NATIONAL ACADEMY of EDUCATION

Workshop to Examine Current and Potential Uses of
NCES Longitudinal Surveys
by the Education Research Community

Collecting Evidence of Instruction with Video and Observation Data in NCES Surveys

Pam Grossman
Stanford University

National Academy of Education

Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Education Research Community

Collecting Evidence of Instruction with Video and Observation Data in NCES Surveys

Pam Grossman
Stanford University

December 2014

This paper was prepared for the National Academy of Education's *Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Education Research Community*, held on November 5-6, 2013, in Washington, DC. The workshop was organized to provide input to the National Center for Education Statistics (NCES) on how its longitudinal survey program could be enhanced to serve the changing needs of the education research community in light of (1) changing technological opportunities, (2) changing data availability (with an increasing focus on the classroom and learning outcomes), and (3) a changing U.S. population. The collection of commissioned papers reflects the views of the authors and not necessarily those of the National Academy of Education or the U.S. Department of Education. This paper was supported with funding from the Institute of Education Sciences.

OVERVIEW OF USE OF VIDEO AND OBSERVATION DATA FOR RESEARCH ON TEACHING AND LEARNING

Although videos and observation instruments have a long history in research on teaching, relatively few National Center for Educational Statistics (NCES) surveys take advantage of these methodologies for data collection. In this paper, I provide a short overview of how these methods have been used both by NCES and in other large-scale studies of teaching and learning and explore how such data could be useful for researchers in the context of NCES surveys, with a particular focus on longitudinal surveys.

Observation Instruments

Classroom observation enjoys a long history in research on teaching. In fact, classroom observations of many kinds, from longitudinal ethnographic studies to more quantitative observations of teacher–student interaction, have been a primary method for studies of instruction (Evertson & Green, 1986). The instruments used for classroom observation differ dramatically along a number of dimensions, including the grain size of teaching practices captured, the level of inference required by observers, and the focus on particular aspects of teaching practice. Early studies of teaching, including the process-product research of the 1960s (e.g., Dunkin & Biddle, 1974) made use of a number of structured observation instruments, including the Flanders' Interaction Analysis Categories, which coded teacher–student interactions every 30 seconds. The categories captured both teacher talk and student talk and included such elements as teacher praise, teacher questions, and student response or confusion. In contrast, more recent observation protocols, such as those used in the Measures of Effective Teaching study, capture a broader set of teacher–student interactions, including emotional support captured by the CLASS protocol (LaParo et al., 2004), teacher questioning as captured by the Framework for Teaching (Danielson, 2007), the precision of mathematical concepts presented to students as captured by the Mathematical Quality of Instruction (Hill et al., 2008), or the intellectual demand of classroom talk and tasks, as captured by the Protocol for Language Arts Teaching Observation (Grossman et al., 2013).

Observation protocols differ by purpose, scope, and function. Some observation protocols are designed to capture instruction across grade levels and subject areas (e.g., Danielson Framework), whereas others are designed for specific subject matters (MQI, PLATO, etc.). Some were designed initially for purposes of professional development (Framework), while others were initially designed for research on the quality of instruction, from early childhood (CLASS) to middle school English and Language Arts (ELA), (PLATO). The observation protocols also differ with regard to the grain size of the elements of interaction being observed and the sampling of segments for scoring. In the Flanders scale, for example, raters score interactions every 30 seconds, coding for both teacher and student behaviors. In contrast, the CLASS originally was designed to score 20-minute segments, coding more generally for elements such as positive or negative climate in a classroom. Each of these features of an observation system affects the nature of the data to be collected and the inferences that can be made from those data.

Although current discussions of classroom observation instruments are heavily framed by the goal of teacher evaluation, most of the original work on structured classroom observation was designed to make inferences about the qualities of *teaching* rather than to make judgments about individual teachers. In other words, observation protocols are often focused on a variety of pedagogical interactions between teachers and students, rather than on specific characteristics of an individual teacher (Gitomer & Bell, 2013). Researchers (e.g., Pianta, Hamre) seek both to describe the quality of instruction experienced by children and the relationships between elements of instruction and student outcomes, including both achievement and engagement. Such an approach would seem to be quite congruent with many of the NCES surveys that try to capture the qualities of educational environments experienced by American youth.

Use of Observations by NCES Studies

Despite the ubiquity of classroom observation in research on teaching, there are relatively few examples of the use of observations in NCES studies, and most of these occur in the context of studies of early childhood. The Early Childhood Longitudinal Birth Study (ECLS-B) survey, a nationally representative longitudinal study of approximately 11,000 children born in 2001, included an observational component in which a trained observer rated the quality of children's nonparental care and education settings before the children reached kindergarten. These observations focused on the quality of settings for child care rather than on instruction, even though one of the constructs on which education settings were rated was learning activities. The study included two waves of observation, once when children were 2 years old and a second wave when they were in preschool. Trained raters observed a subsample of settings using several instruments: The Family Day Care Rating Scale for home-based settings, the Infant-Toddler Environment Rating Scale for center-based settings, and the Early Childhood Environment Rating Scale (ECERS) for center-based settings when children were preschool aged. Observers also counted children and providers on multiple occasions to get a reliable estimate of the ratio of children to child care providers. Additionally, observers assessed the quality of interactions between children and adults using the Arnett Caregiver Interaction scale. These observations were used to provide relatively comprehensive data about children's experiences in their first 5 years of life, with a focus on the environments in which they spend their time and their interactions with caregivers. It is also worth noting that the observation instruments were designed to capture the quality of environments, rather than specific qualities of individual caregivers.

The SWEEP study assessed the pre-K quality of statewide early childhood education programs in 11 states. For this study, researchers collected data on both structural and procedural qualities of early childhood education using three different observation protocols, the CLASS, ECERS, and Snapshot. CLASS has been widely used in studies of early childhood to look at interactions between children and their teachers. ECERS measures qualities of the child care environment, while Snapshot measures the proportion of class time spent on a variety of different activities by focusing on target children and the approach to instruction during that time over a series of 10-minute intervals. Each of these instruments captures different qualities of the settings in which children spend their time and the nature of their interactions with caregivers and teachers. The purpose of the study was to provide detailed information about the quality of pre-K programs; having reliable instruments that could be used across settings and across states was a critical component of the research.

Uses of Video in Large-Scale Studies

While many observation instruments require raters to actually visit the settings they are rating, the increasing use of video recordings has enabled researchers to collect video data that can be coded using multiple instruments and at multiple points in time. As video recordings have become increasingly cost effective, the use of video for studies of instruction has increased. One of the best known examples of the use of video collection on a large scale was the TIMSS Video Study, part of the Trends in International Mathematics and Science Study (TIMSS) that provides information on student mathematics and science achievement for fourth- and eighth-grade students in the United States and a range of other countries. In 1995, TIMSS launched a Video Study of eighth-grade classrooms in three countries, the first time that video data were incorporated into an NCES Achievement Study. In 1999, this study was then repeated and expanded to include a study of classroom teaching practices in the different participating countries by collecting videotapes of eighth-grade lessons in mathematics and science in 7 countries: the United States, Australia, the Czech Republic, Hong Kong SAR, Japan, the Netherlands, and Switzerland (Stigler et al., 2000). The TIMSS Video Study was designed to understand more deeply how mathematics and science instruction might differ across the United States and high-achieving countries and to identify any patterns of instruction that might be associated with higher performance.

The mathematics part of the TIMSS Video Study included 638 eighth-grade lessons collected from all participating countries. Lessons were randomly selected and the teacher was videotaped for only one lesson, and lessons were videotaped across the school year, to capture the variety of topics and activities that might occur across the full year.

To make inferences about an individual teacher would require more than a single instance of instruction. However, it is important to note that in this study the unit of analysis was not the individual teacher, but rather *teaching* in a national context. Researchers looked for patterns of instruction by country, rather than looking for variations among teachers within the same country. Researchers hoped to make inferences about how the kind of teaching mathematics students received in a particular country might be related to overall patterns of achievement. The video records of lessons enabled researchers to look for more subtle patterns in classrooms that differentiated the quality of instruction students actually received. For example, one analysis, that would not have been possible without such video data, found that even when American teachers started with a challenging mathematical task, they rather quickly moved to decrease the intellectual challenge of the task (Stigler et al., 2000). Because teachers were unlikely to be aware of these small adaptations noted in the videos, surveys of teachers' instructional practice would be unlikely to surface this kind of consequential difference in instruction. Live observation using structured observation protocols would also miss this aspect of instruction if researchers did not know ahead of time to focus on the maintenance of cognitive demand in ratings on instructional quality.

The TIMMS study of mathematics instruction in a variety of countries demonstrates the utility of capturing videos of teaching and learning in mathematics in a cross-national study. The systematic collection of such video data enables researchers to examine closely the differences in mathematics instruction across cultures through fine-grained video analysis and to employ multiple analytic frameworks to analyze video records. The TIMSS study also demonstrates the value of making videos accessible to other researchers and teacher educators. Of the more than 800 videos collected for this study, 53 were designated as public use lessons, available for other researchers and educators to use. These videos have been widely used not only for secondary analyses of the data, but for use in teacher education and professional development.

This example suggests the value of video recordings as data, as compared with live observations using structured protocols. Video recordings provide a more open archive of data that are available for subsequent analyses and reanalyses. Researchers using different conceptual frameworks and perspectives can use the same data to surface different aspects of instruction.

Other than the TIMSS study, the use of video within NCES surveys has been primarily confined to the study of early childhood, as part of the ECLS-B. In this study, researchers video-recorded parent–child interactions to supplement their understanding of the home environments. Video recordings were collected in children's homes when the children were 9 months, 2 years, and 4 years of age.

Measures of Effective Teaching

The Measures of Effective Teaching (MET) project provides a more recent example of the collection of video records of instruction across 3,000 teachers in grades 4–9 in six districts across the country (Kane & Staiger, 2012). This project explicitly set out to look at different measures of teaching and the relationships among these measures. The primary purpose for this study had to do with developing better information for policy makers regarding the use of data for teacher evaluation and how multiple measures of teacher and teaching quality could be combined for purposes of evaluation. Although data were collected across two years in many classrooms, there was no intent to use the data for longitudinal purposes.

Participating teachers were videotaped for four lessons, two on focal topics selected by researchers and two on topics of the teacher's choice. The study used panoramic cameras that captured a 360-degree view of classrooms and an additional camera to capture work on the board or overhead. The camera setup was designed to be easily used by school personnel. The data from this study include both the original video recordings of lessons and raters' scores of these lessons generated by multiple observation protocols.

Although most of the analyses of the MET data use these scores without reference to the videos from which they were generated, the fact that the original videos are archived provides opportunities for researchers to return to the video for secondary analyses or to understand more qualitatively the instruction that led to these scores. As one example, Cohen (2013) recoded all of the fourth-grade lessons in a single district to see how the instructional practices captured by one of the protocols, PLATO, differed in math and ELA lessons. As math lessons were not originally coded using PLATO, which was designed primarily for ELA, such an analysis would not have been possible without access to the original videos.

Although observations and videos have not been widely used within NCES surveys, they are an important component of other large-scale studies of instruction and the quality of educational settings, as discussed above. The technology for employing video capture and storage has improved enormously over the past decade, making it more feasible to conduct large-scale studies using video. In the sections that follow, I explore the kinds of questions researchers might be able to ask if observation and video data were incorporated into NCES surveys.

THE QUESTIONS WE COULD ASK

The longitudinal studies of youth conducted by NCES collect substantial data about the experiences of children and youth from early childhood onward. The data allow researchers both to document the quality of children's experiences in and out of school and to investigate a range of questions about the factors that influence children's cognitive, emotional, and social development. The data include information about schools and teachers, collected primarily through questionnaires sent to teachers and school administrators. The teacher questionnaires ask teachers to comment on the particular focal child, as well as to respond to questions about the nature of instructional activities in the classroom.

A great deal of large-scale observation data is currently being collected on early childhood education, including an extensive study of Head Start classrooms using the CLASS protocol. Ironically we know more about the quality of teacher–student interactions in preschool than we do in the much longer time span from kindergarten through college. In the section that follows, I raise important questions regarding teaching and learning that we are currently unable to answer and suggest how incorporating video or observational data into NCES surveys might help fill these gaps in our knowledge.

Questions About Teaching

Research in teaching is increasingly interested in looking at the mechanisms that link the quality of teaching to student learning. While an impressive body of evidence suggests that teachers matter in their impact on student achievement, there is less evidence about the particular characteristics of teachers or teaching that account for differential impact on student gains.

Currently, the data about instruction collected in NCES surveys is quite thin, consisting almost entirely of teacher responses to survey items. Collecting observational or video data in a large, representative sample of classrooms would enable researchers to explore more deeply the instructional factors that matter most in helping to explain differences in student achievement. Such data would also provide a snapshot of the quality of instruction experienced by students, and how instructional quality might vary by grade level, content, school context, and region.

If we are to improve the settings in which children spend significant amounts of time, we would also need longitudinal data on these settings and the factors that might help improve the quality of interactions between children and their teachers. However, we currently lack longitudinal data that would enable us to look at how and if teachers improve over time. Large-scale quantitative research suggests that there are student gains to teacher experience in the early years of teaching, but there are few data to suggest whether this increased impact is accompanied by changes in classroom teaching. Does teacher impact on student achievement improve in these early years because teachers actually improve the quality of their practice? Having video or observational data on instruction accompanied by the kind of data collected by the Schools and Staffing Survey (SASS), for example, would enable researchers to explore how classroom instruction changes over time, if at all, and what factors might help explain change or lack of change. For example, the SASS collected information about teachers' involvement in professional development, but there is little opportunity to investigate how such involvement might affect the nature of instruction. By incorporating video or observation data from classrooms into such surveys for a carefully sampled population of teachers, researchers could inquire in the relationship between professional development (PD) and instruction. Is involvement in professional development accompanied by changes in classroom teaching? How long do such changes take? Do teachers who report greater involvement in PD differ in instructional practice? If we hope not just to document but to improve the quality of instruction, we would need answers to such questions.

Questions About Policy and Teaching

As part of school reform efforts, policy makers mandate new initiatives, including new standards for student learning, new assessments, reformed curriculum, and other policy tools. The current effort to implement the Common Core State Standards (CCSS) and new CCSS-linked assessments is a good example of such a policy initiative. While such initiatives are intended to improve the quality of student experiences and learning in school, we have relatively few data to suggest whether and how instruction or teacher–student interactions change in response to such initiatives. In order to investigate this question, we would need longitudinal observation or video data on classroom instruction in a representative sample of classrooms. Efforts to improve the quality of math and science instruction, for example, might benefit from longitudinal data on classroom teaching, building, perhaps, on the TIMSS studies. In this case, the effort is not to make inferences about an individual teacher but on the nature of science, technology, engineering, and math teaching over time and link these, if possible, to changes in policy.

We are also entering an era in which districts are investing heavily in different forms of technology, from iPads to Smart Boards, to blended learning. While some see such tools as the way to individualize and differentiate instruction, others have expressed concerns about how these new technologies will affect the quality of teacher–student interactions. Investing in classroom observations of instruction over a particular time period in a carefully selected sample of classrooms would help answer questions related to the ubiquity of these technologies and how they affect classroom interactions. All of these examples suggest the value of investing in longitudinal studies of classroom instruction that incorporate video or classroom observations.

Questions About Schools

Schools represent one of the most important settings that impact youth development, yet we know relatively little about how the quality of instruction is related to school-level factors. For example, researchers might want to explore whether instruction varies more within or across schools. Do schools with greater collegiality and strong administration decrease the variability in classroom teaching? Studies of teacher effectiveness most often include either school fixed effects or controls for school characteristics, in the belief that school factors influence the quality of teaching, but controlling for school-level factors is not the same as understanding how such factors influence teaching and learning. Observations at the school level could supplement administrative or survey data with additional data about how adults and children interact within the school. Such observations would allow researchers to ask questions about currently "unobservable" factors that account for variation among schools. One example of this might be identifying characteristics of high-poverty schools in which teachers stay, in contrast to high attrition experienced by most such schools. By observing the school, as a unit of analysis, researchers could try to identify features of interaction—among teachers, students, administrators, and staff—that might help explain differential patterns of retention.

Methodological Questions

NCES surveys, including the ECLS-K survey, include teacher questionnaires that ask teachers about content coverage and instructional practice in their classrooms. Including video or observational data of selected classrooms would enable researchers to compare self-reports of instructional practice to researcher ratings of the same lessons. Such a study could answer methodological questions about documenting classroom experiences and lead to more cost-effective and efficient approaches to data collection. If, for example, we knew that teachers are accurate at reporting their use of particular practices, then surveys are a much more cost-effective way to collect data on instruction (cf. Rowan & Correnti, 2009). If, however, teachers are less than accurate at reporting their use of instructional activities, then it makes more sense to invest in video or observation data when the quality of instruction is an important variable of the research.

Another important methodological question has to do with relative scoring of classroom environments when live observations are conducted versus the scoring of videos of those same lessons. Sending cadres of observers into educational settings is expensive and logistically complicated; with new video technology, it is more efficient to capture video of classroom settings and have raters then score the videos. This was the decision made in the Measures of Effective Teaching project, which trained hundreds of raters to score classroom videos from their homes. However, one of the unanswered questions of the MET study had to do with how live scoring of teaching might differ from video scoring. It may be that some of the nuances of teacher–student interaction are difficult to detect from video, especially if the student voices are not audible or nonverbal interactions are not clearly visible. Similarly, some of the informal interactions between teachers and students that happen at the beginning and end of lessons may provide additional information that may be missing from a video but captured during live observations. Although there is beginning work in this area (e.g., Casabianca et al., 2013), additional studies that compare scores generated from live versus video coding would help researchers better understand systematic differences in the data.

Methodological Considerations

If NCES is to undertake the collection of observation or video data to supplement their surveys, there are a number of questions to consider. First, researchers will need to be clear on the unit of analysis they are most interested in. Because the focus of many of the NCES surveys is on the settings in which children spend their time, the unit of analysis for studies of classrooms is more likely to be the quality of teaching, rather than the quality of the individual teacher. This is an important distinction for sampling purposes. If the aim of including observational data is to estimate the quality of an individual teacher that a focal child might have, then the number of lessons required to get a reliable estimate of that teacher's practice might be as many as four to six. However, if the goal is to get a reliable estimate of the quality of *teaching* in a school or even a district, then the number of observations required in any individual teacher's classroom would decrease.

A different unit of analysis might be the classroom or school as a setting, rather than the teacher or caregiver. In this case, setting would encompass not only the interactions between children and the caregivers and teachers, but also the interactions among children; the interactions among adults; the resources included in the setting, such as the number of books in a classroom; or the policies around parent interaction at the school level. Some observation protocols are designed to provide a wider-angle lens that captures not just teacher–student interactions but the material resources of the setting. For example, the TEX-IN3 (Hoffman et al., 2004), an observation protocol designed for elementary literacy classrooms, includes information about all the textual resources in a classroom that might impact literacy development. Such a protocol demands live observers, as the kind of audit they include requires the ability to capture all of the print materials available in a classroom. If such features of settings are important to document, this will certainly have implications for how observations are conducted. If settings are the unit of analysis, data would need to be collected in ways that enable researchers to look at similar features across settings. For example, in early childhood, current methods of data collection enable researchers to look at the continuity or discontinuity of care across home and school. At the K-12 level, we might be interested in exploring how the nature of educational settings varies by grade level, for example, or when children change schools, or when children move from classrooms to after-care settings.

A second thorny issue to consider when exploring the use of video or observation data is the sampling of time. Teaching may vary in predictable and less predictable ways by time of year. For example, teachers must do specific things at the beginning of the year to build classroom culture and teach routines that they may do less of later in the year. Teaching may also vary by time of the day; if it is important to capture instruction in a particular subject, for example, observers would need to ensure that they knew when that subject is likely to be taught in an elementary classroom. Instruction may also vary depending on the particular phase of the lesson; establishing the purpose of a lesson is most likely to be visible at the beginning and end of the lesson, whereas classroom discussion is most likely to occur somewhere in the middle. In developing a sampling strategy, researchers would need to think through the timing of observations or video collection depending on the particular questions motivating the research. If the goal is to build a longitudinal database of classroom data, sampling instruction at roughly the same time of year across years of the survey would be important.

Part of the goal of collecting video or observation data of teaching would be to build a national database of instruction that could be used to look at trends over time, or relationships between various features of the policy environment and classroom instruction. The TIMSS provides an example of a study that tries to look at the quality of teaching at the national level. Such a strategy might involve first sampling schools and then sampling teachers within schools. SASS, for example, sampled three to eight teachers per school to complete the teacher questionnaires; it might be possible to follow a similar sampling strategy to collect data on the quality of instruction within schools, depending on grade level and subject-matter focus. At the elementary level, sampling five to eight classrooms per school might be sufficient to develop a snapshot of instruction, but at the secondary level, sampling decisions would also need to take subject matter into consideration.

It would also be possible to build upon the National Assessment of Educational Progress (NAEP) to provide additional information about instruction by drawing a random subsample of classrooms at the same time as the survey sample to provide a representative sample of classroom instruction. Such a strategy might provide a national snapshot of instruction at a particular time that could be repeated on a regular cycle to investigate changes over time. Because states differ in their policies, such a sampling strategy would also generate data about differences in instruction that might be related to policy environments and to NAEP outcome data.

CONCLUSIONS AND RECOMMENDATIONS

Although observation and video data are costly to collect, the benefits of having more accurate data on qualities of teaching and classrooms may well be worth the investment. We have learned an enormous amount from studies such as the TIMSS and the Measures of Effective Teaching project about the challenges of collecting video and observation data at scale and making such data available to a wider community of researchers. Integrating such data into longitudinal surveys would allow researchers to investigate a wide range of questions about teaching and learning of great interest to both policy makers and practitioners. By integrating these data into existing NCES surveys, researchers would also be able to explore further the relationships among qualities of teaching and youth development.

If NCES invests in such a strategy, I would generally recommend the collection of video data over live observation as part of NCES surveys. As the technology continues to improve, collecting and storing video data is likely to become easier (although human-subject constraints will still loom large). The primary advantage of video data over observational data is that the dataset is available for a wider array of secondary analyses. In studies using live observation, once observers have coded their observations into a structured observation instrument, they have constrained the kinds of analyses that can be conducted to the categories included in that system. In contrast, video data can be coded according to multiple observation protocols, as was done in the Measures of Effective Teaching project, or used to generate new questions or ways of coding the data. For these reasons, video data would seem to be the better option for future NCES investments.

In addition, NCES is much better positioned to conduct studies that provide data on national trends in classroom instruction over time, rather than studies of individual teachers. Studies that attempt to provide a snapshot of instructional trends over time would require careful sampling, but would not necessarily require the multiple observations in a single teacher's classroom required to make stable inferences about individual teachers. In collecting such national data, it would also be valuable to sample for different student populations in order to explore how the quality of instruction differs depending on the composition of students in the classroom. Do classes in which there are large proportions of English learners, for example, receive differential kinds of instruction?

Sampling decisions will also need to take subject matter into consideration. Focusing on math and ELA instruction, as the MET study did, has the advantage of targeting core components of the curriculum that are also captured in student achievement data. The disadvantage of privileging these two subject areas is that it sends a signal that other subjects are somewhat less consequential and limits the kinds of questions that researchers can ask. Using sampling strategies linked to NAEP would help target grade levels, states, and subject areas for inclusion in video or observational studies of instruction and would also provide a link to data on student achievement.

Developing a plan to incorporate video or observational data into NCES longitudinal surveys would represent a strategic investment in the generation of knowledge about some of the interactions that matter most in the lives of children. NCES could capitalize on advances in technology and knowledge regarding the capture and storage of observational and video data to design a long-term strategy for supplementing their existing surveys with these rich sources of data.

REFERENCES

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783.

Cohen, J. C. (2013). *Practices that cross disciplines?: A closer look at instruction in elementary math and English language arts* (Unpublished doctoral dissertation). Stanford University, Stanford, CA.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dunkin, M. J. & Biddle, B. J. (1974). *The study of teaching.* Oxford, U.K.: Holt, Rhinehart, & Winston.

Evertson, C. M. & Green, J. L. (1986). Observation as inquiry and method. In M. C. Wittrock, *Handbook of research on teaching* (3rd ed., pp. 162–213). New York: Macmillan.

Gitomer, D. H. & Bell, C. A. (2013). Evaluating teachers and teaching. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 415–444). Washington, DC: American Psychological Association.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J. C., Phelps, G., & Ball, D. C. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26(*4), 430–511.

Hoffman, J. V., Sailors, M., Duffy, G., & Beretvas, N. (2004). The effective elementary classroom literacy environment: Examining the validity of the TEX-IN3 observation system. *Journal of Literacy Research, 36*(3), 303–334.

Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

La Paro, K. M., Pianta, R. C. & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal 104*(5), 409–426.

Rowan, B. & Correnti, R. (2009). Studying reading instruction with teacher logs. Lessons from the Study of Instructional Improvement. *Educational Researcher, 38*(2), 120–131.

Stigler, J., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist*, *35*(2), 87–100.