Brief Report

Generalizing From Observations of Mathematics Teachers' Instructional Practice Using the Instructional Quality Assessment

Anne Garrison Wilhelm
Southern Methodist University

Sungyeun Kim
Seoul National University

One crucial question for researchers who study teachers' classroom practice is how to maximize information about what is happening in classrooms while minimizing costs. This report extends prior studies of the reliability of the Instructional Quality Assessment (IQA), a widely used classroom observation toolkit, and offers insight into the often asked question: "What is the number of observations required to reliably measure a teacher's instructional practice using the IQA?" We found that in some situations, as few as three observations are needed to reliably measure a teacher's instructional practice using the IQA. However, that result depends on a variety of other factors.

Key words: Generalizability; Observation; Reliability

Being able to reliably measure teachers' instructional practice is a fundamental issue for practitioners and researchers (Kane & Staiger, 2012). The use of classroom observations as part of district teacher evaluation systems continues to grow with an emphasis on making their use more rigorous (Herlihy et al., 2014). Further, researchers need to know what takes place in classrooms when teachers' instructional practice is being measured as a dependent or independent variable. Repeated classroom observations provide information about classroom events, but they can be costly. Therefore, the crucial question is how to maximize

The first author was supported by the Institute of Education Sciences predoctoral research training program, Grant No. R305B080025. The second author was supported by a National Research Foundation of Korea grant, funded by the Korean Government (NRF-2013S1A3A2055007). The data come from a larger study, supported by the National Science Foundation under Grant Nos. ESI-0554535 and DRL-0830029. The opinions expressed do not necessarily reflect the views of the U.S. Department of Education or the National Science Foundation. We thank members of the research team for the larger project from which these data come: Paul Cobb, Tom Smith, Kara Jackson, Erin Henrick, Chuck Munter, Glenn Colby, Lynsey Gibbons, Charlotte Dunlap, Rebecca Schmidt, Jonee Wilson, Christy Larson, Dan Berebitsky, and Adrian Larbi-Cherif. We also thank Melissa Boston and Paul Yovanoff for comments on earlier drafts of this article.

information about what is happening in classrooms while minimizing costs.

We address that question by examining the reliability of a widely used classroom observation toolkit, the Instructional Quality Assessment (IQA; Boston, 2012; Matsumura, Garnier, Slater, & Boston, 2008). The IQA mathematics toolkit was designed to assess several elements of ambitious instruction in mathematics (Boston, 2012). It consists of 11 rubrics pertaining to the cognitive demand of the mathematical activity in the classroom (e.g., Doyle, 1988; Stein, Grover, & Henningsen, 1996), the quality of classroom discussion (e.g., Boaler & Staples, 2008; O'Connor & Michaels, 1996; Stein, Engle, Smith, & Hughes, 2008), and the teachers' expectations for students (e.g., Remillard, 1999). These rubrics have been used to assess the quality of instruction using classroom artifacts (e.g., samples of student work) and classroom observation. (For more information about the IQA instrument and rubrics, see Boston, 2012.)

The IQA developers conducted generalizability and decision studies to describe the optimal conditions for obtaining reliable scores for teachers' instructional practices (Matsumura et al., 2008). Using a sample of middle school mathematics teachers in an urban district, Matsumura, Garnier, Slater, and Boston (2008) found that when they limited their analysis to the 11 teachers who complied with the requirements of data collection, as few as two observations yielded a reliable estimate of instructional practice, $\Phi = .86$ (p. 279). Although their study provides some information about optimal conditions and generalizability, with such a small sample, the study merits replication, especially given that the IQA is widely used (e.g., Quint, Akey, Rappaport, & Willner, 2007; Sztajn, Wilson, Edgington, & Confrey, 2011). Thus, we conduced a replication-type study to answer the following question: How many observations are required to reliably estimate teachers' instructional practice using the IQA?

Generalizability Theory

Hill, Charalambos, and Kraft (2012) made a strong argument for the need for more information pertaining to observational instruments, raters, and observations to understand the reliability of teachers' scores. They argued that generalizability studies and decision studies are crucial for gathering information about what they call *observational systems* in an attempt to move beyond just thinking about the instrument itself to also considering the number of observations and the number of raters.

Generalizability theory provides a theoretical basis for understanding sources of variation in scores (Brennan, 2001). For example, is the variation in teachers' scores for a particular instrument attributable to true variation in teachers, items within the instrument, or the day the observation occurs? Central to generalizability theory is the notion of dependability, which pertains to

> the accuracy of generalizing from a person's observed score on a test or other measure . . . to the average score that person would have received under all possible conditions that the test user would be equally willing to accept. (Shavelson & Webb, 1991, p. 1)

When applied to the context of teachers' instructional practice, one might ask whether the score produced by an observational system is representative of that teacher's instructional practice.

Generalizability (G) studies estimate the sources of variability (e.g., items, raters, observations) within a particular observational system. Decision (D) studies help us to understand the conditions (e.g., number of observations) under which the score produced by the observational system will reliably represent the teacher's practice more generally. D studies produce two different coefficients that represent reliability: the generalizability coefficient, $E\rho^2$, which represents the relative reliability, and the dependability coefficient, Φ, which represents the absolute reliability. The former is appropriate for relative comparisons—for example, which teachers are stronger than their colleagues—but the latter is more appropriate for absolute decisions—for example, whether teachers are above a certain threshold (Hill, Charalambos, & Kraft, 2012).

Method

We drew on data collected in the course of a 4-year study that sought to address the question of what is needed to improve the quality of middle grades mathematics teaching at the scale of a large urban district (Cobb & Jackson, 2011; Cobb & Smith, 2008). The research team collaborated with the leaders of four large, urban districts that were attempting to achieve a vision of high-quality mathematics instruction that was compatible with the National Council of Teachers of Mathematics' (2000) *Principles and Standards for School Mathematics*. In each of the four districts, the research team selected a sample of 6 to 10 middle grades schools that reflected variation in student performance and in capacity for improvement in the quality of instruction across the district. Within each school, up to five mathematics teachers were randomly selected to participate in the study, a total of approximately 30 teachers per district.

For the analyses reported in this article, we used data from years 3 and 4 of the project (the 2009–2010 and 2010–2011 school years) because we had created several additional rubrics to assess the quality of the task setup and used them across the entire sample of teachers in those years. It is important to note that we used only 8 of the 11 rubrics that Matsumura et al. (2008) used in their study (three were omitted because of low interrater reliability in previous studies): Task Potential, Implementation, Rigor of the Discussion, Participation, Teacher Linking, Student Linking, Teacher's Press, and Student Providing. We also utilized rubrics developed by a subset of our research team to assess the quality of the task setup: Contextual Features, Mathematical Relationships, and Maintenance of the Cognitive Demand. (For more information about the setup rubrics, see Jackson, Garrison, Wilson, Gibbons, & Shahan, 2013.) We grouped the rubrics into three domains based on factor analyses of the original IQA data from our study and the addition of the setup measures: (a) Cognitive Demand (Task Potential, Implementation), (b) Discussion (Rigor of the Discussion, Participation, Teacher Linking, Student Linking, Teacher's Press, and Student Providing), and

(c) Setup (Contextual Features, Mathematical Relationships, and Maintenance of the Cognitive Demand). We refer to the set of eight Cognitive Demand and Discussion rubrics from the original IQA instrument as the *Original IQA*, and we refer to the instrument that includes those original eight rubrics and the Setup rubrics as the *Expanded IQA*.

Sample

Our final sample included 150 teachers, and data were collected over 2 years with two observations per teacher. As in the Matsumura et al. (2008) study, we asked teachers to engage students in a problem-solving activity with a related whole-class discussion. Of those 150 teachers, 96 held a whole-class discussion during both observations, 41 held a whole-class discussion during only one of the observations, and 13 did not hold a whole-class discussion on either day they were observed. To understand the importance of the match between sample and instrument, and to be consistent with the Matsumura et al. (2008) study, we considered the *full sample* of teachers and a *restricted sample*, which included only teachers who complied with the requirements of data collection and had a whole-class discussion on both days they were observed. Figure 1 shows a comparison between the Matsumura et al. (2008) study and this study on a number of key dimensions. The primary differences between the two studies are the number of teachers in the study, the study time span, and the sample of rubrics used.

	Matsumura et al. (2008)	This study
Number of teachers	11 teachers	150 teachers
District contexts	1 urban district	4 urban districts
Observation timing and subjects	2 consecutive days, with same class period	2 consecutive days, with same class period
Study time span	2 weeks	2 years
Rubrics used	11 original IQA rubrics	8 of 11 original IQA rubrics plus additional setup rubrics
Rater training	4–5 days of training by developer	4–5 days of training by developer
Interrater agreement prior to coding	Exceeded 80% agreement prior to beginning coding	Exceeded 80% agreement prior to beginning coding
G Study Design	Univariate	Multivariate

Figure 1. Comparison of previous and current study.

Analyses

Generalizability theory provides a framework to disentangle multiple sources of error in a measurement system (Brennan, 2001). For the Original IQA and the Expanded IQA with teachers (*t*) as the object of measurement, three facets contribute to the score variability: domains (*d*), items (*i*), and observations (*o*). In our case, we considered Cognitive Demand, Discussion, and Setup as domains. We assumed that the items within each domain were fixed, and teachers were rated using all items. With the IQA, scores measured by different domains are likely to be correlated. For example, more cognitively demanding tasks have the potential for better whole-class discussion, and hence the Cognitive Demand and Discussion domains are correlated. Given these assumptions, a multivariate random facet $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ G study design was the most appropriate (Brennan, 2001). The superscript filled circle ($^{\bullet}$) designates that the facet is crossed with the fixed domains, and the superscript empty circle ($^{\circ}$) designates that the facet is nested within the fixed domains. In this case, we took the items as nested within the fixed domains (e.g., Task Potential is an item within the Cognitive Demand domain, and we did not allow for it to vary in the model). No superscripts are used for univariate designs.

The IQA developers conducted a univariate G study rather than a multivariate G study. Ignoring the correlation between domains can result in a biased estimate of the reliability coefficients (Clauser, Harik, & Margolis, 2006). In particular, the univariate design considers only the overall IQA score, whereas the multivariate design considers the domains and items within the IQA. Although we believe that a multivariate approach is better suited to the IQA, in order to compare to the results from the previous IQA generalizability study and better understand the reliability of the IQA, we conducted both univariate and multivariate analyses. We performed eight total analyses, combining the Original IQA or Expanded IQA with the full sample or restricted sample in (a) univariate $t \times o$ analyses that ignored both the item facet and the correlations between domains and (b) multivariate $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ analyses that took into account the item facet and the correlations between domains. We report only on the D study results below, but G studies were first conducted to understand sources of variation. In G studies, the estimated variance components are used to calculate the effects of various hypothetical measurement designs (i.e., the D study results)—in this case, whether those designs varied in the number of observations conducted.

With respect to interpretation of results, there is substantial variation in what is considered an acceptable level for $E\rho^2$ and Φ. Generalizability theory developed out of classical test theory, which has generally considered a Cronbach's alpha coefficient of .80 sufficient reliability for making decisions about individuals (Webb, Shavelson, & Haertel, 2007). Hence, the .80 level is a generally acceptable guideline for $E\rho^2$ and Φ. There is good reason and precedent to believe that .80 may be a stringent requirement for observational measures of teachers. In the Measures of Effective Teaching (MET) study, a level of .65 was deemed acceptable when determining what procedures within the observational system were necessary for reliably measuring teachers' instructional practice

(Kane & Staiger, 2012). Although precedent may exist, we know of no empirical work determining appropriate levels of reliability for observational instruments of this complexity; therefore, the decision regarding an appropriate reliability level is a decision left up to the researcher. More work is needed in this area.

Results

The D study results of the univariate $t \times o$ design, including both $E\rho^2$ and Φ, are reported for different numbers of observations—from 1 to 7—in Table 1. The values in the shaded region of Table 1 represent the actual study data with two observations per teacher, but the other values are hypothetical. It is possible that error from G study estimates of variance components influences the accuracy of the reliabilities produced for hypothetical designs, but it is common practice to use the known variance components to estimate reliabilities for hypothetical designs, and we know of no reason why there would be error in the variance estimates based on two observations. The reliability coefficients were uniformly higher for the Expanded IQA than for the Original IQA. In addition, the reliability coefficients were uniformly higher for the restricted sample than for the full sample. Also, the reliability coefficients for just two observations are lower than the standard cutoff, .80, regardless of the sample and the instrument used. The column with $\Phi = .652$ for the restricted sample with two observations is the best for comparison to the findings of Matsumura et al.'s (2008) previous IQA G study with $\Phi = .86$; it is clear that the Φ for our sample is considerably lower. Further, the result is even lower for the Original IQA and the full sample, $\Phi = .522$.

The multivariate $t^\bullet \times o^\bullet \times i^\circ$ D study results, based on the composite scores, are reported in Table 2. Because we assume that observational instruments are more likely to be used for relative evaluation than absolute evaluation, we report only $E\rho^2$ for all subsequent analysis. The values in the shaded region of Table 2 represent the actual situation (i.e., two observations), other values are based on obtaining $E\rho^2 = .65$ or $E\rho^2 = .80$. For the restricted sample, three observations with either instrument yield a generalizability coefficient meeting the .65 cutoff, but for the full sample four observations are required to meet this threshold. The differences between the samples are the opposite for the .80 cutoff: $E\rho^2$ for the full sample is above .80 with nine observations, but for the restricted sample it is achieved with 10 observations. Possible interpretations and implications of these results are discussed in the next section.

Discussion

We sought to add to the information about generalizability and reliability of the Instructional Quality Assessment (IQA) because of its growing use. In particular, we wanted to better understand how many observations are necessary to reliably measure teachers' instructional practice with the IQA. We found that the number of observations required depends on the group (i.e., universe) of teachers to which you want to generalize (e.g., limited to those with whole-class discussion or the

Table 1
D Study Results of the Univariate t × o Design

Number of observations	Original IQA				Expanded IQA			
	n_t = 150		n_t = 96		n_t = 150		n_t = 96	
	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$
1	.353	.375	.484	.495	.397	.427	.563	.587
2	.522	.546	.652	.663	.568	.599	.720	.740
3	.621	.643	.738	.747	.664	.691	.794	.810
4	.686	.706	.789	.797	.725	.749	.838	.850
5	.732	.750	.824	.831	.767	.789	.866	.877
6	.766	.783	.849	.855	.798	.817	.885	.895
7	.793	.808	.868	.873	.822	.839	.900	.909

Note. The Original IQA used in this study includes only 8 of the 11 rubrics used in the Matsumura et al. (2008) study.

Table 2
D Study $E\rho^2$ Results of the Multivariate $t^\bullet \times o^\bullet \times i^\circ$ Design

Number of Observations	Original IQA		Expanded IQA	
	n_t = 150	n_t = 96	n_t = 150	n_t = 96
2	.519	.576	.554	.626
3	.611	.650	.639	.686
4	.671	.693	.692	.720
9	.802	.781	.804	.785
10	.815	.790	.816	.791

full sample), the modeling approach used (e.g., univariate or multivariate), the instrument used (e.g., the original or the expanded IQA), and the level of reliability desired (e.g., .65 and higher). In the case when the sample of teachers consistently has a whole-class discussion and the level of reliability is set to .65, three observations are needed to reliably measure a teacher's instructional practice using the IQA. There are a number of broader conclusions and implications for the design of studies with respect to sample, instrument, and number of observations.

First, in extending Matsumura et al.'s (2008) study, our univariate results with our restricted sample of 96 teachers were both similar to and different from the results of Matsumura et al.'s analysis with 11 teachers. We found that, in general, the dependability coefficients were lower than those of the prior study. A possible reason for this is that our sample is larger and potentially less homogeneous than

their sample of 11 teachers. It is important to acknowledge that we used only 8 of the 11 instruments they used, and this could have contributed to some additional unexplained variance. One finding that was similar between the two studies was that the inclusion of teachers who did not regularly have a whole-class discussion following work on the task decreased the reliability of the instrument. In our study, we investigated this result by comparing the reliability for the full sample and the restricted sample. This finding has important implications for the use of the IQA: For teachers who do not regularly have a whole-class discussion following work on the task, it generally requires more observations to reliably assess their instructional quality using the IQA. This finding likely has implications for the use of other instruments that similarly focus on specific aspects of instruction: The match between the instrument and the sample will influence the number of observations it takes to reliably assess instructional practice.

Second, with respect to our differing results between the univariate and multivariate studies, in general, reliability coefficients were higher for the univariate studies than the multivariate studies. This is likely because the univariate analyses are a special case of the multivariate analyses that do not consider the correlations between domains (Brennan, 2001). We believe that the multivariate results are more accurate for the IQA because they account for correlations between domains.

Third, the inclusion of the setup rubrics did improve the reliability, yet the marginal benefit tends to decrease as the number of observations increases. This result for the addition of the setup rubrics is consistent with an interpretation that more information about what happened in the classroom is better when you have a limited number of observations. Further, as the number of observations increase, the information about the cognitive demand and discussion might be sufficient to generalize about a teacher's practice. This suggests that when you are planning to do observations, you can make trade-offs between attending to more aspects of instruction (with additional rubrics) and the number of observations. It is likely that adding the rubrics specific to the setup would be less expensive than adding more observations. Further, across the analyses, it is clear that by adding more observations the reliability of the IQA in assessing teachers' typical instructional practice improves. Using generalizability theory, the goal is to be able to generalize from a limited set of observations to the universe of observations (i.e., teachers' typical practice). Recall that these districts were focusing on high quality mathematics instruction in a way that is not typical of urban districts. This suggests that, given the large number of observations required, the IQA might not be an efficient choice for reliably assessing teachers' typical practice in most urban districts. As suggested by Boston (2012),

> The focus of the IQA Mathematics rubrics dictates that they are best suited for assessing reform-oriented instructional practices for use in implementation studies of curriculum or professional development, or to identify changes in the nature of school- or district-wide instructional practice over time. (pp. 96–97)

Lastly, our results demonstrate that the level of reliability deemed sufficient (i.e., the reliability cutoff) will greatly influence the number of minimally necessary observations. The relationship is in fact nonlinear with diminishing returns as the number of observations increases. For example, going from three observations to four observations has a greater payoff with respect to reliability than going from nine observations to 10 observations. These are the types of considerations that will be important when deciding how to balance research costs with reliability desires.

It is important to consider one key limitation of this analysis. We did not consider rater effects, which is another important source of error variation in teachers' scores from observational instruments. Given the rigorous rater training procedures, we believe that the rater effects are likely to be small. However, ideally, future studies will consider this source of variation.

References

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record, 110*(3), 608–645.

Boston, M. (2012). Assessing instructional quality in mathematics. *Elementary School Journal, 113*(1), 76–104. doi:10.1086/666387

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Clauser, B. E., Harik, P., & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicans' clinical skills. *Journal of Educational Measurement, 43*(3), 173–191. doi:10.1111/j.1745-3984.2006.00012.x

Cobb, P., & Jackson, K. J. (2011). Towards an empricially grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development, 13*(1), 6–33.

Cobb, P., & Smith, T. M. (2008). District development as a means of improving mathematics teaching and learning at scale. In K. Krainer & T. Wood (Eds.), *International Handbook of Mathematics Teacher Education: Vol. 3. Participants in mathematics teacher education: Individuals, teams, communities and networks* (pp. 231–254). Rotterdam, the Netherlands: Sense Publishers.

Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational Psychologist, 23*(2), 167–180. doi:10.1207/s15326985ep2302_6

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/content.asp?contentid=17292

Hill, H. C., Charalambos, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. doi:10.3102/0013189X12437203

Jackson, K. J., Garrison, A. L., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education, 44*(4), 646–682. doi:10.5951/jresematheduc.44.4.0646

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment, 13*(4), 267–300. doi:10.1080/10627190802602541

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

O'Connor, M. C., & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussion. In D. Hichs (Ed.), *Discourse, learning, and schooling* (pp. 63–103). New York, NY: Cambridge Unversity Press.

Quint, J. C., Akey, T. M., Rappaport, S., & Willner, C. J. (2007). Instructional leadership, teaching quality, and student achievement: Suggestive evidence from three urban school districts. Retrieved from http://www.mdrc.org/sites/default/files/full_406.pdf

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry, 29*(3), 315–342. doi:10.1111/0362-6784.00130

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*(4), 313–340. doi:10.1080/10986060802229675

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488. doi:10.3102/00028312033002455

Sztajn, P., Wilson, P. H., Edgington, C., & Confrey, J. (2011). Learning trajectories and key instructional practices. In L. R. Wiest & T. Lamberg (Eds.), *Proceedings of the 33rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 434–442). Reno, NV. Retrieved from http://www.pmena.org/proceedings/PMENA%2033%202011%20Proceedings.pdf

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 81–124). Amsterdam, the Netherlands: Elsevier.

Authors

Anne Garrison Wilhelm, Department of Teaching and Learning, Simmons School of Education and Human Development, Southern Methodist University, P.O. Box 750455, Dallas, TX 75275-0455; awilhelm@smu.edu

Sungyeun Kim, Seoul National University, #611, Education Information Hall (10-1 Dong), 1 Gwanak-ro, Gwanak-gu, Seoul, South Korea 151-742; sykim0401@snu.ac.kr

Submitted June 9, 2014
Accepted October 7, 2014