Creating an Indicator of K–12 Classroom Coverage of Science, Technology, Engineering, and Math (STEM) Content and Practices

Laura S. Hamilton, Brian M. Stecher, Kun Yuan

RAND CORPORATION

For more information on this publication, visit www.rand.org/t/RR1913

Published by the RAND Corporation, Santa Monica, Calif.

Cover: kali9

Preface

In 2013, the National Research Council Committee on the Evaluation Framework for Successful K–12 STEM Education identified 14 indicators for tracking the nation's progress toward improving science, technology, engineering, and mathematics (STEM) education in the United States. This report focuses on one of those indicators: classroom coverage of content and practices in the Common Core State Standards for mathematics and *A Framework for K–12 Science Education* (National Research Council, 2012). It describes the rationale for examining new approaches to measuring students' exposure to standards-aligned content and practices, summarizes what is known about currently available measures, and explores innovative approaches that might be adopted to create new measures. The report should be of interest to policymakers and educators who are developing or implementing measures of students' classroom experiences in STEM classes.

The research reported here was conducted by RAND Education, a unit of the RAND Corporation, and was funded by the National Science Foundation under grant DGE-1445670. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contents

Preface iii
Summary vii
Acknowledgments xiii
Abbreviations xv

CHAPTER ONE
Introduction 1
Methods for Measuring Exposure to STEM Content and Practices 3
Which STEM Learning Experiences Should Be Emphasized? 4
Why Is Measuring STEM Learning Experiences Challenging? 7

CHAPTER TWO
Existing Measures of STEM Instruction 11
Self-Reported Measures 13
Direct Classroom Observation 19
Artifact-Based Methods 21
Summary 25

CHAPTER THREE
Promising Digital Data-Collection and Analysis Methods 27
Audio Recording 27
Video Recording 28
Evidence Collected in Computer-Based Learning Environments 29
Challenges of Incorporating Digital Measures into Indicators 35
Potential Benefits of New Digital Tools 36

CHAPTER FOUR
Conclusions and Recommendations 39
Conclusions 39
Recommendations for Developers and Users of Indicators of Students' Classroom Experiences in STEM 42
Final Thoughts 45

Bibliography 47

Summary

In 2013, the National Research Council Committee on the Evaluation Framework for Successful K–12 STEM Education identified 14 indicators for tracking science, technology, engineering, and mathematics (STEM) education in the United States (National Research Council, 2013). These indicators address a range of topics, including professional development and state assessment systems. One of these indicators, indicator 5, focuses on classroom coverage of content and practices in the Common Core State Standards for mathematics and the Next Generation Science Standards (NGSS). Although existing surveys and other data-collection tools can provide some evidence of students' exposure to standards-aligned content and practices, these methods fail to provide a detailed record of students' experiences. One of the primary limitations of most existing methods is their focus on what the teacher is doing rather than what students are doing. This focus on the teacher can be particularly limiting in classrooms in which different students are engaged in different learning activities simultaneously. These within-classroom differences can result from efforts to differentiate instruction to meet the needs of individual students and are likely to be especially prevalent in classrooms that rely on technology-based, personalized-learning approaches. This diversity of experience within a single classroom complicates the task of documenting the content and practices to which students are exposed.

This report explores the challenges associated with collecting information on students' classroom experiences in STEM to inform indicator 5, particularly in light of the need to address this within-classroom diversity. Although indicator 5, as presented in the National Research Council report, focused on the Common Core for math and the NGSS, in this report, we do not attempt to summarize all of the content and practices described in those standards documents or identify which measurement approaches might be suitable for which standards. Instead, this report focuses on methods for measuring classroom experiences that could be applied to a wide range of STEM content and practices. We summarize evidence on currently available measures of instruction and describe several technology-facilitated approaches that could be adopted to develop new measures, drawing on a review of literature; interviews with experts in education, measurement, and technology; and a May 2016 convening with additional experts. The report ends with a discussion of implications for future research and policy related to indicators.

Existing Measures of STEM Instruction

Efforts to develop measures for indicator 5 can build on existing approaches to capturing information about students' learning experiences. Most of these measures have an explicit focus on the teacher's perspective rather than the student's. Commonly used tools for measuring instruction include methods that rely on teacher self-report, such as surveys, vignette-based measures, and instructional logs; methods that involve direct classroom observation; and methods based on analyzing artifacts derived from the classroom, such as protocols to evaluate curriculum materials or student work. And some of these techniques, such as surveys and logs, are being adapted for delivery using computers, tablets, and smartphones, although the respondents and the core judgments they make remain the same. Most of these measures have been developed and used primarily for research purposes, as a means of collecting evidence of teachers' instructional practices or content coverage. Some of these tools have also been used in teacher evaluation systems and for professional development purposes, including classroom observations and student feedback surveys.

Although these methods have produced high-quality and useful data on teaching, none of them, in their present form, can provide complete information to shape indicator 5. In particular, most measures focus on the perspective of teachers—e.g., how teachers organize the lesson, what activities they assign to students, how they allocate classroom time—rather than trying to measure how individual students' experiences differ within a teacher's class. As noted earlier, this limitation is particularly problematic in personalized-learning environments that rely heavily on technology and that offer different learning experiences to different students. In addition, many available measures lack evidence of technical quality for the purposes of monitoring teaching and learning experiences, some of them are expensive to administer and score, and few measures have been developed to capture instructional experiences in engineering and technology. Clearly, there is a need for a research and development effort focusing on new ways to collect evidence of STEM learning experiences.

Promising New Digital Methods for Collecting Data on Students' STEM Learning Experiences

In the past few years, a great deal of creative development has occurred in computer-based instruction and assessment in the STEM fields. Such digital learning platforms might serve as the basis for gathering evidence of students' STEM learning opportunities. Technology-enhanced, personalized-learning environments retain large amounts of data about students' learning interactions with computerized lessons, their performance on periodic assessments, and their work products retained informally or cataloged in a personal portfolio. They are rich environments for extracting digital information on students' opportunities to learn STEM content and practices. Many researchers think that "log-stream analysis" or data mining of student–computer interactions in technology-based learning environments offers powerful opportunities for measuring more than mere mastery of facts or procedures. If these kinds of instructional software were widely used, the data might support valid inferences about students' learning experiences writ large.

Gaming is another type of computer environment that excites many educators, who see electronic games as potentially potent learning and assessment tools. There are many ways it might be possible to extract information about students' opportunities to learn in an unobtrusive manner based on their game play. Some developers are building larger architectures that game developers can use to more easily track students' learning experiences and assess their understanding while playing games. In addition, some new technology-based data-collection approaches outside the STEM fields could support the collection of data about the teaching and learning process. Examples include analyses of audio recordings using natural language processing and methods for collecting and coding video recordings (e.g., wearable video cameras).

Although technology-based environments provide exciting opportunities for indicator development, both conceptual and practical challenges will have to be overcome before useful measures of STEM learning can be derived using these new measurement approaches. The most obvious challenge is that none of these innovative methods is used widely at present, and it is hard to imagine that any will ever be used universally because of the wide range of curriculum and instructional materials in schools and the large variety of instructional technologies available to educators. There are also concerns about drawing valid inferences from complex student interactions with variable learning environments.

Conclusions

Surveys Appear to Be the Most Plausible Method for Measuring Exposure to STEM Content and Practices Across the United States at This Time

Although surveys provide incomplete information on students' STEM learning experiences, in the near term, it is likely that an indicator system will rely on this method of data collection because surveys are relatively inexpensive and can be deployed fairly easily for large-scale data collection. These surveys might include logs of instructional activity in addition to more-traditional survey items, they could be administered via paper or online, and they could include responses from both educators and students. Surveys could be designed in a way that supports a more detailed understanding of students' classroom experiences than has been available through most existing survey data-collection efforts, such as by incorporating detailed questions about time allocation and by examining variability in responses for students within the same classroom.

Technology-Based Learning Systems, Particularly Simulations, Have Potential to Support Future STEM Indicator Measurement Efforts

New applications of instructional technology offer the possibility of novel data-collection approaches that could gather detailed information on what students are doing and how much time they spend doing it. Although these approaches are not yet ready for broad deployment, additional research on the feasibility of these data-collection efforts and on the technical quality of the information they produce could lead to new ways of capturing information about students' learning experiences. This research and development will benefit from close collaboration among software developments, measurement experts, and educators.

The Use of Technology-Based Learning Systems for Large-Scale Measurement Is Likely to Be Limited Because of Variability Across Schools in the Computer-Based Tools and Other Instructional Materials That Are Adopted and Used

Although some of the new approaches have promise, one major factor that will hinder large-scale adoption is the substantial diversity of software products used across schools, along with inconsistency in how data are collected and how the software reports them. These approaches might be more useful at the school and classroom levels, enabling teachers and other educators to extract customized data that meet their day-to-day planning needs. In addition to variation in software packages, the education system in the United States is characterized by extensive local control that has resulted in a lack of consistency in curriculum and instruction across schools. Even in districts that adopt a common set of curriculum materials, teachers often supplement with materials they develop or find from other sources. Students' STEM experiences are shaped to a significant degree by the materials available to them, and the variety of materials across classrooms presents challenges to designing common surveys or data-collection efforts.

STEM Practices, Such as Those Identified in the Common Core State Standards for Math and the Next Generation Science Standards, Are Enacted Differently in Different Content Areas

Although an indicator system might focus on generic practices, such as scientific inquiry or engineering design, students engage in these practices in the context of a specific content area. For a large-scale indicator system, measures of these practices are likely to require sampling across classrooms or schools to capture different content areas without overburdening respondents.

Some Opportunities to Engage in STEM Content and Practices Occur Outside of Traditional Courses

One overarching question in designing a measure of STEM learning experiences is which settings, both in school and out of school, should be included in data-collection efforts. Out-of-school opportunities to engage in STEM learning are common, particularly in such areas as robotics. Challenges for indicator system developers include the need to determine which of these experiences to provide and to devise a method for identifying the relevant coursework and other opportunities in light of the enormous diversity across schools in naming conventions and data availability.

Recommendations for Developers and Users of Indicators of Students' Classroom Experiences in STEM

Create a Working Group That Includes Key Stakeholder Representatives to Inform Indicator Development

Measures to track students' STEM experiences are likely to be of great interest to educators, policymakers, funders, and the general public. One set of measures almost certainly cannot meet all of the needs and wishes of each group. However, input from representatives of these different groups could help maximize the likelihood that the measures that are eventually collected to support indicator 5 (or other STEM indicators) will be perceived as beneficial and useful to multiple stakeholders and could promote long-term support for the system. One approach would be to create a working group that includes multiple stakeholder representatives who collaborate to plan data collection and monitor the measures over time.

Use Multiple Measures to Collect Evidence Related to Indicator 5

Although any indicator of students' STEM learning experiences that is deployed in the near future will probably rely primarily or exclusively on surveys, it might be feasible to adopt some of the more-novel approaches on a small scale to gather supplemental data in a few jurisdictions. This could help developers and users of measures to gather evidence of validity and reliability of the new measures while also providing data that enrich our understanding of learning opportunities beyond what is available in surveys alone. Even if short-term indicator efforts do not adopt these more-novel approaches, there is value in incorporating multiple measures, such as surveys of both teachers and students, which could be designed to reflect each group's perspective and provide complementary evidence related to a common set of topics. Ideally, these measures would attempt to describe variation within a classroom rather than assuming that all students' experiences in that classroom are identical.

Begin by Building on Existing Data-Collection Tools and Systems

Several of the surveys reviewed in Chapter Two of this report could provide a basis for an indicator; many of the practices measured by those tools are consistent with the higher-order activities supported by the NGSS and the Common Core for math and recommended by our expert advisers. As a next step to build on these existing tools, developers should focus on new items that address domains not currently reflected in existing measures.

Design the Measures to Support Longitudinal Comparisons

The measures should support the ability of policymakers and others to monitor STEM learning experiences over time. This information can be useful for assessing the effects of investments in STEM education for informing future funding and policy decisions. Although it is not essential to follow the same students or teachers over time, there could be a benefit to developing a supplemental data-collection effort that collects longitudinal data for a subset of teachers using some of the more-novel measurement methods explored in this report. These data could support a deeper understanding of how teachers have changed their instructional approaches as the new standards have taken hold.

Consider Incorporating Measures of Student Knowledge into the Broader Indicator System

Although the National Research Council indicator effort does not directly address student achievement measures, a system that could link measures of students' learning experiences (i.e., what happens in classrooms) with achievement data could benefit decisionmakers and contribute to the knowledge base. In some cases, high-quality achievement data could actually be useful for gaining insight into the classroom activities in which students participated. For instance, students' ability to carry out a specific problem-solving approach as evidenced by performance on an assessment might signal that students were exposed to that approach during their instruction. Caution would be required in making such inferences, but the key point is that student achievement data should incorporated into the broader indicator system in some way to support a comprehensive picture of STEM education in the United States.

Avoid Attaching Stakes to the Measures

Research on high-stakes testing shows that undesirable outcomes, such as curriculum narrowing or test-score corruption, often result when systems attach serious consequences to performance. To the extent that the indicator system is intended primarily to monitor what is happening rather than to induce specific changes, performance on the indicator system should not be tied to specific consequences for schools, educators, or students.

Continue to Conduct Research on STEM Teaching and Learning to Inform Future Indicator Efforts

Existing research on instruction provides some guidance regarding the mathematics and science classroom experiences that are likely to promote learning, but the field lacks definitive evidence regarding the specific practices associated with learning the disciplinary core ideas and crosscutting concepts in the standards. More-focused research is needed to inform the development of detailed measures, particularly if the goal of the system is to focus on learning experiences that predict academic achievement.

Acknowledgments

This report reflects the thoughtful contributions of the external experts who generously agreed to talk with us and to participate in an expert convening. Their ideas are reflected throughout the report, and we are extremely grateful that they agreed to participate in this project. The following experts participated in interviews, the convening, or both:

- Andy Calkins, Next Generation Learning Challenges
- Douglas B. Clark, Vanderbilt University
- Shelbi Cole, Smarter Balanced Assessment Consortium
- Daniel Damelin, Concord Consortium
- Christopher Dede, Harvard University
- Susan Fine, New Classrooms
- Janice Gobert, Rutgers University
- Heather C. Hill, Harvard University
- Dalia Hochman, Educause
- Nathan D. Jones, Boston University
- Nicole B. Kersting, Arizona State University
- Marcia C. Linn, University of California, Berkeley
- Danielle S. McNamara, Arizona State University
- Andreas Hugo Oranje, Educational Testing Service
- V. Elizabeth Owen, ABCmouse
- Edys Quellmalz, WestEd
- Chris Quintana, University of Michigan
- Michelle Riconscente, Designs for Learning
- Chris Rogers, Tufts University
- Carri Schneider, Getting Smart
- Elliot Soloway, University of Michigan
- E. Caroline Wiley, Educational Testing Service
- Mary Ann Wolf, Friday Institute.

This project was one of several that were funded by the National Science Foundation and that addressed the National Research Council indicator report. We received extremely valuable feedback and advice from the other grantees and from members of the SRI International team who coordinated the STEM indicator initiative.

Our RAND colleagues Brittany Seymour and Stacey Gallaway provided crucial assistance with the planning of the expert convening, Donna White assisted with formatting, and Lisa Bernard ably edited the document. John Pane and William Johnston made substantive contributions throughout the project and shared their knowledge and experiences at the expert convening. They also provided helpful feedback on early drafts of this report. The report greatly benefited from reviews by Catherine Augustine, Julia Kaufman, and José Felipe Martínez. All remaining errors are our own.

Abbreviations

DRM	Day Reconstruction Model
ELA	English language arts
EMA	ecological momentary assessment
ESM	event sampling method
HRI	Horizon Research, Inc.
IQA	Instructional Quality Assessment
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
NGSS	Next Generation Science Standards
NSF	National Science Foundation
NSSME	National Survey of Science and Mathematics Education
OECD	Organisation for Economic Co-operation and Development
PISA	Program for International Student Assessment
R&D	research and development
STEM	science, technology, engineering, and mathematics
TEL	technology and engineering literacy
TIMSS	Trends in International Mathematics and Science Study
WISE	Web-based Inquiry Science Environment

CHAPTER ONE

Introduction

For the past several decades, educators, business leaders, and policymakers in the United States have been working to improve student performance and promote college and career readiness in fields related to science, technology, engineering, and mathematics (STEM). In 2011, a panel of STEM experts convened by the National Research Council identified three goals to improve STEM education in the United States: expanding the number of students who pursue advanced degrees and careers in STEM fields and broadening the participation of women and minorities in those fields; expanding the STEM-capable workforce and broadening the participation of women and minorities in that workforce; and increasing STEM literacy for all students, including those who do not pursue STEM-related careers or additional study in STEM (National Research Council, 2011). Meeting these three goals will require high-quality instructional resources and practices in STEM classrooms. To that end, the National Research Council Committee on the Evaluation Framework for Successful K–12 STEM Education identified 14 indicators for tracking progress toward the three goals related to improving STEM education in the United States (National Research Council, 2013). These indicators are categorized into three broad groups and listed in Table 1.1.

The National Research Council's emphasis on the need for indicators reflects a lack of high-quality, systematically gathered evidence regarding the quality and availability of STEM education and related resources. A set of indicators like those listed above could serve some important objectives, including informing funding and policymaking decisions, identifying local educational needs related to curriculum and instruction, and generating support for efforts to expand the number of students pursuing STEM education and careers and improve STEM literacy throughout the United States.

Of course, this is not the first instance of large-scale efforts to improve educational outcomes in the United States, nor the first time policymakers and researchers have tried to develop methods for measuring and monitoring the quality of teaching and learning. Although a thorough review of all these efforts is beyond the scope of this report, it is useful to highlight a few of them. For example, more than 30 years ago, the National Science Foundation (NSF) sponsored a study to develop a model for a national indicator system to monitor science and mathematics education (Shavelson et al., 1987). In addition, diverse lines of research have explored the features of effective teaching and teachers, including work to identify specific teaching processes that produce desired student learning (Brophy, 1979; Brophy and Evertson, 1976; Good and Grouws, 1977, 1979), efforts to measure students' opportunity to learn subject-matter content (Husén, 1967; Floden, 2002; McDonnell, 1995; Porter, 1995), investigations of links between standards-based reform policy with classroom practices (e.g., Swanson and Stevenson, 2002) and efforts to examine the use of measures of teaching practice in evaluation systems

Table 1.1
K–12 STEM Education Indicators Recommended by the National Research Council

Category	Indicator
Access to quality STEM learning	1. Number of, and enrollment in, different types of STEM schools and programs in each district 2. Time allocated to teach science in grades K–5[a] 3. Science-related learning opportunities in elementary schools 4. Adoption of instructional materials in grades K–12 that embody the Common Core State Standards for mathematics and *A Framework for K–12 Science Education* (National Research Council, 2012)[a] 5. Classroom coverage of content and practices in the Common Core for mathematics and *A Framework for K–12 Science Education* (National Research Council, 2012)[a]
Educators' capacity	6. Teachers' science and mathematics content knowledge for teaching[a] 7. Teachers' participation in STEM-specific professional development activities 8. Instructional leaders' participation in professional development on creating conditions that support STEM learning
Policy and funding activities	9. Inclusion of science in federal and state accountability systems[a] 10. Inclusion of science in major federal K–12 education initiatives 11. State and district staff dedicated to supporting science instruction 12. States' use of assessments that measure the core concepts and practices of science and mathematics disciplines 13. State and federal expenditures dedicated to improving the K–12 STEM teaching workforce 14. Federal funding for the research identified in *Successful K–12 STEM Education* (National Research Council, 2011)[a]

SOURCE: National Research Council, 2013.

NOTE: The Next Generation Science Standards (NGSS) were not published when the National Research Council report was written, so the committee used *A Framework for K–12 Science Education* (National Research Council, 2012). In this report, we primarily refer to the more-recent NGSS when referring to science standards.

[a] High-priority indicator.

(Kane and Staiger, 2012). The widely varying goals of these lines of research have resulted in important methodological and substantive contributions but have also helped promote a diversity of frameworks and tools for measuring instruction.

The National Research Council committee identified six of the indicators as high priority, as shown in Table 1.1. One of these indicators, indicator 5, focuses on classroom coverage of content and practices in the new generation of STEM standards in grades K–12: specifically, the Common Core for mathematics and the NGSS.[1] Although a variety of tools are available to gather evidence related to instruction and learning environments, there is currently no straightforward approach to collecting the information about students' experiences in STEM classes that would be needed to shape indicator 5. This report explores the challenges associated with collecting such information, summarizes what is known about currently available measures, describes approaches that might be adopted to develop new measures, and discusses implications for future research and policy related to indicator systems. It should be of interest to policymakers seeking to monitor the nation's efforts to promote STEM learning and to educators who are working to develop or implement measures of instruction and opportunity to learn in STEM classes.

[1] As of the fall of 2016, 42 states and the District of Columbia had adopted the Common Core State Standards for math, either verbatim or with minor changes (Korn, Gamboa, and Polikoff, 2016). As of the fall of 2015, 15 states and the District of Columbia had adopted the NGSS (Heitin, 2015).

Methods for Measuring Exposure to STEM Content and Practices

The growing availability of innovative, technology-based curriculum and assessment tools in K–12 schools provides an opportunity to rethink how instruction and student experiences are measured. In this report, we review currently available measures that could be used to support indicator 5, and we describe several promising approaches to collecting information that would shed light on students' STEM learning experiences in new ways. We do not attempt to determine an ideal approach or develop specific plans for implementing that approach in practice. Instead, this report is intended to generate ideas and interest in research and development (R&D) efforts that might result in usable tools over the next several years.

Our purpose in exploring measures for inclusion in an indicator system differs from the purposes of most currently available measures of instruction, many of which were developed to collect detailed information that can inform decisions about instruction and professional development or to collect data for use in research studies. The primary purpose of the indicator system described in the 2013 National Research Council report is to monitor students' access to STEM learning opportunities on a broad scale as a way to inform educators, policymakers, and members of the public about the nation's investments in, and progress toward, providing high-quality STEM education to all students. Measures that are used as monitoring tools need to have well-specified data-collection protocols and must be applied consistently across different contexts, but, unlike measures that are used to make decisions about individual students or teachers, measures used in large-scale monitoring systems do not need to produce scores or ratings that are reliable at the student or teacher level. They also do not need to capture information from every student or every teacher but instead would typically rely on a representative sample of schools or local education agencies. A new approach to gathering information for inclusion in an indicator system therefore could benefit from a rethinking of how we typically gather evidence of what is happening in classrooms, especially because many prior efforts focused on student- or teacher-level measurement. It should also be informed by a recognition that the types of validity and reliability evidence required to support the use of measures in an indicator system are likely to be different from the types required to support other uses.

To identify promising methods, we conducted interviews in 2015 and early 2016 with 17 experts, who represented a variety of roles and institutions (e.g., university faculty, software development company) and who had experience in curriculum, assessment, or technology use in at least one STEM discipline. We identified these experts through a snowball process, in which we began by interviewing a few highly regarded STEM education experts who themselves had received research support from NSF; at the conclusion of each interview, we asked the respondent to recommend other researchers with expertise that was relevant to our inquiry. Our list grew rapidly to about two dozen people who were working on relevant topics. We also began to hear the same names repeatedly, and we stopped the snowball process when the number of new names dwindled and the number of repeats increased. We used a semistructured interview protocol that included questions about STEM content and practices that the interviewee believed should be the focus of data collection; tools from the interviewee's own work that could support measurement of STEM content and practices; recommendations for other related work, including innovative, technology-supported data-collection activities; and recommendations for approaches to addressing within-classroom differences in students' learning experiences.

We also reviewed literature on existing methods for measuring instruction and on technology-based methods of data collection that could be adapted to this context (e.g., event sampling methodology). The literature review was not intended to be systematic or exhaustive; we consulted other recent reviews of instructional practice measures and scanned tables of content in journals that frequently publish articles related to instruction or educational technology. We also reviewed websites for tools about which we learned through the literature or through other interactions with developers and users of educational technology resources. The experts provided additional suggestions for literature and websites.

Informed by these interviews and reviews, we identified several broad categories of measurement methods that could be harnessed to measure STEM learning experiences (including those specified in the Common Core for math and the NGSS). We then convened 16 experts (some of whom participated in interviews) in a one-day meeting to expand on the ideas and explore further options. The conversation covered a wide range of topics but focused on answers to such questions as "What are the highest-priority STEM practices we should measure as part of an indicator system?" and "How might we operationalize each of those?" "How can we leverage new methods to create measures of students' STEM learning experiences?" "What should be the next steps for R&D?" In the rest of this report, we draw on the input from these experts, sometimes providing quotes and other times paraphrasing discussions. We do not associate individual names with the quotes, but we list all of the experts and their affiliations in the acknowledgments at the beginning of this report.

Our focus is on gathering information relevant to understanding the extent to which K–12 students are exposed to instructional content and engaged in practices that are aligned with the Common Core for math and the NGSS. We did not examine the application of any of these methods to the measurement of student achievement, although several of the methods could clearly be used for that purpose.

In addition, although the National Research Council report referred to "quality" of STEM learning experiences, in this report, we generally refer to measures' suitability for capturing information about the extent of exposure or participation rather than quality. Many of the existing measures that we review in Chapter Two were designed to gather evidence about quality of instruction, but, because there are different views among educators and policymakers regarding the desirability of particular classroom activities, it is nearly impossible to specify the features of an activity that all potential users would agree reflect high quality. It is also reasonable to emphasize extent of exposure rather than quality during the early stages of indicator development, particularly given the lack of existing evidence about students' exposure to a broad range of STEM content and practices. At the same time, many of the measures and data-collection approaches that we discuss could be used to measure aspects of instructional quality in some cases, so we are not arguing that quality should be ignored.

Which STEM Learning Experiences Should Be Emphasized?

Like the standards that preceded them, both the Common Core for math and the NGSS cover a wide range of content. It would be impractical for a single set of measures to track student engagement with all of the mathematical clusters and domains contained in the Common Core for math or all of the crosscutting concepts and core ideas reflected in the NGSS. Moreover, the new standards explicitly describe the kinds of mathematical, scientific, and engineering practices in which students are expected to develop proficiency. When we asked our experts which aspects of the standards would be most important to prioritize in indicator 5, most focused on the practice dimensions rather than on content knowledge or specific skills. As one expert told us,

> We have tended to view science as things to know rather than things to do . . . . In science, we try to measure how much people know in different areas of science and [the] NGSS [are] trying to get away from that and measure science behaviors and practices instead. We can't engage with practices absent content—can't measure practices without the context of content—but we are still focusing on too many content areas and not focusing on practices; practices are an afterthought. We should focus on practices and on content in service of those practices; content should not be the primary focus.

Multiple respondents echoed this sentiment, noting that the most-important constructs to measure involved opportunities for "*doing* science or *doing* math rather than *knowing* [science or] math" and "intertwining or layering practice with content." These ideas are not new; they have been explored extensively in prior R&D on STEM curricula. For example, the 2007 National Research Council report *Taking Science to School* includes the following recommendation:

> Developers of curricula and standards should present science as a process of building theories and models using evidence, checking them for internal consistency and coherence, and testing them empirically. Discussions of scientific methodology should be introduced in the context of pursuing specific questions and issues rather than as templates or invariant recipes. (Duschl, Schweingruber, and Shouse, 2007, pp. 5–6)

The following list includes STEM practices that at least some experts argued are important for high-quality STEM education and that are related to practices included in the Common Core for math or the NGSS. This list is offered to illustrate the kinds of practices one might want to capture in indicator 5; it is not complete and, at the same time, is probably too long to be covered completely as part of a data-collection effort in support of that indicator. The point is that it is not enough to know the sorts of things about classroom activities that have been measured historically, such as how classroom time is allocated among a variety of activities, including lecture and demonstration, group activities, and individual work. The interviewees encouraged us to capture information about classroom coverage of the following types of STEM practices:

- Engage in questioning and discussion.
- Engage in modeling and computational thinking.
- Participate in activities that lead to a divergence of student ideas and products rather than a convergence of effort and thinking; as one expert noted, "If answers are uniform, it's a bad engineering problem; if they are diverse, it's better."
- Link and connect new ideas and build integrated knowledge frameworks.
- Participate in inquiry-oriented investigations—a process that can be complex, ill defined, and take multiple directions, some productive and some unproductive.

- Experience learning that connects practices with content and builds students' understanding in ways that are consistent with contemporary cognitive models and learning progressions.
- Interact with teachers who take on the role of mentor, as well as the role of assessor.
- Collaborate to solve problems.
- Make valid arguments from evidence, including critiquing the arguments of others.
- Formulate ideas and opinions.
- Develop solutions to complex problems.
- Realize in materials what one imagines in one's mind (engineering).

Experts pointed out that students' engagement in various STEM practices should reflect coherence rather than an attempt to focus on each practice in isolation. Both the Common Core for math and the NGSS advocate for more cross-disciplinary instruction, both within and beyond the four STEM disciplines. A recent National Research Council report (Honey, Pearson, and Schweingruber, 2014) suggests that, when integrated STEM learning experiences are adopted, student knowledge and skill in each of the individual disciplines should be supported by the teacher's instruction. Therefore, measurement systems might need to address each relevant discipline in a reasonably coherent and comprehensive way. At the same time, a given practice might look very different across different content areas, even within a single STEM discipline. For example, an inquiry-oriented investigation in a biology class will draw on knowledge and skills different from one in a physics class. These practices should not be viewed as isolated, decontextualized activities, divorced from content. These comments are consistent with other published guidance on how the NGSS and the Common Core should be implemented, emphasizing the integration of practices and content and the need for instruction to help students make connections across ideas and topic areas (see, e.g., Pellegrino et al., 2014). Experts also noted that, although there is extensive research linking some types of STEM practices to improved student learning, this body of research is far from complete, and much more investigation is needed to help teachers understand the specific types of learning experiences that are likely to promote student achievement as measured against new, more-rigorous standards.

The growing role of technology in classrooms also has implications for the kinds of experiences that teachers can offer students to support their STEM learning. High-quality technology-based curriculum materials can enable students to generate models, engage in complex problem-solving, and participate in other activities that focus on the STEM practices discussed above. These resources are most likely to be effective when they are implemented by teachers with high levels of skill and knowledge to incorporate them into an inquiry-oriented instructional environment (Gerard, Varma, et al., 2011). Thus, efforts to document students' exposure to high-quality STEM learning needs to attend to the technology and the teaching, as well as their interface.

Later in this report, we share examples of measurement approaches that might allow us to build an indicator to measure at least some of the recommended practices now or in the near future and that can accommodate various degrees of technology use. Before turning to that discussion, we summarize some of the challenges associated with measurement.

Why Is Measuring STEM Learning Experiences Challenging?

Indicator 5 reflects an understanding of the crucial role that K–12 instruction plays in ensuring high-quality STEM education and a well-prepared workforce. This indicator focuses on the learning experiences of students in STEM-related courses—both the content students study and the activities in which they engage. There are two perspectives to take when considering measures of content and practices covered in a class. The first perspective focuses on what teachers are doing in the classroom, whereas the second emphasizes students' experiences. These two perspectives overlap quite a bit; most classroom observation rubrics, for example, focus on documenting teachers' practices but also include attention to what students are doing. Although measures developed under both perspectives provide useful information about the quality of learning that occurs in classrooms, measures that focus on students' experience with desired content and practices might be more closely related to students' learning outcomes and therefore more important for understanding the quality of STEM education and how to improve it (Haystead, 2010; Priest, Rudenstine, and Weisstein, 2012). We consider both perspectives in this report, examining existing measures of instructional practice that focus primarily on teachers and exploring additional measurement approaches that can provide evidence regarding what students are experiencing.

During the past several decades, educators and researchers have engaged in extensive R&D efforts to create measures to document teachers' instruction and students' experiences. Most of these measures were originally intended to be used in a research context, but, more recently, local and state education agencies have adopted new teacher evaluation systems that rely heavily on measures of instructional practice. As a result of these research and policy developments, a wide variety of measures is available for use in STEM classrooms. At the same time, there are several challenges associated with trying to understand students' STEM instructional experiences in a systematic way.

Ambiguity Regarding What Practices to Emphasize

Indicator 5 addresses both STEM content and practices, but we currently have relatively cost-effective, practical approaches only for measuring the former; measuring the classroom opportunities that promote students' proficiency to engage in STEM practices has proven more elusive. Familiar measures of content include analyses of textbooks and other instructional materials combined with surveys that capture information about how teachers allocate class time among various topics. These measures do not perfectly capture evidence of content exposure (e.g., two algebra courses that use the same textbook might vary in their emphases on different parts of that book, in ways that might not be fully captured by teacher reports), but they can provide good estimates of this exposure on a fairly large scale. Textbook analyses and surveys are less useful for understanding the extent to which students engage in mathematical or scientific practices, such as modeling or design. Efforts to understand student exposure to these practices often involve attempts to measure teachers' specific instructional behaviors (sometimes called "teacher moves"), but this too has been challenging, in part because of a lack of agreement among educators and researchers on what moves should be measured. Efforts to teach the Common Core for math and the NGSS almost certainly benefit from the use of certain kinds of instructional approaches, but the standards themselves are largely silent on the question of how teachers should help students master the content and practices included in the standards,[2] and the field lacks solid evidence regarding what types of teacher moves are most effective for teaching the standards. A system to measure students' exposure to standards-aligned STEM content and develop STEM-related practices would need to identify the desired teacher moves and employ techniques to measure them.

Limitations of Existing Measures

As we discuss in greater detail in Chapter Two, measuring instruction in a way that supports valid inferences about classroom activities and students' experiences has proven to be difficult (Goe, Bell, and Little, 2008). Although researchers have developed and used a variety of measures, such as teacher and student surveys, teacher logs of instructional activity, classroom observations based on detailed observation protocols, principal ratings, and instructional artifacts, each option has significant limitations. Thus, a need continues for further development and refinement of these measures.

Lack of Clear Understanding Regarding How Technology and Engineering Instruction Is Delivered

As we discuss in Chapter Two, most existing measures of instruction either are subject-neutral (e.g., the Framework for Teaching [Danielson, 2007]) or were designed for use in mathematics, science, or English language arts (ELA) classes. The inclusion of technology and engineering in the STEM acronym suggests that efforts to document students' exposure to content and practices will need to extend beyond traditional mathematics and science courses if we want to gain a comprehensive understanding of the instructional activities in which students are engaging across the four STEM disciplines. Complicating this task is a lack of clear definitions for what constitutes technology and engineering instruction. We draw on the 2014 National Assessment of Educational Progress (NAEP) technology and engineering literacy (TEL) assessment for definitions of these disciplines; the framework defines technology as encompassing "any modification of the natural world done to fulfill human needs or desires" (National Assessment Governing Board, 2014, p. 3) and engineering as "a systematic and often iterative approach to designing objects, processes, and systems to meet human needs and wants" (National Assessment Governing Board, 2014, p. 3). Educators and policymakers have increasingly advocated for the inclusion of these two disciplines in K–12 instruction to help ensure that students are prepared for the growing number of careers that draw on skills in technology and engineering.

Students might be exposed to STEM content through experiences in non-STEM courses or outside of school, but this challenge is particularly salient for engineering and technology, as the breadth of the definitions provided above makes clear. These disciplines are sometimes addressed through courses that have "engineering" or "technology" in their titles, but they might also be taught in math, science, business, or career and technical education courses. Engineering, for example, might be incorporated into science classes, and this is particularly likely in light of the NGSS emphasis on engineering-related principles (NGSS Lead States, 2013). However, engineering is also sometimes offered as a stand-alone course or sequence of courses.

Measuring students' exposure to standards-aligned instruction in technology is difficult, given the relative lack of explicit guidance in the standards documents and the fact that stu-

[2] Such organizations as Student Achievement Partners have published guidance regarding instructional practices to promote standards-aligned teaching; see Student Achievement Partners, undated.

dents often interact with technology in ways that do not enhance technological competence (e.g., using computers to complete math drills or using presentation software to communicate the results of research). The framework that guided the NAEP TEL assessment development includes a large number of TEL activities that students could experience either inside or outside of school. An indicator of technology and engineering learning experiences would need to account for these outside experiences or would need to be clearly defined as focusing exclusively on within-school opportunities.

Growing Prevalence of Technology-Based, Personalized-Learning Approaches

Most existing measures of instruction were developed and tested in the context of fairly traditional classrooms that include a single teacher who provides instruction to a group of students. In these contexts, documenting the behaviors of the teacher can capture much of what students experience in that classroom. In recent years, however, schools have increasingly adopted instructional approaches, sometimes called personalized learning, that typically rely on a combination of technology-based resources and teacher-provided instruction to vary the instructional experiences among students within the same class (Horn and Staker, 2011). According to one recent report,

> Although there is not yet one shared definition of personalized learning, leading practitioners in the field generally look for the following: (1) systems and approaches that accelerate and deepen student learning by tailoring instruction to each student's individual needs, skills, and interests; (2) a variety of rich learning experiences that collectively prepare students for success in the college and career of their choice; and (3) teachers' integral role in student learning: designing and managing the learning environment, leading instruction, and providing students with expert guidance and support to help them take increasing ownership of their learning. (Pane et al., 2015, pp. 2–3)

Although fully personalized models are relatively rare, they have become more widespread recently (see, e.g., the Next Generation Learning Challenges initiative [Next Generation Learning Challenges, undated]). Moreover, teachers in traditional classrooms and schools have adopted many of the practices that characterize personalized learning—e.g., having some students work on individualized tutoring software while a teacher provides instruction to the other students at the same time. Finding ways to measure these differences is important for understanding the kinds of experiences to which students are exposed and, in particular, for exploring whether students with different backgrounds and achievement levels are given equitable access to high-quality, challenging STEM curriculum and instruction. Schools that implement personalized-learning models also frequently adopt new staffing approaches that break from the traditional model of one teacher per class; various forms of teaming are common.

Because traditional measures of instruction usually capture a single teacher's whole-class practices, most of them are not ideally suited to classrooms in which individual students use different materials and interact with one or more adults in different ways. A large-scale indicator that relies on sampling of students would not necessarily require documenting all of these within-classroom differences, provided that it included an approach to gathering student-specific information about instructional activities rather than relying on whole-class information. This could be done, for instance, by asking the teacher to report on the activities of two or three specific students or by collecting data directly from one or more students in the classroom. However, there are potential benefits to measuring within-classroom variability directly. In particular, this approach to measurement could shed light on possible inequities in instructional opportunities that occur within a classroom and could help us understand how teachers allocate their time across different students and activities. Therefore, it would be beneficial to identify measures that can capture students' varying experiences within the same classroom. Such measures will probably require innovative approaches that are not currently in widespread use for the purpose of measuring students' learning experiences in STEM classrooms, as we discuss in the next section.

In the next chapter, we review the more-prominent measures of instruction that are in current use in research or evaluation contexts. Then, in Chapter Three, we describe innovative methods for data collection and analysis that hold promise for capturing information about students' STEM learning experiences and could be the basis for future measures. The final chapter discusses the implications of these analyses for future STEM indicator R&D efforts.

CHAPTER TWO

Existing Measures of STEM Instruction

In this chapter, we provide an overview of tools that have been developed and are currently being used to capture information about students' classroom learning experiences, all of which can be applied to STEM disciplines. We describe several types of measures that have been used in large- and small-scale research on teaching and learning, including their purposes, format, strengths and limitations, and technical quality. All offer some potential for inclusion in an indicator system as measures of students' opportunities to learn STEM content and practices. Of course, for the purposes of the STEM indicator system discussed in Chapter One, it is essential that the tool, be it a survey, log, portfolio, or other technique, be focused on information that is relevant to the NGSS or Common Core State Standards for math.

Because most existing measures focus on teachers' instructional behaviors, rather than on what students are doing, we use the term *instruction* throughout this chapter when describing the targets of measurement for the methods we review. We argued earlier that a comprehensive set of measures for indicator 5 would need to include more than merely the teacher's actions. However, the most–commonly used measures of teaching tend not to take that broader approach; instead, they are designed to focus on teacher practices and classroom-level coverage of content.

Before we present information about widely used measures of instruction, it is useful to review the standards that are used to judge the technical quality of such measures. Quality is one of the features that will need to be considered when thinking about the applicability of these measures in the context of indicators. The two key dimensions of quality that should be considered when looking measures of instruction are reliability and validity for use in an indicator system. *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) defines reliability as

> [t]he degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group. (pp. 222–223)

Scores on a measure would be considered to have high reliability if someone who completed the measure would receive approximately the same score after completing the measure multiple times under the same conditions (assuming that no learning occurs as the measure is administered). Various sources of error can influence scores and threaten this consistency; these sources of error can include differences in the extent to which items within a measure function similarly and measure the same construct (often documented by an index of internal consistency reliability), differences in how human raters score a performance (interrater reliability), and differences in scores across occasions (test–retest reliability). For example, the reliability of survey responses has been extensively examined, usually as part of the survey development process. Most studies have focused on internal consistency reliability in the form of a Cronbach's alpha coefficient, and, in most cases, the reliability of survey-derived scales that are used for reporting results meets or exceeds an acceptable level based on expert recommendations (e.g., DeVellis, 1991). However, other sources of error should be considered, such as consistency among responses given on different occasions. These other sources of error are generally more difficult to assess than internal consistency and are therefore much less frequently reported but not necessarily less important.

It is also important to investigate the validity measures related to student learning opportunities. According to *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), *validity* refers to

> [t]he degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed. (p. 225)

Thus, validity needs to be examined in the context of a specific purpose, such as portraying the status of STEM education or estimating the number of hours of STEM-related instruction students receive. Evidence regarding the validity of measures for various purposes is less commonly presented than is evidence about reliability. This is true for survey responses, as well as for logs, artifacts, portfolios, and other measures discussed in this chapter. When validity evidence is examined, researchers often investigate the correlation among different measures of instruction—for example, between survey results and classroom observation or between survey measures and measures based on student achievement growth (Banilower et al., 2013; Kane and Staiger, 2012; Blank, Porter, and Smithson, 2001; Burstein et al., 1995; Porter et al., 1993; Levine, Huberman, and Buckner, 2002). Some studies have also investigated the internal structure of the measure (i.e., the way different subscores of a measure function). For instance, Jonathan Schweig conducted multilevel factor analysis on the Tripod student survey and found evidence that supported a different internal structure from what the developers suggested (Schweig, 2014; Tripod, undated). After describing each class of measures, we summarize what is generally known about their reliability and validity.

Because this report focuses on the use of measures to inform a large-scale indicator, the relevant considerations about reliability and validity are somewhat different from those for measures that produce individual-level scores. In particular, measures that fail to produce individual-level scores with high reliability can sometimes do so at an aggregate (e.g., district or state) level. Producing aggregate information that has a high level of reliability and validity for its intended purposes requires thoughtful decisions about how to sample across schools, classrooms, students, and lessons. Learning experiences typically vary extensively across lessons; a single lesson will provide incomplete information about the kinds of learning experiences in which students participate. The system could sample specific types of lessons, such as those in which a concept is introduced and those in which students are applying knowledge to a new type of problem. Much of the research on measures of instruction has emphasized individual-level uses, so readers should keep in mind the ways in which a large-scale indicator that produces aggregate scores might function differently.

A range of different techniques have been used to measure instruction, including methods that rely on teacher self-report, such as surveys, vignette-based measures, and instructional logs; methods that involve direct classroom observation; and methods based on analyzing artifacts derived from the classroom, such as protocols to evaluate curriculum materials or student work. Most of these methods have been used primarily for research purposes, as a means of collecting evidence of teachers' instructional practices or content coverage. Some have been incorporated into tools for teacher evaluation and professional development. In the following sections, we describe each type of measure and provide examples.

Self-Reported Measures

Surveys

Surveys are commonly used to collect information from stakeholders about selected features of teaching. Typically, surveys are requested from teachers themselves but sometimes from other stakeholders, such as students or parents or guardians. Because of their relatively low cost and the large amount of research evidence that has been gathered to inform survey development, surveys outnumber other methods of measuring instruction, and this section of the chapter is significantly longer than the others. In this section, we describe a variety of surveys that have been used at international, national, and local levels. We then briefly describe two alternative self-report approaches: vignettes and logs of instructional activity. The technical quality of self-report methods is discussed at the end of the section.

Several large-scale national and international educational assessment programs, such as NAEP (National Center for Education Statistics [NCES], 2015a; 2015b), the Trends in International Mathematics and Science Study (TIMSS) (NCES, 2014a, 2014c; Schmidt, Raizen, et al., 1997; Schmidt, McKnight, and Raizen, 2002), and the Program for International Student Assessment (PISA) (Organisation for Economic Cooperation and Development [OECD], 2012a, 2012b), use teacher, student, and parent and guardian surveys to collect data on mathematics and science instruction. These large-scale surveys are typically used to measure curriculum content and emphasis, as well as instructional techniques; they are also frequently used to gather information on teacher background, knowledge, and beliefs, as well as stakeholder attitudes about teaching and learning. In some cases, these surveys do include questions about what students do in the classroom, but typically not in a format that allows teachers to report different experiences for different students.

For instance, recent administrations of NAEP and TIMSS have asked teachers to rate how much emphasis they gave to selected topics during instruction (e.g., use models to explain calculations) (NAEP, undated, second item 12, response c) or how often they used certain instructional practices (e.g., "How often do you encourage students to express their ideas in class?") (NCES, 2014b, item 15, response h). PISA surveyed students on the frequency with which they engaged in specific mathematics tasks (e.g., understanding scientific tables presented in an article) (OECD, 2012b, item 46, response d) during class and how often they were asked to solve certain types of problems in their mathematics classes (e.g., "solve 2x + 3 = 7") (OECD, 2012b, item 46, response e). PISA also surveyed parents on their involvement with their children in mathematics-related activities (e.g., "discuss with my child how mathematics can be applied in daily life") (OECD, 2012a, § F, item Q, response g). Similar items have been included in other large-scale surveys fielded by NCES, such as the Early Childhood Longitudinal Study.

NSF also uses surveys to collect information about mathematics and science education nationally. For example, over the past three decades, NSF commissioned a series of National Survey of Science and Mathematics Education (NSSME) administrations. Horizon Research, Inc. (HRI) conducted the fifth national survey of mathematics and science teachers in 2012 (Banilower et al., 2013). This survey was administered to a nationally representative sample of teachers and collected detailed information about "curriculum and instruction in a single randomly selected class" (Banilower et al., 2013, p. 2), in addition to teachers' backgrounds, preparedness, beliefs, professional development, and other factors that might affect their instruction.

By focusing each survey on a single subject, science or mathematics, and by asking teachers to focus on a single, recent lesson they taught in one specific classroom, the national survey has been able to collect more-detailed information about instruction. For example, the survey covers time spent on instruction overall and on specific activities (e.g., lectures, small-group work, taking a test or quiz), objectives of instruction (e.g., the extent to which a certain topic received heavy emphasis during instruction), instructional materials used, use of facilities and instructional technology (e.g., handheld computers), the amount of homework, and assessment activities.

The Surveys of Enacted Curriculum offer another example of how the survey format can be used to examine specific aspects of curriculum and instruction. These surveys were developed jointly by the Council of Chief State School Officers and the Wisconsin Center for Education Research (Blank, 2005). The surveys include a set of tools to collect, analyze, and report data about instructional content and practices in ELA, mathematics, and science at elementary, middle, and high school levels. The teacher and student surveys draw in part on such surveys as those used by NAEP, TIMSS, and NSSME. Teachers are asked to report on a full school year of instruction when completing the survey. In addition, the surveys also include questions gauging the extent to which instruction is aligned to state or local curriculum standards, in terms of both breadth and depth (i.e., cognitive rigor) of the enacted curriculum.

Some school districts also administer surveys to gather information on classroom and school environment and activities. Some districts develop these instruments themselves, whereas others work with partner organizations. One of the most prominent examples of the latter is the University of Chicago Consortium on School Research, which administers surveys in the Chicago Public Schools and uses the data to provide school and district reports, as well as to inform research. The Chicago consortium's surveys of teachers and students capture information about a wide range of factors, including teacher background and experience, professional development opportunities, focus of instruction, and school and classroom climate.

In recent years, researchers have developed surveys to measure teachers' content knowledge for teaching in a variety of fields, including mathematics and science. A survey-based measure of Mathematical Knowledge for Teaching, for example, assesses the extent to which mathematics teachers understand specific aspects of mathematics that are important for effective teaching—e.g., alternative ways to represent mathematics principles and procedures (Hill, Schilling, and Ball, 2004). Although this survey does not provide direct evidence of instructional practices, responses have been shown to relate to the quality of instruction and to student outcomes (Hill, Rowan, and Ball, 2005).

Although most surveys of instruction are completed by teachers, student surveys can also be used to gather information about instruction and other aspects of the classroom experience. The Tripod student surveys (Tripod Education Partners, undated) are among the most–widely used student surveys of instruction (Kane and Staiger, 2012). These surveys ask students to rate their agreement with statements (e.g., "My teacher asks students to explain more about answers they give") about several aspects of classroom climate and instruction. The Chicago consortium has also developed student surveys that include questions about teachers' practices and that are used to inform schoolwide indicators of instruction (University of Chicago Consortium on School Research, undated). Student surveys have the advantage of capturing within-classroom differences in students' learning experiences because each student reports on instruction from his or her own perspective.

Most of the surveys we have described so far were administered to scientifically selected samples of teachers or students so the results could be generalized to a particular district, state, or the nation as a whole. In contrast, Project Tomorrow is a voluntary survey effort open to all schools and districts in the country. In 2015, more than half a million people, including students, parents, teachers, administrators, and community members, responded to Project Tomorrow and provided information about participation in STEM learning experiences, such as STEM academies and computer programming clubs.

Vignette-Based Methods

To reduce the potential error resulting from differences in respondents' understandings of survey items and response scales, vignettes describing mathematics and science instruction have been used to standardize teachers' understanding of the survey questions and use of response scales (Ruiz-Primo and Li, 2002; Stecher, Le, et al., 2006; Stein, Correnti, et al., 2017; Yuan et al., 2014). Vignettes provide concrete scenarios about instructional context, practice, and teacher–student interactions. In some closed-ended vignettes, for instance, teachers are given multiple choices about potential reactions in response to a scenario that describes events that take place during a lesson and are asked to select one or more options to indicate how they would respond. One specific approach, commonly referred to as "anchoring vignettes" (King et al., 2004), presents multiple vignettes that are created to correspond to different levels of performance on specific dimensions of instruction. Teachers rate the extent to which each vignette represents a certain level of practice (e.g., whether the teacher in the vignette engages in the practice "frequently" or "extensively"), and they respond to a question about their own practice in that same area. Teachers' reports about their own practices can then be statistically adjusted based on how they categorize the vignettes, so that the self-reports of a teacher who has an overall tendency to rate practices highly (in terms of intensity or frequency) can be placed on the same scale as those of a teacher who tends to assign lower scores.

Instructional Logs

One limitation of traditional surveys is that they are generally administered infrequently and therefore often require respondents to recall events over a long period of time. Infrequent administration also results in an inability to capture information about the ways in which instruction and learning experiences vary across lessons. One approach that addresses these limitations involves the administration of instructional logs, which typically take the form of shorter surveys that are administered at frequent intervals (e.g., each day for a two-week period). Logs have been used to collect detailed information about instructional content and practice in many large-scale studies about teaching and learning, such as the Beginning Teacher Evaluation Study (Fisher et al., 1978); the Reform Up Close study (Smithson and Porter, 1994); the University of California, Los Angeles (UCLA)/RAND Corporation Validating National Curriculum Indicators project (Burstein et al., 1995); the Study of Instructional Improvement (Rowan, Camburn, and Correnti, 2004); the RAND Mosaic II Study (Le et al., 2004); and a recent evaluation of personalized learning (Pane et al., 2015). These instructional logs were often designed for a particular study and focused on specific aspects of instruction. Logs have been used more often to study mathematics and ELA than science (Brandon and Taum, 2005; Le et al., 2004). Logs are administered either in paper-and-pencil format or online. Results provide evidence of the frequency and amount of teachers' coverage of certain content and use of specific instructional practice over the logged period of time, which can, in some cases, be used to support inferences about the coverage of content and practice over a school year.

Instructional logs have some advantages over the one-time surveys in terms of collecting nuanced data about classroom instruction. For example, instructional logs often include specific questions about instructional practices that are likely to occur frequently in instruction. By asking about these practices across multiple lessons, logs can provide evidence regarding the frequency with which individual teachers engage in them. Moreover, instructional logs provide results for multiple sampled time points to estimate the overall coverage of content and practice; this offers greater generalizability than one-time surveys when drawing inferences about content coverage or practice over a school year. Finally, logs can be designed to allow teachers to focus on the instruction provided to an individual student or a small group of students rather than to the entire class, varying the specific student or groups across the multiple log administrations. This approach could facilitate the collection of evidence of varying learning experiences that might occur in a personalized-learning environment.

With the advent of computers, tablets, and smartphones, electronic logs have appeared that might be more useful in an indicator system. For example, MyiLOGS is a daily online log that teachers can use to indicate how they spend class time on various features of instruction (MyiLOGS, undated). The software produces reports on content coverage, instructional time, and other factors. Scores derived from the logs can help teachers keep track of their use of time, review their lessons, and develop their own growth plans. They can also reveal how students are using their time and the kinds of instructional activities and experiences in which they are engaged. Developed originally for use with special education students, MyiLOGS is now available more widely. If MyiLOGS use were widespread, it might be possible to extract common measures of student learning opportunities, such as class time devoted to particular topics. More importantly, because the data are already in electronic form, the process of collecting, analyzing, and reporting on the data could be relatively efficient.

Another version of log described by one of our experts is a two-stage written log developed to improve the accuracy of self-reported changes in teacher practice made on the basis of information obtained from formative assessments. Initially, the researchers asked teachers to complete logs that asked about both the kinds of formative assessments that occurred during a lesson and the way they used the information. They described the options in precise behavioral terms hoping to make the reporting more accurate. However, when researchers observed lessons, they found that teachers were making reporting errors. To reduce the number of reporting errors, the research team developed a two-step procedure that placed less cognitive demand on teachers. In step 1, teachers indicated whether a particular practice occurred, and only if the practice occurred were they asked to respond to prompts about the quality of the practice in step 2. For example, the step 1 log might ascertain whether the teacher presented learning goals or success criteria for a particular lesson, and, if that occurred, the step 2 log presented questions about the way the practice was carried out—e.g., "learning standard written on the board" or "learning intentions discussed with students." To the extent that two-stage logs yield more valid information about teaching practice, they could support more-detailed measures of student learning opportunities. Although most of the research using logs has relied on teachers as the respondents, it would also be possible to have middle or high school students complete logs to report on their individual learning experiences.

Technology offers other options for gathering log-like information about learning experiences in ways that do not cause major alteration to instruction but do yield more-extensive information about students' opportunities to learn. For example, smartphones can be used in a variety of ways to collect data in real-world contexts, including classrooms. One example is a technique known as the event sampling method (ESM) or ecological momentary assessment (EMA) (Scharf et al., 2013). Researchers send short prompting messages to subjects' smart devices at different times during a lesson, and the subjects briefly respond with a report on some predetermined aspect of their immediate experience. ESM has been used in many different contexts other than education. For example, it is well suited to measuring people's emotions at work or in family contexts and how attitudes and feelings vary in response to specific activities (e.g., Offer and Schneider, 2011). It is easy to prompt with short, quick questions, such as "What are you doing?" "What are you thinking?" "Where are you?" "Who are you with?" ESM has been used with some success to measure affective variables, such as challenge, positive affect, or activation stress (Csikszentmihalyi and Larson, 1987; Hektner, Schmidt, and Csikszentmihalyi, 2007). The use of smartphones or other personal electronic devices might not be appropriate with younger students in elementary classrooms, but the technique probably can be used with middle school or high school students.

One disadvantage of methods like ESM or EMA is that they can interrupt the normal flow of instruction. According to one of our experts, research shows that, if subjects respond within 15 minutes of being beeped, their responses are fairly accurate (at least in some settings). Therefore, it might be feasible to design a data-collection approach that does not require an immediate response and is less disruptive in school settings, although a 15-minute delay might not be enough to avoid disrupting instruction. In theory, modification of methods like ESM or ELA can better address the challenge of highly differentiated classrooms in which students are engaged in different activities at the same time. Sending a brief common set of prompts with a range of response options could make it possible to measure variation across students in their learning experiences. ESM and EMA also have limitations; in particular, they are not ideal for capturing extensive details about learning situations because both the prompt and response must be brief. Similarly, classroom actions and reactions can happen too fast for this approach to yield good information about features of moment-to-moment teacher–student interactions.

The day reconstruction method (DRM) was developed as an alternative to ESM in situations in which frequent responses are not possible or would impose too great a burden (e.g., they would interfere with the activity they are trying to measure). In DRM, the responses are gathered retrospectively at the end of the day. For example, respondents might be asked to list all the activities in which they engaged during the day and then to annotate the list with different kinds of recollections, such as the time devoted to each activity or their attitudes or feelings about each. Thus, DRM is essentially a form of log designed to capture information about affective factors, such as engagement. Although these affective factors might not be central to indicator 5, this information can be useful for understanding the extent to which students are highly engaged and interested in the activities, which, in turn, might be informative for assessing the broader STEM learning environment. For some uses, DRM has been found to produce results that have a degree of validity similar to ESM. One natural way to apply this method in an educational setting would be a lesson reconstruction approach, asking teachers to list all the activities that occurred during a lesson and then comment on some aspect of each activity—e.g., their perceptions of student engagement. This could even be mapped onto experiences measured directly from students. Together, such techniques could support accurate measures of students' focused learning time and using individual student reports, possibly even variation in learning time among students in a given classroom.

Limitations of Self-Report Methods

Surveys, vignettes, and logs have some limitations as methods for gathering information about students' exposure to STEM content and practices. Classroom surveys are usually administered on one occasion to a sample of respondents to obtain a static snapshot of classroom teaching and learning experiences. They usually require respondents to recall and report on prior events. The results collected from the sampled stakeholders are often used to make inferences about all students' learning experiences over an extended period of time, up to a whole school year. Both generalizing beyond the sampled respondents and relying on memory of prior events can introduce errors into survey findings (Rowan, Jacob, and Correnti, 2009; Stecher, Hamilton, et al., 2002). Surveys are also limited in their ability to capture interactions between teachers and students because they are static and represent the perspective of one actor. Additionally, surveys can be subject to bias because of such factors as social desirability (Grimm, 2010), low response rates, errors in respondents' recollections of their teaching or learning experiences, and differences in respondents' understanding of the items and use of the response options.

Researchers have extensively studied the technical quality of surveys for describing STEM instructional opportunities, but these studies have produced inconsistent findings (Banilower et al., 2013; Kane and Staiger, 2012; Blank, Porter, and Smithson, 2001; Burstein et al., 1995; Porter et al., 1993; Schweig, 2014; Levine, Huberman, and Buckner, 2002). One common finding is that teachers' survey reports of certain practices often overstate the frequency or extent of those practices when compared with other sources, such as direct observations (Kaufman, Stein, and Junker, 2016; Mayer, 1999; Ross et al., 2003; Spillane and Zeuli, 1999). Different findings might reflect actual differences in the technical quality of the surveys that were examined, or they might reflect differences in the sources of reliability and validity evidence that were used. In particular, low correlations between survey results and a student achievement measure might reflect the fact that the survey was not designed to measure practices that are likely to promote the specific skills and knowledge covered in the achievement measure or that the achievement measure lacks adequate reliability or sensitivity to instruction. Similarly, weak relationships between different measures of instructional practice might reflect differences in the specific types of instruction that each measure is intended to capture.

Instructional logs and related methods, such as ESM and DRM, have limitations that are similar to those for surveys. Because logs are typically short and rely on closed-ended questions, they are limited in their ability to collect data about social interactions and differences in students' individual learning experiences. Instructional logs also suffer from potential bias resulting from such factors as social desirability, memory error, and differences in respondents' understanding of the log questions and use of response scales.

Moreover, to produce scores with adequate reliability, logs need to be administered over a period of at least several days, which creates a high response burden for teachers. According to Rowan, Camburn, and Correnti, 2004, roughly 20 logs per year are needed from teachers to reliably discriminate among teachers' instructional content and practice, although the number of logs required to achieve this goal can vary, depending on the constructs to be measured and the reliability and generalizability of scores for a particular construct. The need to administer multiple logs creates challenges for data collection. Logs also impose added costs related to monitoring of responses and the need for financial incentives to promote high response rates. Prior research on the technical quality of instructional logs provided some support for using instructional logs to collect information about classroom instruction but also identified issues that threaten the validity of log results, such as differences in teachers' understanding of the log questions, the lack of occurrence of certain instructional activities covered by log questions, and inconsistent technical properties when teachers use rating scales with different levels to answer the same log questions (Ball et al., 1999; Brandon and Taum, 2005; Camburn and Barnes, 2004; Rowan, Camburn, and Correnti, 2004; Rowan, Harrison, and Hayes, 2004; Le et al., 2004). These concerns also apply to surveys more generally.

Research on the technical properties of results from vignette responses suggests that anchoring vignettes have the potential to correct for self-report bias in surveys. However, additional research is needed to examine their technical quality before using them to collect data for high-stakes decisionmaking (Ruiz-Primo and Li, 2002; Stecher, Wood, et al., 2005; Yuan et al., 2014).

Direct Classroom Observation

Classroom observation protocols are also commonly used to collect detailed data about instruction. Several subject-specific protocols have been developed to evaluate STEM instruction, such as the Reformed Teaching Observation Protocol for mathematics and science (Piburn and Sawada, 2000), the Mathematical Quality of Instruction (Learning Mathematics for Teaching Project, 2011), Quality Science Teaching (Schultz and Pecheone, 2014), Inside the Classroom Observation and Analytic Protocol (HRI, 2002), and UTeach Observation Protocol (Walkington and Marder, 2014). Other non–subject-specific classroom observation protocols, such as the Classroom Assessment Scoring System (Pianta et al., 2009) and the Framework for Teaching (Danielson, 2007) have also been used to evaluate STEM instruction (Kane and Staiger, 2012). For each of the protocols listed, we indicate the subjects for which it has been developed, the domains for which scores are provided, the scoring approach, and, where relevant, the standards that were used to inform development of the protocol. When using these observation protocols, trained raters observe live or video-recorded classroom instruction and rate teachers on a multilevel, multidomain protocol. The observation protocol developers identified key aspects of instruction, typically based on research and theories of teaching and learning, as well as existing educational standards in specific discipline areas. The raters, whether they are school administrators or others, usually need extensive training to understand how to apply the observation protocol in a way that leads to accurate ratings. Many researchers have found that it is difficult for raters to apply the protocols in a consistent manner, and they require that potential observers pass an initial rating calibration test and accept continuing monitoring of their scoring and, where necessary, retraining (e.g., Cash et al., 2012).

Although live observations are probably most common, new techniques for video recording are making it easier to obtain high-quality recordings of classroom activities that permit the observation protocol to be applied at another time or location. For example, the camera and audio input on a tablet computer or laptop can be used to record audio and video during a lesson or interaction. In addition to allowing for ratings to occur in a remote location and at a different time from the instruction that is being observed, these tools also allow recordings to be used for professional development purposes. BetterLesson is a provider that is building a data bank of lesson plans and videos from master teachers that can be used as training tools (BetterLesson, undated). Online platforms, such as BetterLesson, can also be used by teachers to monitor and improve their own performance.

One of the primary advantages of using classroom observations to measure instruction is that the approach provides authentic, moment-by-moment evidence about instructional content and practice. Depending on how the observation protocols are structured, they can also capture information about individual students' learning experiences for a select group of students in a classroom, although it is challenging to capture individual students' learning experiences for everyone in the classroom. An additional potential benefit to observations is that they are not subject to the self-report biases that can influence responses to surveys.

Limitations of Direct Classroom Observation

Compared with surveys and logs, classroom observation can be very expensive and might not be feasible for large-scale data collection. In addition, although some studies have found that structured observation protocols can be applied with adequate reliability, others have raised questions about the consistency of ratings and the validity of scores obtained from these tools for different purposes (Harris and Sass, 2007; Henry, Murray, and Phillips, 2007; HRI, 2000; Jacob and Lefgren, 2008; Kane, Kerr, and Pianta, 2014; Learning Mathematics for Teaching Project, 2011; Medley and Coker, 1987; Piburn and Sawada, 2000; Schultz and Pecheone, 2014; Walkington and Marder, 2014). Rater error is a major threat to the reliability and validity of classroom observation scores and principal ratings, despite great effort to train raters and calibrate their scoring (Casabianca, Lockwood, and McCaffrey, 2015; Myford, 2012; Whitehurst, Chingos, and Lindquist, 2014). Using Classroom Assessment Scoring System data from the recent Measures of Effective Teaching project, Drew Gitomer and his colleagues found that observers had difficulty agreeing on ratings of certain dimensions of teaching, particularly those for which teacher performance was generally relatively weak (Gitomer et al., 2014).

Studies also found varied associations between observation scores and other measures, such as student achievement and student surveys, with most studies finding weak to modest relationships (Garrett and Steinberg, 2015; Banilower, 2005). However, as noted above, lack of strong relationships with these measures does not necessarily indicate lack of validity because the measures might be capturing different aspects of instruction. In addition, some studies that examined the internal structure of existing observation protocols did not find strong evidence to support the predetermined structure of these protocols (McCaffrey et al., 2015). Moreover, some analyses have uncovered a lack of variation in ratings on certain dimensions of the observation protocols they studied, which makes it difficult to study the validity of rating scores by comparing them with other measures (Schultz and Pecheone, 2014). All these findings suggest the need to continue investigating the technical quality of classroom observation protocols and how to better train raters to use them.

Artifact-Based Methods

The third major approach that has been used to try to measure instruction focuses on the materials that are used or generated as part of the instructional process. These artifact-based approaches can focus on published curriculum materials, artifacts generated in the classroom (such as worksheets, assignments, and student work products), and portfolios, which are organized collections of materials of many types.

Published Curriculum Materials

Curriculum materials, including textbooks and other supplemental instructional materials, have substantial influence on what and how teachers teach in the classroom. Previous research found high correlations between topics covered in textbooks and what teachers actually taught (Schmidt, Houang, and Cogan, 2002). Although there is a vast amount of variation in the enacted curriculum among teachers who use the same curriculum materials (Remillard, 2005), analysis of such curriculum materials can provide important information about the learning opportunities that a student might have for exploring different types of content and cognitive activities.

For instance, Robert Reys and his colleagues evaluated the impact of three standards-based mathematics curricula funded by NSF—Connected Mathematics, Mathematics in Context, and MathThematics—on middle school students (Reys et al., 2006). Their review of the NSF-funded and publisher-generated mathematics textbooks found a significant difference in the extent to which these curriculum materials emphasized major content strands. Curriculum materials generated by publishers spent significantly more time on the numbers and operations strand and significantly less time on the geometry and measurement and the data analysis and probability strands, than the NSF-funded curriculum materials. Another study, by Mary Kay Stein and Gooyeon Kim, found that certain textbooks placed higher instructional demands on teachers than others did (Stein and Kim, 2009).

They also found that teachers teaching both types of mathematics curricula did not differ in terms of the extent to which their instruction covered the content in the textbook. Both groups taught about 60 to 70 percent of the textbook lessons, including about 80 percent of the lessons related to numbers and operations and 60 to 70 percent related to data analysis and probability. However, teachers who used the NSF-funded curriculum placed greater emphasis on algebra than teachers who used the publisher-generated curriculum.

Jitendra et al., 2005, reviews five third-grade mathematics textbooks to examine the extent to which these textbooks provided opportunities for students to learn problem-solving, reasoning, communicating, connecting, and representing mathematical content, as identified by *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989). The authors used a researcher-developed protocol and trained raters to code word problems in lessons on addition and subtraction of whole numbers in the selected five textbooks. Results showed that these textbooks varied substantially within and across each standard examined. Although all five textbooks presented a reasonable number of opportunities for problem-solving, most of them did not provide many opportunities to learn about reasoning, communicating, connecting, and representation.

Another example is a study by Morgan Polikoff in which he used the Surveys of Enacted Curriculum to examine the alignment between four popular fourth-grade mathematics textbooks and standards, including the Next Generation Sunshine State Standards and the

Common Core State Standards for mathematics (Polikoff, 2015). Trained analysts coded both the standards and the selected textbooks, and indices were generated to show the alignment between textbooks and standards. Results showed that the alignment between the standards and selected textbooks was not ideal. Moreover, the majority (ranging from 87 to 93 percent of the total textbook content) of all four textbooks emphasized memorization and procedures. Less than 15 percent of the textbook content required students to demonstrate understanding. Four textbooks had almost no content that required students to conjecture, generalize, prove, solve nonroutine problems, and make connections—activities that reflect high levels of cognitive demand. These examples demonstrate how reviews of curriculum materials can provide evidence of learning opportunities for students.

Classroom Artifacts

Classroom artifacts, such as teachers' lesson plans, assignments, assessments, student work, and scoring rubrics, have been used to gather information about instructional content and practice. Often, teachers are asked to collect specific kinds of artifacts as part of a portfolio to describe their work. Prominent examples of protocols developed to analyze classroom artifacts include the Intellectual Demand Assignment Protocol (Newmann, Bryk, and Nagaoka, 2001), the Instructional Quality Assessment (IQA) (Matsumura, Slater, et al., 2006), the Scoop Notebook (Borko, Stecher, and Kuffner, 2007), and the Quality Assessment in Science Notebook (Martínez, Borko, Stecher, et al., 2012).

The Intellectual Demand Assignment Protocol was developed to examine the authenticity and intellectual demand of classroom assignments in writing and mathematics (Newmann, Bryk, and Nagaoka, 2001). For each subject, the developers identified three standards for assignments and student work, respectively. These standards focused on the construction of knowledge, disciplined inquiry, and value beyond school within each subject. Each standard was translated into more-detailed scoring rubrics. Trained raters evaluated teacher assignments and student work on a four-point scale using the detailed scoring rubrics for each subject.

Another way to use student work as the basis for an indicator is through an online digital repository or archive for storing and reviewing student work. Many schools now use technology, such as Epsilen or Moodle, to allow students and teachers to create online portfolios that document their work in a manner that others can review and assess (Texas Education Agency, 2013). As one expert told us, the NGSS call for "students to do projects, and technologies are available to collect portfolios and document student project work. It would be highly valuable to show what teachers are doing to make science accessible to students."

Another artifact-based approach for measuring instruction is the IQA. The IQA examines three aspects of classroom instruction in reading comprehension and mathematics: level of cognitive demand of tasks and activities, classroom talk, and expectations communicated to students for the quality of their work (Matsumura, Garnier, Slater, et al., 2008). The level of cognitive demand of tasks and activities was evaluated based on subject-specific evidence. For reading comprehension, it was evaluated based on the potential of the task to support high-level engagement with a text, the intellectual demand of the task or discussion, and the guidance students received on how to write extended responses and use evidence to support their arguments. For mathematics, it was based on the potential of a task to support high-level conceptual thinking and the cognitive demand of enacted learning tasks.

The IQA also includes observation rubrics for classroom talks and teachers' expectations, which are similar across the two subjects (Matsumura, Garnier, Slater, et al., 2008). The rubric for classroom talks focuses on the proportion of students who participated in a discussion, the extent to which a teacher presses students to explain their thinking and use ideas and concepts, and a teacher's use of specific "talk moves" to help all students understand and reason. The rubric for teachers' expectations focuses on the content of instruction teachers provide to students and how they communicate to students about what "good" student work should look like. Teacher assignment and student work are evaluated on a five-point scale using the IQA rubrics.

The Scoop Notebook protocol was initially developed for mathematics (Borko, Stecher, and Kuffner, 2007) but has also been modified for use in science classrooms (Martínez, Borko, and Stecher, 2012). Its developers identified 11 dimensions of mathematical instruction based on such documents as *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 2000). Multiple classroom artifacts can be analyzed using the Scoop Notebook (Borko, Stecher, and Kuffner, 2007). For example, the instructions ask teachers to provide three types of artifacts: materials generated prior to class (e.g., lesson plans and handouts), materials generated during class (e.g., writing on the board and student in-class work), and materials generated outside of class. All materials are rated on a five-point scale.

Technology is also being used to enhance the capability of artifact methods. The electronic Quality Instruction in Science (e-QIS) tool (University of California, undated) is a tablet-based portfolio application derived from the Scoop Notebook in science and aligned to the 11 dimensions of practice aligned to the NGSS. E-QIS allows for the collection and scoring of an enhanced range of classroom artifacts, including not only teacher-generated materials and student work products but also photographs of activities and experiments and short videos of presentations, discussions, and other classroom interactions. The materials are uploaded through a Wi-Fi connection to the project data archive, where they can be viewed and scored (Martínez, Kloser, et al., 2016).

Portfolios of Classroom Materials

As noted above, artifacts and other resources are sometimes combined into portfolios that are used for teacher licensure, certification, or evaluation. Teachers select a collection of materials to show evidence of their teaching practice, school activities, and student progress. This evidence might include lesson plans, assignments, assessments, student work samples, videos of classroom instruction, reflective writing, notes from parents, and special awards or recognitions (Goe, Bell, and Little, 2008). Teachers also provide explanations of how these materials show that their performance meets the targeted performance standards for high-quality teaching.

Perhaps the most prominent example of a teacher portfolio program is the National Board for Professional Teaching Standards certification. The National Board certification recognizes accomplished teachers with high teaching performance (National Board for Professional Teaching Standards, 2002). It includes an assessment of subject knowledge and a portfolio assessment of teaching practice. Teachers provide four different types of materials for their portfolios, with three based on classroom teaching and one based on their work with families, the community, colleagues, and the larger profession. Required evidence includes videos of instructional practice and teacher–student interaction, as well as student work samples. Teachers also need to provide detailed reflection and analysis notes of the materials submitted. Trained assessors evaluate whether teaching practice demonstrated in the submitted portfolio materials meet the rigorous standards set by the National Board.

The Stanford Center for Assessment, Learning and Equity developed edTPA, another portfolio-like process that builds on the growing ease of videotaping and the growing capacity of networks to transfer large digital files electronically. According to its website, edTPA is a "performance-based, subject-specific assessment and support system used by teacher preparation programs throughout the United States" (edTPA, undated). Teachers submit portfolios to edTPA that include lesson plans, videos of their teaching, and assessments, and trained teaching experts judge this material. Many teacher preparation programs also use this portfolio system to assess preservice teachers.

Compared with surveys and logs, artifacts and portfolios have the advantage of being able to provide richer data about teacher–student interactions (either through video or through written feedback teachers provide on students' work) and a better picture of the enacted curriculum. And compared with classroom observations, artifacts and portfolios can draw on a broad range of evidence regarding student–teacher interactions and instructional activities rather than being limited to one or a few specific lessons. Evidence collected in portfolios is sometimes considered as providing an "authentic assessment" of classroom instruction (Goe, Bell, and Little, 2008).

Limitations of Artifact-Based Methods

Results from research on the technical quality of curriculum materials, classroom artifacts, and portfolio ratings suggest that these are promising methods for collecting detailed data about instructional content and practice (Borko, Stecher, and Kuffner, 2007; Johnson, McDaniel, and Willeke, 2000; Koretz et al., 1994; Matsumura, Garnier, Slater, et al., 2008; Newmann, Bryk, and Nagaoka, 2001; Tucker et al., 2003). However, there are limits with each method. For example, teachers vary in the extent to which they adhere to the curriculum materials, and many teachers draw on supplemental materials from a variety of sources. Opfer, Kaufman, and Thompson, 2016, for instance, reports that teachers in states that had adopted the Common Core State Standards for math took advantage of a large trove of online resources to supplement or replace lessons in their schools' adopted curricula. Therefore, an accurate understanding of what teachers and students are doing requires that reviews of curriculum materials be supplemented with evidence from other sources, such as instructional logs or reviews of classroom artifacts and portfolios to obtain a comprehensive understanding of students' learning experiences in the classroom. However, rater error can threaten the reliability and validity of artifact and portfolio ratings (Stecher, Wood, et al., 2005; Junker et al., 2006). These studies also suggested that a large number of raters would be required to ensure the quality of ratings of teaching portfolios (Johnson, McDaniel, and Willeke, 2000). The completeness of the artifacts and portfolio entries collected also affect the quality of ratings (Stecher, Wood, et al., 2005). In addition, some studies also found a lack of variation in artifact ratings that raises concerns about validity and makes it difficult to compare scores with other measures (Matsumura, Garnier, Pascal, et al., 2002). A final limitation is that the process of collecting, organizing, and scoring artifacts and other materials for portfolios is time-consuming and expensive compared with some of the other alternatives described above.

Summary

This brief review indicates that, although a wide variety of measures of teachers' instruction and students' learning environments are available and have produced high-quality and useful data for research and evaluation purposes, none of them alone provides adequate information to inform indicator 5. And, although measures from multiple categories in this chapter (e.g., teacher and student surveys along with classroom observations) have been combined for some purposes, such as teacher evaluation, even the combination is likely to be inadequate for addressing indicator 5. In particular, most measures are designed to capture information about what the teacher is doing and therefore do not provide evidence regarding the ways in which individual students' experiences differ within that teacher's class. As noted earlier, this limitation is particularly problematic in personalized-learning environments that rely heavily on technology and that offer different learning experiences to different students. In addition, many available measures lack evidence of technical quality for the purposes of monitoring teaching and learning experiences, some of them are expensive to administer and score, and few measures have been developed to capture instructional experiences in engineering and technology. Developers could adapt some of these approaches in ways that would allow them to examine within-classroom differences, such as through logs that ask teachers to focus on a different student during each lesson, but this approach would provide only minimal information about each student's experiences and therefore would support only limited inferences about those experiences. Clearly, there could be benefits to an R&D effort focusing on new ways to collect evidence of STEM learning experiences. In the next chapter, we discuss several promising methods that could eventually serve this purpose.

CHAPTER THREE

Promising Digital Data-Collection and Analysis Methods

In this chapter, we explore innovative data-collection methods for capturing evidence relevant to indicator 5 that could address some of the shortcomings of the commonly used measures identified in Chapter Two. In particular, we describe several digital data-collection tools that might be utilized to measure STEM learning experiences from the students' perspective. We also discuss the extent to which these methods are suitable for use in personalized-learning environments—i.e., whether they can be adapted to classrooms in which exposure to content and instructional practices varies across students.

To identify promising new methods, we draw on the interviews we conducted with educational researchers and instructional product developers, described in Chapter One. All of the data-collection and analysis methods mentioned in this chapter are being deployed in some manner in schools, although many are not yet ready to be implemented on a large scale. Furthermore, many of these methods are associated with computer-based instructional delivery platforms that are far from universal in K–12 public schools. Nevertheless, the rapidly changing landscape for technological applications in education suggests that the effort to explore approaches to creating an indicator of classroom coverage of STEM content and practices should take a forward-looking view, despite the uncertainty and risks associated with that approach. Today's experimental, technology-based approaches to measurement might support tomorrow's widely used indicator. In this chapter, we examine technological data-collection options involving audio and video recording in traditional classrooms, computer-based learning environments and learning management systems, and computer gaming for learning. In each case, we explore the extent to which the technology affords improved methods for measuring students' exposure to STEM content and practices, particularly in highly individualized classes in which students are working on different material.

Audio Recording

Audio recording has been used in a variety of ways to monitor teacher and student spoken interactions during lessons. Even digital recording is relatively "low-tech" by today's standards, but it is effective at capturing interactions and making them available for detailed analysis at a later time. Audio recording might be employed in a variety of ways. For example, one expert suggested a simple model for gathering information on students' science investigations. A digital recorder is given to one student or a group of students who are asked to narrate a description of their experience. Then it is passed to the next student or group of students until all have described their efforts. Listening to the recording provides a lot of information on the activities in which the students engaged, as well as their understanding of the underlying scientific ideas. At the other extreme, digital audio can be used to keep a record of a whole lesson in real time.

One drawback of audio recording, particularly whole-class recording, is that analysis of such recordings can be both time-consuming and complex. However, researchers interested in instructional improvement have recently begun to approach the analysis of classroom audio with more "high-tech" methods that use natural language processing software to provide teachers with rapid feedback based on an analysis of classroom audio recordings. These techniques include the use of discourse analysis to identify features of speech, such as the number of interactions initiated by the teacher or by students, the amount of discourse occurring during a lesson, and the quality of the discourse (Zastavker, Darer, and Kessler, 2013). For example, researchers can calculate the balance between teacher talk and student talk during a lesson and use a discourse visualization tool to produce visual displays that highlight features of classroom interactions, such as turn taking and length of utterances (Clarke et al., 2015; Chen, Clarke, and Resnick, 2015). These structural features of the learning experience could add to our understanding of students' opportunities to engage in standards-aligned practices in particular, although they might not provide complete information about classroom discourse, particularly when multiple groups are engaged in discussions simultaneously.

Video Recording

Video cameras have been used for years to capture classroom interactions that are subsequently rated using such rubrics as the ones described in Chapter Two, but the use of cameras was typically restricted to large-scale, externally funded research projects, and they were not widely used by local educators on a regular basis. Video remained limited as a research activity because the cameras were awkward to use and expensive, produced large amounts of information that was difficult to analyze, and were invasive and disruptive of normal teaching and learning. Various new kinds of cameras can be more easily incorporated into classrooms, providing insights into students' engagement with content in relatively inexpensive and noninvasive manners. These devices include small, portable video recorders; wearable cameras; and mini-cameras built into smartphones, tablets, and laptop computers. These devices can capture audio and video directly or stream it over the web to observers or to central data storage for later review. For example, 360-degree cameras can capture multiple concurrent activities, which are common in personalized-learning environments and STEM classes. BetterLesson outfits master teachers with wearable cameras to build a library of good lessons and accompanying lesson plans; they use these resources to provide virtual coaching and professional development for teachers. One of our interviewees has found that cameras worn by students can also produce good sound and picture quality, and it is possible to capture almost everything that is occurring in a classroom. Equally importantly, new software can make the analysis of video files more efficient. Although video recording has most often been used in research to study teacher practice and learn about effective pedagogy, it also provides a source of information that could be tapped to study students' engagement in learning activities and exposure to content.

There are other ways to capture and use images that are not as burdensome as extended classroom video. For example, web cameras found in many laptop computers and tablets can be used to capture students' facial expressions, which potentially offer information about students' engagement with the lesson. Eye-tracking software is becoming available in small devices, which offer the possibility of directly monitoring students' interaction with computer-delivered lessons (e.g., MangoldVision eye-tracking solutions) (Mangold International, undated). Such information could be used to measure the extent to which students focus their attention on computerized learning tasks. Other types of sensors, such as measures of galvanic skin response (see, e.g., Affective Computing, undated), could be used to gain insight into students' classroom experiences, including their affective responses to lessons (e.g., engagement), although efforts to use such measures have been met with resistance in the past (see, for example, Kroll, 2012).

Evidence Collected in Computer-Based Learning Environments

Many of the experts we contacted are working on measurement methods that are embedded in computer-based instructional programs of one type or another. Such measurement strategies have clear advantages in a personalized-learning environment in which each student has input into his or her own learning priorities, progresses through topics at his or her own pace, and has proficiency measured when relevant instructional activities are completed. A primary advantage is that embedded approaches are likely to be less disruptive to instruction than stand-alone measures. Our interviewees were also enthusiastic about embedded measurement because they saw other advantages in using digital technology to enhance teaching and learning; they mentioned in particular the power of digital simulations to expand the scope of student learning experience (e.g., collecting evidence in a simulated undersea environment) and the customized feedback that digital tutors can offer to students. Technology-enhanced, personalized-learning environments can retain large amounts of data about students' learning interactions with computerized lessons and games, their performance on periodic assessments, and their work products (whether retained informally or cataloged in a personal portfolio). They can be rich environments for extracting digital information. As one respondent explained, "In personalized learning, on any given day, there should be lots of different kinds of instruction . . . online, face-to-face, tutoring, small groups, whatever works." In such a context, it would be almost impossible to use a stand-alone strategy to measure the kinds of learning occurring for every student. Yet, if much of the learning is being guided or supported or monitored electronically, the task of tracking each student's learning activities is potentially more feasible. For example, the Web-based Inquiry Science Environment (WISE) is an online virtual learning environment that can be used to develop inquiry projects that involve interactive simulations, graphs, and models. Researchers studying WISE have found that teachers used evidence from student work and from embedded formative assessments to revise their curricula and rethink their instructional plans (Gerard, Spitulnik, and Linn, 2010). Information about student performance captured in WISE enabled teachers to customize curriculum and pedagogy in ways that improved both student and teacher learning. Teachers found that the information contained in student responses to formative assessments was detailed enough to support targeted instructional improvements. Furthermore, evidence about students' assessment responses and teachers' efforts to address student needs revealed in these responses can itself inform our knowledge of students' learning opportunities.

One significant limitation of most currently used methods for extracting data from instructional technologies is that they tend to focus on measuring content-specific student achievement and related student outcomes rather than the instructional experiences that might promote these outcomes. The promise of new teaching and learning technologies for measuring students' opportunities to learn is unknown, but the methods we selected for inclusion in this chapter are well suited to capturing information about the learning environment. Thus, with adequate R&D, they might have potential to support measures that go beyond student outcomes. And as we noted previously, to be consistent with the NGSS, it is important to document students' opportunities to "do" science rather than just learn about it.

Keystroke Mapping and Log-Stream Analysis

Computer-based learning environments afford numerous opportunities to gather information about teaching and learning activities. In addition, they are quite varied, presenting different contexts with different opportunities for measuring classroom coverage of practices aligned with the Common Core State Standards for math and the NGSS. For example, virtual environments are one of the most powerful ways in which computers have been used to enhance teaching and learning. Some of the most interesting applications are in the form of simulated multiuser virtual environments, but the potential for gaining insight into student learning experiences is present in most simulations. In EcoMUVE (EcoLearn, undated), the student is placed in a simulated ecological environment that is under stress, such as a pond losing aquatic species. The user commands an avatar who can move about in the environment; use scientific tools to measure key features, such as water pH and temperature; and is asked to report on conditions, develop hypotheses about causes, and more. While users interact within virtual simulations, such as this one, their actions can all be tracked in the system's electronic record—not just their answers to prompts or questions but actions, such as their direction of movement, choice of tools, and speed of response.

Many researchers think that log-stream analysis or data mining of student–computer interactions in technology-based learning environments offers powerful opportunities for measuring more than mere mastery of facts or procedures. For example, tracking patterns of movement and action in EcoMUVE can yield information about a student's level of participation in instruction and his or her prior exposure to relevant constructs, which can shed light on what activities are occurring in the classroom. Similarly, although the Cognitive Tutor (Carnegie Learning, undated) software is designed to diagnose student understanding based on responses to questions and offer customized support and feedback, the log-stream data can help researchers and practitioners understand how much time students spend working with the software each day, how they spend this time (e.g., what materials and resources they access), and the kinds of problems on which they work. These data might also reveal whether certain resources (such as help, exercises, or quizzes) are being accessed as frequently as intended, what kinds of trajectories students take through the learning materials, and which trajectories are more or less productive for learning. In science, Janice Gobert and colleagues have used data-mining techniques on log data from scientific microworlds to assess student inquiry skills (Gobert et al., 2013); these same data might also provide evidence regarding students' opportunities to learn these inquiry skills.

If these kinds of instructional software were widely used, the data might support valid inferences about students' learning experiences writ large. For example, Lee, Penfield, and Maerten-Rivera, 2009, uses log-stream analysis from science simulations to tabulate the proportion of responses generated by student partners and to create a class-level measure of student participation. The authors found that this measure of implementation predicted class learning and mediated relationships between other teaching variables and student learning.

This finding suggests that log-stream data are potentially powerful measures of students' learning experiences. Moreover, most learning systems have the ability to track student clicks and keystrokes and develop an extensive stream of evidence about a student's pattern of interaction, providing a potentially rich source of evidence to learn about the kinds of opportunities students have to engage in activities that might support attainment of the skills and knowledge specified in the Common Core State Standards for math and the NGSS. Eventually, this capability could extend to other interaction modes as the environments become more sophisticated. Possibilities include students controlling the environment through voice commands, gestures, or, in the case of virtual reality, moving through the environment.

Currently, however, most software packages do not collect data in a form that easily lends itself to the construction of an indicator. For instance, information is sometimes logged in an incomplete way or is recorded at a low level that does not easily translate into meaningful measures of learning or instruction. Furthermore, although it is possible to collect objective measures of student interaction with the system (e.g., keystroke counts, pathways taken, and time of engagement), it is much harder to interpret this information and draw a valid inference about some aspect of student learning. For example, Common Online Data Analysis Platform and SageModeler are online data analysis and visualization platforms that students can use with a wide range of data to explore patterns, visualize relationships, and more. However, students use these tools in exploratory ways, and, although their explorations can be tracked, it is difficult to make any sense of the trace because each student or team can be exploring a different part of the data, examining a different relationship, and using the tool in a different way.

Researchers are exploring ways to analyze log-stream data for instructional uses, including knowledge engineering, machine learning, and expert models, that could also support the development of an indicator. These approaches fall into two classes: post hoc methods that try to extract meaning from existing data by looking for patterns and associations and built-in methods that try to build meaning into the data by imposing structure on the system at the design stage.

Post Hoc Log-Stream Analyses

Some of the researchers with whom we spoke expressed great faith that the first type of brute-force analytic methods (i.e., methods that look for mathematical patterns in large data sets without predetermined hypotheses or expected relationships) can be used with logs of student–computer interactions to tease out information about a student's mastery of scientific practices, as well as insights into students' intrapersonal competencies, such as engagement or persistence in the face of challenge. As one respondent said,

> I think that online environments can measure student persistence and willingness to review materials that improve their answers, help-seeking behavior, and how they use the helping material. [The] ability to recognize that you need help and that you should air your own ideas [is] perhaps more important than anything you can teach.

In theory, these insights could be gleaned from students' interactions with online simulations, such as Gizmos (ExploreLearning, undated), or online interactive textbooks, such as Techbook (Discovery Education, undated). In one specific example, log data from intelligent tutoring systems, such as Cognitive Tutor, have been analyzed to determine whether students are permitted to progress through mastery for the entire school year or whether teachers interrupt this process to move students onto more-advanced material, perhaps because of a need to cover material that is included in standards or tests.

Built-In Log-Stream Methods

Other researchers are less sanguine about brute-force data-mining methods, and some expressed a worry that we might make inappropriate inferences from the data just because they are available. However, many experts were confident that it is possible to draw insights about instruction from simulations if they have inferential intentions designed within the simulation. They argued that, to be useful, student participation data need to be collected purposefully, based on a theory of learning or specific design principles. As one respondent told us,

> You can't just collect all keystrokes and throw them into a Bayesian net model and explain what kids are learning. You need to understand the context of clicks. You need to create the game so you can code the environment in a way that allows you to make inferences.

Furthermore, to interpret log-stream evidence meaningfully, users need to understand the instructional context in which the simulation or other technology is being used, e.g., what activities preceded it, what goals did the teacher set for it, and what prior experience have students had with similar activities?

Simulations that are built using principles of evidence-centered design are an example of this type of approach. They yield much more direct insights into targeted constructs, such as student understanding, because the virtual world and the options available to students have been built around an understanding of specific scientific principles and the way they normally develop in students. In building the SimScientists simulation, researchers at WestEd developed design guidelines so the inquiry experience would generate interpretable diagnostic information about student understanding. They also developed ordered continua to characterize the complexity of scientific phenomena and the complexity of scientific practices, which they use to help in interpreting student interactions with the simulated environment. For example, the continuum for complexity runs from recognition at one end to developing arguments at the other (Quellmalz et al., 2012). The developers used evidence-centered design to incorporate levels of these continua into the design of the simulations.

Such approaches are not limited to the STEM fields. Writing Pal is an interactive writing training system that teaches writing strategy and instruction for persuasive writing and offers monitored practice. Students can write and revise whole essays within Writing Pal and receive feedback that is informed by the underlying strategic approach to writing. Multiple scoring algorithms focus on different components of writing, such as length, structural elements, and paragraph quality, and provide appropriate feedback to students. (Jacovina and McNamara, 2016). By building these features into the system, developers ensured that students' progress through the system would yield interpretable, actionable data.

In a similar manner, it should be possible to design instructional environments that collect evidence about a student's learning experience, the kinds of materials presented, the nature of the options selected, the time spent in the environment, and other indications of each student's learning activities. Were such design features incorporated into the computer-based learning activity, it would be possible to extract measures of opportunities to learn at the individual student level, which could support an aggregate indicator.

In some cases, students are given the ability to make annotations, choosing when to insert a note into the interaction stream. One of our respondents described a useful tool that is provided in some virtual learning systems: a record button that allows the student to take a snapshot of his or her position in a given simulation and annotate it with comments about his or her thinking, problems encountered, or frustration with the challenge. Student annotations like this should trigger responses, either from the teacher or the system itself, customizing instruction to meet the needs of the student. As a by-product, information about such queries and responses might be used to develop an indicator of specific learning opportunities.

Many of these computer-based learning systems also are designed to provide feedback that teachers can use to improve instruction outside the context of the computer environment. For example, one of our respondents described a system that uses log files from scientific simulation-based software and provides real-time diagnostics in the form of metrics on which teachers can act—e.g., how many students are failing to form a testable hypothesis? Are students changing too many variables at one time? By monitoring when this diagnostic information is provided and how teachers use it, researchers can learn more about the learning environments in which STEM instruction takes place.

Gaming

Gaming is another type of computer environment that excites many educators, who see electronic games as potentially potent learning and assessment tools. The term *gaming* describes a wide range of interactive software delivered on smartphones, tablets, and computers. Games usually involve a virtual environment with rich graphics, obstacles or challenges, an end goal, rules of play that delineate allowable actions on the part of the gamer, and rapid or continuous feedback on the effects of one's actions (Clark, Tanner-Smith, and Killingsworth, 2016). Educators are trying to take advantage of the enthusiasm that students show for electronic games and use this engaging medium for teaching and learning purposes. Although it is not a primary focus of educational gaming at present, it is not difficult to imagine how student interaction with games could also provide information to support measures of many facets of instruction. For example, science inquiry games, such as Quest Atlantis, Atlantis Remixed, and Newton's Playground, engage middle school students in scientific discovery (Atlantis Remixed, undated; Educade, undated). In Argument Wars, students argue historical Supreme Court cases to develop their ability to understand valid and invalid arguments (iCivics, undated). Zoo U is an interactive game designed to assess and improve students' social and emotional skills, such as impulse control, emotion regulation, and cooperating with others (Centervention, undated).

These games are designed to improve students' academic skills, as well as their inter- and intrapersonal skills, and playing the game is a learning opportunity. At a minimum, exposure to the game could be measured as one important learning activity. More-sophisticated information about pathways through these kinds of games might reveal more-nuanced insights into the kinds of challenges each student confronts as part of his or her individualized learning pathway.

Another type of game is primarily designed for assessment. For example, Project Gemini is a game whose primary purpose is assessing skills. It is designed to measure collaborative problem-solving while a pair of students works together to solve Rube Goldberg–style puzzles. Assessment games seem less relevant to measuring opportunity to learn, but, in a game-rich environment, assessment games might be used to assign students to subsequent games based on their assessment performance. In combination, students' trajectories through a sequence of games might provide an indirect indication of the skills they have had opportunities to learn. Games also allow students to experiment with situations that are too dangerous or time-consuming in real life, giving them simulated experiences that could not be included in classroom learning. These experiences can be made more realistic using virtual-reality simulations.

Another way to use games to learn about classroom coverage of STEM practices, as well as student performance, is to integrate higher-order learning experiences into game navigation and play. The Scaffolding Understanding by Redesigning Games for Education series attempts to do this by helping students build on the intuitive understandings that emerge from game play to "develop more formal representations, concepts and processes" (Scaffolding Understanding by Redesigning Games for Education, undated [a]). In the Fuzzy Chronicles, one of the SURGE games, players navigate an avatar to collect treasures while avoiding obstacles; they are asked to create a navigation plan to improve their actions, and there is evidence that this approach transfers into improved model-based thinking in other contexts (Scaffolding Understanding by Redesigning Games for Education, undated [b]; Clark, Tanner-Smith, and Killingsworth, 2016). Monitoring the steps students take to navigate through the game could yield a direct measure of their exposure to learning experience embedded in game navigation.

The previous examples illustrate possibilities for extracting information about students' opportunities to learn in an unobtrusive manner based on their game play. Such data synthesis might be done at the point when a student finishes interacting with the game, or it might be done in real time. One example of real-time processing of this type has been described as "stealth assessment," which is invisible to the student but can provide teachers with immediate feedback on the development of student competencies (Shute, 2011). For example, researchers were able to conduct stealth analyses of differences in students' learning behaviors while interacting with an adaptive reading learning environment, Interactive Strategy Training for Active Reading and Thinking—Motivationally Enhanced (Snow et al., 2016). It might be possible to embed "stealth" measures of learning opportunities into games as well.

Finally, some developers are building larger architectures to more easily track students' learning experiences within and across game contexts for a variety of purposes. Although most of these efforts are currently focused on tracking student learning, it does not seem far-fetched to imagine that they could be adapted to track students' participation in STEM practices as well. For example, the Tin Can application program interface works across multiple platforms and allows people to link game play information with other evidence of student performance (see Tin Can Application Program Interface, undated). The GlassLab Game Services platform can be overlaid on any game to extract and report data so student progress can be more visible to teachers. The GlassLab teacher dashboard provides measures of student progress across five performance levels and produces alerts to teachers in the form of "shout outs" and "watch outs," as well as suggestions for future instruction in the form of "what now?" reports (GlassLab Services, undated). A similar approach might be used to derive measures of students' exposure to different content and different learning experiences.

Although there is quite a lot of optimism among developers and educators regarding the potential of the methods described in this chapter to improve teaching, learning, and assessment, there are important questions about how any of these techniques could be used to support national indicators of students' STEM learning experiences.

Challenges of Incorporating Digital Measures into Indicators

There are both conceptual and practical challenges that will have to be overcome before useful measures of STEM learning can be derived using these newer measurement approaches. Our conversations with experts raised several conceptual concerns. The primary concern when it comes to educational measurement is always validity; that is, can we make an evidence-based case that the information extracted from logs of clicks and keystrokes, segments of video, analyses of discourse, responses to formative questions, patterns of game interactions, or sets of online responses accurately portray the targeted aspect of the learning environment? This is not a trivial matter, particularly when information comes from unfamiliar sources. One usual aspect of making an argument to support a validity claim is to show patterns of convergence and divergence among measures of similar and dissimilar concepts. Relevant evidence is likely to be harder to find when the construct of interest—in this case, classroom coverage of STEM practices—has not been measured extensively before, and few alternative means exist to measure it.

The situation becomes even more complex with methods that draw on multiple sources of data. Although triangulating among methods (e.g., surveys and observations) is a standard technique in qualitative analysis and in some assessment and evaluation contexts, research is needed to assess the quality of the results when combining multiple measures derived from various technology-based and non–technology-based tools. Moreover, it might be difficult to show that measures derived through complex algorithms that act on streams of information produce reliable information across settings, individuals, and approaches.

Similarly, although these systems demonstrate exciting potential for providing detailed information about student learning experiences, many of them are built on very weak conceptual models—i.e., they do not have an explicit theory that connects the learning environment, the choices given to students, and the pattern of student responses to a claim about students' knowledge and capacity. This is not surprising because the primary goal of most developers is instruction rather than measurement, but it is disappointing because most software entails some degree of assessment of student understanding as a basis for moving the student through the options. Yet, these assessments are often unsophisticated because they need to provide only enough information for the system to assign the next available activity not to make a judgment about the student's knowledge or performance that is valid outside the context of the system.

An additional challenge is identifying which aspect of student STEM learning activities should be included in an indicator and making them the priority for measurement. It is tempting, when working with new tools, to put the emphasis on those things the tools are best at producing and those concepts or ideas the tools are best able to measure. There is a natural excitement about using a new tool to measure something we were not able to measure before. However, as an example, the fact that eye-tracking software can judge whether a student is focused on a laptop and is engaged with a computer-based activity rather than glancing around the room does not mean that steely-eyed concentration is an important indicator of a desirable learning situation. At the macro level, eye-tracking software might reveal students who spend little or no time engaged with a particularly activity, but, at the micro level, differences in patterns of glances are likely to reveal little that is important about a student's learning experience. There might be temptation with embedded measures to put the measurement cart before the learning-domain horse, but it should be resisted. As one of our experts noted, the choice of measures that are included in an indicator will send messages to educators about what is important, and those should be aspects of STEM instruction that are represented in the standards.

There are also practical challenges associated with using these measurement techniques in a national indicator. The most obvious challenge is that none of these innovative methods is used widely at present, and it is hard to imagine that any will ever be used universally. New products are constantly being developed, and existing products are frequently subject to revisions. A national indicator is useful only if it represents features of the educational system on a wide scale. Said another way, users of a national indicator system are likely to want to infer what conditions would be found in a typical educational setting anywhere in the United States. At the moment, none of the methods highlighted in this chapter is being used throughout the country. The limited current usage does not mean that an indicator derived from these measures cannot be employed at some point in time. It will be important to monitor the uptake of these capabilities and not attempt to create new measures before the capacity to implement them is widespread.

The lack of uniform uptake of these new technology-rich teaching and learning tools is due not only to the diversity of products available but also to practical concerns, including lack of broadband access in schools and lack of adequate computers to give all students sufficient access to employ some of these tools. These concerns might be minimized by the large growth in the number of students with smartphones and the migration of much software to tablets and smartphones. Beyond the challenges associated with computing capacity, there are also practical challenges associated with teacher capacity. Teachers' comfort with new technologies would seem to be a necessary condition for widespread use of the kinds of digital tools described above.

Of course, it is not necessary to achieve universal application. Indicators almost always draw on information from a sample of schools or classrooms, so it might be possible to add new measures once the necessary tools are in place in a representative slice of classrooms or schools. For example, scenario-based interactive computer tasks are now being used in NAEP, administered to a representative national sample of students. That such tasks have been successfully deployed in NAEP suggests that there is promise for measures that are derived from computer-based activities to be included in an indicator system.

Potential Benefits of New Digital Tools

The focus of this report is on measures that might be useful for developing an indicator of STEM learning opportunities, but it should be obvious that the tools we described offer other educational benefits. Most of the technologies were developed to improve teaching and learning directly or to provide better insight into student understanding. For example, simulated learning environments are designed primarily to help students think like scientists, improve their reasoning ability, or enhance their social and emotional skills. Many of these systems also provide teachers with useful feedback on student mastery that can serve as the basis for instructional planning. Teacher dashboards that track student progress through games or curriculum units are examples of tools designed to provide helpful formative feedback that teachers (and students) can use to improve learning. Similarly, many of the archival tools (e.g., video systems and electronic portfolios) are designed to stockpile information that teachers can use to gain insights into their own practice. Providing data that could serve as an indicator of STEM content and practices was probably not on the minds of any of the developers of the tools described in this chapter; all exist primarily to affect classroom practice more directly. Their potential role as the basis for an indicator of STEM learning opportunities is a secondary or tertiary benefit.

CHAPTER FOUR

Conclusions and Recommendations

A confluence of factors, including the increasing availability of information technology resources in schools and a continued interest in preparing students for careers in STEM fields, has created an opportunity to rethink how we monitor students' access to the instructional opportunities that will position them well for rewarding postsecondary opportunities in STEM. This report is the result of an initiative to develop broad indicators of STEM education that could be adopted in the near future; yet, as we pursued that objective, we also identified several approaches to measurement that are not yet ready for deployment but that have potential for improving our understanding of K–12 coverage of STEM content and practices (i.e., indicator 5) in future systems of measurement. Informed by this work, we offer some broad conclusions and present several recommendations for next steps. We also discuss some challenges that will need to be confronted to move this work forward.

Although it might be obvious from the fact that the National Research Council is conducting this initiative to develop indicators related to STEM education, it seems worth noting that none of the current national efforts to collect information on teaching and learning in the STEM fields (described in Chapter Two) provides adequate evidence about students' opportunities to learn STEM content and practices. Resources, such as NAEP, NSSME, and other large-scale surveys, provide useful data to monitor some aspects of curriculum and instruction in the STEM disciplines, but these surveys do not fully capture the kinds of instructional activities (e.g., inquiry and modeling) that our experts argued are essential for preparing students for college and careers in STEM fields. Even the highest-quality survey items often fail to measure the extent to which teachers and students engage in these activities, and they also do not provide detailed data on how students' experiences vary within the same classrooms. Because of these limitations, an effort to develop new approaches to measurement that can be deployed in either the near or long term seems worthwhile.

Conclusions

Surveys Appear to Be the Most Plausible Method for Measuring Exposure to STEM Content and Practices Across the United States at This Time

Despite their limitations, for the immediate future, any national indicator system that is intended to track information about students' STEM learning experiences will almost certainly use surveys to do so because of their widespread use, familiarity, and known characteristics. These surveys can take a variety of forms and could include instructional logs in addition to more-traditional survey items, and they could be administered via paper or online. They are likely to be completed by teachers but could also be administered to students. Among the more novel measurement methods we explored in this report, ESM might be the closest to being ready for inclusion in an indicator system in future years and have the potential to expand our understanding of how individual students' experiences differ within classrooms, but the technical and logistical challenges associated with ESM suggest that it is unlikely that this method will be widely adopted in the near term.

Nevertheless, traditional surveys could support more-detailed data than are reported at present. Once the data are collected, the choices regarding how to analyze and present the data are numerous. One possible approach would be to identify a set of key practices for a given grade level or range, develop items to measure the amount of instructional time devoted to those practices (which could include the number of lessons and the number of minutes per lesson that focused on those practices), and gather data via teacher or student surveys to calculate the percentages of students who engaged in the relevant instructional activities at various time points during the year. This information would provide a means to estimate in rough terms how much exposure students have to those practices and how this exposure varies among students and over time. This type of data would naturally have a fair amount of measurement error at the individual student level but could provide reliable information at an aggregate level. It would also likely produce more-nuanced evidence than is currently available using such surveys as those administered through NAEP (that is, single-point-in-time surveys that ask teachers to provide information about practices over the course of a full school year).

Technology-Based Learning Systems, Particularly Simulations, Have the Potential to Support Future STEM Indicator Measurement Efforts

Researchers and educators are involved in a wide variety of efforts to harness technology to improve instruction, and many of these efforts have the potential to support better measurement of students' opportunities to engage with specific STEM content and practices. Many of these systems collect extensive data that could be used to document what students are doing and how much time they spend doing it. These data could be analyzed and reported in a way that indicates the kinds of STEM practices in which students engage and how much of their classroom time is devoted to those practices. Additional work will be needed to assess the feasibility of these approaches and to document their reliability, validity, and fairness for the purpose of monitoring students' STEM learning experiences.

The Use of Technology-Based Learning Systems for Large-Scale Measurement Is Likely to Be Limited Because of Variability Across Schools in the Computer-Based Tools and Other Instructional Materials That Are Adopted and Used

Despite the promise of many of the methods described in Chapter Three, one barrier that will almost certainly hinder their use in an indicator system is the diversity of products used across schools and the lack of consistency in how data are gathered and made accessible. Software providers do not all design their log files in a way that facilitates use of data for instructional decisionmaking or for monitoring students' experiences, and it would be unreasonable to expect uniformity in these features across products and providers. District-wide implementation of specific technology-based curricula or learning management systems might give central office leaders an opportunity to create measures of content coverage or student participation that could be deployed district-wide, and these systems might be even more useful at the school and classroom levels, at which teachers and school leaders can extract customized data to inform their day-to-day decisionmaking.

It is not just software that varies. The decentralized nature of the U.S. educational system has resulted in a wide variety of curriculum materials in use across schools. Even when districts mandate a common curriculum, many teachers supplement with materials they develop themselves or find via such resources as web-based lesson plan banks or even general-interest sites, such as Pinterest (Opfer, Kaufman, and Thompson, 2016). And many teachers use a combination of software-based tools and more-traditional materials, such as textbooks. The activities in which students engage in STEM classes are inevitably shaped by the materials available to them. This poses challenges for developing detailed measures because the materials are unique to a given teacher or classroom. The variation in materials means that content might be organized differently, topics might be connected differently, and students might perform different experimental activities, all of which present challenges to designing common surveys or data-collection efforts. To the extent that new standards, such as the Common Core State Standards for math and the NGSS, become more widely adopted and teachers begin enacting the standards in common ways, this challenge might diminish over time.

STEM Practices, Such as Those Identified in the Common Core State Standards for Mathematics and the Next Generation Science Standards, Are Enacted Differently in Different Content Areas

Although an indicator system might be designed to measure students' exposure to broad, discipline-based practices, such as scientific inquiry, engineering design, or mathematical modeling, students' engagement in these practices occurs in the context of a specific content area (such as earth science or biological science), and efforts to measure the practices without taking the content into consideration are likely to produce results that are incomplete at best and misleading at worst. A useful indicator of students' participation in activities related to scientific inquiry, for example, would need to examine inquiry activities across a range of science content areas, an approach that could be feasible within a large-scale indicator system but would require careful attention to sampling.

Some Opportunities to Engage in STEM Content and Practices Occur Outside of Traditional Courses

Regardless of whether the indicator relies on old-fashioned surveys or novel analysis of software log data, one overarching question is which settings should be the focus of data collection. Students participate in STEM-related instructional activities in their mathematics and science courses, but many schools also offer courses in engineering, robotics, and other related topics, and they also provide out-of-school opportunities for students to engage in STEM activities. It is likely that, for many students, participation in inquiry, engineering design, and other activities that promote sophisticated understanding of STEM will be more likely to occur in these settings outside mathematics and science courses. To include these settings in a national indicator system, developers will need to devise a method for identifying the relevant coursework and other opportunities in light of the enormous diversity across schools in naming conventions and data availability. The indicator system might need to rely primarily on traditionally named courses at first, but a thorough understanding of how and the extent to which students are participating in STEM instruction will require a more comprehensive approach.

Recommendations for Developers and Users of Indicators of Students' Classroom Experiences in STEM

Although it is not currently feasible to develop a system that would provide comprehensive evidence to inform indicator 5, our review does offer guidance to inform R&D efforts that could eventually lead to usable measures that could be incorporated into an indicator system.

Create a Working Group to Inform Indicator Development

A large number of groups have some stake in any national indicator system like the one described by the National Research Council. Policymakers, funders, and the general public need high-quality information to inform decisionmaking; educators at all levels also need good information to help them adjust practice; educators and students need assurance that the system will not be overly burdensome, that individual data will be protected, and that consequences associated with the system are reasonable and appropriate. One approach to gathering stakeholder input would be to form a working group that would jointly develop a plan for data collection and would monitor the system over time, adjusting it as needed in response to feedback from the field and evidence regarding the quality and utility of information it produces.

This group's decisions would need to be informed by input from experts in instruction, measurement, statistics, and other fields. There are clear trade-offs between cost and complexity. Trade-offs also exist between uniformity and comparability on the one hand and the rich information that can be obtained by examining specific, unique instructional events on the other; and incorporating a variety of voices into these decisions could contribute to a more robust and useful system.

Use Multiple Measures to Collect Evidence Related to Indicator 5

Although we have concluded that large-scale data collection for an indicator of students' exposure to STEM content and practices will probably rely on surveys for the near future at least, some of the novel approaches described in Chapter Three could provide an opportunity to gather supplemental data in a small number of jurisdictions using more-sophisticated (and more-expensive) methods. Doing so could enable developers of new measures to gather evidence of reliability and validity while also helping to advance the broader educational field's understanding of what STEM learning experiences look like in practice. Eventually, these new approaches might be incorporated in some form into an indicator system, which would provide a more comprehensive picture of what is happening with STEM education nationally and help to inform future policy and practice.

Our reviews and interactions with expert panelists identified several potential approaches. One suggested coupling software log data or event sampling methodology data with survey questions to teachers and students about what they are doing and the extent of students' engagement. Bringing student- and teacher-collected data together should provide a fuller picture than relying on either source alone. Software logs, for instance, can tell us something about what is happening in class (e.g., are students making graphs?) but not necessarily what the teacher is doing pedagogically.

In the short term, although these more-sophisticated methods are still in development, thoughtful design and use of both teacher and student surveys could contribute to a comprehensive, multiple-measure approach to collecting data on indicator 5. Survey questions would not need to capture identical information but should be designed to take advantage of each group's unique perspectives while providing complementary evidence related to a common set of topics. To the extent that students have different learning experiences, an indicator of STEM content and practices should describe variation within a classroom rather than just the median or typical experience.

Begin by Building on Existing Data-Collection Tools and Systems

Because the system will probably need to rely on surveys in the near future, developers should draw on existing resources, such as the surveys developed for NSSME, to find measures of practices consistent with the NGSS and the Common Core State Standards for math (discussed in Chapter Two). Developers should also start developing new survey items and logs geared specifically toward underrepresented domains. Similarly, developers should look to existing national data-collection systems as potential carriers of new items. As part of an earlier NSF study that explored the feasibility of a national indicator system for monitoring science and mathematics education (Shavelson et al., 1987), researchers recommended that NSF begin by undertaking three activities:

> (1) initiation of efforts to develop better indicators of science and mathematics education using existing data, i.e., building a patchwork system; (2) initiation of studies to develop better measures in areas where current measures are non-existent or inadequate; and (3) sponsoring of exploratory studies of how indicator data might be made more useful to policymakers. (p. 52)

Those recommendations seem equally appropriate today when it comes to an indicator of students' opportunities to learn STEM content and practices. A hybrid approach would mean negotiating agreements to embed a few new items into existing regular data-collection efforts, such as NAEP and NSSME, while conducting research on the development of more-sensitive measures, such as those discussed in this report. Their third recommendation remains valid as well; it is essential to think about indicator use when making plans for indicator development and reporting.

Design the System to Support Longitudinal Comparisons

One important benefit of a national indicator system is its ability to enable policymakers, educators, and members of the public to track experiences over time. This information can help funders and policymakers track the outcomes of investments in STEM education and can be useful for informing future priorities. Measuring change could be especially informative in light of the heavy demands placed on many teachers to improve their skills and change their practices in response to the rigorous expectations embodied in the Common Core State Standards for math and the NGSS. Monitoring the extent to which classroom activities change over time could provide evidence regarding how teachers are modifying their practices and how these changes align with the ambitious learning goals in the standards. One of the experts suggested a strategy that might be tried to measure teachers' incorporation of teaching and learning tools: Provide new tools to a set of teachers in a given year and track their use of the tools in subsequent years to see whether they grow more sophisticated as they learn the capabilities of the tools.

The system could produce valuable longitudinal information even if it does not track the same teachers or students over time. However, as a supplement to the national system, it might be valuable to collect longitudinal data for a subset of teachers using some of the more-sophisticated measurement methods. These tools could be implemented in a sample of classrooms across multiple years to provide richer information on instructional change than is available via surveys.

Consider Incorporating Measures of Student Knowledge into the Broader Indicator System

Although the National Research Council indicator effort does not focus on measures of student achievement, ideally, data from a STEM indicator system could be linked to achievement data to provide new evidence regarding the relationships between classroom practices and student learning. In particular, measures of student achievement that capture the more-sophisticated skills, such as inquiry and design, could provide valuable information to help validate novel measures of STEM learning experiences. One of our expert panelists argued that high-quality achievement data could actually be useful for understanding the types of classroom activities to which students were exposed; e.g., if students demonstrate the ability to engage in a specific problem-solving approach, these data could perhaps be used to infer that students were exposed to that approach during their instruction. Such inferences would need to be carefully considered and, to the extent possible, validated with other data, but the general point is that evidence of student learning should be incorporated into the system in some way to provide the most-comprehensive information possible on STEM learning experiences.

Avoid Attaching Stakes to the Indicator

Extensive research on high-stakes measurement in education and other fields reveals a high likelihood of undesirable consequences resulting from high stakes. In education, these outcomes include a narrowing of curriculum and instruction to emphasize what is measured (and how it is measured) and reduced attention paid to other material, and this narrowing can lead to inflation of scores (see Hamilton, Stecher, and Yuan, 2012, for a review). This is not to suggest that measures should never have high stakes, but, if the central purpose of the indicator system is to monitor what is happening rather than to induce specific changes, it will be important to refrain from linking the indicator system to specific consequences for schools, educators, or students. Lack of stakes is especially crucial for newer measures that are not yet well tested and for which it is difficult to predict all of the ways in which practices could be negatively influenced or data corrupted. In addition, lack of stakes will increase the probability that teachers will feel safe sharing the details of their practice and allowing new measurement methods to be deployed in their classrooms.

To Inform Future Measurement Efforts, Continue to Conduct Research on STEM Teaching and Learning

Although the experts and the literature suggest that such strategies as project-based learning or simulated investigations conducted in rich, computer-based environments can engage students in STEM practices (e.g., “model[ing] with mathematics” or “planning and carrying out investigations”), the field lacks definitive evidence regarding exactly what kinds of projects or simulated investigations foster which practices. Educators and researchers are working to develop curriculum materials and instructional guidelines designed to promote the goals of the Common Core State Standards for math and the NGSS. For example, *Taking Science to School* (Duschl, Schweingruber, and Shouse, 2007) describes the kinds of activities that help students develop scientific practices associated with specific disciplinary core ideas and crosscutting concepts. Although this guidance suggests the kinds of instructional behaviors that are consistent with the goals of the NGSS, the descriptions are still too general and abstract to serve as the bases for teacher surveys or observational protocols. Making valid judgments about specific instructional choices requires a finer level of detail that describes more-explicit actions on the part of teachers and students. Efforts like these to improve curriculum and instruction in the STEM fields help to identify the kinds of classroom activities that are desirable. However, more work in this area is needed to move from general descriptions of instructional goals or classes of behaviors to the level of description needed as a basis for an indicator.

Final Thoughts

One of the important things we learned during the course of this study is the potential benefit of including measures of opportunity to learn STEM content and practices in a national indicator system. By signaling to educators the kinds of activities that are aligned with standards and by broadening the range of learning experiences that are measured, an indicator can help to inform STEM educational policy and practice far more directly than measures of student performance that are typically used for monitoring and accountability. We are cognizant of the testing burden that current accountability systems are placing on students and schools, and we do not recommend lightly the addition of new data-collection efforts. Nevertheless, the potential value of direct measures of students' classroom experiences more than justifies their inclusion in an indicator system. In fact, we might value them above some of the extended measures of knowledge that are currently being used, especially in light of their potential to help policymakers and educators understand, and therefore improve, teaching. We also understand that an indicator system is, of necessity, a "lean" set of measures, focusing on a few choice variables, gathered from a sample of classrooms, so the results will have to be interpreted with caution. Although there is some risk of overinterpretation initially, we expect that this will be overcome with careful presentation and familiarity over time. The primary conclusions from this work are that one could start collecting measures of STEM learning opportunities relatively easily in the short term, conduct research to improve them in the intermediate term, and broaden their scope using more-sophisticated methods in the long term.

Bibliography

Affective Computing, "Research on Sensing Human Affect," undated. As of February 14, 2017:
http://affect.media.mit.edu/areas.php?id=sensing

American Association for the Advancement of Science, *Science for All Americans*, Washington, D.C., 1989. As of December 22, 2016:
http://www.project2061.org/publications/sfaa/online/sfaatoc.htm

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 2014.

Atlantis Remixed, "Educators," undated. As of February 14, 2017:
http://atlantisremixed.org/site/view/Educators

Ball, Deborah Loewenberg, Eric Camburn, Richard Correnti, Geoffrey Phelps, and Raven Wallace, *New Tools for Research on Instruction and Instructional Policy: A Web-Based Teacher Log*, Seattle, Wash.: University of Washington, Center for Teaching and Policy working paper, December 1999.

Banilower, Eric R., *A Study of the Predictive Validity of the LSC Classroom Observation Protocol*, Chapel Hill, N.C.: Horizon Research, April 2005. As of February 11, 2017:
http://www.horizon-research.com/a-study-of-the-predictive-validity-of-the-lsc-classroom-observation-protocol/

Banilower, Eric R., P. Sean Smith, Iris R. Weiss, Kristen A. Malzahn, Kiira M. Campbell, and Aaron M. Weis, *Report of the 2012 National Survey of Science and Mathematics Education*, Chapel Hill, N.C.: Horizon Research, February 2013. As of February 11, 2017:
http://www.horizon-research.com/2012nssme/research-products/reports/technical-report/

BetterHelp, home page, undated. As of February 13, 2017:
http://betterlesson.com/home

Blank, Rolf K., *Surveys of Enacted Curriculum: Tools and Services to Assist Educators*, Washington, D.C.: Council of Chief State School Officers, 2005. As of February 11, 2017:
https://eric.ed.gov/?id=ED484708

Blank, Rolf K., Andrew Porter, and John Smithson, *New Tools for Analyzing Teaching, Curriculum and Standards in Mathematics and Science*: *Results from Survey of Enacted Curriculum Project—Final Report*, Washington, D.C.: Council of Chief State School Officers, July 2001. As of April 20, 2016:
http://jsmithson.wceruw.org/reference/secnewtoolsreport01.pdf

Borko, Hilda, Brian M. Stecher, and Karin Kuffner, *Using Artifacts to Characterize Reform-Oriented Instruction: The Scoop Notebook and Rating Guide*, Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate School of Education and Information Studies, University of California, Los Angeles, Center for the Study of Evaluation Technical Report 707, February 2007. As of February 11, 2017:
https://eric.ed.gov/?id=ED495853

Brandon, Paul R., and Alice K. H. Taum, *Development and Validation of the Inquiry Science Teacher Log and the Inquiry Science Teacher Questionnaire*, presented at the annual meeting of the American Evaluation Association, Toronto, October 2005. As of February 11, 2017:
http://manoa.hawaii.edu/crdg/wp-content/uploads/AEA_05_PB__AT.pdf

Brophy, Jere E., *Teacher Behavior and Its Effects*, East Lansing, Mich.: Institute for Research on Teaching, Michigan State University, Occasional Paper 25, 1979.

Brophy, Jere E., and Carolyn M. Evertson, *Learning from Teaching: A Developmental Perspective*, Boston, Mass.: Allyn and Bacon, 1976.

Burstein, Leigh, Lorraine M. McDonnell, Jeannette Van Winkle, Tor Ormseth, Jim Mirocha, and Gretchen Guiton, *Validating National Curriculum Indicators*, Santa Monica, Calif.: RAND Corporation, MR-658-NSF, 1995. As of February 11, 2017:
http://www.rand.org/pubs/monograph_reports/MR658.html

Camburn, Eric, and Carol A. Barnes, "Assessing the Validity of a Language Arts Instruction Log Through Triangulation," *Elementary School Journal*, Vol. 105, No. 1, September 2004, pp. 49–74.

Carnegie Learning, "Cognitive Tutor for Students Grades 9–12," undated.

Casabianca, Jodi M., J. R. Lockwood, and Daniel F. McCaffrey, "Trends in Classroom Observation Scores," *Educational and Psychological Measurement*, Vol. 75, No. 2, 2015, pp. 311–337.

Cash, Anne H., Bridget K. Hamre, Robert C. Pianta, and Sonya S. Myers, "Rater Calibration When Observational Assessment Occurs at Large Scale: Degree of Calibration and Characteristics of Raters Associated with Calibration," *Early Childhood Research Quarterly*, Vol. 27, No. 3, 3rd Quarter 2012, pp. 529–542.

Centervention, "Elementary School," undated. As of February 14, 2017:
https://www.centervention.com/zoo-u/

Chen, Gaowei, Sherice N. Clarke, and Lauren B. Resnick, "Classroom Discourse Analyzer (CDA): A Discourse Analytic Tool for Teachers," *Technology, Instruction, Cognition and Learning*, Vol. 10, No. 2, April 2015, pp. 85–105.

Clark, Douglas B., Emily E. Tanner-Smith, and Stephen S. Killingsworth, "Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis," *Review of Educational Research*, Vol. 86, No. 1, March 2016, pp. 79–122.

Clarke, Sherice N., Gaowei Chen, Donna D. Bickel, Jennifer Z. Sherer, and Lauren B. Resnick, "Through the Looking Glass: Using a Classroom Discourse Visualizer to Support Teacher Reflection on Practice," paper presented for the 11th International Conference on Computer Supported Collaborative Learning, Gothenburg, Sweden, June 7–11, 2015.

Connecticut State Department of Education, *A Guide to the BEST Program for Beginning Teachers: 2007–2008*, Hartford, Conn., 2007.

Csikszentmihalyi, Mihaly, and Reed Larson, "Validity and Reliability of the Experience-Sampling Method," *Journal of Nervous and Mental Diseases*, Vol. 175, No. 9, September 1987, pp. 526–536.

Danielson, Charlotte, *Enhancing Professional Practice: A Framework for Teaching*, Alexandria, Va.: Association for Supervision and Curriculum Development, 2007.

DeVellis, Robert F., *Scale Development: Theory and Applications*, Newbury Park, Calif.: Sage, 1991.

Discovery Education, home page, undated. As of February 14, 2017:
http://www.discoveryeducation.com/

Duschl, Richard A., Heidi A. Schweingruber, and Andrew W. Shouse, eds., *Taking Science to School: Learning and Teaching Science in Grades K–8*, Washington, D.C.: National Academies Press, 2007. As of February 12, 2017:
https://www.nap.edu/catalog/11625/taking-science-to-school-learning-and-teaching-science-in-grades

EcoLearn, "ecoMUVE," undated. As of February 14, 2017:
http://ecolearn.gse.harvard.edu/ecoMUVE/overview.php

edTPA, "About edTPA," undated. As of February 14, 2017:
http://www.edtpa.com/PageView.aspx?f=GEN_AboutEdTPA.html

Educade, "Newton's Playground," undated. As of February 14, 2017:
http://www.educade.org/teaching_tools/newtons-playground

ExploreLearning, home page, undated. As of February 14, 2017:
https://www.explorelearning.com/

Fisher, Charles W., Nikola N. Filby, Richard S. Marliave, Leonard S. Cahen, Marilyn M. Dishaw, Jeffrey E. Moore, and David C. Berliner, *Teaching Behaviors, Academic Learning Time and Student Achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study*, San Francisco, Calif.: Far West Laboratory for Educational Research and Development, Technical Report V-1, 1978.

Floden, Robert E., "The Measurement of Opportunity to Learn," in Andrew C. Porter and Adam Gamoran, eds., *Methodological Advances in Cross-National Surveys of Educational Achievement*, Washington, D.C.: National Research Council, Division of Behavioral and Social Sciences and Education, Board on International Comparative Studies in Education, Board on Testing and Assessment, 2002, pp. 229–266. As of February 11, 2017:
https://www.nap.edu/catalog/10322/
methodological-advances-in-cross-national-surveys-of-educational-achievement

Garrett, Rachel, and Matthew P. Steinberg, "Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence from the Randomization of Teachers to Students," *Educational Evaluation and Policy Analysis*, Vol. 37, No. 2, 2015, pp. 224–242.

Gerard, Libby F., Michele Spitulnik, and Marcia C. Linn, "Teacher Use of Evidence to Customize Inquiry Science Instruction," *Journal of Research in Science Teaching*, Vol. 47, No. 9, November 2010, pp. 1037–1063.

Gerard, Libby F., Keisha Varma, Stephanie B. Corliss, and Marcia C. Linn, "Professional Development for Technology-Enhanced Inquiry Science," *Review of Educational Research*, Vol. 81, No. 3, 2011, pp. 408–448.

Gitomer, Drew, Courtney Bell, Yi Qi, Daniel McCaffrey, Bridget K. Hamre, and Robert C. Pianta, "The Instructional Challenge in Improving Teaching Quality: Lessons from a Classroom Observation Protocol," *Teachers College Record*, Vol. 116, No. 6, 2014, pp. 1–32.

GlassLab Services, "Shout Out! Watch Out! What Now?" undated.

Gobert, Janice D., Michael Sao Pedro, M., Juelaila Raziuddin, and Ryan S. Baker, "From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining," *Journal of the Learning Sciences*, Vol. 22, No. 4, 2013, pp. 521–563.

Goe, Laura, Courtney Bell, and Olivia Little, *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*, Washington, D.C.: National Comprehensive Center for Teacher Quality, 2008.

Good, Thomas L., and Douglas A. Grouws, "Teaching Effects: A Process–Product Study in Fourth Grade Mathematics Classrooms," *Journal of Teacher Education*, Vol. 28, No. 3, May–June 1977, pp. 49–54.

———, "The Missouri Mathematics Effectiveness Project: An Experimental Study in Fourth-Grade Classrooms," *Journal of Educational Psychology*, Vol. 71, No. 3, June 1979, pp. 355–362.

Grimm, Pamela, "Social Desirability Bias," in *Wiley International Encyclopedia of Marketing*, Part 2: *Marketing Research*, Hoboken, N.J.: John Wiley and Sons, 2010.

Hamilton, Laura S., Brian M. Stecher, and Kun Yuan, "Standards-Based Accountability in the United States: Lessons Learned and Future Directions," *Education Inquiry*, Vol. 3, No. 2, June 2012, pp. 149–170.

Harris, Douglas N., and Tim R. Sass, *Teacher Training, Teacher Quality, and Student Achievement*, Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, Urban Institute, Working Paper 3, March 2007. As of February 11, 2017:
https://eric.ed.gov/?id=ED509656

Haystead, Mark W., *RISC vs. Non-RISC Schools: A Comparison of Student Proficiencies for Reading, Writing, and Mathematics*, Bloomington, Ind.: Marzano Research Laboratory, April 2010.

Heitin, Liana, "Updated Map: Which States Have Adopted the Next Generation Science Standards?" *Curriculum Matters*, August 31, 2015. As of February 13, 2017:
http://blogs.edweek.org/edweek/curriculum/2015/08/
updated_map_which_states_have_adopted_the_next_generation_science_standards.html

Hektner, Joel M., Jennifer A. Schmidt, and Mihaly Csikszentmihalyi, *Experience Sampling Method: Measuring the Quality of Everyday Life*, Thousand Oaks, Calif.: Sage Publications, 2007.

Henry, Martha A., Keith S. Murray, and Katherine A. Phillips, *Meeting the Challenge of STEM Classroom Observation in Evaluating Teacher Development Projects: A Comparison of Two Widely Used Instruments*, Fairfax, Va.: M. A. Henry Consulting, 2007.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball, "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement," *American Educational Research Journal*, Vol. 42, No. 2, 2005, pp. 371–406.

Hill, Heather C., Stephen G. Schilling, and Deborah Loewenberg Ball, "Developing Measures of Teachers' Mathematics Knowledge for Teaching," *Elementary School Journal*, Vol. 105, No. 1, September 2004, pp. 11–30.

Honey, Margaret, Greg Pearson, and Heidi Schweingruber, eds., *STEM Integration in K–12 Education: Status, Prospects, and an Agenda for Research*, Washington, D.C.: National Academy of Engineering, National Research Council, Committee on Integrated STEM Education, 2014. As of February 11, 2017:
https://www.nap.edu/catalog/18612/stem-integration-in-k-12-education-status-prospects-and-an

Horizon Research, Inc., *Inside the Classroom Interview and Analytic Protocol*, Chapel Hill, N.C., 2000. As of February 20, 2017:
http://www.horizon-research.com/inside-the-classroom-observation-and-analytic-protocol/

Horn, Michael B., and Heather Staker, *The Rise of K–12 Blended Learning*, Lexington, Mass.: Innosight Institute, January 2011. As of February 11, 2017:
http://www.innosightinstitute.org/innosight/wp-content/uploads/2011/01/The-Rise-of-K-12-Blended-Learning.pdf

HRI—*See* Horizon Research, Inc.

Husén, Torsten, ed., *International Study of Achievement in Mathematics: A Comparison of Twelve Countries*, Vol. I, New York: Wiley, 1967.

iCivics, "Argument Wars," undated. As of February 14, 2017:
https://www.icivics.org/games/argument-wars

Jacob, Brian A., and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, Vol. 26, No. 1, January 2008, pp. 101–136.

Jacovina, Matthew E., and Danielle S. McNamara, "Intelligent Tutoring Systems for Literacy: Existing Technologies and Continuing Challenges," in Robert Kenneth Atkinson, ed., *Intelligent Tutoring Systems: Structure, Applications and Challenges*, Hauppauge, N.Y.: Nova Science Publishers, 2016, pp. 153–174.

Jitendra, Asha K., Cynthia Griffin, Andria Deatline-Buchman, Caroline Dipipi-Hoy, Edward Sczesniak, Natalie G. Sokol, and Yan Ping Xin, "Adherence to Mathematics Professional Standards and Instructional Design Criteria for Problem-Solving in Mathematics," *Exceptional Children*, Vol. 71, No. 3, 2005, pp. 319–337.

Johnson, Robert L., Fred McDaniel II, and Marjorie J. Willeke, "Using Portfolios in Program Evaluation: An Investigation of Interrater Reliability," *American Journal of Evaluation*, Vol. 21, No. 1, Winter 2000, pp. 65–80.

Junker, Brian, Yanna Weisberg, Lindsay Clare Matsumura, Amy Crosson, Mikyung Kim Wolf, Allison Levison, and Lauren Resnick, *Overview of the Instructional Quality Assessment*, Los Angeles, Calif.: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Center for the Study of Evaluation Technical Report 671, 2006. As of February 12, 2017:
https://cxarchive.gseis.ucla.edu/xchange/multiple-measures-of-good-teaching/xpress/overview-of-the-instructional-quality-assessment

Kane, Thomas J., Kerri A. Kerr, and Robert C. Pianta, eds., *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, San Francisco, Calif.: Jossey-Bass, 2014. As of April 21, 2016:
http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Kane, Thomas J., and Douglas O. Staiger, *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*, Seattle, Wash.: Bill and Melinda Gates Foundation, Measures of Effective Teaching Project, Research Paper, January 2012. As of February 12, 2017: https://eric.ed.gov/?id=ED540960

Kaufman, Julia H., Mary Key Stein, and Brian W. Junker, "Factors Associated with Alignment Between Teacher Survey Reports and Classroom Observation Ratings of Mathematics Instruction," *Elementary School Journal*, Vol. 116, No. 3, March 2016, pp. 339–364.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon, "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research," *American Political Science Review*, Vol. 98, No. 1, February 2004, pp. 191–207. As of April 21, 2016: https://dash.harvard.edu/bitstream/handle/1/3965182/King_EnhancingtheValidity.pdf

Koretz, Daniel, Brian Stecher, Stephen Klein, and Daniel McCaffrey, "The Vermont Portfolio Assessment Program: Findings and Implications," *Educational Measurement: Issues and Practice*, Vol. 13, No. 3, September 1994, pp. 5–16.

Korn, Shira, Martin Gamboa, and Morgan Polikoff, "Just How Common Are the Standards in Common Core States?" Center on Standards, Alignment, Instruction, and Learning, November 3, 2016. As of February 13, 2017:
http://c-sail.org/resources/blog/just-how-common-are-standards-common-core-states

Kroll, Luisa, "Gates Foundation Responds to GSR Bracelets Controversy," *Forbes*, June 13, 2012. As of February 14, 2017:
http://www.forbes.com/sites/luisakroll/2012/06/13/gates-foundation-responds-to-gsr-bracelets-controversy/#1e1b5a2f526f

Le, Vi-Nhuan, Brian M. Stecher, Laura S. Hamilton, Gery W. Ryan, Valerie L. Williams, Abby Robyn, and Alicia Alonzo, *Vignette-Based Surveys and the Mosaic II Project*, Santa Monica, Calif.: RAND Corporation, WR-165-EDU, 2004. As of February 12, 2017:
http://www.rand.org/pubs/working_papers/WR165.html

Learning Mathematics for Teaching Project, "Measuring the Mathematical Quality of Instruction," *Journal of Mathematics Teacher Education*, Vol. 14, No. 1, February 2011, pp. 25–47.

Lee, Okhee, Randall Penfield, and Jaime Maerten-Rivera, "Effects of Fidelity of Implementation on Science Achievement Gains Among English Language Learners," *Journal of Research in Science Teaching*, Vol. 46, No. 7, September 2009, pp. 836–859.

Levine, Roger, Mette Huberman, and Kathryn Buckner, *The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Grade Students and Teachers to Questionnaire Items*, Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education, Working Paper 2002-06, August 2002. As of April 21, 2016:
http://nces.ed.gov/pubs2002/200206.pdf

Mangold International, home page, undated. As of February 14, 2017:
https://www.mangold-international.com/

Martínez, José Felipe, Hilda Borko, and Brian M. Stecher, "Measuring Instructional Practice in Science Using Classroom Artifacts: Lessons Learned from Two Validation Studies," *Journal of Research in Science Teaching*, Vol. 49, No. 1, January 2012, pp. 38–67.

Martínez, José Felipe, Hilda Borko, Brian Stecher, Rebecca Luskin, and Matt Kloser, "Measuring Classroom Assessment Practice Using Instructional Artifacts: A Validation Study of the QAS Notebook," *Educational Assessment*, Vol. 17, No. 2–3, 2012, pp. 107–131.

Martínez, José Felipe, Matthew J. Kloser, Jayashri Srinivasan, Kate Riedell, Brian Stecher, Rose Rocchio, Matthew Wilsey, and Hongsuda Tangmunarunkit, *Next-Generation Tablet-Based Electronic Teacher Portfolio for Measuring and Reflecting on Next Generation Science Standards Science Instruction*, paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 10, 2016.

Matsumura, Lindsay Clare, Helen Garnier, Jenny Pascal, and Rosa Valdés, "Measuring Instructional Quality in Accountability Systems: Classroom Assignments and Student Achievement," *Educational Assessment*, Vol. 8, No. 3, 2002, pp. 207–229.

Matsumura, Lindsay Clare, Helen E. Garnier, Sharon Cadman Slater, and Melissa D. Boston, "Toward Measuring Instructional Interactions 'At-Scale,'" *Educational Assessment*, Vol. 13, No. 4, 2008, pp. 267–300.

Matsumura, Lindsay Clare, Sharon Cadman Slater, Brian Junker, Maureen Peterson, Melissa Boston, Michael Steele, and Lauren Resnick, *Measuring Reading Comprehension and Mathematics Instruction in Urban Middle Schools: A Pilot Study of the Instructional Quality Assessment*, Los Angeles, Calif.: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles, Center for the Study of Evaluation Technical Report 681, May 2006. As of February 12, 2017:
https://eric.ed.gov/?id=ED492885

Mayer, Daniel P., "Measuring Instructional Practice: Can Policymakers Trust Survey Data?" *Educational Evaluation and Policy Analysis*, Vol. 21, No. 1, Spring 1999, pp. 29–45.

McCaffrey, Daniel F., Kun Yuan, Terrance D. Savitsky, J. R. Lockwood, and Maria O. Edelen, "Uncovering Multivariate Structure in Classroom Observations in the Presence of Rater Errors," *Educational Measurement: Issues and Practice*, Vol. 34, No. 2, Summer 2015, pp. 34–46.

McDonnell, Lorraine M., "Opportunity to Learn as a Research Concept and a Policy Instrument," *Educational Evaluation and Policy Analysis*, Vol. 17, No. 3, 1995, pp. 305–322.

Medley, Donald M., and Homer Coker, "The Accuracy of Principals' Judgments of Teacher Performance," *Journal of Educational Research*, Vol. 80, No. 4, March–April 1987, pp. 242–247.

Myford, Carol M., "Rater Cognition Research: Some Possible Directions for the Future," *Educational Measurement: Issues and Practice*, Vol. 31, No. 3, Fall 2012, pp. 48–49.

MyiLOGS, home page, undated. As of February 13, 2017:
https://www.myilogs.com/rtdev/myilogs/public/index.php

NAEP—*See* National Assessment of Educational Progress.

National Assessment Governing Board, *Technology and Engineering Literacy Framework for the 2014 National Assessment of Educational Progress*, 2014. As of February 12, 2017:
https://www.nagb.org/publications/frameworks/technology/2014-technology-framework.html

National Assessment of Educational Progress, *Mathematics Teacher Questionnaire: 2017 Grade 8*, undated. As of February 21, 2017:
https://nces.ed.gov/nationsreportcard/subject/about/pdf/bgq/teacher/2017_sq_teacher_math_g8.pdf

National Board for Professional Teaching Standards, *What Teachers Should Know and Be Able to Do*, Arlington, Va., August 2002. As of April 21, 2016:
http://www.nbpts.org/sites/default/files/what_teachers_should_know.pdf

National Center for Education Statistics, *Science in Action: Hands-On and Interactive Computer Tasks from the 2009 Science Assessment*, Washington, D.C.: U.S. Department of Education, June 2012. As of February 12, 2017:
https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012468

———, *Trends in International Mathematics and Science Study Student Questionnaire, Grade 8*, Washington, D.C.: International Association for the Evaluation of Educational Achievement, 2014a. As of April 21, 2016:
https://nces.ed.gov/timss/pdf/2015_8th_grade_Student_Questionnaire.pdf

———, *Trends in International Mathematics and Science Study Teacher Questionnaire, Grade 4*, Washington, D.C.: International Association for the Evaluation of Educational Achievement, 2014b. As of February 13, 2017:
https://nces.ed.gov/timss/pdf/2015_4th_grade_Teacher_Questionnaire.pdf

———, *Trends in International Mathematics and Science Study Teacher Questionnaire, Science, Grade 8*, Washington, D.C.: International Association for the Evaluation of Educational Achievement, 2014c. As of April 21, 2016:
https://nces.ed.gov/timss/pdf/2015_8th_grade_Teacher_Questionnaire_Science.pdf

———, *Mathematics Teacher Questionnaire 2015, Grade 8*, Washington, D.C.: U.S. Department of Education, c. 2015a. As of April 21, 2016:
https://nces.ed.gov/nationsreportcard/subject/about/pdf/bgq/teacher/2015_bq_teacher_g08_m.pdf

———, *Science Student Questionnaire 2015, Grade 8*, Washington, D.C.: U.S. Department of Education, c. 2015b.

National Council of Teachers of Mathematics, Commission on Standards for School Mathematics, *Curriculum and Evaluation Standards for School Mathematics*, Reston, Va.: National Council of Teachers of Mathematics, 1989.

———, *Principles and Standards for School Mathematics*, Reston, Va.: National Council of Teachers of Mathematics, 2000.

National Research Council, Division of Behavioral and Social Sciences and Education, Board on Science Education, National Committee on Science Education Standards and Assessment, *National Science Education Standards*, Washington, D.C.: National Academies Press, 1996. As of February 12, 2017:
https://www.nap.edu/catalog/4962/national-science-education-standards

———, Division of Behavioral and Social Sciences and Education, Board on Science Education, Board on Testing and Assessment, Committee on Highly Successful Schools or Programs for K–12 STEM Education, *Successful K–12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*, Washington, D.C.: National Academies Press, 2011. As of April 29, 2016:
http://www.nap.edu/catalog/13158/successful-k-12-stem-education-identifying-effective-approaches-in-science

———, Division of Behavioral and Social Sciences and Education, Board on Science Education, Committee on a Conceptual Framework for New K-12 Science Education Standards, *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, Washington, D.C.: National Academies Press, 2012. As of February 12, 2017:
https://www.nap.edu/catalog/13165/a-framework-for-k-12-science-education-practices-crosscutting-concepts

———, Division of Behavioral and Social Sciences and Education, Board on Science Education, Board on Testing and Assessment, Committee on the Evaluation Framework for Successful K–12 STEM Education, *Monitoring Progress Toward Successful K–12 STEM Education: A Nation Advancing?* Washington, D.C.: National Academies Press, 2013. As of February 12, 2017:
https://www.nap.edu/catalog/13509/monitoring-progress-toward-successful-k-12-stem-education-a-nation

NCES—*See* National Center for Education Statistics.

Newmann, Fred M., Anthony S. Bryk, and Jenny Nagaoka, *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago, Ill.: Consortium on School Research, January 2001. As of February 12, 2017:
https://consortium.uchicago.edu/publications/
authentic-intellectual-work-and-standardized-tests-conflict-or-coexistence

Next Generation Learning Challenges, home page, undated. As of February 13, 2017:
http://nextgenlearning.org/

Next Generation Science Standards Lead States, *Next Generation Science Standards: For States, by States*, Washington, D.C.: National Academies Press, 2013. As of February 12, 2017:
https://www.nap.edu/catalog/18290/next-generation-science-standards-for-states-by-states

NGSS Lead States—*See* Next Generation Science Standards Lead States.

OECD—*See* Organisation for Economic Co-operation and Development.

Offer, Shira, and Barbara Schneider, "Revisiting the Gender Gap in Time-Use Patterns: Multitasking and Well-Being Among Mothers and Fathers in Dual-Earner Families," *American Sociological Review*, Vol. 76, No. 6, 2011, pp. 809–833.

Opfer, V. Darleen, Julia H. Kaufman, and Lindsey E. Thompson, *Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy: Findings from the American Teacher Panel*, Santa Monica, Calif.: RAND Corporation, RR-1529-HCT, 2016. As of December 20, 2016:
http://www.rand.org/pubs/research_reports/RR1529.html

Organisation for Economic Co-operation and Development, *OECD Program for International Student Assessment 2012 Parent Questionnaire*, Paris, 2012a.

———, *OECD Program for International Student Assessment 2012 Student Questionnaire*, Paris, 2012b. As of April 21, 2016:
https://nces.ed.gov/surveys/pisa/pdf/MS12_StQ_FormB_ENG_USA_final.pdf

Pane, John F., Elizabeth D. Steiner, Matthew Baird, and Laura S. Hamilton, *Continued Progress: Promising Evidence on Personalized Learning*, Santa Monica, Calif.: RAND Corporation, RR-1365-BMGF, 2015. As of February 12, 2017:
http://www.rand.org/pubs/research_reports/RR1365.html

Pellegrino, James W., Mark R. Wilson, Judith A. Koenig, and Alexandra S. Beatty, eds., *Developing Assessments for the Next Generation Science Standards*, Washington, D.C.: National Academies Press, 2014. As of February 12, 2017:
https://www.nap.edu/catalog/18409/developing-assessments-for-the-next-generation-science-standards

Pianta, Robert C., Bridget K. Hamre, Nancy J. Haynes, Susan L. Mintz, and Karen M. La Paro, *Classroom Assessment Scoring System (CLASS): Secondary Manual*, Charlottesville, Va.: University of Virginia Center for Advanced Study of Teaching and Learning, 2009.

Piburn, Michael, and Daiyo Sawada, *Reformed Teaching Observation Protocol (RTOP): Reference Manual*, Tempe, Ariz.: Arizona State University, Arizona Collaborative for Excellence in the Preparation of Teachers Technical Report IN00-3, 2000. As of February 12, 2017:
http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Biology%20Assessments/RTOP%20Reference%20Manual.pdf

Polikoff, Morgan S., "How Well Aligned Are Textbooks to the Common Core Standards in Mathematics?" *American Educational Research Journal*, Vol. 52, No. 6, 2015, pp. 1185–1211.

Porter, Andrew, "The Uses and Misuses of Opportunity-to-Learn Standards," *Educational Researcher*, Vol. 24, No. 1, January–February 1995, pp. 21–27.

Porter, Andrew C., Michael W. Kirst, Eric J. Osthoff, John L. Smithson, and Steven A. Schneider, *Reform Up Close: An Analysis of High School Mathematics and Science Classrooms*, final report to the National Science Foundation to the Consortium for Policy Research in Education, Madison, Wis.: Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison, October 1993. As of February 12, 2017:
https://eric.ed.gov/?id=ED364429

Priest, Nora, Antonia Rudenstine, and Ephraim Weisstein, *Making Mastery Work: A Close-Up View of Competency Education*, Quincy, Mass.: Nellie Mae Education Foundation, November 2012. As of February 12, 2017:
http://www.competencyworks.org/resources/making-mastery-work/

Quellmalz, Edys S., Michael J. Timms, Matt D. Silberglitt, and Barbara C. Buckley, "Science Assessments for All: Integrating Science Simulations into Balanced State Science Assessment Systems," *Journal of Research in Science Teaching*, Vol. 49, No. 3, March 2012, pp. 363–393.

Remillard, Janine T., "Examining Key Concepts in Research on Teachers' Use of Mathematics Curricula," *Review of Educational Research*, Vol. 75, No. 2, Summer 2005, pp. 211–246.

Reys, Robert, Barbara Reys, James Tarr, and Óscar Chávez, *Assessing the Impact of Standards-Based Middle School Mathematics Curricula on Student Achievement and the Classroom Learning Environment*, Washington, D.C.: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education, March 2006. As of August 4, 2016:
http://mathcurriculumcenter.org/MS2_report.pdf

Ross, John A., Douglas McDougall, Anne Hogaboam-Gray, and Ann LeSage, "A Survey Measuring Elementary Teachers' Implementation of Standards-Based Mathematics Teaching," *Journal for Research in Mathematics Education*, Vol. 34, No. 4, July 2003, pp. 344–363.

Rowan, Brian, Eric Camburn, and Richard Correnti, "Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms," *Elementary School Journal*, Vol. 105, No. 1, September 2004, pp. 75–102.

Rowan, Brian, Delena M. Harrison, and Andrew Hayes, "Using Instructional Logs to Study Mathematics Curriculum and Teaching in the Early Grades," *Elementary School Journal*, Vol. 105, No. 1, September 2004, pp. 103–128.

Rowan, Brian, Robin Jacob, and Richard Correnti, "Using Instructional Logs to Identify Quality in Educational Settings," *New Directions for Youth Development*, Vol. 121, No. 2009, Spring 2009, pp. 13–31.

Ruiz-Primo, Maria Araceli, and Min Li, "Vignettes as an Alternative Teacher Evaluation Instrument: An Exploratory Study," paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., April 1–5, 2002.

Scaffolding Understanding by Redesigning Games for Education, home page, undated (a). As of February 14, 2017:
http://www.surgeuniverse.com/home-1

———, "SURGE: Fuzzy Chronicles," undated (b). As of February 14, 2017:
http://www.surgeuniverse.com/home/game/fuzzy-chronicles-1

Scharf, Deborah M., Steven C. Martino, Claude M. Setodji, B. Lynette Staplefoote, and William G. Shadel, "Middle and High School Students' Exposure to Alcohol- and Smoking-Related Media: A Pilot Study Using Ecological Momentary Assessment," *Psychology of Addictive Behaviors*, Vol. 27, No. 4, December 2013, pp. 1201–1206.

Schmidt, William, Richard Houang, and Leland Cogan, "A Coherent Curriculum: The Case of Mathematics," *American Educator*, Vol. 26, No. 2, Summer 2002, pp. 47–48.

Schmidt, William H., Curtis C. McKnight, Richard T. Houang, HsingChi Wang, David E. Wiley, Leland S. Cogan, and Richard G. Wolfe, *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning*, San Francisco, Calif.: Jossey-Bass, 2001.

Schmidt, William H., Curtis C. McKnight, and Senta H. Raizen, eds., *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education*, New York: Kluwer Academic Publishers, 2002.

Schmidt, William H., Senta H. Raizen, Edward. D. Britton, L. J. Bianchi, and Richard G. Wolfe, eds., *Many Visions, Many Aims*, Vol. 2: *A Cross-National Investigation of Curricular Intensions in School Science*, Boston, Mass.: Kluwer, 1997.

Schultz, Susan E., and Raymond L. Pecheone, "Assessing Quality Teaching in Science," in Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta, eds., *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, San Francisco, Calif.: Jossey-Bass, 2014, pp. 444–492.

Schweig, Jonathan, "Cross-Level Measurement Invariance in School and Classroom Environment Surveys: Implications for Policy and Practice," *Educational Evaluation and Policy Analysis*, Vol. 36, No. 3, 2014, pp. 259–280.

Shavelson, Richard J., Lorraine M. McDonnell, Jeannie Oakes, Neil Brian Carey, and Larry Picus, *Indicator Systems for Monitoring Mathematics and Science Education*, Santa Monica, Calif.: RAND Corporation, R-3570-NSF, 1987. As of February 12, 2017:
http://www.rand.org/pubs/reports/R3570.html

Shute, Valerie J., "Stealth Assessment in Computer-Based Games to Support Learning," in Sigmund Tobias and J. D. Fletcher, eds., *Computer Games and Instruction*, Charlotte, N.C.: Information Age Publishing, 2011, pp. 503–523.

Smithson, John L., and Andrew C. Porter, *Measuring Classroom Practice: Lessons Learned from Efforts to Describe the Enacted Curriculum—The Reform Up Close Study*, New Brunswick, N.J.: Consortium for Policy Research in Education Research Report Series Report 31, October 1994. As of February 12, 2017:
http://files.eric.ed.gov/fulltext/ED377563.pdf

Snow, Erica L., Aaron D. Likens, Laura K. Allen, and Danielle S. McNamara, "Taking Control: Stealth Assessment of Deterministic Behaviors Within a Game-Based System," *International Journal of Artificial Intelligence in Education*, Vol. 26, No. 4, December 2016, pp. 1011–1032.

Spillane, James P., and John S. Zeuli, "Reform and Teaching: Exploring Patterns of Practice in the Context of National and State Mathematics Reforms," *Educational Evaluation and Policy Analysis*, Vol. 21, No. 1, Spring 1999, pp. 1–27.

Stecher, Brian M., Laura S. Hamilton, Gery W. Ryan, Vi-Nhuan Le, Valerie L. Williams, Abby Robyn, and Alicia Alonzo, *Measuring Reform-Oriented Instructional Practices in Mathematics and Science*, Santa Monica, Calif.: RAND Corporation, DRU-2787-EDU, 2002. As of February 12, 2017:
http://www.rand.org/pubs/drafts/DRU2787.html

Stecher, Brian, Vi-Nhuan Le, Laura Hamilton, Gery Ryan, Abby Robyn, and J. R. Lockwood, "Using Structured Classroom Vignettes to Measure Instructional Practices in Mathematics," *Educational Evaluation and Policy Analysis*, Vol. 28, No. 2, Summer 2006, pp. 101–130.

Stecher, Brian M., Alice C. Wood, Mary Lou Gilbert, Hilda Borko, Karin L. Kuffner, Suzanne C. Arnold, and Elizabeth H. Dorman, *Using Classroom Artifacts to Measure Instructional Practices in Middle School Mathematics: A Two-State Field Test*, Los Angeles, Calif.: National Center for Research on Evaluation, Standards and Student Testing, University of California, CSE Report 662, 2005. As of February 12, 2017:
https://eric.ed.gov/?id=ED492892

Stein, Mary Kay, Richard Correnti, Debra Moore, Jennifer Lin Russell, and Katelynn Kelly, "Using Theory and Measurement to Sharpen Conceptualizations of Mathematics Teaching in the Common Core Era," *American Educational Research Association Open*, Vol. 3, No. 1, 2017.

Stein, Mary Kay, and Gooyeon Kim, "The Role of Mathematics Curriculum Materials in Large-Scale Urban Reform: An Analysis of Demands and Opportunities for Teacher Learning," in Janine T. Remillard, Beth A. Herbel-Eisenmann, and Gwendolyn M. Lloyd, eds., *Mathematics Teachers at Work: Connecting Curriculum Materials and Classroom Instruction*, New York: Routledge, 2009, pp. 37–55.

Student Achievement Partners, "Instructional Practice for the CCSS," undated. As of February 13, 2017:
http://achievethecore.org/page/2730/instructional-practice-for-the-ccss

Swanson, Christopher B., and David Lee Stevenson, "Standards-Based Reform in Practice: Evidence on State Policy and Classroom Instruction from the NAEP State Assessments," *Educational Evaluation and Policy Analysis*, Vol. 24, No. 1, Spring 2002, pp. 1–27.

Texas Education Agency, *A Guide to Understanding Student E-Portfolios: K–12 Digital Portfolio Programs for College and Career Readiness*, Austin, Texas: University of Texas at Austin, 2013. As of February 12, 2017:
http://www.esc20.net/users/0026/docs/ProjectShare/
K-12%20Digital%20Portfolio%20Programs%20for%20CCR%20in%20Project%20Share.pdf

Tin Can Application Program Interface, "What Is the Tin Can API?" undated. As of February 14, 2017:
https://tincanapi.com/overview/

Tripod Education Partners, "Learn About Tripod," undated. As of February 13, 2017:
http://tripoded.com/about-us-2/

Tucker, Pamela D., James H. Stronge, Christopher R. Gareis, and Carol S. Beers, "The Efficacy of Portfolios for Teacher Evaluation and Professional Development: Do They Make a Difference?" *Educational Administration Quarterly*, Vol. 39, No. 5, December 2003, pp. 572–602.

University of California, Los Angeles, Graduate School of Education and Information Studies, "e-QIS Tablet Portfolio," undated. As of February 13, 2017:
https://eqis.gseis.ucla.edu/

University of Chicago Consortium on School Research, "Surveys of CPS Schools," undated. As of February 13, 2017:
https://consortium.uchicago.edu/surveys

Walkington, Candace, and Michael Marder, "Classroom Observation and Value-Added Models Give Complementary Information About Quality of Mathematics Teaching," in Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta, eds., *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, San Francisco, Calif.: Jossey-Bass, 2014, pp. 234–277.

Whitehurst, Grover J., Matthew M. Chingos, and Katharine M. Lindquist, *Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts*, Washington, D.C.: Brookings Institution, 2014. As of February 12, 2017:
https://www.brookings.edu/research/
evaluating-teachers-with-classroom-observations-lessons-learned-in-four-districts/

Yuan, Kun, John Engberg, Julia Kaufman, Laura Hamilton, Heather Hill, Kristin Umland, and Daniel McCaffrey, "Using Anchoring Vignettes to Calibrate Teachers' Self-Assessment of Teaching," paper presented at the spring meeting of the Society for Research on Educational Effectiveness, Evanston, Ill., March 2014. As of February 12, 2017:
https://eric.ed.gov/?id=ED562857

Zastavker, Yevgeniya, Veronica Darer, and Alexander Kessler, *Improving STEM Classroom Culture: Discourse Analysis*, conference proceedings, Franklin W. Olin College of Engineering, October 2013. As of May 1, 2016:
http://digitalcommons.olin.edu/facpres_2013/3