Abstract Title Page

Title:
Measurement of Classroom Teaching Quality with Item Response Theory

Ben Kelcey
Wayne State University
ben.kelcey@gmail.com

Daniel McGinn
Harvard University
daniel_mcginn@gse.harvard.edu

Heather Hill
Harvard University
heather_hill@gse.harvard.edu

Background / Context:

Recent policy has charged schools and districts with maintaining highly qualified teachers and differentiating among teachers in terms of their effectiveness (U.S. Department of Education, 2009). This emphasis has driven the development and implementation of teacher quality measures which are increasingly being used to evaluate teachers with important consequences (Schochet & Chiang, 2010). One increasingly common component of these evaluations is the direct observation of teachers in their classrooms. Classroom observations have been long viewed as a promising way to evaluate and develop teachers because they anchor assessments in specific and observable criteria (Gitomer, 2009).

Despite the potential of classroom observations to identify strengths and address specific weaknesses in teachers' practices (MET, n.d.), the systems used to conduct classroom observations tend to be influenced by aspects of the observational environment beyond the teaching quality (Kennedy, 2010). For instance, many observation systems evaluate teaching using fallible indicators and raters and generally draw inferences using only a small sample of teachers' lessons. Such features potentially introduce bias and imprecision into teaching quality assessments. Combined with the fact that these measures often form the basis for many high stakes decisions, the robustness of teaching quality scores to these features has taken on increasing importance. Yet despite this importance, relatively little is known about the accuracy and precision of these scores amidst construct-irrelevant variation and the extent to which different treatments of this variation arrive at similar scores and precision levels.

Two approaches to scoring classroom observations of teaching quality have largely dominated this literature: classical test theory (CTT) and generalizability theory (GT). In this proposal, we developed a third alternative approach based on item response theory (IRT). Each approach incorporates measurement error into their framework; however, they do so in distinctly different ways. We very briefly outline the features of these approaches as they apply to treatments of construct-irrelevant sources of variation.

Both CTT and GT construct estimates of teaching quality by summing or averaging across all items and observations. As a result, CTT/GT assumes that the original (ordinal) teaching ratings all hold equal amounts of information and are continuously scaled such that scores are created by averaging over any construct-irrelevant variation (e.g., raters). CTT then estimates a single level of reliability and measurement error for each teacher by simply decomposing observed variance into true and error score variance. GT refines this approach by further decomposing the error variance among sources (e.g., raters, occasions). GT then estimates a single level of reliability and precision for all indices by assuming that units within each source of variation (both current and future) are exchangeable (e.g., teachers or raters are exchangeable). As a result, although CTT/GT acknowledges the influence of construct-irrelevant variance on the scores, their reliability and their precision, the reported scores make no adjustments for these errors and allow construct-irrelevant variance to accumulate as measurement error.

In contrast to these methods, an IRT based approach does not assume raw ratings are continuously scaled and hold equal information and instead estimates the latent trait theoretically underlying teachers' observed rating patterns by postulating a probabilistic response model. Similar to GT, our extension of the common IRT model to classroom observations recognizes the influence of multiple sources of construct-irrelevant variation. However, in contrast to GT, our IRT based approach adjusts for construct-irrelevant variation to provide a measure of teaching quality that is as independent as possible of the sources of construct-irrelevant variation.

Purpose / Objective / Research Question / Focus of Study:

In this proposal, we investigated the robustness of classroom observation scores to three measurement approaches: CTT, GT, and IRT. We investigated the extent to which choices among these approaches lead to indeterminacies in conclusions regarding teaching quality, the precision with which we can index this quality, and the relation of this quality to student achievement. We then provide insights delineating the reasons for the discrepancies among methods and explore the extent to which the observed differences are indicative of true differences among teachers in their teaching quality.

Setting & Population / Participants / Subjects & Intervention / Program / Practice:

This study takes place within the larger Developing Measures of Effective Mathematics Teaching study which focused on developing identifying practices and characteristics that distinguish between more and less effective teachers. The sample includes 250 teachers from 40 schools and 4 districts. Table 1 presents a few basic descriptive statistics. Our study drew on the Mathematical Quality of Instruction (MQI) classroom observation system (Hill et al., 2008; (Hill, Charalambous, & Kraft, 2012). The MQI system was designed to provide assessments for teachers on important dimensions of classroom mathematics instruction. The structure of this system was developed to provide a multidimensional and balanced view of mathematics instruction (Hill et al., 2008). In the current investigation we studied each of the system's dimensions but present only the general dimension for proposal brevity.

Statistical, Measurement, or Econometric Model:

To index teaching quality using CTT, we averaged ratings across items, chapters, and raters thus collapsing across all sources of construct-irrelevant variation. Let

$$\theta_t^T = \frac{1}{IC_tR_c}\sum_{i=1}^{I}\sum_{c=1}^{C_t}\sum_{r=1}^{R_c} Y_{ictr} \tag{1}$$

where θ_t^T is the estimate of teaching quality for teacher *t* across all items, chapters, and raters. Y_{ictr} is the score for item *i* in chapter *c* for teacher *t* given by rater *r,* R_c is the total number of raters for chapter *c*, C_t is the number of chapters observed for teacher *t,* and *I* is the total number of items. To describe the uncertainty associated with CTT scores of teaching quality, we used CTT's concept of the standard error of measurement. Specifically, define coefficient alpha as

$$\alpha = (\frac{I}{I-1})(1-\frac{\sum_{i=1}^{I}\sigma_i^2}{\sigma_Y^2}) \tag{2}$$

where *I* is the number of items, σ_i^2 is the variance of item *i* across teachers and σ_Y^2 is the variance of the observed total scores. Standard errors were obtained using

$$\text{Standard Error of Measurement (SEM)} = \sigma_Y\sqrt{1-\alpha} \tag{3}$$

Confidence intervals were formed using each teacher's score plus or minus double the *SEM*.

To describe teaching quality using GT we also scored teachers using equation (1). Subsequently, we constructed standard errors using the SEM (3) replacing α with

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \frac{\sigma_r^2}{R} + \frac{\sigma_c^2}{C}} \tag{4}$$

were σ_t^2 is the teacher variability, σ_r^2 is rater variability with R (=2) as the number of raters/observation, and σ_c^2 is the chapter variability with C as the average number of observed chapters for teacher t. Teacher variation, σ_t^2, stems from stable differences among teachers in terms of their consistent quality; rater variation (σ_r^2) arises from differences among observers in their severity; chapter variation (σ_c^2) manifests when teachers' quality varies across chapters. Our partially crossed design largely precludes estimation of remaining interactions among these components, however, we did also consider the rater-by-teacher interaction.

To describe teaching using IRT, we developed a multilevel graded response model

$$P(Y_{ictr}=k\mid\theta)=P(Y_{ictr}\geq k\mid\theta)-P(Y_{ictr}\geq k+1\mid\theta)=\frac{1}{1+\exp[-[a_i(\theta_t+\gamma_r+\alpha_{ctr}-d_i^k)]]}-\frac{1}{1+\exp[-[a_i(\theta_t+\gamma_r+\alpha_{ctr}-d_i^{k+1})]]} \quad (5)$$

Here Y_{ictr} is the score for item i in chapter c for teacher t rated by rater r, a_i represents the discrimination parameter for item i, θ_t represents teacher t's stable level of teacher quality and is assumed to be normally distributed, γ_r is a fixed effect for rater r's level of leniency and α_{ctr} is the deviation specific to chapter c for teacher t rated by rater r with assumed normal distribution $N(0,\sigma_c^2)$. Further, let K represent the number of categories items are graded on (three with MQI) with k as a specific category and let $d_i^{(1)},\ldots,d_i^{(K-1)}$ be a set of K-1 ordered item difficulty intercepts. To identify the scale, we set θ to have a normal distribution with mean zero and unit variance. Estimation was carried out using maximum likelihood with corresponding quality levels estimated using an expected a posteriori approach. Finally, the standard errors of scores were estimated using the posterior standard deviations obtained from the second derivative of the log-likelihood function. Symmetric confidence intervals were formed using each teacher's score plus or minus double the posterior standard deviation.

Research Design:

Using 40 raters, the study observed 250 teachers across three time points. Observations were broken into several chapters of about seven minutes in length and two raters provided ordinal ratings (scores of 1, 2, or 3) of teachers' instruction along 21 different items/indicators. Below we describe the results for a single dimension, the general or overall quality of teaching.

Findings / Results:

Overall there were significant discrepancies among results suggesting that estimates of teaching quality and their precision were sensitive to scoring approach. We very briefly highlight only a few findings. Indices constructed using CTT indicated the (alpha) reliability of teachers' scores was on the order of 0.48 whereas GT results suggested that the reliability of the average was 0.60 (Table 2). Our application of IRT both acknowledged and adjusted for the presence of construct-irrelevant variance, and as a result, it recorded higher levels of reliability despite the presence of a substantial amount of construct-irrelevant variance. Because information/reliability is a function of the latent trait in IRT, we described the reliability of IRT scores by presenting the average reliability for teachers at each level of the continuum (Figure 1). Overall, the average reliability of IRT scores fluctuated between 0.86 for low quality teachers and 0.93 for high quality teachers, although for specific teachers the reliabilities ranged from 0.72 to 0.95.

In assessing the extent to which the methods agreed on the actual values of teachers' scores, we observed a correlation between CTT/GT scores and IRT scores of 0.82. A scatter plot of the scores indicated that the standardized CTT/GT scores generally had a wider range than the IRT scores (Figure 2). CTT/GT scores ranged from four SDs below the mean to three SDs above

the mean whereas IRT scores were shrunken toward the mean with a range between negative three SDs and positive two SDs.

Corresponding to the aforementioned discrepant reliabilities, we also found substantial differences in the precisions with which the methods could index teaching quality scores (Figure 3). In particular, because CTT does not discriminate among sources of variation with regard to indexing reliability and GT acknowledges them, GT standard errors of scores tended to be smaller (0.72 for CTT and 0.63 for GT). The IRT adjustments for construct-irrelevant variation further reduced the size of standard errors for IRT scores by about half. Our results generally indicated that the width of the 95% confidence intervals for IRT based scores was substantially tighter than their CTT and GT counterparts (Figure 4). For instance, for an average teacher the width of the confidence interval for IRT based scores spanned 1.2 standard deviations (e.g., 0±0.6) whereas the widths of confidence intervals for CTT/GT averages were 2.52 standard deviations for GT (e.g., 0±1.26) and 2.88 standard deviations for CTT (e.g., 0±1.44).

Despite the potential for item response models to improve the accuracy and precision of teacher scores, there is some question as to the validity of the adjustments made by IRT. Among other important assumptions (e.g., unidimensionality, invariance), our IRT approach assumes that the model based adjustments we made for construct-irrelevant variance are valid and accurate. Two sources of construct-irrelevant variation that our IRT model adjusted for were differences among rater severities and atypical chapters. To examine the validity of these adjustments, we correlated teacher value-added scores with CTT/GT and IRT scores with and without these adjustments.

Our results suggested that our adjustments were of mixed value (Table 3). Use of an IRT model that did not adjust for construct-irrelevant variance (i.e., $\gamma_r = 0$ and $\alpha_{ctr} = 0$ in equation (5)) shared nearly the identical relation with value-added scores as did CTT/GT scores (0.16). Adjustments for atypical chapters but not raters (i.e., $\gamma_r = 0$ equation (5)) improved the IRT score correlation by 25% to 0.20 and pushed it under the nominal *p*-value cutoff of 0.05. In contrast, similar adjustments for raters diminished this relationship to 0.14 suggesting our simple adjustments for rater severities might be insufficient in describing the complex variation among raters.

Conclusions:

Overall the results suggest that construct-irrelevant variance is sizeable in classroom observations and that treatment of this variance had significant implications for the resulting scores. Although the authority of correlating teaching observation scores with value-added scores in validating the appropriateness of each method is unclear, the results suggested that there is much to be gained from methods which directly address construct-irrelevant variation. Specifically, our results suggested that we might be able to create more reliable, more precise and more differentiated indices of teaching quality as they relate to students' achievement by estimating the impact of different sources of construct-irrelevant variance. At the same time, our empirical application also highlighted the potential for erroneous adjustments. However, because classroom observations are potentially attached to high stakes decisions, ignoring construct-irrelevant variation does not seem like a viable option. More specifically, because the reliabilities of the averages under CTT and GT are so low and the confidence intervals they produce are so wide, it seems unlikely that decision makers will be willing to make evaluations amidst so much uncertainty. To this end, our results suggest that empirically based adjustments for construct-irrelevant variance are a promising, albeit complex, approach to understanding teaching quality through classroom observations.

Appendices

Appendix A. References

Gitomer, D. (2009). Measurement Issues and Assessment for Teaching Quality. London: Sage Publications.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H., Charlambous, C., & Kraft, M. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher, 41, 2, pp. 56-64.*

Kennedy, M. M. (2010). Attribution error and the quest for teaching quality. *Educational Researcher, 39*, 591–598.

Measures of Effective Teaching. (n.d.). Retrieved from the Bill and Melinda Gates Foundation website: http://www.gatesfoundation.org/united -states/Pages/measures-of-effective-teaching-fact-sheet.aspx

Schochet, P., & Chiang, H. (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. U.S. Dept of Education report NCEE 2010- 4004.

U.S. Department of Education (2009). Race to the Top Fund – Executive Summary: Notice of Proposed Priorities, Requirements, Definitions, and Selection Criteria. Washington, DC: U.S. Department of Education.

Appendix B. Tables and Figures

Table 1: Selected descriptive statistics of teachers

Teacher Variable	Mean/Percent
Number of Mathematics courses taken	
None	3.50
One	4.70
Two	9.90
Three or more	80.8
White	0.66
Years of experience	9.86
Majored/minored in mathematics	0.06

Table 2: Proportion of variance attributable to source

	Variance
Teacher	0.11
Raters	0.06
Chapter	0.83

Table 3: Standardized regression coefficient of teachers' observation scores predicting their value-added scores

	Standardized coefficient (standard error)	*t*-value
CTT/GT	0.15 (0.10)	1.65
IRT without chapter or rater adjustments	0.16 (0.10)	1.65
IRT with chapter but without rater adjustments	0.20* (0.10)	2.03
IRT with chapter and rater adjustments	0.14 (0.10)	1.44

*$p<0.05$

Figure 1: Reliability of scores by method

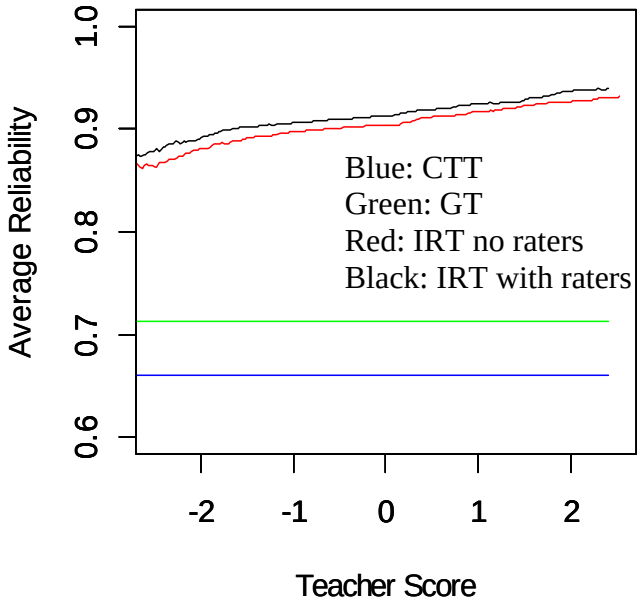


Figure 2: Scatter plot of IRT and CTT/GT scores

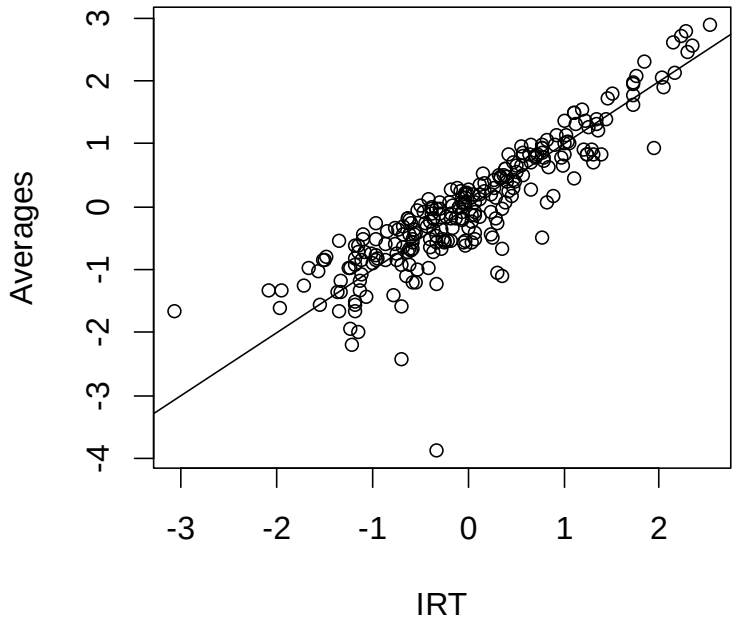


Figure 3: Size of standard error by method

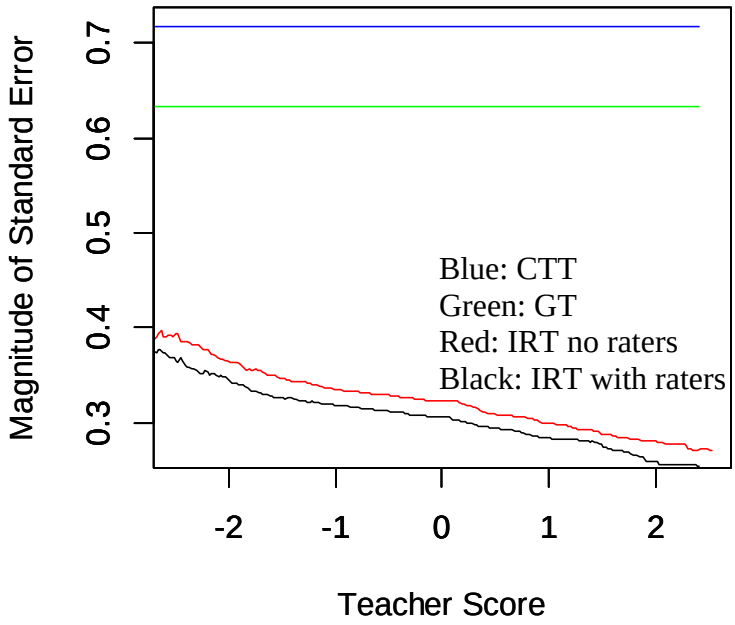
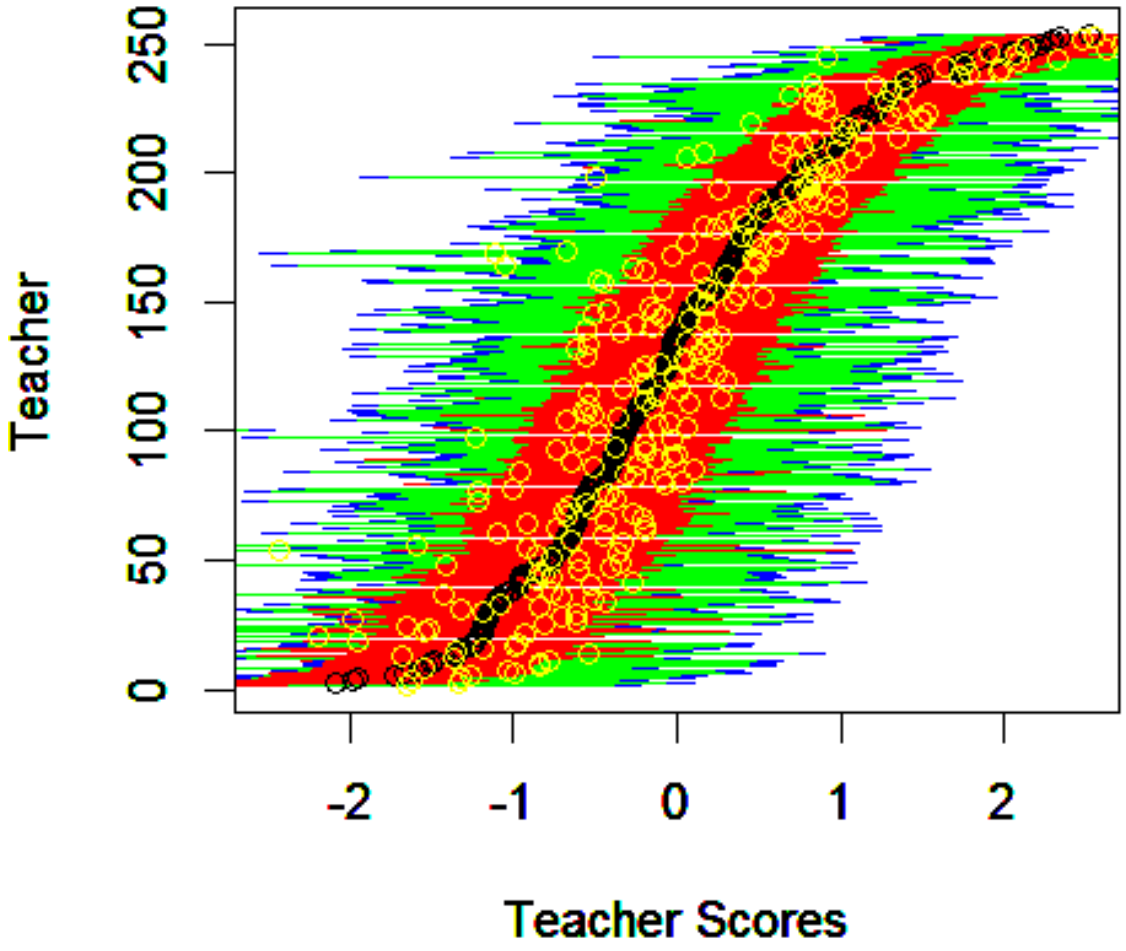


Figure 4: Plot of teachers scores surrounded by 95% confidence intervals by scoring method



Note:
Orange dots: Scores based on CTT
Blue bands: Confidence intervals based on CTT reliability
Green bands: Confidence intervals based on GT reliability
Black dots: Scores based on IRT
Red: Confidence intervals based on IRT