threshold 3:* As presented in Figure 3 countries in this group include, for example, Lithuania, Luxembourg, Turkey, Tunisia, Chile, Iceland, Switzerland and Costa Rica. Students in these countries tend to use the end-points "strongly agree" and "strongly disagree" more than students in other countries, when evaluating their teachers' management skills.

h) *Low Threshold 1 and High Threshold 3*: As presented in Figure 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 Viet Nam. Students in these country-areas tend to use more the mid-points "agree" and "disagree", when evaluating their teachers, than students in other countries.

In addition, it is also of interest to study whether there are countries which estimated thresholds are all high and so, whose reporting scale is shifted to the right. Students in this type of 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 which estimated thresholds are all low, indicating their reporting scale is shifted to the left. That is, students in these countries would have lower standards or be generally more optimistic when evaluating their teachers.

i) *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. The U.S., the American states of Florida, Massachusetts, Connecticut, the United Kingdom, Iceland, and to some extend also Austria are examples of other country-regions in this group. Students in these countries tend to use the reporting category "strongly agree" for a narrower range of teachers' classroom management levels while they use the reporting category "strongly disagree" for a wider range of levels. They are in general more pessimistic when evaluating their teachers.

j) *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. Students in these countries show a general optimism when evaluating their teachers' classroom management behavior. They tend to use the reporting category ("strongly agree") for a wider range of classroom management levels and the reporting category ("strongly disagree") for a narrower range of teachers' classroom management levels.

Finally, it should be also 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.

4.1.2 Adjusted vs. unadjusted levels of teachers' classroom management skills

Previous section showed a great amount of heterogeneity among students' use of reporting scales across different countries. We found countries whose students preferred to use the end-points of

the scale when evaluating their teachers, countries whose students preferred the use of mid-points, countries whose scale was shifted to the right, countries whose scale was shifted to the left, along with other countries that did not use the scale in any extreme manner. Given this 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.

Figure 4 presents the adjusted and unadjusted distributions of teachers' classroom management skills for a selected group of countries. Both distributions are predicted using the CHOPIT model estimates[11]. For the adjusted distribution, which results are shown on the right of Figure 4, we chose the U.S. scale (i.e. the estimated thresholds for the U.S.) as our reference scale and expressed the distribution for other country-areas as if they used the scale of reference instead of their own estimated thresholds. The numbers preceding the adjusted distribution represent the order of the country in a ranking based on the adjusted average teachers' classroom management skills, such as the lowest numbers represent top positions. The unadjusted distribution, which results are shown on the left of Figure 4, is based on predictions when using each countries own estimated thresholds. Also in this case countries were ordered by decreasing averaged predicted level of teachers' classroom management skills. The predicted unadjusted distributions of teachers' classroom management skills are very similar to the ones one would obtain looking at the raw data distribution. This result is reassuring as it is an indication that our estimated CHOPIT model performs well.

---

[11] Alternatively one could use a regular ordered probit for obtaining the unadjusted distribution. Results based in this approach are similar to the ones presented above, although the CHOPIT models does a slightly better job approximating the observed distributions of student's assessments.

As illustrated in Figure 4, we observe large differences between the ranking of countries when based on the adjusted versus unadjusted distribution of teachers' classroom management skills. This result is not surprising given the observed differences in scale usage across countries discussed above. In particular, some countries significantly improve their position in the ranking when based on adjusted average teachers' classroom management skill. This is the case for several countries, including Norway, the Netherlands, Denmark, New Zealand, Australia, the U.S., and Great Britain. For example, Norway moved up from 57th position to 22th after the adjustments. In contrast, other countries' position considerably worsened after the adjustments. This is the case, for example, of Jordan, Indonesia, Romania, Malaysia, and United Arab Emirates, among others. For instance, in the extreme case, Jordan moved from 6th position to 64th position after adjustments.[12]

4.2 Determinants of heterogeneity in reporting behavior and variation in teachers' classroom management performance across PISA countries

Previous sections showed a significant amount of heterogeneity in the students' use of response scales across countries, and the importance of correcting for it to be able to arise at more comparable measures of students' perceptions on teacher's classroom management performance. The remaining question is whether our adjustments, based on the parametric anchoring vignettes method, move us closer to the actual levels of teachers' classroom management skills. In this section we try to shed some light on this question by studying how heterogeneity in the use of response scales and student's perceptions on teacher's performance in the classroom covary with other information at the country level. The results of the simple specifications presented in this

---

[12] The full set of country-average estimates and rankings can be found in Appendix B, Table B.2.

section are not meant to provide a complete picture of the determinants of differences in reporting styles or actual classroom management quality across countries, and should not be interpreted as causal. However, we still believe these results are of interest, as they allow us to study whether we find that reporting styles and student's perceptions on classroom management covary in a plausible manner with external variables and how these correlations are affected by our adjustments using the parametric vignettes method.

First, we study correlations of the heterogeneity in student's use of reporting scales and country-specific variables, such as public expenditure in secondary education, per capita gross domestic product[13] and percentage of private schools among the country's participating schools in the PISA study. In addition, we also introduced a dummy variable indicating whether the country had a curriculum-based external high school exit exam (CBEEE)[14]. Many countries have adopted CBEEEs as a means to ensure that the high school diploma serves as an indicator of basic skills. However, another driving force behind the use of CBEEEs is the idea that they will motivate students to work harder and thus, increase the overall academic performance of all

---

[13] Information for these variables was obtained from the World Bank data indicators (http://data.worldbank.org). The GDP per capita is PPP adjusted and measured in constant 2011 international Dollars. Public expenditure in secondary education is measured as a percentage of GDP and per capita and corresponds to the year 2010. Note that we lost a total of 26 countries (areas) in the analysis for which we could not find comparable GDP or public expenditure information. These countries (areas) include: ALB, ARE, BRA, CAN, CRI, DNK, GRC, HRV, KAZ, LIE, MAC, MNE, QAT, QCN, QRS, QUA, QUB, QUC, ROU, RUS, TAP, TUN, TUR, URY, VNM.

[14] We followed the CBEEE definition put forth by Bishop (1997) in order to develop our inclusion criteria. Under Bishop's definition, an exit exam is considered to be a CBEEE if it meets the following criteria: Holds real consequences for students, measures achievement using a central, externally defined standard, is organized by subject area and administered after the student has completed the appropriate course sequence, covers a large portion of content within the subject area, and is mandatory of all secondary school students. 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 provided from the Encyclopedia of the TIMSS 2011 assessment to identify countries that met our CBEEE criteria; however, for countries in our sample that were not also participants in TIMSS 2011, we relied on country-level education system reports produced by UNESCO.

students in an education system. Although not conclusive, there is some evidence of the potential effectiveness of exit exams at increasing student performance. For instance, using data from the 1994 TIMSS study, Bishop (1998, 2000) found that higher levels of performance on TIMSS in both science and math were more likely in countries that had a CBEEE in the subject. Given this potential, we study whether the existence of a CBEEE is related with different student standards for evaluating their teachers. Under the hypothesis that exit exams help raise standards and motivate students to work harder, one would think that they could make students demand more of their teachers. However, CBEEEs place consequences directly on the individual student, thus transferring much of the responsibility away from the teacher and to the student. This shift in responsibility can possibly change students' vision of the teacher, causing them to view the teacher as important in helping them prepare for the exit exam, but recognizing that their performance on the assessment is an individual responsibility. In this sense, students may be more likely to hold one another accountable for how they behave in the classroom, and thus be less critical when evaluating teachers' classroom management.

Results of these regressions are shown in Table 5. Our results show that students from countries with higher levels of public expenditure per student tend to have higher standards for their teachers' classroom management skills, as we observe a significant and positive correlation with thresholds 2 and 3. Note this is still the case even when we control for country's wealth, introducing GDP per capita measures. On the other hand, countries with higher proportion of private schools participating in the PISA study do not present higher or lower levels of the thresholds, once we control for public expenditure in education and GDP. Finally, with regards to CBEEEs, we see that the variable measuring compulsory exit exams in mathematics has a negative and significant effect and it is a statistically significant predictor of thresholds 1 and 2.

The negative coefficient for math exit exam indicates that countries with mandatory CBEEEs in math tend to have lower threshold values. In this sense, we can say that students in countries with compulsory exit exams in math present lower standards and are less critical and demanding when evaluating their teachers.

After studying potential sources of a country's heterogeneity in the use of scale, an important remaining question is whether adjustments based on the parametric vignettes method lead us to an improvement in our measures of teachers' classroom management skills. To do so, we also studied the correlations of country average levels of teachers' classroom management skills, before and after adjustments for heterogeneity in the use of reporting scales, and country level variables. 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 6 shows simple correlations among adjusted and unadjusted average teachers' classroom management skills and average math and reading test scores at the country level. Again, it is worth noting that we recoded the responses to questions 1 to 3 so the higher the estimated country effects the higher the level of teachers' classroom management skills. Keeping this in mind, it is very interesting to observe that unadjusted measures of teachers' 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 levels of average teachers' management skills are those that have higher average performance on reading and math. These results are totally reversed when we look at correlations based on adjusted teachers' classroom management measures. In this

case, as one would expect, we observe that countries with higher levels of teachers' classroom management skills are those with higher average math and reading scores.

We also obtained correlations of adjusted and unadjusted teachers' classroom management skills and other external country level variables. These results are presented in Table 7. An interesting pattern is observed in this table. Average teachers' 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. This correlation could be, however, just a representation of the higher pessimism levels in grading teacher's performance found in these countries, as it was discussed above. In fact, once we control for such 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 no 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 capital 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 CBEEEs, with either unadjusted or adjusted teacher's classroom management skills measures. These results stress the importance of correcting for potential differences in differential scale usage when making cross-country comparisons.

5 Conclusions

Self-reports are a central source of information for education research. Data from student surveys is increasingly being used to inform personnel and policy decisions. 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 problem has been called reference group bias. 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, this approach is relatively new in education and more research is needed to study its validity.

In this paper, we use data from PISA 2012 to study the heterogeneity in student's assessments of teacher performance. In particular we examine differences in reporting behavior across countries. We use a parametric anchoring vignettes method as a way to correct for this heterogeneity in reporting behavior. 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.

Increasingly, in the United States in particular, teacher's classroom management skills are one of the dimensions often measured in student perception surveys. Previous research has found this dimension of teacher quality to be predictive of student achievement gains, and as a result many policymakers are calling for student reports to be used in official teacher evaluations. However, it is possible that student standards of teacher quality are impacted by social or classroom conditions that are outside of a teacher's control. Our analysis of PISA 2012 was designed to examine whether there was truly evidence of this problem, and whether new methods could be used to address it. To our knowledge this is the first study of the use of the parametric anchoring vignettes method to correct student's perceptions on a dimension of teacher quality.

Our results show significant differences between the adjusted and unadjusted distributions of teachers' classroom management skills across countries. These are a result of differences in scale usage that are found across countries. For instance, countries like Norway, the Netherlands, Denmark, New Zealand, Australia, United Kingdom or the U.S., highly improve their relative position in the ranking of teachers' classroom management skills across countries after adjusting for heterogeneity in scale usage. This suggests that these are countries that have higher standards for judging teacher behavior (they tend to perceive a given level of teachers' classroom management worse than other countries) and their actual classroom management is better than they initially reported. On the other hand, there are countries whose position highly worsens after such adjustments. Countries in this group are, for example, Jordan, Indonesia, Romania, Malaysia and United Arab Emirates. We also show that these differences in scale usage might be geographically related – for example, students in some Northern and Central European countries tend to have higher standards while 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. So too are the apparent associations between student-reported teacher quality and certain policy variables. 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.

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 towards more intuitive values lead us to think that these adjustments are changing the results in the right direction. Therefore, we conclude, that the parametric anchoring vignettes method shows a good potential to enhance the validity and international comparability of self-reported measures in education.

From the point of view of public policy, our results warn of the potential for false inferences when exploring factors associated with student-reported teacher quality measures, due to the differential use and interpretation of the scales in the given questions. Our findings suggest strongly that any student surveys of teacher quality should include anchoring vignettes. Our future research will further study the presence of reference group bias in student reports of teacher quality, within a given country, across schools with different compositions, and the potential of anchoring vignettes methods to adjust for such bias.

References

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2012). Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics*, 76, 643-666. Doi: 10.1111/obes.12039

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2014). Age, health and life satisfaction among older Europeans. *Social Indicators Research*, 105, 293-308. Doi: 10.1007%2Fs11205-011-9882-x

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. Doi: 10.1002/hec.1269

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17, 48-62. (see www.iselt.org/ijthle)

Bill and Melinda Gates foundation (2012). *Asking students about teaching: Student perception surveys and their implementation*. MET project Policy and Practice Brief.

Bill and Melinda Gates foundation (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. MET project Policy and Practice Brief.

Bishop J.H. (2000). Curriculum-based external exit exam systems: Do students learn more? How? *Psychology Public Policy and Law,* 6, 199-215. Doi: 10.1037/1076-8971.6.1.199

Bishop, J. H. (1998). The effect of curriculum-based external exit exams systems on student achievement. *The Journal of Economic Education,* 29, 171-82. Doi: 10.1080/00220489809597951

Buckley, J. (2008). *Survey Context Effects in Anchoring Vignettes*. New York University. Retrieved from: http://polmeth.wustl.edu/media/Paper/surveyartifacts.pdf

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

Bonsang, E., & van Soest, A. (2012). Satisfaction with social contacts of older Europeans. *Social indicators research,* 105, 273-292. Doi: 10.1007/s11205-011-9886-6

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of

    rating scales among East Asian and North American students. *Psychological Science*, 6,

    170–175. Doi: 10.1111/j.1467-9280.1995.tb00327

Fergurson, R. F., & Danielson, C. (2014). *How Framework for Teaching and Tripod 7Cs*

    Evidence Distinguish Key Components of Effective Teaching. Designing Teacher

    Evaluation Systems. New Guidance from the Measures of Effective Teaching Project.

    Kane, J.R., Keer, K.A., & Pianta R.C. editors. Jossey –Bass, Chapter 4, 98-143.

Glazerman, S., Goldhaber, D., Loeb, S., Staiger, D., Raudenbush, S., & Whitehurst, G. J. (2011).

    Passing muster: Evaluating teacher evaluation systems. Washington, DC: The Brookings

    Institution.

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, 246-241. Doi: 10.1177/0022146510396713

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. Doi:10.1.1.333.4536

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An*

    experimental evaluation. NBER Working Paper 14607.National Bureau of Economic

    Research.

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. Doi:

    10.1257/aer.97.1.461

Kapteyn, A., Smith, J. P., & van Soest, A. (2010). Life satisfaction. In E. Diener, J. E. Helliwell, & D. Kahneman (Eds.), *International differences in well-being* (pp. 70–104). Oxford: Oxford University Press.

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. Doi: 10.1017/S000305540400108X

King, G., & Wand, J. (2007). Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes. *Political Analysis,* 15, 46-66. Doi: 10.1093/pan/mpl011

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15, 96–117. Doi: 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*, *Chapter 12*, (pp. 277–285).

McCaffrey, D. F, Lockwood J.R., Koretz, D., Thomas, A. L., & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29, 67-101. Doi: 10.3102/10769986029001067

Peng, K., Nisbett, R. E., & Wong, N. Y. C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods*, 2, 329–344. Doi: 10B2.989X

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

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

Vonkova, H. (2013). Subjektivní hodnocení problémů s pohybem: Užití parametrického modelu metody ukotvujících vinět (Subjective assessment of problems with moving around: Use of the parametric model of the anchoring vignette method). *Orbis Scholae*, 7(1), 49-66.

Vonkova, H., Bendl, S., Papajoanu, (2015). O. How students report dishonest behavior in school: Self-assessment and anchoring vignettes. *Journal of Experimental Education*, forthcoming.

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. Doi: 10.1016/j.compedu.2015.03.003

Vonkova, H., & Hullegie, P. (2011). Is the anchoring vignettes method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society: Series A*, 174, 597-620. Doi: 10.1111/j.1467-985X.2011.00704.x

West, M., Kraft, M. A., Finn, A. S., Martin, R., Duckworth, A. L., Gabrieli, C. F. O. and Gabrieli. J. D. E. (2014). "Promise and Paradox: Measuring Students' Non-cognitive Skills and the Impact of Schooling". Cambridge, MA: Center for Education Policy Research. Retrieved from http://www.gse.harvard.edu/cepr-resources/files/news-events/cepr-promise-paradox.pdf

Figure 1: Comparison of Classroom Management in Two Hypothetical Countries with Different
Usage of Scale

country A

1 strongly disagree 2 disagree   3 agree   4 strongly agree

country B

1 strongly disagree   2 disagree   3 agree   4 strongly agree
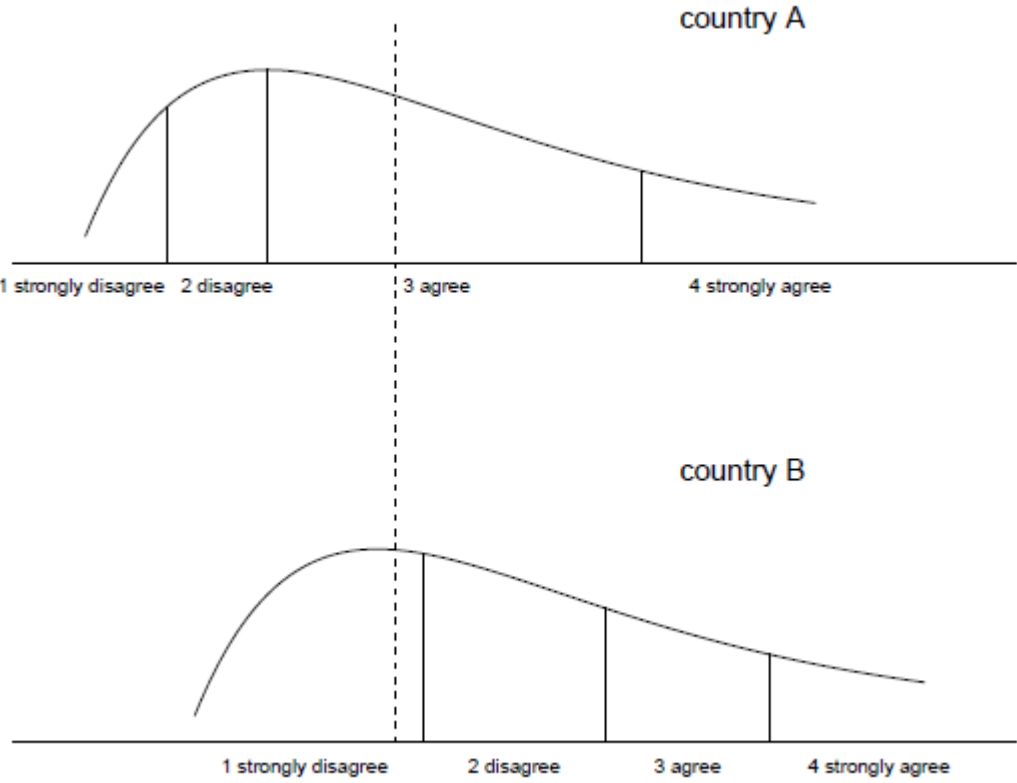
Figure 2: Geographic Distribution of Estimated thresholds using CHOPIT model

Threshold 1

Threshold 2
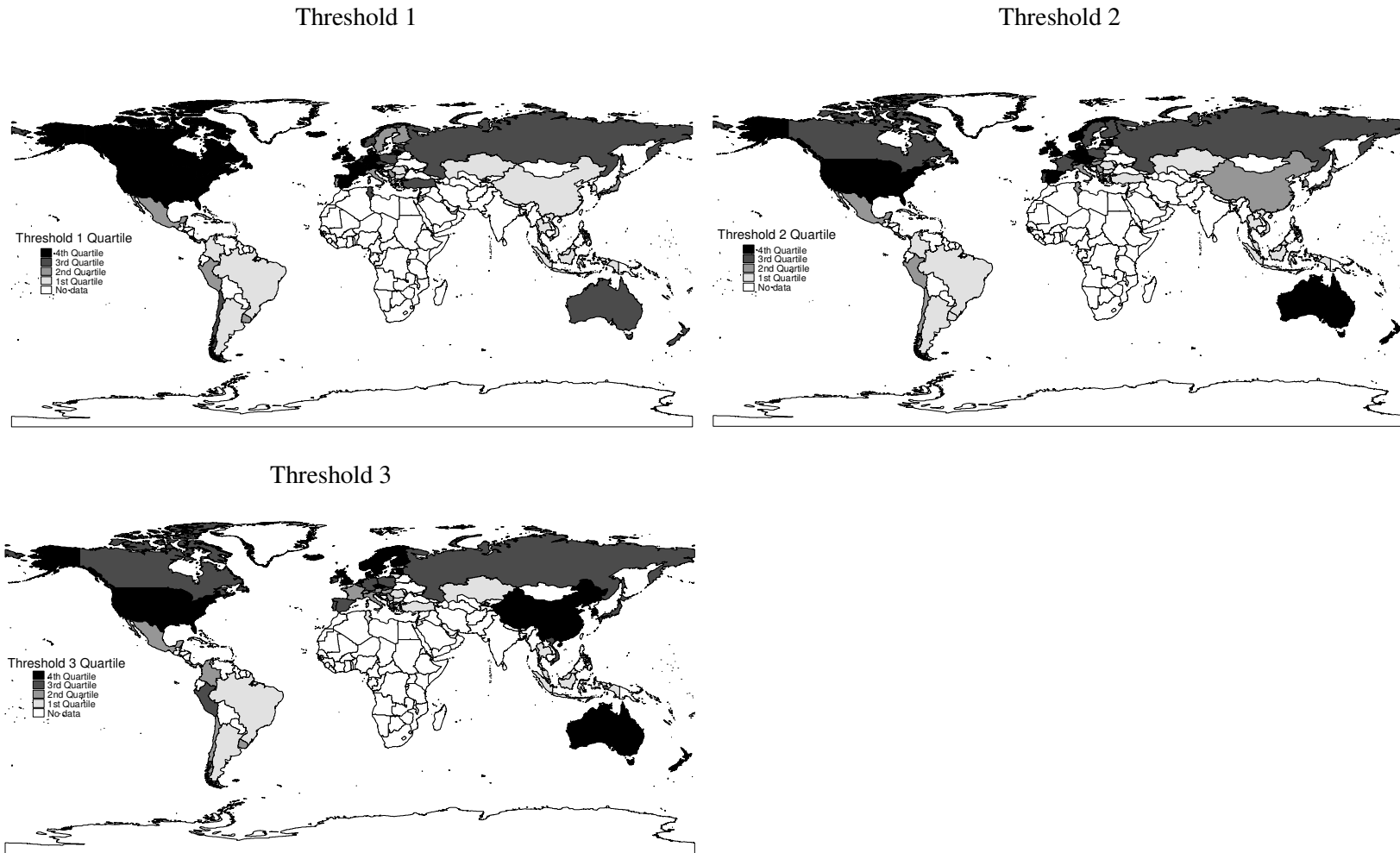
Threshold 3

Note: Threshold 1 delimits the decision between the categories "strongly disagree" and "disagree", Threshold 2 delimits the decision between the categories "disagree" and "agree", and Threshold 3 delimits the decision among the categories "strongly agree" and "agree" with the statements in recoded Questions 1, 2, 3 and Question 4. For China, only results for Shanghai are presented.

Figure 3: Relationship between estimated threshold 1 and threshold 3
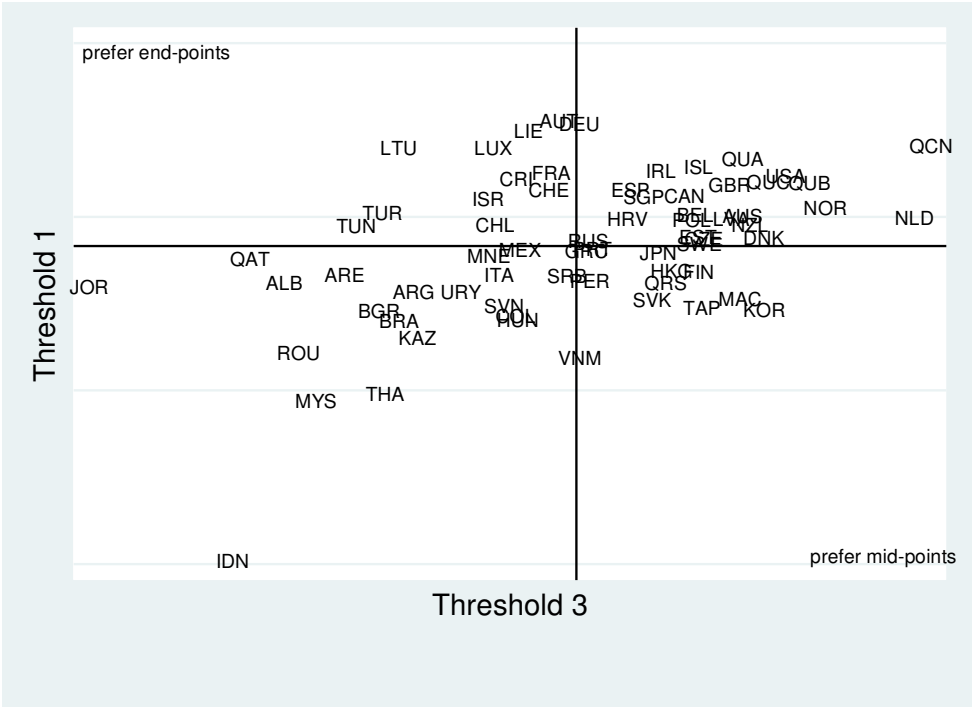
Note: Solid lines represent the mean of the distribution of threshold 1 and threshold 3 values.

Figure 4: Adjusted and unadjusted distributions of teachers' classroom management skills

Differential Scale Usage and Ranking by
Levels of Teacher Classroom Management
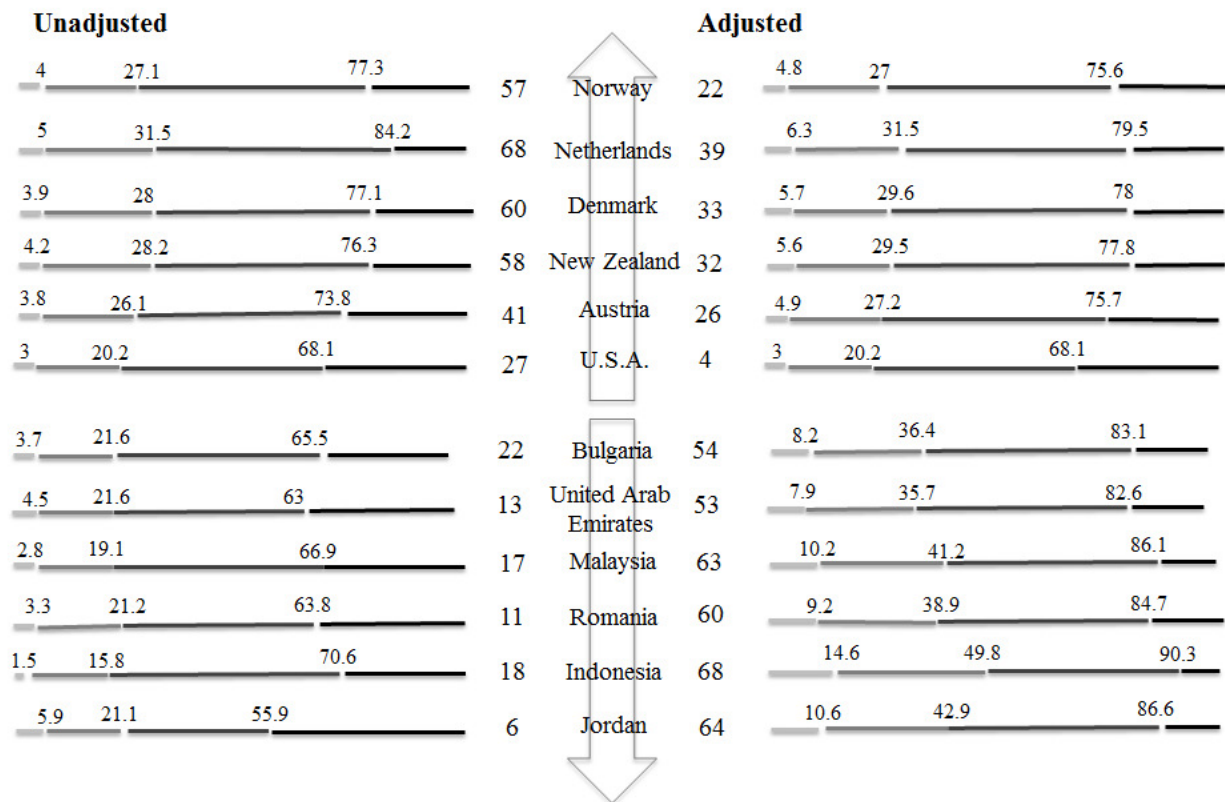
Table 1: Classroom Management Questions, Summary Statistics

Question	4. Strongly Agree	3.Agree	2.Disagree	1.Strongly Disagree	Mean	SD
1	34.4%	52.6%	10.6%	2.5%	3.19	0.72
2	31.2%	52.0%	14.0%	2.8%	3.11	0.74
3	36.5%	47.3%	13.7%	2.5%	3.18	0.76
4	10.3%	26.8%	42.3%	20.7%	2.73	0.90
Note: All tabulations and statistics calculated using final student weights						

Table 2: Descriptive responses to Question 2: My teacher keeps the class orderly

	N	% Strongly Agree	% Agree	% Disagree	% Strongly Disagree
U.S.A	**3,226**	**30.13**	**54.82**	**12.50**	**2.56**
North America					
Canada	13,898	30.46	51.66	14.48	3.40
Connecticut (USA)	1,079	32.04	54.25	12.14	1.58
Florida (USA)	1,241	30.54	52.97	13.86	2.63
Massachusetts (USA)	1,120	31.15	55.54	11.37	1.93
Central America					
Costa Rica	2,861	43.87	43.78	10.43	1.93
Mexico	22,134	37.73	48.52	11.79	1.96
South America					
Argentina	3,689	24.52	49.61	20.65	5.22
Brazil	11,964	30.02	48.20	18.82	2.95
Chile	4,490	26.20	48.76	21.37	3.67
Colombia	5,443	35.21	50.99	11.84	1.96
Peru	3,636	33.46	53.45	11.90	1.19
Uruguay	3,299	29.31	50.66	17.20	2.82
Northern Europe					
Denmark	4,740	19.93	60.60	17.03	2.45
Estonia	3,162	30.50	52.59	14.69	2.21
Finland	5,652	17.96	54.46	23.26	4.32
Iceland	2,240	32.19	49.00	15.78	3.03
Ireland	3,313	31.24	47.71	16.95	4.10
Latvia	2,816	30.70	54.44	12.78	2.08
Lithuania	3,071	48.14	37.34	11.28	3.25
Norway	2,959	18.66	59.44	19.16	2.74
Sweden	2,985	19.44	54.03	22.38	4.15
United Kingdom	8,240	28.49	52.38	16.08	3.05
Southern Europe					
Albania	2,671	57.13	37.77	4.28	0.82
Croatia	3,312	26.43	49.61	19.07	4.88
Greece	3,366	20.65	45.94	26.73	6.68
Italy	20,424	26.40	50.16	17.89	5.54
Montenegro	3,054	34.34	49.18	12.84	3.64

Portugal	3,711	28.04	49.49	18.50	3.98
Serbia	3,029	27.73	51.04	17.43	3.80
Slovenia	3,775	29.42	49.71	18.50	2.36
Spain	16,518	25.37	50.00	19.99	4.64
Eastern Europe					
Bulgaria	3,335	36.23	46.06	14.43	3.28
Czech Republic	3,431	26.75	53.12	16.44	3.70
Hungary	3,157	26.90	47.13	21.16	4.80
Perm (Russian Fed.)	1,157	35.84	52.84	10.40	0.93
Poland	3,028	22.72	51.47	21.08	4.73
Romania	3,340	42.88	44.46	10.20	2.47
Russian Federation	3,451	42.78	47.00	8.93	1.29
Slovak Republic	3,025	23.63	59.05	14.81	2.52
Western Europe					
Austria	3,087	32.23	42.86	19.84	5.06
Belgium	5,405	27.38	54.75	14.32	3.56
France	2,978	23.88	45.65	22.32	8.15
Germany	2,737	28.76	43.43	22.45	5.36
Liechtenstein	188	38.77	43.84	14.63	2.77
Luxembourg	3,385	30.94	42.88	18.94	7.24
Netherlands	2,799	16.82	55.14	22.24	5.80
Switzerland	7,342	30.83	47.54	17.63	4.00
Middle East					
Israel	3,149	40.54	44.49	12.07	2.90
Jordan	4,495	50.59	38.00	8.27	3.14
Qatar	6,575	36.12	44.43	13.76	5.70
Tunisia	2,743	40.49	42.19	11.99	5.33
Turkey	3,175	39.59	45.13	11.23	4.05
United Arab Emirates	7,293	40.94	44.54	11.59	2.93
Central Asia					
Kazakhstan	3,830	54.35	42.95	2.40	0.30
Eastern Asia					
Chinese Taipei	4,007	21.46	56.15	18.94	3.46
Hong Kong-China	3,016	20.34	59.88	16.19	3.58
Japan	4,131	21.70	48.82	24.84	4.64
Korea	3,356	14.17	62.46	19.92	3.45
Macao-China	3,527	18.26	60.91	17.94	2.89
Shanghai-China	3,456	29.56	54.76	13.80	1.88
South-Eastern Asia					
Indonesia	3,663	37.63	57.51	4.31	0.55
Malaysia	3,358	39.14	53.00	6.67	1.19
Singapore	3,660	36.04	53.53	9.16	1.28
Thailand	4,378	32.52	59.88	7.04	0.56
Viet Nam	3,299	26.53	63.38	9.15	0.93
Oceania					
Australia	9,293	24.78	52.33	19.46	3.43
New Zealand	2,753	21.83	52.89	21.70	3.59

Note: All tabulations and statistics calculated using final student weights

Table 3: Vignettes Questions, Summary Statistics

Vignette	4.Strongly Agree	3.Agree	2.Disagree	1.Strongly Disagree	Mean	SD
1	53.2%	39.4%	5.9%	1.4%	3.44	0.67
2	14.2%	31.7%	41.8%	12.3%	2.48	0.88
3	5.9%	14.1%	37.4%	42.6%	1.83	0.88
Note: All tabulations and statistics calculated using final student weights						

Table 4: High Vignette: Teacher Classroom Management

	N (actual)	% Strongly Agree	% Agree	% Disagree	% Strongly Disagree
U.S.A	**3,216**	**53.36**	**41.56**	**0.43**	**0.75**
North America					
Canada	13,872	60.50	35.39	3.32	0.80
Connecticut (USA)	1,070	51.98	42.98	4.35	0.70
Florida (USA)	1,240	54.11	41.12	4.12	0.65
Massachusetts (USA)	1,120	57.75	38.43	3.13	0.68
Central America					
Costa Rica	2,865	62.87	32.72	3.24	1.17
Mexico	22,130	60.67	32.18	5.56	1.59
South America					
Argentina	3,637	48.36	39.69	8.79	3.16
Brazil	11,965	47.48	41.37	9.20	1.94
Chile	4,476	64.27	29.55	4.85	1.33
Colombia	5,432	52.55	40.10	5.97	1.37
Peru	3,633	52.36	41.16	5.24	1.23
Uruguay	3,256	56.73	36.60	4.84	1.83
Northern Europe					
Denmark	4,746	60.98	31.76	5.86	1.41
Estonia	3,154	69.37	27.05	2.92	0.65
Finland	5,662	64.88	30.32	3.75	1.04
Iceland	2,245	63.78	28.96	5.27	1.99
Ireland	3,314	66.49	31.03	1.96	0.52
Latvia	2,815	62.89	31.09	5.21	0.81
Lithuania	3,049	67.36	22.76	7.35	2.53
Norway	2,968	46.26	35.44	14.05	4.26
Sweden	2,975	45.10	40.22	11.27	3.41
United Kingdom	8,225	59.03	36.41	3.84	0.71
Southern Europe					
Albania	2,663	68.15	26.33	4.08	1.44
Croatia	3,292	60.11	34.69	4.30	0.91
Greece	3,364	50.75	38.90	8.24	2.11
Italy	20,383	58.44	34.75	5.10	1.70
Montenegro	2,985	51.75	37.86	7.06	3.32

Portugal	3,705	63.14	33.23	2.82	0.82
Serbia	2,998	48.90	39.69	8.40	3.01
Slovenia	3,758	57.47	34.14	6.85	1.54
Spain	16,519	64.89	28.22	5.07	1.83
Eastern Europe					
Bulgaria	3,319	48.04	40.55	9.02	2.39
Czech Republic	3,429	57.40	36.15	4.89	1.56
Hungary	3,151	57.20	36.05	5.42	1.33
Perm (Russian Fed.)	1,155	50.46	40.09	8.70	0.75
Poland	3,028	55.35	35.22	7.07	2.36
Romania	3,342	49.84	38.03	9.36	2.77
Russian Federation	3,446	54.45	35.58	8.52	1.45
Slovak Republic	3,010	45.63	43.00	9.22	2.15
Western Europe					
Austria	3,079	69.46	17.99	8.13	4.42
Belgium	5,370	56.47	36.99	4.91	1.63
France	2,953	66.32	27.31	4.38	2.00
Germany	2,719	68.89	19.72	8.34	3.06
Liechtenstein	190	66.21	22.29	7.21	4.29
Luxembourg	3,369	62.03	24.53	8.85	4.59
Netherlands	2,805	35.93	56.60	6.40	1.07
Switzerland	7,323	64.22	24.98	8.01	2.78
Middle East					
Israel	3,147	63.85	28.73	5.06	2.37
Jordan	4,463	60.74	29.22	7.68	2.36
Qatar	6,549	47.47	36.01	12.72	3.80
Tunisia	2,714	53.48	33.04	9.46	4.01
Turkey	3,164	59.27	31.37	7.27	2.09
United Arab Emirates	7,262	58.72	32.66	6.63	2.00
Central Asia					
Kazakhstan	3,823	58.44	35.91	4.65	1.00
Eastern Asia					
Chinese Taipei	4,007	42.28	49.66	6.30	1.76
Hong Kong-China	3,014	34.12	56.63	8.20	1.05
Japan	4,157	40.81	46.22	10.71	2.26
Korea	3,327	34.02	57.68	6.74	1.55
Macao-China	3,528	45.79	47.32	5.92	0.96
Shanghai-China	3,456	49.36	46.57	3.40	0.67
South-Eastern Asia					
Indonesia	3,675	46.70	49.51	3.34	0.45
Malaysia	3,366	54.80	38.03	6.27	0.90
Singapore	3,653	55.81	39.60	3.13	1.46
Thailand	4,378	54.61	40.13	4.82	0.44
Viet Nam	3,295	55.50	40.73	3.33	0.44
Oceania					
Australia	9,246	57.12	38.22	3.81	0.84
New Zealand	2,755	52.33	41.90	4.74	1.03

Note: All tabulations and statistics calculated using final student weights

Table 5: Regression estimates of estimated Thresholds from CHOPIT on country level characteristics

| | Threshold | | |
	1	2	3
Public expenditure per pupil	0.0077	0.0113**	0.0091**
	(0.0049)	(0.0043)	(0.0037)
GDP per capita-1000$	0.0074***	0.0059***	0.0047*
	(0.0013)	(0.0015)	(0.0025)
% Private schools	-0.0011	-0.0007	0.0017
	(0.0011)	(0.0010)	(0.0011)
Math Exit Exam	-0.1193**	-0.1280***	-0.0753
	(0.0487)	(0.0445)	(0.0560)
Constant	-2.4089***	-1.3835***	-0.1516
	(0.1362)	(0.1238)	(0.1068)

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.

Table 6: Correlation among estimated adjusted and unadjusted averaged teachers' classroom management levels at the country level and averaged test scores

	Ordered Probit	CHOPIT Model
Average Math Scores	-0.3254	0.4518
Average Reading Scores	-0.3725	0.4768

Note: Number of Observations: 68 country-regions.

Table 7: Regression estimates of Ordered Probit and CHOPIT estimated Country-Level Effects on country level characteristics

| | Ordered Probit | CHOPIT |
	Q1-Q4	Q1-Q4
Public expenditure per pupil	-0.0051*	0.0043
	(0.0026)	(0.0044)
GDP per capita- 1000$	-0.0017	0.0038**
	(0.0011)	(0.0018)
% Private schools	-0.0013	-0.0009
	(0.0090)	(0.0009)
Math Exit Exam	0.0294	-0.0758
	(0.0403)	(0.0510)
Constant	0.0807	-0.4999***
	(0.0774)	(0.1161)

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.

Appendix A

Table A1: Country names and abbreviations

Abbre.	Country name	Abbre.	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	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
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	Viet Nam

Appendix B:

This appendix presents the estimated thresholds using CHOPIT Model in Table B.1 and the estimated country effects and ranking of countries using both ordered probit and CHOPIT models in Table B.2.

Table B.1: Estimated Thresholds from CHOPIT model

Country	Threshold 1	Threshold 2	Threshold 3
Shanghai-China	0.087***	0.074***	0.024***
	(0.0206)	(0.0160)	(0.0099)
Connecticut (USA)	-0.020	0.033*	0.014
	(0.0305)	(0.0247)	(0.0142)
Massachusetts (USA)	-0.016	0.008	-0.013
	(0.0315)	(0.0246)	(0.0140)
United States of America (Reference)	-1.882***	0.046***	0.266***
	(0.0121)	(0.0116)	(0.0069)
Costa Rica	-0.008	-0.148***	-0.183***
	(0.0213)	(0.0182)	(0.0112)
Russian Federation	-0.188***	0.045***	-0.106***
	(0.0203)	(0.0159)	(0.0102)
Florida (USA)	0.048**	-0.051***	-0.043***
	(0.0273)	(0.0216)	(0.0136)
Perm (Russian Federation)	-0.306***	0.143***	-0.015
	(0.0302)	(0.0224)	(0.0142)
United Kingdom	-0.026*	0.023**	-0.060***
	(0.0176)	(0.0136)	(0.0083)
Latvia	-0.123***	0.087***	-0.037***
	(0.0216)	(0.0168)	(0.0106)
Iceland	0.026	0.001	-0.120***
	(0.0225)	(0.0175)	(0.0115)
Kazakhstan	-0.465***	-0.005	-0.028***
	(0.0206)	(0.0168)	(0.0102)
Liechtenstein	0.129**	-0.151***	-0.298***
	(0.0615)	(0.0546)	(0.0363)
Canada	-0.057***	-0.003	-0.062***
	(0.0166)	(0.0128)	(0.0078)
Estonia	-0.176***	0.128***	-0.069***
	(0.0211)	(0.0157)	(0.0102)
Singapore	-0.061***	-0.098***	-0.028***
	(0.0205)	(0.0169)	(0.0098)
Ireland	0.015***	0.018*	-0.172***
	(0.0211)	(0.0158)	(0.0104)
Lithuania	0.080***	-0.302***	-0.300***
	(0.0205)	(0.0177)	(0.0113)
Japan	-0.221***	0.032**	0.009
	(0.0192)	(0.0153)	(0.0094)
Belgium	-0.110***	0.051***	-0.054***
	(0.0185)	(0.0140)	(0.0089)

Albania	-0.309***	-0.133***	-0.212***
	(0.0221)	(0.0186)	(0.0121)
Norway	-0.091***	0.086***	0.039***
	(0.0207)	(0.0152)	(0.0098)
Australia	-0.116***	0.078***	-0.021***
	(0.0175)	(0.0131)	(0.0081)
Mexico	-0.213***	-0.038***	-0.091***
	(0.0159)	(0.0123)	(0.0075)
Peru	-0.302***	0.064***	-0.029***
	(0.0201)	(0.0157)	(0.0096)
Austria	0.159***	-0.085***	-0.349***
	(0.0210)	(0.0160)	(0.0116)
Czech Republic	-0.180***	0.125***	-0.056***
	(0.0204)	(0.0148)	(0.0099)
Switzerland	-0.039***	-0.042***	-0.206***
	(0.0177)	(0.0135)	(0.0087)
Israel	-0.066***	-0.061***	-0.243***
	(0.0204)	(0.0160)	(0.0109)
Germany	0.152***	-0.080***	-0.317***
	(0.0221)	(0.0165)	(0.0119)
Colombia	-0.404***	0.056***	-0.021***
	(0.0187)	(0.0145)	(0.0088)
New Zealand	-0.141***	0.094***	-0.009
	(0.0222)	(0.0161)	(0.0103)
Denmark	-0.178***	0.116***	0.015*
	(0.0194)	(0.0143)	(0.0091)
Viet Nam	-0.523***	0.053***	0.131***
	(0.0223)	(0.0178)	(0.0096)
Montenegro	-0.229***	-0.043***	-0.109***
	(0.0198)	(0.0157)	(0.0102)
Macao-China	-0.352***	0.078***	0.145***
	(0.0204)	(0.0155)	(0.0091)
Portugal	-0.209***	0.062***	-0.099***
	(0.0199)	(0.0149)	(0.0096)
Spain	-0.039***	-0.005	-0.138***
	(0.0162)	(0.0123)	(0.0077)
Netherlands	-0.120***	0.107***	0.128***
	(0.0208)	(0.0153)	(0.0097)
Croatia	-0.122***	0.031**	-0.102***
	(0.0199)	(0.0149)	(0.0098)
Luxembourg	0.079***	-0.143***	-0.303***
	(0.0196)	(0.0153)	(0.0106)
Poland	-0.127***	0.065***	-0.057***
	(0.0204)	(0.0149)	(0.0100)
Sweden	-0.194***	0.073***	-0.002
	(0.0205)	(0.0153)	(0.0099)
Turkey	-0.106***	-0.203***	-0.217***
	(0.0197)	(0.0162)	(0.0106)
Slovak Republic	-0.355***	0.136***	0.015*

	(0.0210)	(0.0152)	(0.0100)
Hong Kong-China	-0.273***	-0.052***	0.122***
	(0.0215)	(0.0163)	(0.0091)
Serbia	-0.286***	0.065***	-0.067***
	(0.0199)	(0.0148)	(0.0098)
Chinese Taipei	-0.380***	0.142***	0.078***
	(0.0189)	(0.0140)	(0.0089)
Finland	-0.277***	0.163***	-0.022***
	(0.0188)	(0.0135)	(0.0088)
Tunisia	-0.143***	-0.158***	-0.254***
	(0.0201)	(0.0167)	(0.0111)
France	0.009***	-0.054***	-0.237***
	(0.0216)	(0.0153)	(0.0107)
Chile	-0.139***	-0.035***	-0.189***
	(0.0191)	(0.0147)	(0.0096)
United Arab Emirates	-0.283***	-0.140***	-0.154***
	(0.0176)	(0.0140)	(0.0086)
Bulgaria	-0.390***	-0.047***	-0.097***
	(0.0202)	(0.0156)	(0.0096)
Slovenia	-0.374***	0.070***	-0.070***
	(0.0194)	(0.0148)	(0.0095)
Uruguay	-0.332***	0.030**	-0.120***
	(0.0202)	(0.0155)	(0.0101)
Hungary	-0.414***	0.154***	-0.103***
	(0.0205)	(0.0145)	(0.0100)
Greece	-0.214***	0.071***	-0.111***
	(0.0200)	(0.0145)	(0.0098)
Italy	-0.285***	-0.001	-0.086***
	(0.0160)	(0.0122)	(0.0075)
Romania	-0.508***	-0.010	-0.123***
	(0.0206)	(0.0159)	(0.0099)
Korea	-0.386***	0.087***	0.182***
	(0.0199)	(0.0150)	(0.0090)
Brazil	-0.417***	0.008	-0.098***
	(0.0168)	(0.0128)	(0.0079)
Malaysia	-0.647***	-0.008	0.007
	(0.0212)	(0.0169)	(0.0096)
Jordan	-0.319***	-0.316***	-0.315***
	(0.0187)	(0.0160)	(0.0101)
Thailand	-0.627***	-0.028**	0.078***
	(0.0207)	(0.0162)	(0.0089)
Qatar	-0.240***	-0.242***	-0.233***
	(0.0172)	(0.0139)	(0.0087)
Argentina	-0.332***	-0.034***	-0.116***
	(0.0193)	(0.0147)	(0.0095)
Indonesia	-1.107***	0.100***	0.167***
	(0.0229)	(0.0170)	(0.0093)

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.

Table B.2: Estimated Country Effects and Ranking of Countries using CHOPIT and Ordered Probit Models

Country	CHOPIT	Ordered Probit	CHOPIT	Ordered Probit
Shanghai-China	1	19	0.209***	0.028***
			(0.0212)	(0.0096)
Connecticut (USA)	2	16	0.064**	0.039***
			(0.0293)	(0.0139)
Massachusetts (USA)	3	15	0.023	0.040***
			(0.0294)	(0.0136)
United States of America	4	27	0.000	0.000
			Reference	Reference
Costa Rica	5	3	-0.022	0.256***
			(0.0213)	(0.0098)
Russian Federation	6	5	-0.026	0.196***
			(0.0200)	(0.0099)
Florida (USA)	7	26	-0.031	0.000
			(0.0273)	(0.0125)
Perm (Russian Federation)	8	7	-0.059**	0.105***
			(0.0281)	(0.0146)
United Kingdom	9	32	-0.064***	-0.018***
			(0.0176)	(0.0080)
Latvia	10	29	-0.068***	-0.007
			(0.0221)	(0.0105)
Iceland	11	31	-0.071***	-0.016*
			(0.0235)	(0.0106)
Kazakhstan	12	1	-0.083***	0.404***
			(0.0193)	(0.0097)
Liechtenstein	13	9	-0.114**	0.081***
			(0.0714)	(0.0272)
Canada	14	30	-0.120***	-0.014**
			(0.0164)	(0.0075)
Estonia	15	33	-0.124***	-0.032***
			(0.0217)	(0.0097)
Singapore	16	14	-0.125***	0.046***
			(0.0201)	(0.0095)
Ireland	17	36	-0.140***	-0.059***
			(0.0213)	(0.0091)
Lithuania	18	4	-0.141***	0.236***
			(0.0214)	(0.0095)
Japan	19	28	-0.189***	-0.004
			(0.0200)	(0.0103)
Belgium	20	38	-0.204***	-0.103***
			(0.0188)	(0.0085)
Albania	21	2	-0.220***	0.375***
			(0.0221)	(0.0109)
Norway	22	57	-0.222***	-0.231***
			(0.0215)	(0.0104)

Australia	23	48	-0.227***	-0.166***
			(0.0172)	(0.0079)
Mexico	24	8	-0.228***	0.087***
			(0.0157)	(0.0073)
Peru	25	23	-0.245***	0.013*
			(0.0198)	(0.0099)
Austria	26	41	-0.250***	-0.136***
			(0.0222)	(0.0093)
Czech Republic	27	47	-0.256***	-0.158***
			(0.0209)	(0.0093)
Switzerland	28	34	-0.261***	-0.050***
			(0.0179)	(0.0080)
Israel	29	24	-0.267***	0.013*
			(0.0206)	(0.0094)
Germany	30	49	-0.272***	-0.172***
			(0.0232)	(0.0096)
Colombia	31	10	-0.288***	0.075***
			(0.0180)	(0.0088)
New Zealand	32	58	-0.296***	-0.232***
			(0.0218)	90.0102)
Denmark	33	60	-0.300***	-0.239***
			(0.0196)	(0.0092)
Viet Nam	34	12	-0.307***	0.051***
			(0.0213)	(0.0108)
Montenegro	35	20	-0.321***	0.027***
			(0.0204)	(0.0098)
Macao-China	36	43	-0.324***	-0.148***
			(0.0200)	(0.0099)
Portugal	37	39	-0.346***	-0.128***
			(0.0199)	(0.0092)
Spain	38	56	-0.352***	-0.220***
			(0.0161)	(0.0074)
Netherlands	39	68	-0.354***	-0.390***
			(0.0208)	(0.0104)
Croatia	40	53	-0.361***	-0.197***
			(0.0203)	(0.0093)
Luxembourg	41	50	-0.394***	-0.176***
			(0.0207)	(0.0090)
Poland	42	64	-0.395***	-0.285***
			(0.0209)	(0.0096)
Sweden	43	61	-0.396***	-0.259***
			(0.0210)	(0.0101)
Turkey	44	21	-0.400***	0.024***
			(0.0203)	(0.0098)
Slovak Republic	45	55	-0.423***	-0.204***
			(0.0207)	(0.0102)
Hong Kong-China	46	51	-0.429***	-0.184***
			(0.0207)	(0.0099)
Serbia	47	46	-0.431***	-0.157***
			(0.0204)	(0.0096)
Chinese Taipei	48	59	-0.436***	-0.237***

			(0.0188)	(0.0092)
Finland	49	65	-0.436***	-0.299***
			(0.0188)	(0.0087)
Tunisia	50	25	-0.437***	0.009
			(0.0209)	(0.0099)
France	51	62	-0.443***	-0.266***
			(0.0220)	(0.0091)
Chile	52	45	-0.447***	-0.157***
			(0.0192)	(0.0088)
United Arab Emirates	53	13	-0.469***	0.046***
			(0.0174)	(0.0082)
Bulgaria	54	22	-0.487***	0.017***
			(0.0198)	(0.0094)
Slovenia	55	44	-0.507***	-0.149**
			(0.0197)	(0.0094)
Uruguay	56	40	-0.517***	-0.131***
			(0.0203)	(0.0099)
Hungary	57	52	-0.525***	-0.193***
			(0.0204)	(0.0095)
Greece	58	66	-0.526***	-0.304***
			(0.0200)	(0.0093)
Italy	59	54	-0.546***	-0.199***
			(0.0158)	(0.0073)
Romania	60	11	-0.552***	0.055***
			(0.0199)	(0.0096)
Korea	61	67	-0.557***	-0.347***
			(0.0203)	(0.0106)
Brazil	62	37	-0.576***	-0.097***
			(0.0164)	(0.0077)
Malaysia	63	17	-0.613***	0.035***
			(0.0208)	(0.0103)
Jordan	64	6	-0.635***	0.148***
			(0.0191)	(0.0090)
Thailand	65	35	-0.657***	-0.058***
			(0.0195)	(0.0100)
Qatar	66	42	-0.717***	-0.139***
			(0.0177)	(0.0084)
Argentina	67	63	-0.723***	-0.283***
			(0.0196)	(0.0096)
Indonesia	68	18	-0.830***	0.035***
			(0.0211)	(0.0107)

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10%, respectively.