

HOW MANY CLASSROOM OBSERVATIONS ARE SUFFICIENT?

Empirical Findings in the Context of a Longitudinal Study

Jeffrey C. Shih

University of Nevada, Las Vegas

Marsha Ing

University of California, Riverside

James E. Tarr

University of Missouri

One method to investigate classroom quality is for a person to observe what is happening in the classroom. However, this method raises practical and technical concerns such as how many observations to collect, when to collect these observations and who should collect these observations. The purpose of this study is to provide empirical evidence to address these concerns using a particular middle school mathematics classroom observation tool. Data included multiple ratings of mathematics classrooms from 69 middle school teachers within 11 districts that were analyzed using generalizability theory. Findings suggest that raters trained to use this particular measure required 3 observations to consistently capture habitual classroom environments. Implications for investigating classroom quality using this and other classroom observation tools should be guided by decisions about the specific purpose of the observation tool, as well as budget and practical considerations.

The Association for Middle Level Education (2010) offers Curriculum, Instruction, and Assessment as one of the three areas used to organize research-based characteristics of successful middle level education. These characteristics describe the learning opportunities that should be provided for middle grade students, among them engagement in “active, purposeful learning” and a “challenging, exploratory, integrative, and relevant” curriculum. Teachers should be prepared to “use mul-

iple learning and teaching approaches.” Teachers are asked to move beyond superficial coverage of content to using curriculum that “addresses substantive issues and skills, is geared to (students’) levels of understanding, and increasingly enables them to assume control of their own learning” (p. 18).

The challenges as well as the need to engage in research that reflects the reality of classroom teaching and learning are well documented (Confrey et al., 2008). Documenting

• **Jeffrey C. Shih**, Associate Professor, The Department of Teaching and Learning, University of Nevada, Las Vegas, NV.
E-mail: jshih@unlv.nevada.edu

Middle Grades Research Journal, Volume 8(2), 2013, pp. 41–49
Copyright © 2013 Information Age Publishing, Inc.

ISSN 1937-0814
All rights of reproduction in any form reserved.

actual teaching practices requires reliable observational data of teachers in their classrooms but such data are not easy to obtain for many reasons (Chval, Reys, Reys, Tarr, & Chávez, 2006). First, external observers represent an intrusion into classrooms and can disrupt the regular classroom routine. Second, the measurement challenges of classroom data raise concerns about how accurately one is able to capture classroom teaching and learning (e.g., Hiebert & Grouws, 2007). For example, training different observers to focus on the same particular teacher behaviors and characterize those behaviors in the same way can be challenging. Furthermore, observing teachers and coding information related to the teaching practices and the learning environment is both time consuming and costly. Due to budget and time constraints, much research in schools is carried out without capturing detailed information about what teachers actually do within their classrooms (National Research Council, 2004). Ideally, the goal of conducting classroom observations is to collect sufficient data to characterize the teaching and learning practices, but not intrude upon teachers or spend more observational time than is necessary (e.g., Marcoulides, 1997). To accomplish the goal of conducting classroom observations in a way that balances measurement and practical concerns, one might assume that including more observations of teachers in their classrooms provide more accurate profiles of their teaching and learning environments. While this may be the case, nevertheless, there are tradeoffs in terms of the measurement and cost to consider.

This study compares the consistency of observational data characterizing teaching and learning environments across different observations. This study is based on data collected from a research study conducted at the middle school level that focused on teachers' use of different mathematics curricula and their impact on student learning, the Middle School Mathematics Study (Tarr et al., 2008). The research study documented teaching practices and curriculum implementation via multiple

classroom observations over 2 years using a variety of measures, including the Middle School Mathematics Observation Tool, which was adapted from an observation tool used in the Wisconsin Longitudinal Study (Romberg & Shaffer, 2004). Observations were made during the fall, winter, and spring of each year. As researchers, we wanted to know how many observations per teacher would be sufficient to accurately characterize teaching. Suppose we had used only the fall observation: would this have yielded a markedly different profile of teaching? What if only two observations (fall and winter, fall and spring, or winter and spring) were used? This study addresses these practical questions by providing an empirical example to illustrate the quality of inferences drawn from fewer or greater number of classroom observations.

CONCEPTUAL FRAMEWORK

The conceptual framework for the development of the observational tool comes from the Curriculum Research Framework of the Center for the Study of Mathematics Curriculum (Center for the Study of Mathematics Curriculum, 2012), which captures the interconnectedness of various types of curriculum. Porter and Smithson (2001) distinguish the intended curriculum (local, state and/or national curriculum standards) from the assessed curriculum (represented by high-stakes tests), and the enacted (implemented) from the learned curriculum (the content that students learn).

Our conceptual framework focuses primarily on links between the textbook curriculum (curriculum materials provided to the students and teachers by schools), implemented curriculum, and learned curriculum. Specifically, we examine the impact of two distinct types of textbook curriculum (National Science Foundation funded and publisher developed) on the implemented curriculum and, in turn, on the learned curriculum by looking at the relationship between two distinct types of textbooks

(as implemented by teachers), the classroom learning environment, and student learning outcomes. It was assumed that capturing the implementation of the classroom learning environment would be fairly stable or consistent, independent of both curriculum type and content focus throughout the year (Ing & Shih, this issue). Given the assumptions of the phenomena being measured, the purpose of this study is to explore a particular measurement concern related to this observational tool. More specifically, this present study was guided by the research question: using our observation tool, how many classroom observations are necessary to accurately characterize teaching? This research question is specific to our particular observation tool. Findings are not meant to apply to other observation tools with different assumptions about the phenomena being measured. For example, for another observation tool, the purpose might be to capture aspects of instruction that are not stable over time. In this other situation, it might be the case that the number of observations required for an acceptable level of reliability differs from the observation tool described in this study.

METHOD

Participants

The design for the overall study included both quantitative and qualitative methodology. Initially over 30 districts were contacted about participating in the research study. The solicitation of districts was based on a number of criteria: number of years since adoption of the current mathematics textbook series, organizational structure of the middle school, and size of the community in which the school was located. Districts and schools were not considered if they were in their first year of implementation of any textbook. While districts and students were not offered any compensation for their participation, teachers were offered a

\$495 stipend for their time (e.g., attending an orientation meeting, completing surveys, allowing classroom observations, administering and returning project-related exams). In addition, a \$250 school stipend was offered for provision of student achievement data, class rosters, and ongoing communication with project staff, including local data collectors in accordance with internal review board guidelines and approval. The principal investigators of National Science Foundation (NSF) Instructional Materials Development grants were contacted to provide names of school districts that were past the first year of implementation. Once six schools/districts using the NSF-funded materials were selected, schools that used published-developed mathematics textbooks with similar demographics and a significant market share based on the 2000 Mathematics and Science Education Survey (Weiss, Banilower, McMahon, & Smith, 2001) were identified.

The eleven middle schools (Grades 6-8) within each of the eleven districts also reflected a range of student demographic characteristics. One mathematics teacher from each school was selected to participate based on the criteria described earlier. Six districts were using the same NSF-funded textbooks and represented a range of student demographics, particularly with regards to ethnicity and percentage of students qualifying for free or reduced lunch. The other five districts used non-NSF-funded textbooks but had similar student demographics as the other six schools. As part of the data collection for the overall study, teachers completed a textbook-use-diary during three 10-day intervals (October, January, and March). These intervals were selected to avoid testing days, field trips, or other potential conflicts with regular instructional days. Classroom observations were scheduled to occur on days where teachers made entries in their textbook-use-diaries, allowing for corroboration of the teacher self-reported data in textbook-use-diaries.

Middle School Mathematics Study Observation Tool

The Middle School Mathematics Study Observation Tool was used to measure teaching strategies related to the National Council of Teachers of Mathematics (NCTM) process standards (NCTM, 2000). The observation tool and its training manual were adapted from a tool developed at the University of Wisconsin (Romberg & Shafer, 2004). The practices measured by this observational tool are assumed to be habitual. In other words, the authors did not expect that these practices would vary due to the particular mathematical content or the types of students in the classroom and none of the adaptations interfered with this expectation. Each teacher was scored on the following five items:

1. The lesson provided opportunities for students to make conjectures about mathematical ideas.
2. The lesson fostered the development of conceptual understanding.
3. Students explained their responses or solution strategies.
4. Multiple strategies were encouraged and valued.
5. The teacher valued students' statements about mathematics and used them to build discussion or work toward shared understanding for the class.

Each item was scored using a 3-point Likert scale. For each teacher and observation, scores across these five items were averaged (Table 1).

Raters

There were eight raters who participated in this study. The group of raters included experienced researchers, advanced graduate research assistants, and retired teachers. All raters had ample classroom experience and were familiar with the curriculum materials used by the participants. Prior to data collection, all raters were trained together in the use of the tool. Specifically, raters discussed what the tool was designed to measure and talked through examples of each type of event. Observers watched tapes of different classrooms (both from teachers using publisher-generated textbooks and NSF-funded textbooks) and independently coded each lesson. Their codes were compared and differences were discussed. This process was iterated several times, using different tapes, until all observers had equal or almost equal codes for each classroom event. At the end of the training session, raters independently watched a sample of classroom videos and achieved exact agreement on over 90% across all five items. After the training session, raters independently observed teachers on three occasions. For the most part, different raters observed teachers and there was a significant time lapse between observations (2-3 months). For each teacher, the first observation occurred in October, the second in January or February, and the final observation occurred in April.

The observation tool was designed to document specific elements of classroom instruction that can reasonably be expected to be fairly constant for a given teacher. For example, one of the classroom events is "Students explained their responses or solution strate-

TABLE 1
Descriptive Statistics for Range of Scores From Composite Measure of Teaching Practice

<i>Observation</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Minimum</i>	<i>Maximum</i>
1	69	1.61	0.59	1	3
2	69	1.58	0.59	1	3
3	69	1.70	0.60	1	3

gies.” It is conceivable that a teacher may do this only once during the one time he or she is being observed (or not do it that one time, and do it for the rest of the school year). In this scenario, one observation might not capture this teacher’s practice. Two or three (or more) observations, conducted in the span of the entire school year, gives a better picture of whether this is a stable practice or one that is dependent on the topic being taught. If the observations were collected on adjacent days, it is likely that these observations are similar due to the fact that the teacher is teaching the same content or lesson. Therefore, spacing the observations throughout the entire school year addressed the goals of this particular observation tool.

The specific purpose of this observation tool guided decisions on how many observations to collect, when to collect these observations and who should collect these observations. For example, the decision was made to conduct observations at three different times throughout the school year rather than three observations that covered the same lesson. This was an attempt to document a sample of instruction throughout the school year on various lessons. It does not provide specific information about any particular lesson but rather provides information about instructional practices that may be habitual for the teacher under observation. Other studies may be interested in research questions that would require a finer-grained observation for which it is important to document even smaller variations in, for example, certain teacher behaviors. For example, researchers interested in investigating changes in novice teacher practice during his or her first year of teaching might need more than three evenly spaced observations throughout the year to provide information about the specific behaviors under study.

Analysis

We investigated the dependability of the observational measure using generalizability theory (Brennan, 2002; Cronbach, Gleser,

Nanda, & Rajaratnam, 1972; Macroulides, 2000; Shavelson & Webb, 1991). Generalizability theory was used to estimate the magnitude of multiple sources of error and provided a reliability (generalizability) coefficient for the proposed use of the observational measure. It helped isolate the sources of variation to think through designing, developing and administering an observational tool that was cost and time efficient. For the data used in this study, we investigated teachers as the object of measurement and one source of variation that contribute to errors in measurement: observations. All of the teachers in the selected sample were observed for the same number of observations, referred to as a one-facet design with person-crossed-with observations ($p \times o$). Observations were the only potential source of error included in this particular measurement procedure. The variance components were calculated using mean squares from an analysis of variance, equating these to their expected values and solving a set of linear equations. The advantage of using generalizability theory was to interpret variance components, particularly, the main effect difference from one occasion to another. Generalizability theory also allowed us to estimate the overall reliability of the measure with three observations and allowed us to calculate how this reliability index varied if the number of observations were changed (cf. Brennan, 2002; Shavelson & Webb, 1991 for greater detail on these calculations). These coefficients ranged from 0 to 1.0, with higher values reflecting more dependable procedures. To improve dependability of the measure, results from a generalizability study were used to calculate a “what-if” decision study (d-study). This provided information about what could be done differently if different sources of error were modified. Through a d-study, information about potential tradeoffs in future implementation of the measure were determined. In this particular study, we only had one source of error, occasions, so the d-study provided information about how the reliability coefficients changed with different numbers of observations.

TABLE 2
Analysis of Variance Mean Squares and Estimates of Variance Components

<i>Sources of Variation</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Squares</i>	<i>Estimated Variance Components</i>	<i>Percentage of Total Variance</i>
Teacher	54.60	68	0.80	0.23	63.70
Observation	0.57	2	0.28	0.00	0.60
Residual	17.14	136	0.13	0.13	35.60
Total	72.31	206			

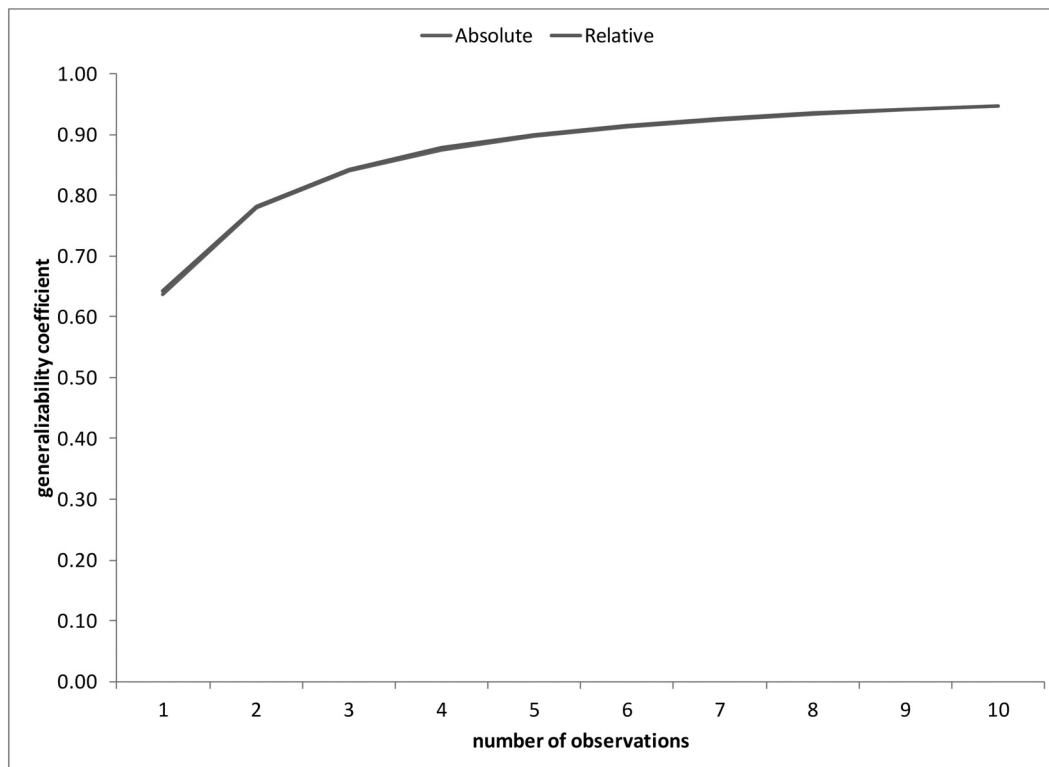


FIGURE 1
Absolute and Relative Coefficients

RESULTS

Findings from the generalizability theory model confirmed the lack of variability due to observations (Table 2). Less than 1% of the variability was due to differences between observations. The lack of variability indicated

that on average, teachers received the same scores regardless of whether they were observed at the beginning, middle or end of the school year. Most of the variation (64%) was due to differences between teachers. Approximately 36% of the variation was due to the interaction between teachers and occasions.

The relative generalizability coefficients which provide information about relative or comparative decisions was high (.84) for three observations. This suggests that this composite measure of teaching practices provides a sufficient sample of classroom events with three observations. The d-study varying the number of observations suggests an acceptable level of reliability with three observations (Figure 1).

DISCUSSION

The purpose of this study raises issues that advocate for a closer examination of the measurement qualities of the numerous and widely available classroom observation tools. We explored the measurement qualities of one particular classroom observation tool, the Middle School Mathematics Study Observation Tool, and found evidence to suggest that trained raters using this particular observation tool required only three observations to capture seemingly habitual teaching strategies. It is important to note that these findings cannot be generalized beyond the five items included in this observational tool and the particular construct measured by these five items. Other constructs might require additional items or additional occasions. For example, less stable teaching practices such as a characterizing how a teacher interacts with students in his or her classroom around misconceptions might require a greater number of observations to capture the variation. In other words, a teacher's interactions might vary considerably with different students depending on the content area. Thus, in some circumstances more observations might be needed to capture the range of ways in which teachers interact with students in their classrooms for that particular construct.

While we emphasize that these results are applicable to this specific research tool, considering the time and cost associated with increasing the number of observations has a high practical value. These results provide a significant starting point for a much needed

discussion among mathematics education researchers involved in large-scale studies about the measures that are currently being used for classroom observation. These results should inform discussion about both the scale of the measures and the stability of the constructs over time.

Although this study provides evidence that trained raters only require three observations when using this particular classroom observation tool, it also raises additional questions and suggests future steps for the use of other observational tools. First, we are concerned with the characteristics and preparation experience of raters who would carry out these observations. In the original study, raters were closely associated with the development of the observation tool. Furthermore, all of the raters were experienced researchers and each had extensive (more than 15 years) teaching experience. It is not clear whether raters without such characteristics would be able to carry out the classroom observations with the same degree of reliability. We were unable to examine raters as a source of variability with this particular study. From a measurement and study design perspective, teachers were not observed by the same rater for each occasion. Future research should examine the extent to which raters are a source of variation, particularly the extent to which rater characteristics such as mathematical knowledge and mathematics teaching knowledge or experience interact with the mathematics demands or focus of the observational tool. In addition, the research should examine the extent to which training and calibration may be able to overcome the lack of mathematical knowledge or experience for particular observations tools but perhaps not for other tools.

Second, we are concerned about the concurrent validity of this classroom observation tool with other measures of classroom quality. This measure is one particular way of defining and measuring a particular construct of classroom quality but there are others such as COSMIC (McNaught, Tarr, & Grouws, 2008), LieCAL (Cai, Moyer, Wang & Nie, 2011), SCALE

(Stein & Kaufman, 2010), and Inside the Classroom: Observation and Analytic Protocol (Horizon Research, 2009). Furthermore, these results were based solely on the five-classroom environment items described above; the complete Middle School Mathematics Study Observation Tool includes other attributes of the classroom learning environment that were not analyzed. Next steps will include investigations of the extent to which our particular definition of teaching practices relates to other measures of classroom quality. We are particularly interested in learning more about the characterizations of teachers that vary considerably from one observation tool to another.

Finally, we are also concerned about both the practical and conceptual issues in using classroom level observational measures that do not explicitly capture individual student participation (Ing & Webb, 2012). Some observation tools focused on students provide an aggregate picture of student participation by sampling a few students to be observed while others focus more exclusively around teacher behaviors. For practical reasons, it is important to weigh the costs associated with collecting data at the student level compared with classroom-level observations. Moreover, an important measurement question that we propose to investigate further is the extent to which classroom-level measurement of student participation adequately represents the possible range of students' experiences in the classroom. This question of validity will be investigated further by comparing classroom-level codes like those currently used in the Middle School Mathematics Observation Tool to student-level codes where individual students were identified and their participation measured.

We hope that researchers continue to explore the magnitude of multiple sources of variation for the proposed use of their observational measures (e.g., Borko et al., 2006; Hill, Charalambous, & Kraft, 2012; Martinez, Borko, & Stecher, 2012; Matsumura, Garnier, Pascal, & Valdés, 2002; Newton, 2010; Pianta, Hamre, Haynes, Mintz, & La Paro, 2007). We also hope that researchers continue to collabo-

rate with each other to develop these multiple measures and that these measures help inform and improve teacher instructional practices and the quality of instructional opportunities for all students. Moreover, we think that our approach suggests a model to investigate similar questions related to other similar classroom observation data. This could provide researchers a way to compare the results we present in this study with those from other studies.

REFERENCES

- Association for Middle Level Education. (2010). *This we believe: Keys to educating young adolescents*. Westerville, OH: Author.
- Borko, H., Stecher, B. M., Martinez, F., Kuffner, K. L., Barnes, D., Arnold, S. C., ... Lou, M. (2006). *Using classroom artifacts to measure instructional practice in middle school science: A two-state field test* (CSE Technical Report 690). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards and Student Testing.
- Brennan, R. L. (2002). *Generalizability theory*. New York, NY: Springer.
- Cai, J., Moyer, J. C., Wang, N., & Nie, B. (2011). Examining students' algebraic thinking in a curricular context: A longitudinal study. In J. Cai & E. Knuth (Eds.), *Early algebraization: A global dialogue from multiple perspectives* (pp. 161-186). New York, NY: Springer.
- Center for the Study of Mathematics Curriculum. (2012). Curriculum research framework. Retrieved from http://mathcurriculumcenter.org/research_framework.php
- Chval, K. B., Reys, R. E., Reys, B. J., Tarr, J. E., & Chávez, O. (2006). Pressures to improve student performance: A context that both urges and impedes school-based research. *Journal for Research in Mathematics Education*, 37(3), 158-166.
- Confrey, J., Battista, M. T., Boerst, T. A., King, K. D., Reed, J., Smith, M. S., ... Sutton, J. T. (2008). Situating research on curricular change. *Journal for Research in Mathematics Education*, 39(2), 102-112.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

- Hiebert, J. S., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester, Jr. (Ed.), *Second handbook on mathematics learning and teaching* (pp. 371-404). Charlotte, NC: Information Age.
- Hill, H. C., Charalambous, 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.
- Horizon Research. (2009). *Inside the classroom: Observation and analytic protocol*. Retrieved from <http://www.horizon-research.com/instruments/clas/cop.php>
- Ing, M., & Shih, J. C. (2013). Using generalizability theory as a framework for informing measurement issues in middle school settings. *Middle School Research Journal*, 8(2), 25-40.
- Ing, M., & Webb, N. M. (2012). Characterizing mathematics classroom practice: Impact of observation and coding choices. *Educational Measurement: Issues and Practice*, 31(1), 14-26.
- Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement*, 57(5), 808-812.
- Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527-551). San Diego, CA: Academic Press.
- Martinez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, 49(1), 38-67.
- Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, 8(3), 207-229.
- McNaught, M., Tarr, J., & Grouws, D. (2008, March). *Assessment curriculum implementation: Insights from the Comparing Options in Secondary Mathematics: Investigating curriculum (COSMIC) project*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 evaluations*. Washington, DC: National Academy Press.
- Newton, X. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation*, 36, 1-13.
- Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). *Classroom assessment scoring system manual, middle/secondary version*. Charlottesville, VA: University of Virginia.
- Porter, A. C., & Smithson, J. L. (2001, December). *Defining, developing, and using curriculum indicators* (CPRE Research Report No. RR-048). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education. Available: <http://www.cpre.org/Publications/rr48.pdf>
- Romberg, T. A., & Shafer, M. C. (2004). *Purpose, plans, goals, and conduct of the study* (Mathematics in Context Longitudinal/Cross-Sectional Study Monograph 1). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.
- Stein, M. K., & Kaufman, J. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal*, 47(3), 663-693.
- Tarr, J., Reys, R., Reys, B., Chávez, O., Shih, J., & Osterlind, S. (2008). The impact of middle grades mathematics curricula on student achievement and the classroom learning environment. *Journal for Research in Mathematics Education*, 39(3), 247-280.
- Weiss, I. R., Banilower, E., McMahon, K., & Smith, P. S. (2001). *Report of the 2000 National survey of science and mathematics education*. Chapel Hill, NC: Horizon Research.