



William T. Grant
FOUNDATION



BILL & MELINDA
GATES *foundation*

Measuring Instruction in Higher Education

Summary of
a Convening

Measuring Instruction in Higher Education

Summary of a Convening Held in Chicago, Illinois,
November 17-18, 2014

Organized by:

William T. Grant Foundation

Spencer Foundation

Bill & Melinda Gates Foundation

The William T. Grant Foundation invests in high-quality research to ensure that young people from diverse backgrounds reach their fullest potential.

The Spencer Foundation investigates ways in which education, broadly conceived, can be improved around the world.

The Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives.

Participants at the Convening

Richard Arum, New York University/Bill & Melinda Gates Foundation

Courtney Bell, Educational Testing Service

Daniel Bernstein, University of Kansas

Rebecca Blank, University of Wisconsin–Madison

Andrea Bueschel, Spencer Foundation

Matthew Chingos, Brookings Institution

Charles Clotfelter, Duke University

Erin Driver-Linn, Harvard University

John Easton, Spencer Foundation

Peter Ewell, National Center for Higher Education Management Systems

Adam Gamoran, William T. Grant Foundation

Drew Gitomer, Rutgers, The State University of New Jersey

Daniel Greenstein, Bill & Melinda Gates Foundation

Pamela Grossman, Stanford University

Karen Inkelas, University of Virginia

Robert Mathieu, University of Wisconsin–Madison

Mike McPherson, Spencer Foundation

Steve Olson, Freelance Writer

Amy Proger, Spencer Foundation

Josipa Roksa, University of Virginia

Susan Singer, National Science Foundation

Carl Wieman, Stanford University

Executive Summary

The William T. Grant Foundation, the Spencer Foundation, and the Bill & Melinda Gates Foundation all have supported efforts to improve instruction and student learning in K-12 education. On November 17–18, 2014, the three foundations sponsored a convening in Chicago to explore the possibility of extending those efforts to higher education. The meeting brought together twenty-two experts on education and the learning sciences to discuss a specific aspect of teaching and learning in colleges and universities—the measurement of instructional quality—to guide possible future initiatives by the foundations.

Teaching and learning in college have many parallels with K-12 education, but the two levels of education also have critical differences. Most important, students in colleges and universities learn much more outside the classroom than inside, though the context of education is also crucial in K-12 education. But in college, students interact in new ways with other students and with the resources that higher education makes available to them, deepening and enriching the knowledge they acquire in classrooms. Thus, measuring instruction in higher education requires evaluating the entire student learning experience, not just what happens in classrooms.

Earlier research and practice-based projects have created a solid foundation of knowledge and experience for more in-depth examinations of instructional measurement. For example, the Social Science Research Institute has studied the acquisition of generic skills in college, including critical thinking and complex reasoning, and is currently studying subject-specific learning in higher education. Observational protocols have been developed to assess what happens in classrooms. Procedures used in the peer review of research have been applied to the review of instructional quality. The federal government, as part of its five-year strategic plan in STEM education, is seeking to increase the number of students who succeed in STEM fields, particularly among groups historically underrepresented in those fields. However, none of these initiatives is focused directly on the measurement of instruction to improve student learning in higher education.

The participants at the convening provided a wide range of input to the foundations in considering possible future initiatives in instructional

measurement. Broadly speaking, instructional measurement can be focused on describing instruction (generally in the context of research), improving instruction, or evaluating instruction (for example, to make high-stakes hiring and promotion decisions). Several participants at the convening, including representatives of the three foundations, said that evaluating instruction for the purposes of high-stakes decisions would not be an appropriate focus for an initiative. However, one option for the foundations would be to invest not only in the development of measurement tools but also in building the knowledge needed for implementing those tools in a program of improvement or even evaluation. Still, the intended use of tools needs to be made clear, several participants said, to provide direction for their development.

Another and more immediate option for the foundations would be to support the creation of a taxonomy of instruction—a snapshot of what is happening in higher education today. In particular, an initiative to develop such a taxonomy could look at digital modalities of instruction, either in association with conventional classes or in formats that are entirely online.

An initiative could encompass one or a handful of disciplines, or it could explicitly foster cross-disciplinary interactions. A cross-disciplinary initiative could extend work done in one discipline or group of disciplines to others. Alternately, deep study of a small number of disciplines could yield a conceptual framework that ultimately benefits other disciplines as well as the discipline being studied.

With regard to the conditions and structure of instruction, specific targets for investigation include large introductory courses, developmental courses, the alignment of courses and programs within and across institutions, or alternatives to lecturing. Another target could be student learning outcomes, especially for introductory courses and core courses where desired learning trajectories can be identified. Measures of instruction differ from measures of student learning, though they are often conflated. But tools to measure instruction could track student behavioral responses that are associated with learning, including behaviors that occur outside the classroom.

Given the uneven distribution of capacity in higher education to carry out this work and implement the

results of research, an initiative could be directed at capacity building or include capacity building as a substantial component of any grant. The development of instructional measures requires collaboration among people with different expertise, which increases the challenge of building capacity. But many postdoctoral fellows and new faculty members are very interested in education research and in applying the results of this research, and additional grant money directed toward those purposes could attract considerable interest.

An initiative could make special arrangements for institution involvement—for example, by providing incentives for institutions to use the results of research to improve instruction. An initiative also could support the development of networks of investigators within an institution or across institutions, with convenings being held to jumpstart work in specific areas.

Understanding instruction and learning how to change it are complex problems. But a tremendous opportunity currently exists, participants at the convening agreed, to change long-standing practices by bringing new knowledge to bear on these problems.

Goals of the Convening

Higher education has come to be seen as a prerequisite for success in a technologically sophisticated and rapidly changing world. But a troubling question is associated with the experiences many students have at colleges and universities. How much are they learning from the instruction they receive in college-level classes? Are their experiences inside and outside the classroom preparing them adequately for the challenges they will face in the workforce and the broader society?

To explore one important aspect of this issue, the William T. Grant Foundation, the Spencer Foundation, and the Bill & Melinda Gates Foundation sponsored a meeting in Chicago on November 17–18, 2014, focused on ways to measure the quality of instruction in higher education. Twenty-two experts on education and the learning sciences discussed existing instructional measures in higher education, the link between instruction and learning, and ways of improving instructional measurement. The convening was designed to provide the three foundations with the information they need to consider future investments in this area.

This summary of the convening has been prepared to involve a broader audience in the conversation. By capturing the major observations and conclusions made at the convening, it provides a reference point for future discussion and scholarship on the measurement of instruction in higher education.

An Expansive View of Instruction

In the opening session of the convening, representatives of the three foundations that sponsored the convening described their goals for the event.

One way to think about instruction in higher education is to consider the conditions that teachers lay out for students—the materials, objectives, and activities of instruction, observed Adam Gamoran. But that would give only a partial picture of instruction. Gamoran’s research on K-12 education has demonstrated that what students bring to the instructional context is as important as what teachers

bring. In that respect, instruction should be seen as what teachers and students do together.

This perspective implies that measuring instruction requires as much attention to how students respond to the materials and activities of teaching as to those materials and activities themselves. Students’ mastery of content is elevated when they are engaged both cognitively and affectively, said Gamoran, when students are not just checking off boxes because they know the answers. Measuring instruction at the K-12 level therefore requires measuring how students respond to the conditions created by teachers.

This is even more the case in higher education. As was pointed out throughout the convening, college students learn much more outside than inside the classroom. They interact with other students and with the resources that higher education makes available to them, deepening and enriching the knowledge they acquire in classrooms. Thus, measuring instruction in higher education requires evaluating the entire student learning experience, not just what happens in the classroom. This learning experience is affected by the population of students in an institution, the culture and climate of the institution, the institution’s expectations and goals for its students, the goals that students have for their own learning, and the goals that instructors have for student learning. All these factors influence what students get out of a particular class, with associated effects on measures of instructional quality.

In addition, the William T. Grant Foundation recently launched an initiative to support research on programs, practices, and policies that reduce inequality among young people. Who succeeds and who fails in higher education is a major source of the inequality that exists in the United States. Furthermore, prior research suggests that more engaging instruction may be especially beneficial for students who are less well prepared and feel a sense of isolation in college. In that respect, focusing on instruction in higher education also addresses issues of inequality. “Approaches to improving instruction that emerge from this effort could also reduce gaps in students college performance and completion,”

Gamoran said. “From the standpoint of my foundation, I’d like to see that remain part of our focus.”

Measures of Student Success

Richard Arum agreed that student success in college depends on many factors, including the institution, the field of study, and a student’s goals. Nevertheless, student success can be measured in terms of specific goals. The first is obtaining a certification, qualification, or degree from a college or university. The second is engaging deeply with college-level material. The third is developing the ability to see the world with a critical eye, to ask good questions, and to know how to seek answers to those questions. The fourth is obtaining the cognitive, academic, and social skills needed to lay the groundwork for a future occupation.

With the partial exception of science, technology, engineering, and mathematics (STEM) fields, higher education does not have standardized measures of instructional improvement and student learning, Arum noted. However, progress in STEM fields and in K-12 education has demonstrated that significant progress is possible. For example, the Measures of Effective Teaching project sponsored by the Bill & Melinda Gates Foundation has found significant correlations among instructional observations, student surveys, and student test score gains.

No measurement is better than poor measurement, and poor measurement already exists in higher education. For example, student surveys at the end of a class are today the de facto major measure of instruction, yet a solid body of evidence points to the inadequacies of such measures. Many steps will need to be taken to produce measures of instructional improvement and student learning in colleges and universities. Still, said Arum, “that is no excuse for not attempting to start to put the building blocks in place to get there.”

Higher education is currently in a period of tumultuous change, observed Daniel Greenstein. Colleges and universities are under pressure from students and families who ask questions about the value of a college education. They are under pressure from policy makers facing restricted budgets, competing demands, and calls for fiscal accountability. They are under financial pressure from increasing costs and constrained revenues. As a result of these

pressures, higher education will look much different in 15 years than it does today, Greenstein said. Measures of instructional quality could help guide these changes in such as way as to improve student outcomes.

Greenstein pointed to other measures that have been applied or are being developed to measure instruction in higher education, many of which are not just poor measures but potentially damaging. “Absent a response from inside the academic community, that trend will continue, and we will be drowned out by the noise,” he said. “We have to try to put out some alternative means of capturing information about learning.”

For its part, the Gates Foundation is interested in making higher education work for more rather than fewer students in the United States, Greenstein added. To lead healthy, productive, and sustaining lives, young people increasingly need college credentials. The Gates Foundation wants to know how it can work with and support institutions of higher education so that more students can acquire the credentials they need to succeed. “We’re very keen to know whether or not the institutions that we’re looking at deliver the kinds of results that we’re interested in.”

A Knowledge Building Exercise

McPherson said that the convening should be seen as a knowledge building exercise. How do the choices made by instructors, administrators, and policy makers affect student outcomes? Without this knowledge, taking actions to improve higher education is “like operating in the dark.”

He also issued several cautions. First, initiatives to measure instruction have the potential to do harm as well as good. The No Child Left Behind initiative at the federal level had positive outcomes, but it also had many negative consequences. “It’s very important that we take the time and the thoughtfulness to examine the risks in the measurement efforts that we undertake.”

In addition, measurements have value only to the extent that their purpose is known. Measures are tools that are intended to be put to use. “If we can’t spell out who’s going to use those measures to do what, . . . then we have to go back and think again. You can’t talk about the validity of a measurement without knowing its purpose.” Discussions of measures tend to focus on the hows, he said, but attention also needs to be devoted to the whys.

Organization of the Report

Chapter 2 of this summary of the convening describes two earlier projects that had goals comparable to the goals of the convening.

Chapter 3 looks at the incentive structure in colleges and universities and examines the various goals toward which instructional measurement can be directed.

Chapter 4 describes past and current research on instructional measurement, including the development of such tools as observational protocols, teaching inventories, and peer review of teaching.

Finally, Chapter 5 summarizes the input and options provided by participants at the convening on possible initiatives that the three foundations could undertake.

Review of Earlier Projects

At the beginning of the convening, the participants discussed two earlier projects supported by the Grant, Spencer, and Gates foundations that focused on the classroom environment and student learning. These projects resulted in tools, insights, and follow-on initiatives that can inform any future effort in the area of instructional measurement.

The Joint Project for the Development and Improvement of the Measurement of Classroom Quality

Michael McPherson and Pamela Grossman briefly described an earlier project on classroom measurement supported by the Spencer and William T. Grant foundations. In 2007 the two foundations embarked on an effort to develop tools that could measure the effectiveness of teaching in K-12 classrooms. Existing evidence indicated that the large variation in student learning across classrooms depended to a significant extent on the experiences students had in those classrooms, including which teachers they had. The Joint Project for the Development and Improvement of the Measurement of Classroom Quality sought to develop ways of generating accessible and reliable data that describe the classroom environment and the instructional activities students experience.¹

Through requests for proposals, the foundations supported a variety of research teams to explore the development of tools that could reliably describe and quantify what happens in the classroom while being suitable for practical work in schools. The members of these teams had diverse backgrounds and interests, and the foundations encouraged them to form a learning community so that they could build on each other's expertise. The project considered a variety of possible tools for learning about what goes on in classrooms, including in-person observation, video observation, teacher logs, student logs, and classroom artifacts. The intention was not to develop evaluative

tools but to understand what was happening in classrooms by developing efficient, reliable, and valid measures of instruction.

This project led to several major conclusions:

- Measures of teaching effectiveness need to be well developed before scaling up a measurement program. Measuring instruction is hard to do at scale, which requires that measurement tools be carefully thought out before being applied in such a context.
- The development of reliable and valid measurement tools requires blended expertise, including people who deeply understand psychometrics, teaching, and the subject matter of a class.
- Targeting a few subject areas and having researchers share their work within and across those subjects is preferable to a more broadly based effort.
- Multiple measures of student learning can assess different kinds of learning outcomes and the contributions of different classroom activities to those outcomes.
- Different kinds of instruction have varying degrees of effectiveness with different students. Some forms of instruction benefit some students but not others.
- Measures of instruction and of outcomes are often conflated. The two need to be disentangled and closely analyzed to understand the complex relationship between them.
- Instruments used to measure and describe instruction are not necessarily designed to help instructors improve.

Many of the tools and understandings developed in this project were later adopted by the Measures of Effective Teaching project supported by the Bill & Melinda Gates Foundation.² However, as Drew Gitomer pointed out, the initial pilot projects sponsored by the Spencer and William T. Grant foundations were very different than the large-scale evaluations done

¹ More information is available from <http://www.spencer.org/content.cfm/measurement-of-classroom-quality>.

² More information is available from <http://www.metproject.org>.

as part of the MET project. When applied in practice, instruments developed in research can become “very different entities than what we developed,” he said.

The Measuring College Learning Project

In their book *Academically Adrift*, Richard Arum and Josipa Roksa demonstrated that a large proportion of college students make little or no progress on measures of critical thinking, complex reasoning, and writing during their first two years in college.³ As they write in their book, “large numbers of U.S. college students can be accurately described as academically adrift. They might graduate, but they are failing to develop the higher-order cognitive skills that it is widely assumed college students should master.”

A common response to their findings, noted Roksa at the convening, has been that colleges do not set out specifically to teach critical thinking, complex reasoning, and writing. Rather, they teach particular subject matter and assume that more generic skills will develop in the process.

To improve understanding of subject-specific learning in higher education, the Social Science Research Council has undertaken a project that builds on the earlier research on generic skills.⁴ This Measuring College Learning (MCL) project, which has been funded by the Bill & Melinda Gates Foundation and the Teagle Foundation, has been bringing together panels of about a dozen faculty members in six fields—biology, business, communications, economics, history, and sociology—to identify the competencies, conceptual knowledge, and practices that students should learn in each of those fields during college. The six faculty panels also have been discussing the principles that underlie assessments in their fields. The overall goal, said Roksa, is to develop representative learning outcomes that can be measured to indicate broader learning in those fields. Pairs of faculty from each field are writing white papers that will synthesize and expand upon the work of each of the panels.

Some of the six fields had already made progress toward identifying key competencies and concepts

before the project began. For example, the Lumina Foundation has been supporting a faculty-led process, known as Tuning, in which a range of stakeholders—including students, employers, and recent graduates—jointly determine the specific learning outcomes required for a student to earn a degree in a certain discipline.⁵ The correspondence between these earlier processes and the MCL project has been “remarkable,” noted Roksa. However, the MCL project is taking the next step of identifying the 21st-century subject-specific skills that students need to acquire during their college years and seeking ways to measure those skills and relate them to instructional quality.

At the time of the convening, the MCL project was in the process of establishing a demonstration project in one of the fields—mostly likely, business, Roksa said. A demonstration project could explore the links between subject-specific learning and the mastery of more generic skills. It also could investigate in more detail the relationship between classroom instruction and learning, along with the effects of institutional policies and climate on student outcomes.

The MCL project is intended to foster discussions about instruction and learning within departments and disciplines in each of the six fields. For example, one important topic of discussion could be whether students who follow particular pathways or have particular experiences show greater learning gains, either on tests of subject-specific skills or generic skills. Another important outcome of the project, Arum added, could be new instruments developed by assessment companies for the 21st-century competencies, knowledge, and practices identified by the project.

³ Richard Arum and Josipa Roksa. (2011.) *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.

⁴ More information is available from <http://www.ssrc.org/programs/measuring-college-learning>.

⁵ More information is available from <http://tuningusa.org>.

Incentives in Higher Education and the Purposes of Instructional Measurement

The improvement of instructional measurement in higher education will take place in a rich, complex, and inertia-bound culture that encompasses many stakeholders and interests. Two participants at the convening described several critical aspects of this culture and the prospects for change.

Incentives in Higher Education

The incentives existing within colleges and universities are a powerful influence on instructional measurement and how it may be used within an institution, said Charles Clotfelter. The mission statements of colleges and universities generally include teaching and learning as essential goals. In practice, however, teaching in universities is often understood to be secondary to the research mission. The realities of the academic evaluation process for promotion and tenure in universities often mean that teaching is mentioned in a committee report out of duty rather than because teaching is receiving substantial weight, though the situation can differ among colleges, Clotfelter acknowledged.

Another feature of higher education is that faculty members often view themselves more as independent contractors than as employees. As one element of this perspective, they tend not to examine disciplines outside their own critically. In that respect, higher education differs from K-12 education, said Clotfelter, where evaluation is more broadly based.

The leaders of universities and colleges also have incentives to take teaching and learning seriously, he added. Institutions compete with each other in attracting future applicants. Government officials may want to do something about low graduation rates or lackluster scores on international tests. Also, innovations in teaching and learning, if properly implemented, may not detract from research or service and may produce savings in the long term, even if they have short-term start-up costs. The question then

becomes how to convince faculty members to learn and adopt new techniques to realize these longer term benefits.

The Purposes of Instructional Measurement

Measuring educational outcomes and environments has many purposes, observed Peter Ewell, and these purposes in turn shape the measurements to be made and how those measurements are made. Among the most prominent of these purposes are:

- To detect unsatisfactory levels of performance. This is the basic purpose of most state- or system-level indicator systems in K-12 education. The major measurement questions are the level of precision needed and an *a priori* judgment of what constitutes a satisfactory level of performance.
- To build a knowledge base, generally for research. This purpose is more general and does not depend on a particular context. The main measurement questions are, again, the level of precision needed and the connection of the domain examined to a candidate hypothesis or the existing research literature.
- To certify or accredit. This purpose centers on conferring a summative value on a measure related to institutions, classrooms, programs, and so on. The main measurement questions are precision and the need to minimize Type II error so that an attribute is not overlooked.
- To decide where or for whom to intervene. This purpose centers on identifying populations, situations, or settings that should be targeted for a program or other intervention. The main measurement questions are precision and the ability to appropriately disaggregate potential targets.
- To monitor the effects of an intervention. This purpose is to determine whether or not a given intervention or program has been successful. The

main measurement questions are precision, an *a priori* criterion of “success,” and a control of some kind (whether a randomized control group or use of a quasi-experimental design).

- To track trends over time. This purpose monitors effects over a longer time period. The main measurement questions are the same as above but include the ability to obtain comparable and valid values of the measure at different points in time.
- To demonstrate an outcome or persuade a decision maker to take a position. This purpose embraces most of the above properties, but it adds the human element of making a persuasive argument. The main measurement questions are the same as above but also include the face validity of the measure for various constituencies and the ability of measured values to be converted to easy-to-understand formats such as graphics.
- To signal that an organization or institution is managed rationally. This purpose also involves a human element. The main measurement questions are face validity for a particular constituency or stakeholder group, but once this condition is met, no other measurement properties matter much, including validity and reliability.

Cutting across all of these purposes are two additional elements, Ewell added. The first is the unit of analysis, which can be individuals, programs, institutions, or populations. The second is change over time, which might be measured simply to examine trends but could also include special cases, such as “value added” in outcome measures or changes in school environments.

Ewell cautioned against making methodologically complex measurements simply because it is possible to do so. “Whenever I hear something getting too fancy, I want to say, ‘Let’s get back to basics.’” He also pointed to the challenge of implementation. “We know a lot more than we do.” A good measurement system may not lead to action.

Ewell acknowledged that the purposes of measurement that he listed are largely for the use of people outside an instructional setting. But measurements also can provide feedback to instructors and enable them to improve. For example, Ewell also has been working with the National Institute for Learning Outcomes Assessment on

ways that academic programs and institutions can use assessment data internally to strengthen undergraduate education.⁶

Description, Improvement, or Evaluation?

Clotfelter’s and Ewell’s comments on institutional incentives and the purposes of measurement sparked a wide-ranging discussion that extended throughout the convening. Gamoran sorted and simplified Ewell’s purposes for instructional measurement into three categories: to describe (usually for research purposes), to improve, and to evaluate (often for the purposes of hiring or promotion decisions). To some extent, these overlap, he said, yet different purposes can call for different tools or for the use of the same tool in different ways.

As Robert Mathieu pointed out, education research does not necessarily lead to instructional change. In higher education, faculty members have to act on the results of measurements for change to occur, which requires thinking about how the results of measurements are going to be used to create change.

Carl Wieman contended that the most important purpose of measurement is to improve instruction and outcomes. In that case, it must be designed to inform instructors about what they can do better. At the same time, it can inform department chairs, deans, and provosts about what they can do to improve instruction, and it can provide information to students about how they can get the education that they will need for success after college. Making those purposes of measurement primary is the way to create long-term change, Wieman said.

⁶ More information is available at <http://learningoutcomeassessment.org>.

Research on Measurement Tools

Several participants at the convening surveyed past and ongoing research in instructional measurement to provide background for the consideration of possible future initiatives.

Lessons from Measurement Research

Past experience with measurement tools at the K-12 level have produced valuable lessons on measuring learning environments, observed Drew Gitomer and Courtney Bell in an overview of current tools for learning about what happens in classrooms. First, achieving measurement with high reliability is very difficult. The question is what to make of this observation. One possibility is that instructional measurement is a technical task that requires a technical solution, such as better instruments, better training and quality control, and increased sample sizes. A second possibility is that the difficulty reflects a lack of shared understanding of the dimensions of instructional practice. If this is the case, this lack of shared understanding could be a target of research.

A third possibility is that the difficulty of instructional measurement is a result of the complexity and context dependence of teaching and learning. Researchers may look for stable traits and for variation, but instruction may not be an activity that lends itself to that sort of approach. For example, researchers may seek variation among individuals and settings, and their observations can affect how the problem is conceptualized. But this process can lead to incomplete understanding and policies that do not fully address the needs that exist.

Educational research has revealed differences among classrooms and some consistent relationships among measures, which contributes to confidence in the validity of the measures. However, these measures also have revealed generally limited levels of performance across most classrooms. Students tend not to be intellectually challenged. They engage in little genuine discourse. Classrooms display a lack of disciplinary practices. And the variation that does occur tends to be at the lower end of scale distributions.

Differential methodologies tend to obscure this homogeneity of practice, which leads to different interpretations of the problems and how to address those problems. Are the problems more systemic in nature? Is the educational system effective in general but simply requires tweaking at the lower end of the distribution? Should goals be defined in a norm-referenced or criterion-referenced manner?

The context of teaching clearly matters a great deal. What happens in the classroom is not just a function of an attribute that might be called a teacher's ability. Classroom interactions are influenced by the composition of students in the classroom, the curriculum, when during the school year data are collected, the particular lessons that are sampled, and many other factors. The context of judging also matters. Scores from research studies often differ substantially from scores emerging from the field in practice, with the scores in research tending to be much lower than those in practice. As a result, the qualitative descriptions of what scores mean are dramatically different, in essence moving from a modal description of practice as relatively limited to descriptions of practice that are strongly positive.

These observations have several implications for the measurement of instruction in higher education, Gitomer observed. First, context and the homogeneity of practice can be seen as a hurdle to good measurement. One way to deal with context is to use powerful statistical tools to adjust for contextual effects, which provides provisional answers, but this may limit understanding of the phenomenon. An alternative approach is to understand the context rather than simply adjusting for it, and this path needs to be pursued, Gitomer said.

Bell called attention to a biological parallel with education. Perhaps the complex act of instruction is less like a biological trait and more like an emergent property. In that case, trying to isolate for teaching quality while controlling for context is the wrong modeling strategy.

Bell elaborated on this point by discussing the contextual factors, constructs, and measures associated with teaching quality (Figure 4-1). Teaching

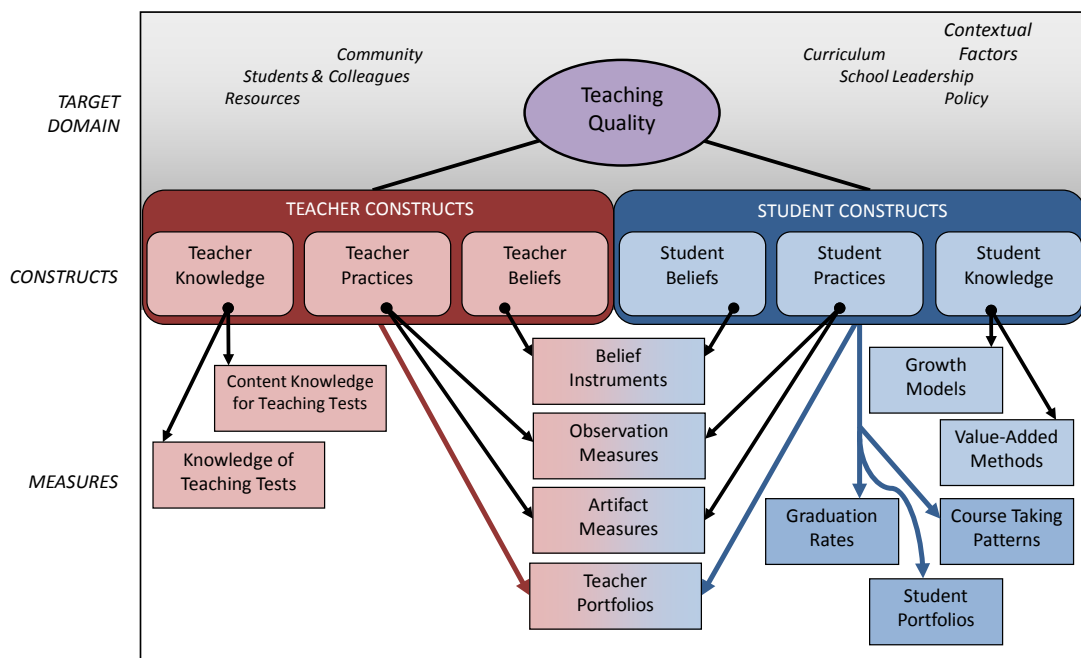


FIGURE 4-1 Teaching quality depends on teacher and student constructs that can be measured using various instruments, as well as on contextual factors. Bell, C.A., Gitomer, D.H., Croft, A.J. (2011). The contextual factors, constructs, and measures associated with teaching quality. (Diagram) Princeton, NJ: Educational Testing Service.

consists of an interactive dialogue among teachers, students, and content. These interactions cannot be completely separated from such contextual factors as the curriculum, school leadership, broader educational policies, resources, and the characteristics of students, teachers, and communities, because the effects of these contextual factors on teaching are not fully understood.

An alternative approach is to think about teacher constructs and student constructs, with each of these sets of constructs divided into knowledge, practices, and beliefs. A wide variety of instruments, models, and methods are used to measure these various constructs. For example, value-added methods generally are based on state tests that measure student knowledge, while an observational measure is based on observations of student or teacher practices. These two measures are related, but there is no reason to think that they would be highly correlated—and in practice, they generally are not.

Any observational system needs to have a theory of use, Bell said. She proposed the following progression:

1. Information: An observational system creates information about the level of teaching skill demonstrated by each teacher.

2. New insights: This information leads to insights about teaching.
3. Practice and learning: These insights lead to new understandings and strategies through professional development, use of new tools, and so on.
4. Changes in teaching: Teachers incorporate understandings and strategies into their practice, making it more effective.
5. Changes in learning: Effective teaching practice results in improved student learning.

In practice, observations of teaching are often assumed to lead directly to changes in teaching and learning. What this assumption overlooks is the generation of new insights and the incorporation of those insights into teaching through practice and learning. Even the observations themselves can be based on mistaken premises, or measures can be poorly conceived or executed. As was observed in response to Bell's presentation, teacher learning is as complex and poorly understood as student learning.

Then again, as Arum pointed out, most of these measures are relatively new, so people are still exploring how to translate information about teaching into insights, teacher learning, and changes in

instruction. Furthermore, added Bell, simply making observations of teaching can change the discourse among teachers and between teachers and principals, which by itself can be a positive influence on teaching practices.

Observational Protocols

Classroom observations sample one important aspect of college teaching—what happens in the classroom. In their presentation at the convening, Carl Wieman and Karen Kurotsuchi Inkelas compared nine observational protocols chosen on the basis of three criteria:

- The work was based on empirical (that is, data based and not experiential or anecdotal) information.
- The work included an actual observation form created to assess classroom teaching.
- The work assessed college-level classrooms, or at least secondary education classrooms (so not primary education).

The titles of the protocols they examined are:

TDOP—Teaching Dimensions Observation Protocol

COP—Classroom Observation Protocol

NxGEN COI—Next Generation Curriculum Observation Instrument

STROBE—Refers to capturing events at regular intervals

RTOP—Reformed Teaching Observation Protocol

CLASS-S—Classroom Assessment Scoring System (secondary education version)

CCCO—Community College Classroom Observation Form

COPUS—Classroom Observation Protocol for Undergraduate STEM

ROCA—Real-time Observation of Classroom Activities

Comparison of the teaching observation protocols revealed common concepts and approaches (Table 4-1). All of the protocols they examined were closed ended, meaning that they did not rely on more open-ended and qualitative written comments. But some of the protocols were subjective, requiring judgments about whether, for example, explanations were clear, while others were objective, such as whether a professor used clickers or commented on

student questions. Seven of the nine protocols required subject matter expertise or extensive training for reliability, while two could be performed by observers without specialized knowledge or training. Many of the protocols were directed toward STEM fields or related disciplines, such as medicine or health sciences.

Descriptions of Observational Protocols

The Teaching Dimensions Observation Protocol (TDOP) was developed to study instruction in STEM fields.⁷ The TDOP contains five categories of codes for classroom observers to monitor every two minutes: 1) teaching methods; 2) pedagogical moves or strategies; 3) teacher–student interactions; 4) cognitive engagement of students; and 5) use of instructional technology. The protocol, while thorough, requires multiday training and practice in order to obtain strong inter-rater reliability and has yet to be used very widely.

The Classroom Observation Protocol addresses a set of key questions organized around classroom activities, the learning expectations for students, and strategies that teachers can use to meet those expectations while individualizing instruction. Observations provide indications of current practice and the conditions under which practice occurs.

The Next Generation Curriculum Observation Instrument gathers demographic information, an inventory of learning practices, observations of technology use, and other features of instruction to describe in-class activities. It directs attention specifically to the goals of active learning and integration as well to evaluation questions about resources and technology use.

STROBE is a classroom observation tool used in the health sciences that assesses in-class student engagement.⁸ Every 5 minutes, an observer records the activity a class is performing and the proportion of the class performing the activity. Then the observer

⁷ Hora, M. T., Ferrare, J. J., & Oleson, A. (2012). *Findings from Classroom Observations of 58 Math and Science Faculty*. Madison, WI: Wisconsin Center for Education Research.

⁸ Kelly, P. A., Haidet, P., Schneider, V., Searle, N., Seidel, C. L., & Richards, B. F. (2005). A comparison of in-class learner engagement across lecture, problem-based learning, and team learning using the STROBE classroom observation tool. *Teaching and Learning in Medicine*, 17(2), 112–118.

TABLE 4-1 Comparison of Teaching Observation Protocols

	TDOP	COP	NxGEN	STROBE	RTOP	CLASS-S	CCCO	COPUS	ROCA
Concepts incorporated into protocol									
Teaching methods	●	●	●	●	●	●		●	●
Pedagogical strategies	●			●			●	●	●
Student–teacher interactions	●				●	●	●	●	●
Student engagement	●	●	●	●	●		●	●	●
Types of student engagement				●			●	●	●
Use of technology	●		●						●
Classroom management		●				●	●	●	●
Classroom climate					●	●	●		
Overall effectiveness		●					●		
Types of and training required for using protocols									
Type: 1. Subjective, closed-ended 2. Objective, closed-ended	1	1	2	2	1	1	1	2	2
Subject matter expertise and extensive training required for reliability	✓	✓	✓	✓	✓	✓	✓		
Disciplines protocol designed to be used with*	STEM	Math, Sci	Med	Med	Math, Sci	HS	CC	STEM	All

* STEM=Science, Technology, Engineering, Mathematics; Math=Mathematics; Sci=Science; Med=Medical or Health Sciences; HS=High School; CC=Community College

selects one student to observe for approximately 10 to 20 seconds, noting the student’s engagement with the activity, including the type of engagement and the object of the student’s engagement. This process is repeated eight to ten times over the course of a single class period. While useful for recording student engagement, STROBE is not as comprehensive in observing instructors’ teaching and pedagogical strategies.

The Reformed Teaching Observation Protocol (RTOP) focuses on observations in high school and college mathematics and science courses regarding: 1) lesson design and implementation; 2) content quality; 3) content engagement; 4) classroom interactions; and 5) student–teacher relationships.⁹ Observers are

required to make interpretive judgments concerning the instructor’s effectiveness in the above five realms, and the codes used can make sharing reports with instructors awkward. These subjective judgments require extensive training and practice in order to obtain high inter-rater reliability.

The Classroom Assessment Scoring System (CLASS) is an observational protocol based on educational and developmental research demonstrating that daily interactions between teachers and students are central to students’ academic and social development. It measures effective student–teacher interactions in prekindergarten through twelfth grade and is aligned with a set of professional development supports that enable teachers to make positive changes in areas of practice. Research has demonstrated that students in classrooms with higher CLASS ratings make greater gains in social skill and academic development than students in classrooms with lower CLASS ratings.

The Community College Classroom Observation Form adapts K-12 classroom observational protocols

⁹ Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.

for community colleges. It directs attention to the preparation and activities of instructors in a variety of categories and the effects of those activities on students.

The Classroom Observation Protocol for Undergraduate Science (COPUS) is designed to reliably characterize how instructors and students are spending their time in science classes, particularly classes with multiple and varied student activities.¹⁰ Every two minutes, observers note which of approximately a dozen activity codes best describes student or teacher activities. Extensive testing and modification went into achieving high inter-rater reliability with use by regular STEM faculty with only 1.5 hours of training, when used across the STEM disciplines.

The Real-time Observation of Classroom Activities (ROCA) is a protocol using a mobile application platform that employs menu screens and options to record the following: teaching methods, pedagogical strategies, student-teacher interactions, student engagement and their types of engagement, uses of technology, and classroom management. The current version of the protocol measures both frequency and duration of classroom activities and is designed to be used in any type of class format (lecture, seminar, flipped, etc.) with classes of any disciplinary focus.

On the basis of their review, Wieman and Inkelaas expressed a strong preference for the objective closed-ended protocols. Such protocols do not require that subjective judgments be made of teaching practices. An observer does not need to be an expert in a discipline or have extensive experience with a wide range of teaching methods. Observers also do not need extensive training to make objective closed-ended evaluations. Previous research has shown high correlations between the use of best instructional practices, which can be measured through objective protocols, and engaged student learning.¹¹ In

particular, COPUS and ROCA, which Wieman and Inkelaas helped develop, are objective protocols that use web-based tools or apps to ease data entry for observers. Widespread use of such protocols could answer some fundamental questions about instruction in higher education that remain unanswered, such as how often professors lecture or how often they ask questions. Furthermore, pilot studies with these protocols have revealed good inter-rater reliability while also showing that classroom structure is fairly homogeneous, even across disciplines.

However, Wieman and Inkelaas also noted that teaching observation protocols offer only one lens into the complex process that encompasses college teaching. If institutions were to announce that all evaluations of college teaching would henceforth be based solely on one or even a few observations of a professor's teaching, there would be widespread resistance—and for good reason, said Inkelaas. Instead, they advocated for a more comprehensive range of assessments of college teaching that follow the natural trajectory of creation, execution, and revision. The University of Virginia Center for Advanced Study of Teaching and Learning in Higher Education and the Teaching Resource Center have collaborated to create a faculty development continuum of the path to effective teaching, as shown in Figure 4-2. The triangles in the figure represent actions by instructors; the circles represent assessments of both instruction and learning. In this progression, an instructor initially decides to work on improving his or her teaching, seeks assistance, and subsequently makes changes to a course and the instruction in that course. Assessments of these changes; of an instructors' beliefs, knowledge, practices, and intentions; of teaching; and of student's learning provide feedback to guide continuing changes. Elements of this approach could be scaled up to aggregate levels and guide much more widespread improvements in college-level instruction.

Another more comprehensive assessment of college teaching is the Teaching Practices Inventory (TPI), which is a 72-item objective inventory of the use of teaching practices across eight domains (Table 4-2). It characterizes the extent of use by the instructor of practices that research has shown can achieve improved student outcomes, though it has been tested for validity only in STEM disciplines. The inventory takes only about ten minutes to fill out yet provides extensive information on in-class activities, supporting

¹⁰ Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, 12, 618–627.

¹¹ Wieman, C., and Gilbert, S. L. (2014, fall). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE-Life Sciences Education*, 13(3), 552–569.

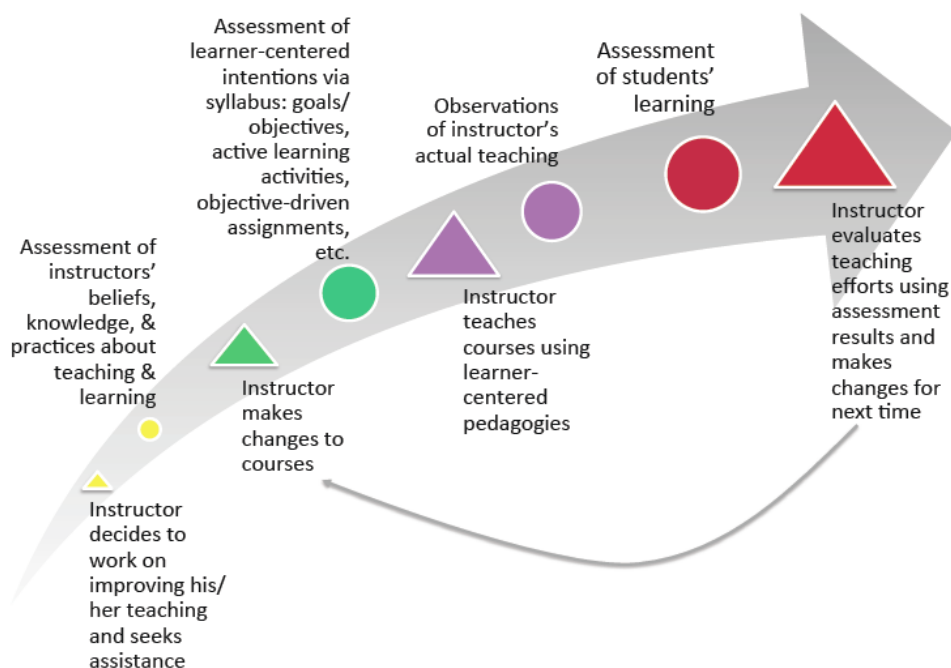


FIGURE 4-2 Instructional changes and assessments of those changes can increase the impact of college teaching on student learning. Adapted from Kreber, C., & Brook, P. (2001). Impact evaluation of educational development programmes. *International Journal for Academic Development*, 6(2), 96–108.

TABLE 4-2 Teaching Practices Inventory (TPI) Categories

I.	Course information provided <i>Information about the course, such as a list of the topics and organization of the course and learning goals/objectives</i>
II.	Supporting materials provided <i>Materials that support learning of the course content, such as notes, videos, and targeted references or readings</i>
III.	In-class features and activities <i>What is done in the classroom, including different types of activities that the instructor might do or have the students do</i>
IV.	Assignments <i>The nature and frequency of homework assignments in the course</i>
V.	Feedback and testing <i>Testing and grading in the course, as well as the feedback from instructor to students and from students to instructor</i>
VI.	Other <i>Assorted items covering diagnostics, assessment, new methods, and student choice and reflection</i>
VII.	The training and guidance of teaching assistants <i>The selection criteria and training used for course teaching assistants and how their efforts are coordinated with other aspects of the course</i>
VIII.	Collaboration <i>Collaboration with other faculty, use of relevant education research literature, and use of educational materials from other sources</i>

materials, student expectations, assignments, instructor feedback, the use of teaching assistants, collaboration, and other aspects of instruction.

The participants at the convening discussed the terms *objective* and *subjective* and whether a more accurate description might be judgments that involve a low degree of inference or a high degree of inference, with the latter requiring much more observer training than the former. They also discussed whether more active forms of instruction, as opposed to straightforward lectures, might be more beneficial for the members of groups that tend to drop out of STEM fields at an elevated rate, as some research has suggested.

Observation protocols can continue to be refined to document which teaching practices are and are not being used in college classrooms across instructors, courses, departments, and institutions. They also can be extended to study all of the different aspects of college teaching, including course planning, course objectives, instructional delivery, and student outcomes. Such tools could guide and drive improvements in instruction.

Participants at the convening also discussed technological improvements in making observations. As Gamoran pointed out, classroom observations used to be made with pencils and clipboards. The development of laptop computers and other small digital devices made it possible to record observations electronically, even if the basic process was the same. More recently, classrooms have been videotaped, either by a camera crew or through automated processes. The next step is automated analysis of classrooms through speech recognition software and artificial intelligence, either using videotapes or real-time observations.

Human expertise could be combined with these powerful technologies in innovative ways. For example, teachers could participate in the selection of classroom recordings for analysis, and the act of selection could help them focus on teaching practices. Teachers could even carry miniaturized video cameras and choose when to turn them on.

Other Evaluative Tools

Many forms of instructional measurement exist besides direct observations. Harvard University, for example, has used a portfolio model to catalyze

activities among faculty members, students, and staff members, explained Erin Driver-Linn, many of which involve measuring the quality of instruction. Individuals or groups can apply for grants to catalyze improvements in teaching and learning. From 600 applications, about 60 grants have been made to support such activities as research on instruction, increases in discourse on instructional topics, and initiatives to find out what others on campus are doing to improve instruction. After an initial phase of building momentum, the initiative has broadened to emphasize the building and deepening of networks of people. It also has been supporting events on both small scales and a campus-wide scale to foster discussion and instructional improvement.

As another part of this initiative, a multidisciplinary research team has been conducting projects on teaching and learning. For example, one project looked at the differences in course evaluations depending on whether students are asked to give an overall rating to the course first or are asked first to provide comments on the course. It found that asking for comments first produces longer, richer, and more meaningful sets of comments. However, students are then less likely to complete the entire evaluation.

Another project used small portable cameras to determine how many of the students who are registered for a course attended the lectures. It found that attendance rates vary widely depending on such factors as whether attendance is measured, whether clickers are used, and whether students are taking the course to fulfill pre-med requirements. It also found that some courses have very high attendance despite the absence of these incentives to attend, though the exact reasons for this high attendance have yet to be determined.

This initiative at Harvard has led to discussions of whether a high-quality prestigious journal could be established to examine the development of instructional measures. Another point of discussion is whether an alternative to Bloom's taxonomy could be developed that is inductively generated from observations of teaching. In addition, instructors at Harvard have been discussing whether learning management systems could promote the adoption of evidence-based best practices in teaching and then demonstrate evidence of how these practices improve learning.

One topic discussed at the convening was whether instruments could be developed to measure the learning that takes place outside of the classroom and the major contributors to that learning. For example, do students get as much from lectures observed online compared with attending the lectures in person? One response to this question was that students learn relatively little from straightforward lectures regardless of how they watch those lectures. The answer to this question could also vary by institutions; for example, many students at community colleges are extremely busy outside of their classes and may have little time to watch online lectures. And the answer could vary by disciplines, with some subjects requiring greater attendance in classes.

Peer Review of Teaching

Another approach that combines observations of teaching with other measures is to conduct peer review of instruction, noted Daniel Bernstein. Such an approach could draw on three useful voices: those of students, the instructor, and peers. Students can both provide ratings of instruction and demonstrate understanding of what is being taught through their performance on assessments. The instructor can describe the course design and intentions as well as whether he or she is satisfied with what students are learning. And peers can judge such factors as the quality of teaching, assignments, and learning and whether instructors are using the feedback they receive to improve their instruction.

At the University of Kansas, the promotion and tenure process and the rewarding of distinguished teaching awards requires that instructors be judged on the quality of intellectual content, the quality of teaching practices, the quality of student understanding, and evidence of reflective consideration and development. Instructors gather artifacts of their course goals, assignments, and examples of student work. They can describe whether students are actively engaged in material, how contact time is being used creatively and effectively, and whether students are developing a deep understanding of the course material.

One drawback of this approach is that it can take a lot of time. But faculty members already devote substantial amounts of time to reviewing research, whether for grant proposals or publication, Bernstein

observed. If they devoted even a small fraction of that time to the review of instruction, teaching could receive much more scrutiny than it does today. In the past, teaching has been viewed as largely a private activity, but this perspective has been changing in K-12 education and could change in higher education as well.

The IDEA Center has been working to develop methods that could be used on a large scale to do peer review of instruction. Review also could be done selectively of instructors and of individual instructors and classes over time.

Participants at the convening discussed the extent to which peer review of instruction is analogous to the peer review of research. For example, peer review of research often relies on the combination of multiple perspectives, and it tries to develop a collective understanding rather than complete reliability. Bernstein countered, however, that the same approach could lead to a much richer understanding of teaching and learning among the reviewers of instruction, as they discussed their collective judgments. Another difference between the peer review of research and peer review of teaching is that the former relies on a well-defined and widely accepted definition of who is an expert in a subject. To this, Bernstein pointed out that examination of teaching can be a scholarly activity that requires well-defined expertise, though it is not necessarily a research activity, and that treating such work as scholarship is “a welcoming metaphor for colleagues.” Such scholarship requires higher level rather than lower level inferences, so it differs from closed-ended objective observations of teaching. But more quantifiable elements could be included in the peer review of instruction along with richer descriptions of an instructor’s teaching.

Federal Investments in Measures of Instructional Effectiveness

Ongoing activities across a variety of federal agencies are related to instructional measurement in higher education even if they are not directly devoted to that issue, said Susan Singer and John Easton in their description of federal investments in instructional measurement. First, the federal

government has a five-year strategic plan in STEM education that has five priority investment areas:¹²

Improve STEM Instruction: Prepare 100,000 excellent new K-12 STEM teachers by 2020 and support the existing STEM teacher workforce.

Increase and Sustain Youth and Public Engagement in STEM: Support a 50 percent increase in the number of U.S. youth who have authentic STEM experiences each year prior to completing high school.

Enhance STEM Experience of Undergraduate Students:

Graduate one million additional students with degrees in STEM fields over a decade.

Better Serve Groups Historically Underrepresented in STEM Fields: Increase the number of students from groups that have been underrepresented in STEM fields who graduate with STEM degrees and improve women's participation in areas of STEM where they are significantly underrepresented.

Design Graduate Education for Tomorrow's STEM

Workforce: Provide basic and applied research expertise, professional development, and specialized skills development of graduate-trained STEM professionals.

All of these goals involve higher education, either in the preparation of K-12 teachers, the provision of research experiences for high school students, the instruction of undergraduates, or the preparation of graduate students for careers in which STEM subjects play a role.

The third goal is the one most related to instructional measurement. In this area, the five-year plan includes an implementation roadmap with four components.

Evidence-Based Practices: Identify and broaden implementation of evidence-based instructional practices and innovations to improve undergraduate learning and retention in STEM and develop a national architecture to improve empirical understanding of how these changes relate to key student outcomes.

Community Colleges: Improve support of STEM education at two-year colleges and create bridges between two- and four-year postsecondary institutions.

Research Experiences: Support and incentivize the development of university–industry partnerships, and partnerships with federally supported entities, to provide relevant and authentic STEM learning and research experiences for undergraduate students, particularly in their first two years.

Mathematics Success: Address the problem of excessively high failure rates in introductory mathematics courses at the undergraduate level to open pathways to more advanced STEM courses.

Throughout these goals, the focus is on high-impact learning environments, whether in the classroom, in the laboratory, online, or in other settings. These goals also apply to all institutions, including two-year institutions, where 60 to 70 percent of incoming students require developmental mathematics and 80 percent of incoming students who need a developmental mathematics course do not complete any college-level course within three years.

An interagency group chaired by Singer has been working on metrics to track progress toward the goals in the five-year plan. For example, the National Center for Education Statistics has been integrating questions on instructional practices into a longitudinal survey that included undergraduates. The National Science Foundation also has been working with the Department of Education to include more learning metrics into the indicators the two agencies compile.

Within the National Science Foundation, the Division of Undergraduate Education has focused on institutional transformation through the Widening Implementation and Dissemination of Evidence-based Reforms (WIDER) program. The foundation also has been requiring institutions to baseline their teaching practices in proposals so that improvements can be measured. These initiatives dovetail with others by nongovernmental organizations. For example, the Association of American Universities is funding the development of structural practice measurements at eight projects sites, and the American Association for the Advancement of Science has issued a report entitled *Describing and Measuring Undergraduate STEM Teaching Practices*, which identified four basic measurement techniques (surveys, interviews, observations, and portfolios), provides an overview of the strengths and weaknesses of each, identifies and summarizes specific protocols and measurement tools

¹² More information is available at <http://www.ed.gov/stem>.

within each technique, and provides references for further details.¹³

Within the Department of Education, the Institute of Education Sciences also has a portfolio around the topic of instructional measurement, though the work is embedded within other activities at its four centers. This work extends well beyond the STEM fields to other subject areas, from K-12 to undergraduate and graduate education to adult education. About 140 separate goals, across ten topic areas, are related to instructional measurement, but relatively few of these are focused directly on measuring instruction.

The work supported by IES offers a tremendous opportunity for in-depth research on instructional measurement, said Easton. He also noted that the staff members at IES are very amenable to outside suggestions on the kind of research they should fund.

Also in the Department of Education, the National Center for Education Statistics makes measurements related to instruction in its assessment work, including its surveys of teacher and students. This aspect of the assessments could be emphasized much more, and surveys at the K-12 level could be extended to higher education, Easton and Singer reported.

The Diversity of Contexts in Education

The convening participants spent some time considering the range of contexts in K-12 education as opposed to higher education. In some respects, K-12 classrooms are more uniform. The structure of classrooms and the interactions between teachers and students tend to be similar. The content of instruction is more uniform than in college, even in subjects where statewide or nationwide standards do not exist. The tests students take can be essentially identical across districts, states, or even the entire country, which creates an expectation of standardized preparation for those tests.

But in other respects, K-12 education is less homogeneous than higher education. The students in many K-12 classrooms have very different backgrounds and abilities. As Gitomer pointed out, the variation among students in a given classroom is typically much greater than the average growth in

student learning over a year. Also, research has shown that different teachers produce more or less learning in different students over the course of a year, though the extent to which this differs from higher education is unknown.

In higher education, in contrast, students tend to be more homogeneous within each institution because of the sorting process associated with applications and admissions. In addition, teaching practices within individual disciplines tend to be remarkably standardized, as Wieman pointed out, because of shared histories, experiences, and cultures within disciplines. Classroom management generally is not an issue in higher education, so classes are less differentiated in that regard. Even across types of institutions, including two-year colleges and four-year colleges, instruction can be very similar, though a lack of data on instructional practices in higher education makes it difficult to know the exact degree of similarity. The most influential differences in college classrooms for learning, said Wieman, involve the type of instruction that takes place.

¹³ AAAS (American Association for the Advancement of Sciences). (2013). *Describing and Measuring Undergraduate STEM Teaching Practices*. Washington, DC: AAAS.

Considerations for Future Initiatives

In the final session of the convening, participants turned to the kinds of future initiatives on instructional measurement that the three foundations might consider supporting. Adam Gamoran introduced the session by laying out several proposed criteria for any such initiative. It should cover the full range of institutions in higher education. It should focus on description and improvement rather than evaluation, with improvement measured by increased student success (including degrees, certifications, and qualifications), enhanced learning, and reduction of gaps among population groups. And it should be scalable so that it can have widespread effects.

A point made by several participants that extends across all possible initiatives is that reducing inequities among groups in higher education requires directing attention specifically to that issue. Initiatives that do not explicitly address inequities are unlikely to close the gaps that exist.

The Development of Instruments

As described in Chapter 3, an initiative could focus on describing instruction, improving instruction, or evaluating instruction. Representatives of the three foundations agreed that evaluating instruction for the purposes of high-stakes decisions would not be an appropriate focus for an initiative. An instrument developed to describe instruction, generally in the context of research on teaching and learning, may provide an empirical basis for practice-based tools. However, describing instruction and managing the performance of instructors are very different goals. Applying a research tool for evaluation may be inappropriate or even counterproductive. A research tool may not even be useful in improving practice if it is too unwieldy to use on a large scale or outside of a research setting.

One option for the foundations would be to invest not only in the development of tools but also in the development of the knowledge needed for implementing those tools into a program of improvement or even evaluation. In this way, the

foundations could help produce increasingly reliable and sophisticated tools that have multiple uses. However, the intended use of tools needs to be made clear, several participants said, given past experience with the inappropriate application of research tools to evaluation. Instrument development should occur in the context of a specific goal to provide direction to that development, they said.

Several participants argued for an open-ended approach to the development of instruments rather than a more prescriptive approach. Different tools capture different aspects of instruction and lend themselves to different purposes. Student reports, observational protocols, teaching inventories, analyses of curricula, and student assessments all reflect aspects of instruction but in different ways. Furthermore, the expanding use of digital technologies in instruction will continue to offer new ways of analyzing instruction.

Tools that provide longitudinal data are especially useful since a baseline can be established and change monitored. Scalability and adaptability are also important considerations so that tools can be used in larger or different settings. NSF is developing indicators for undergraduate instruction in STEM fields to determine whether progress is being made. This approach could be extended to other fields to provide a framework for both description and improvement. In addition, the private sector is investing in the development of tools to measure various outcomes of higher education.

A Taxonomy of Instruction

Another and more immediate option for the foundations would be to support the creation of a taxonomy of instruction—a snapshot of what is happening in higher education today. Today, relatively little is known about such fundamental issues as how much professors lecture, how they structure their courses, or what kinds of interactions they have with students. A description of these attributes of instruction could provide a baseline against which

changes could be measured. It also could range across types of institutions and across the variety of research being carried out today, yielding a comprehensive picture of ongoing activities and the potential for change.

An initiative to develop a taxonomy of instruction could look in particular at digital modalities of instruction, either in association with conventional classes or in formats that are entirely online. As in other areas of instruction in higher education, simply describing the digital modalities being used today would be useful information.

Disciplinary and Cross-Disciplinary Analyses

An initiative could encompass one or a handful of disciplines or a grouping of disciplines, such as the STEM fields, the humanities, or the social sciences. Alternately, it could explicitly foster cross-disciplinary interactions. Both approaches have advantages and disadvantages.

Different disciplines have contrasting ways of knowing, and a cross-disciplinary initiative could explore and take advantage of these different approaches to knowledge. A cross-disciplinary initiative also could extend work done in one discipline or group of disciplines to other disciplines, though that work may need to be adapted to be applicable. Simply providing a venue for faculty members from different disciplines to talk and compare approaches can be valuable, participants said.

Alternately, deep study of one or a handful of disciplines could yield a conceptual framework that ultimately benefits others. It also can greatly benefit that discipline. For instance, a recent emphasis on teaching and learning in undergraduate mathematics has led to increased research on the subject, the founding of a new journal, and new initiatives to increase the quality of undergraduate mathematics instruction. Limiting an initiative to a small number of disciplines and funding multiple projects within those disciplines would also enable the development of different kinds of measures around the same content.

One open question is the extent to which skills learned in one discipline transfer into others or represent more generic skills. Studies of the nature of disciplinary learning, such as the ones being conducted by the Social Science Research Council (see Chapter 2),

could shed light on this issue and have widespread benefits.

The Conditions of Instruction

Particular aspects of higher education are appealing as targets for an initiative because of their prominence. For example, large introductory courses often act as gatekeepers for later courses, and students take them during a period when their attrition from certain majors and from higher education in general is highest. These courses also can contribute to gaps in achievement if some groups of students have a more difficult time getting through them than do others. Examination of models of teaching and learning in these introductory courses could have a major effect on instruction in higher education.

One pressing question is whether an initiative would encompass developmental courses required for students to do college-level work. These courses constitute a substantial part of higher education, several participants pointed out, and are worthy of examination.

Another possible topic for an initiative would be the ways that courses connect or fail to connect across the curriculum. The salience of this issue varies across disciplines, but it has relevance for all departments. An initiative on this topic could result in tools for departments or institutions, as opposed to individual faculty members, to rationalize and align a course of study. It also could help align programs across institutions, given that almost half of students who earn a bachelor's degree have attended two or more institutions of higher education. Such a tool would have to consider the intended curriculum and the enacted curriculum, since the two can vary considerably in any given course.

However, some participants expressed concern about focusing on the curriculum, saying that it could detract from the need to direct attention to instructional measurement. A better approach, they said, would be to think about the teaching and learning of core concepts and how they fit together to create learning progressions over time. In turn, understanding how much students learn in a given course can feed back into curricular decisions, since measures of learning may reveal that students are not mastering the content they need to do well in future courses.

A particular topic for an initiative on instructional measurement would be to explore alternatives to lecturing in college classrooms. Many of these alternatives will require greater management of classroom activities, which has parallels in K-12 instruction but also has distinct features in colleges and universities. For example, higher education has traditionally had a greater emphasis on the creation of knowledge, as exemplified by the widespread participation of undergraduates in research and other forms of scholarly work. Such participation varies by disciplines and by institutions, but some colleges and universities, including two-year colleges, are moving to have even first- and second-year students participate in research as a way of building both discipline-specific and more generic learning.

Learning Outcomes

K-12 education has an infrastructure for evaluation of learning that extends across classrooms, but in higher education most assessment is conducted by the individual instructor, and comparisons across instructors are difficult or impossible. Especially for introductory courses and core requirements, collaboration by faculty members could yield more wide-ranging competencies or learning outcomes that could be assessed, and an initiative could seek to foster or study such collaboration and the resulting learning objectives. However, some participants warned that students can progress rapidly and far in college courses, and the establishment of competencies should not limit their learning. Also, learning in higher education has long-term consequences for future achievement, and some way to measure these outcomes longitudinally would give a more complete picture of learning outcomes.

Measures of instruction differ from measures of student learning, though they are often conflated. However, tools to measure instruction could track student behavioral responses that are associated with learning, including behaviors that occur outside the classroom. In this way, an initiative could measure how instruction induces people to learn without directly measuring student performance.

Alternately, an initiative could specifically examine the learning that occurs outside the classroom and how that learning meshes with what is taught inside the classroom. Such research could take advantage of learning management systems to analyze homework,

interactions among students, and other learning that takes place outside class. It also could explore the differences between residential and non-residential students, who can have very different experiences.

Building Capacity

A major issue in considering initiatives on instructional measurement in higher education is the capacity to do this kind of work. Today, the capacity exists in some places but is weak or nonexistent in others. An initiative could be directed specifically at building capacity, or capacity building could be required as a substantial component of any grant.

The development of instructional measures requires collaboration among people with different expertise, which increases the challenge of building capacity. But people from many different fields can be involved in this work, including sociology, anthropology, social psychology, cognitive psychology, cognitive science, and many other fields. Some investigators have been receiving large grants to do educational research, which can attract the interest of potential collaborators and of institutions.

Several participants made the point that many postdoctoral fellows and new faculty members are very interested in educational research. These individuals were largely trained in leading universities, but they have taken jobs throughout the higher education system. These younger faculty members bring a capacity not only to do research but also to interest institutions in changing educational practices on the basis of that research. They can also help change attitudes among more established faculty members through their efforts and passion for the issue.

Another issue related to capacity is whether to make a request for proposals relatively open ended or more prescriptive. An open-ended request can attract new people to a field who bring with them unexpected ideas and key insights.

The Role of Institutions

Institutions have the potential to play major roles in future research on instructional measurement in higher education, both by serving as venues for that research and by enabling the results of research to be applied within the institutions. An initiative therefore could make special arrangements for institutional involvement.

Grants can be structured in such a way as to provide incentives for institutions to use the results of research to improve instruction. For example, a grant could provide funds for institutional change based on a research project. Another possibility is that a request for proposals could require evidence of institutional support at all levels, thus laying the groundwork for the application of research findings. This approach would require giving reviewers of the proposals strict instructions to observe this criterion.

As part of an initiative, institutions could form groups of instructors within an institution who would work on related problems, or they could form collaborations among institutions to compare experiences. Another possibility would be for an initiative to make arrangements with institutions in advance to serve as partners for researchers. This could result in networks of institutions that are ready to work on instructional measurement with individual research teams, which would give researchers the ability to test ideas on a much broader and more diverse scale than would otherwise be the case. This large-scale testing of research ideas also would be more likely to interest other researchers, journals, and institutions in supporting and extending the work.

The Value of Convenings

The three foundations that sponsored the convening have all had great success with previous meetings that have brought together groups of researchers to discuss joint problems and build collaborations. One or more meetings focused on specific aspects of instructional measurement in higher education could be equally successful. These meetings could look in depth at any of the issues raised in the Chicago convening as a way of jumpstarting work in that area.

Convenings could be held in association with other meetings, whether disciplinary or interdisciplinary, so that people are already gathered in one place. They also could take place before an initiative launches to lay the groundwork for collaborative work or once an initiative is under way to compare problems and progress.

Several participants pointed to the importance of involving junior faculty and young researchers in convenings, since they can bring with them valuable new perspectives. Paying for travel expenses separately from a research grant can enable greater attendance at such meetings.

Capturing the Moment

Higher education is rapidly changing in response to the demographic, technological, and societal changes going on around it. This change is disorienting, but it also has created a unique moment in the history of higher education. An opportunity has opened up to change long-standing practices by bringing new knowledge to bear on existing problems.

Better instruction can make a difference in students' learning, in their persistence in college, and in their subsequent lives. As more is learned about instruction, more students will benefit. Furthermore, given the number of students who drop out of college, even incremental improvements can make a major difference to the nation.

Understanding instruction and learning how to change it are complex problems. But the knowledge and tools exist or are rapidly being developed to make substantial progress on these problems. The opportunities are great, the participants at the convening agreed, if the means and the resolve can be found to grasp them.