WCER Working Paper No. 2013-1
February 2013

A Review of Classroom Observation Techniques in Postsecondary Settings

Matthew T. Hora
Researcher
Wisconsin Center for Education Research
University of Wisconsin–Madison
hora@wisc.edu

Joseph J. Ferrare
Associate Researcher
Wisconsin Center for Education Research
University of Wisconsin–Madison
ferrare@wisc.edu

WCER

Wisconsin Center for Education Research

School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

Copyright © 2013 by Matthew T. Hora and Joseph J. Ferrare
All rights reserved.

Readers may make verbatim copies of this document for noncommercial purposes by any means, provided that the above copyright notice appears on all copies.

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Hora, M. T., & Ferrare, J. J. (2013). *A review of classroom observation techniques in postsecondary settings* (WCER Working Paper 2013-1). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Any opinions, findings, or conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

A Review of Classroom Observation Techniques in Postsecondary Settings

Matthew T. Hora and Joseph J. Ferrare

Classroom observation is a method of directly observing teaching practice as it unfolds in real time, with the observer or analyst taking notes and/or coding instructional behaviors in the classroom or from videoed lessons. Though widely used across the educational spectrum, the technique is far more common—and the methodological sophistication more pronounced—in K–12 schools, where protocols such as the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2007) and the Framework for Teaching (Danielson, 2007) are used for teacher evaluation.

In postsecondary settings, observations are typically less well developed in terms of psychometric testing and conceptual development. Currently, much recent protocol development and use in colleges and universities centers on studying and/or evaluating science, technology, engineering, and mathematics (STEM) teaching. In this paper, we review the basic situations where observation protocols are used in postsecondary settings, key characteristics of these protocols, examples of those that are commonly used, and their strengths and weaknesses. We end with suggestions for a future research agenda.

Use of the Technique

Two applications of classroom observations are the most common: to support professional development, and to assess and/or evaluate teaching quality.

Professional Development

Faculty developers often use classroom observations for coaching and mentoring. Some teaching and learning centers offer services where a trained faculty developer observes a class, often with a structured protocol, and then meets one-on-one with the teacher. Often, faculty developers will not simply conduct a single observation but will integrate pre- and postclass interviews or coaching sessions and provide the instructor targeted feedback. Importantly, this type of coaching can be conducted with a variety of instructional roles, including graduate teaching assistants, tenure-track faculty, and contract lecturers (Sorcinelli & Sorcinelli, 1988). An important aspect of using observations for faculty development is to develop a sense of mutual trust between faculty and analysts (Millis, 1992; Chism, 2007). For instance, the University of Washington Center for Instructional Development and Research (2012) emphasizes the point by stating that center consultants "can provide [instructors] with a neutral, non-threatening perspective on [their] teaching and help [them] reflect on whether, where and how to make changes." Fostering trust is important given evidence that some faculty resent being observed, coached, and generally told how to teach (Eison, 1988; Millis, 1992).

Campus teaching and learning centers often use *unstructured* protocols—where observers take notes during class with no specified direction about what behaviors or facets of teaching to record and in what fashion. This common approach generally yields rich contextualized information about the observed class, yet the variability and lack of standardized data collection procedures negate the possibility of comparing data across raters or even across cases rated by the same observer. In cases where faculty developers use *structured* protocols, those that do not require an evaluative judgment on the part of the observer regarding the quality of instruction and instead describe teaching in concrete terms may aid instructors interested in improving their teaching because "it is much more helpful to identify what [instructors] are doing instead of what [they] should be." (Calvin College, 2012). Additionally, judging instructional quality through the use of protocols that purport to measure teaching efficacy may engender resistance by faculty, particularly if the observers are not disciplinary experts in their field (Centra, 1979).

Assessment and/or Evaluation of Teaching

Assessment and/or evaluation of teaching for the purposes of reviewing employee performance (often called peer evaluations) are commonly conducted for peers within a given academic department (Braskamp & Ory, 1994). One argument in favor of peer observations rather than bringing in an outside evaluator is that only colleagues are in a position to judge the instructor's mastery of the content and the appropriateness of the pedagogy for different types of students (Cashin, 1989; Brent & Felder, 2004). When observation data are used for performance evaluations they are commonly integrated with other types of data as part of a multifaceted evaluation process. As with professional development applications, the protocols used in evaluations vary from unstructured note taking to structured protocols.

Using classroom observations as a form of "high-stakes" assessment and/or evaluation presents a number of problems, a lesson that has been learned in K–12 settings. In these settings, classroom observations have often been used to complement value-added metrics. However, recent research indicates that while value-added is a limited and variable measure of teacher quality (e.g., Rothstein, 2010), classroom observations can be equally unreliable (Guarino & Stacy, 2012). Thus, some have argued that no single data source should be used to evaluate instructional quality (e.g., Shulman, 1987). However, it is not entirely clear what the "value-added" of multiple measures is if such measures are unreliable. At the very least, as higher education moves towards establishing procedures for assessing instructional quality, it is important to learn from the ways in which such procedures have had mixed results in K–12 settings, and in some ways have led to undesirable consequences for teachers' professional autonomy.

Key Characteristics of Observation Protocols

In this section we review the characteristics that distinguish the classroom observation protocols available to researchers and evaluators.

1. Does the protocol evaluate the quality of teaching or merely describe it?

Many observation protocols used in both K–12 and postsecondary settings are designed to evaluate the quality of classroom teaching. This judgment can take the form of underlying scales that purport to capture key aspects of high-quality instruction, which are based on either external criteria or latent variables viewed as proxy measures for student achievement. This approach is attractive for those wanting a single measure of teaching quality, or to determine whether or not teaching measures up to particular standards or expectations (e.g., degree of "reformed-ness"). Thus, in a single measure, analysts can determine whether or not a teachers' practice is aligned with the best practices, as specified by particular criteria (Sawada et al, 2002). Yet caution should be exercised in the use of evaluative protocols for two reasons. First, evaluative measures may turn off or alienate faculty and thus may be of limited utility for professional development purposes. Second, reliability is difficult to attain when analysts are required to not only describe teaching but also judge its quality in real time. A recent review of reliability of evaluative protocols used in K–12 settings found that ratings varied considerably (Guarino & Stacy, 2012), and that rater bias (i.e., pre-existing beliefs about what constitutes high-quality teaching) is a major reason for the high degree of variability observed in the use of these protocols (Cash, Hamre, Pianta, & Meyers, 2012).

With descriptive protocols, the aim is to describe concrete behaviors with no judgment regarding quality or efficacy. In these instances, protocols are designed to document teaching behaviors as they unfold in the classroom with no subjective judgment regarding quality and/or the potential impact on student learning. However, it is inaccurate to say that descriptive protocols do not reflect a priori judgments about instruction, and instrument designers must decide in advance which aspects of teaching to focus upon (e.g., student-teacher interactions, use of instructional technology, etc.).

2. At what level of granularity does the protocol measure classroom teaching?

Another characteristic that distinguishes observation protocols from one another is the level of granularity that is measured in the classroom. Granularity pertains to the level at which teaching behaviors are conceptualized and the temporal frame with which these behaviors are measured. First, any type of teaching behavior can be viewed at different levels of specificity. For example, instructor question asking can be measured at a broad level that simply captures the presence or absence of question asking, or, at a more finely-grained level, that captures specific types of questions that could be posed (e.g., rhetorical questions, algorithmic questions, comprehension questions, etc.). Second, granularity can refer to the temporal frame at which classroom behaviors are measured. At a more coarsely-grained level of temporality, a protocol may require observers to make global judgments at the conclusion of a class period about a particular behavior. At a more finely-grained level of temporality, a protocol may require the analyst to record observed behaviors at regular intervals throughout a class. In the case of question asking, such an approach would entail marking or scoring the presence of questions as they are observed.

3. Does the protocol focus on the instructor, the student, or both?

A key decision for protocol designers is whether to focus on the behaviors of the instructor, the students, or both. In creating their observation protocol, Wainwright, Flick, and Morrell (2003) argue that a focus on only one of the parties will necessarily ignore a key partner in the teaching and learning dynamic (see the Examples of Use section for a description of this and other protocols). In many cases, the student remains invisible during classroom observations. As Good and Brophy (2000) noted in their discussion of effective teaching in K–12 schools:

> "[O]bservers often try to reduce the complexity of classroom coding by focusing their attention exclusively on the teacher … but it is misplaced emphasis. The key to thorough classroom observation is student response. If students are actively engaged in worthwhile learning activities, it makes little difference whether the teacher is lecturing, using discovery techniques, or using small-group activities for independent study." (p. 47)

It is important to note, however, that including both instructor and student behaviors within a single protocol adds to the demands placed on the observer, particularly if the behaviors are conceptualized and measured at a finely-grained level.

4. Does the protocol take the subject matter into account?

Many conceptualizations of the teaching and learning dynamic place a strong influence on the subject matter (e.g., Cohen & Ball, 1999; Neumann, 2012). The gist of this argument is that the nature of the material being taught is a crucial aspect of classroom dynamics without which a complete understanding of the teaching and learning process is not possible. Thus, classroom observations that focus on the outward behaviors of instructors and/or students without capturing the content being taught in a particular class may result in decontextualized accounts of teaching and learning.

5. What degree of inference is required by observers?

In developing the widely used Teaching Behaviors Inventory (TBI), Murray (1997) distinguished between low-inference categories that do not require of the analyst much inference while collecting data (e.g., use of instructional technology), and high-inference categories (e.g., quality of teaching, cognitive engagement of students). Murray noted that low-inference categories are preferable for two reasons: They make data collection more reliable and feasible, and, for professional development applications, teachers may find low-inference data easier to understand and translate into changes in practice.

6. What degree of structure is built into the protocol?

As mentioned, for professional development purposes, some protocols may simply ask the observer to take notes about classroom practices or incidents that catch their attention. Unstructured protocols of this sort may not even indicate to the observer which types of practices or incidents to pay attention to, and in many cases there are no fixed types of response options. In contrast, more structured protocols will include predetermined categories of classroom dynamics that require the observer to pay attention to specific types of teacher and/or student behaviors.

7. Is the observation protocol integrated with other data sources?

Finally, observation protocols can be distinguished by whether or not they are intended to be used in combination with other types of data (e.g., pre-observation instructor interviews or post-observation surveys of student learning), or as a stand-alone instrument. Pre-observation interviews are the most common type of data collected in conjunction with classroom observations, and some argue that it is imperative that the instructors' goals and specific plans be used to help interpret observation data (Chism, 2007).

Examples of Use

In this section we describe protocols that are widely used in postsecondary settings. Protocols developed for professional development applications, of which there are many, often designed for internal use within particular institutions, are not reviewed here. Instead, we review only those protocols that are widely cited in the literature and/or whose development has been documented in a detailed fashion. We organize them into three categories: evaluative protocols that are criterion-referenced, other evaluative protocols, and descriptive protocols. While we do not review the details of the psychometric properties for each protocol, we provide citations where interested readers may find this information.

Criterion-Referenced Evaluative Protocols

Reformed Teaching Observation Protocol (RTOP). The RTOP is widely used among postsecondary researchers and program evaluators, particularly those interested in "reformed" teaching practices. While the RTOP was developed for use in K–12 schools, many postsecondary researchers and program evaluators have adopted it.[1] A key feature of the RTOP is its basis in the constructivist literature about teaching and learning, and, more specifically, the standards-based reform movements in science and math education (Sawada et al., 2000; Sawada et al., 2002). Thus, the RTOP focuses on the extent to which instructors adhere to those practices identified with the inquiry and standards-based literature.

The RTOP consists of 25 items subdivided into three categories of measurement: Lesson design and implementation (five items), content (10 items), and classroom culture (10 items). The content items are further divided into "procedural" and "propositional" knowledge (five items each), and the classroom culture items are subdivided into "communicative interactions" and "student/teacher relationships" (five items each). All items are designed to measure the extent to which various practices are observed in the classroom using a five-point scale ranging

[1] Wainwright, Flick, & Morrell (2003) argue that adopting instruments developed for K–12 settings for use in colleges and universities is unadvisable, largely because the types of student-teacher interactions observed in K–12 classrooms are unlikely to be the same in higher education settings.

from “never occurred” to “very descriptive.”[2] Thus, the RTOP captures elements of both instructor and student behaviors but with more focus on the instructor. The content taught in the observed class is not the protocol’s central focus. One critique of the RTOP’s forced choice response options is the lack of a “not applicable” option, which may result in implausible ratings in some cases (Henry, Murray, & Phillips, 2007). Many published studies use the RTOP, especially in K–12 schools (e.g., Adamson et al., 2003; Roehrig & Kruse, 2005).

The UTeach Observation Protocol (UTOP). Developed at the University of Texas–Austin, the UTOP is described as a protocol to assess the overall quality of instruction but without preference or bias toward any particular way of teaching (Walkington, et al., 2011). The instrument is a modified version of protocols developed by Horizon Research (e.g., the Inside the Classroom protocols) and was designed to serve as an evaluative component of the UTeach program. The protocol maps to expectations for quality instruction by the UTeach program, and the underlying dimensions of the protocol are based on national reform standards such as the National Council of Teachers of Mathematics and National Research Council standards. Because the protocol is criterion-referenced, the developers argue that observers’ opinions and subjective judgments will play no role in the resulting data.

The UTOP comprises 32 classroom indicators across four sections: classroom environment, lesson structure, implementation, and math/science content. Each indicator is rated on a seven-point scale, with “don’t know” and “not applicable” items available in addition to a 1–5 Likert-type scale. The rating choice options (1–5) include two components: the frequency with which a behavior was observed and then the quality of that behavior. For example, one item states: “The classroom environment encourages students to generate ideas, questions, conjectures, and/or propositions that reflected engagement or exploration with important mathematics or science concepts.” Another is: “The majority of students are on task throughout the class.” At the end of the class, the observer makes global judgments about these items in regards to their frequency and quality.

The Oregon Collaborative for Excellence in the Preparation of Teachers (OCEPT) Classroom Observation Protocol (O-TOP). The O-TOP is an instrument designed to study the effects of an instructional intervention in the state of Oregon. To create the O-TOP the researchers focused on two primary categories: teacher behaviors and student behaviors (Wainwright, Flick, & Morrell, 2003). Only 10 items were included in the protocol in order to reduce analyst burden, and a post-observation interview protocol was designed to complement each observation in order to validate the observation data and elicit instructors’ views on their teaching.

The O-TOP uses a four-point Likert style scoring system and includes a “not observed” category. Analysts are instructed to keep notes about the lesson activities during the class and

[2] For information concerning the psychometric properties of the RTOP, see the RTOP reference manual (Piburn et al. 2000).

complete the protocol at the conclusion of the class period. Observers are instructed to view each item “globally” and as “possible indicators,” as opposed to a “required check-off list.” Examples of items include: “This lesson encouraged students to seek and value various modes of investigation or problem solving,” and “Teacher encouraged students to be reflective about their learning.” Each item is associated with a topical focus (e.g., habits of mind, metacognition) and includes several examples of the item immediately underneath the scoring rubric. The O-TOP has been used in studies focused on the OCEPT, including a paper describing the teaching practices of 12 faculty participants in the OCEPT project (Wainwright, Morrell, Flick, & Schepige, 2004).

Other Evaluative Protocols

Teaching Behaviors Inventory (TBI). The TBI has generated the most empirical research and publications based on its use in postsecondary classrooms. The TBI was developed by Harry Murray in the 1980s in order to capture the key aspects of teaching behavior that are hypothesized to be linked to effective instruction and student learning. The TBI is a 60-item protocol containing eight categories of teaching (i.e., clarity, enthusiasm, interactions, organization, pacing, disclosure, speech, and rapport) that were identified over time through scale development procedures (i.e., factor analysis of observation data) and analyses of the predictive validity of scales for various student outcomes (Murray, 1997). Each category comprises several items rated according to a five-point Likert scale (“almost never observed” to “almost always observed”) that the analyst scores at the end of the class. Some versions also have a “+” or a “-” to denote whether or not the instructor should exhibit more or less of a particular behavior.

An extensive body of research correlated TBI ratings with instructor efficacy as measured by student ratings and learning gains on course assessments (see Murray, 1997 for a review). The TBI combines descriptive and evaluative tools. It is descriptive in that the items do not require the analyst to judge the quality of instruction per se but to simply report whether or not a particular teaching behavior occurred (e.g., instructor writes outline of class on board, instructor gestures frequently). It is evaluative in that researchers have found that the categories that constitute the TBI (e.g., clarity, enthusiasm) are often associated with effective teaching. The TBI is commonly found on websites of teaching and learning centers as an easy-to-use protocol for peer review or professional development, though data using the TBI has not appeared in much empirical research in the past decade.

Flanders Interaction Analysis (FIA). The FIA approach was developed for use in K–12 classrooms but has been adopted by some postsecondary researchers and faculty developers. The FIA is based on the assumption that interactions between students and teachers represent a key aspect of effective classrooms (Amidon & Flanders, 1967). The protocol distinguishes between two types of “talk” in the classroom: (a) teacher talk which is either direct (e.g., giving directions) or indirect (e.g., praising, asking questions), and (b) student talk which is considered either as a “response” (i.e., convergent answers to posed questions) or an “initiation” (i.e., divergent questions, or responses to posed questions that depart from the flow of the conversation). Analysts code each type of student and/or teacher-talk every 3–5 seconds, and the intersection between code types represents the interaction in the classroom. These data are entered into a matrix and analyses include the amount of time each party talks during a class, and the nature of student-teacher interactions. While the FIA has mostly been used in K–12 research, Gilbert and Haley (2010) argue for its use in postsecondary settings because data obtained with the protocol are easy to interpret and thus can be useful for professional development.

Descriptive Protocols

Teaching Dimensions Observation Protocol (TDOP). The TDOP is a descriptive observation protocol designed as part of an NSF REESE grant to study the cognitive, cultural, and organizational factors influencing instructional decision making and classroom practice in STEM departments.[3] The TDOP is based on an instrument designed to study inquiry-based science instruction in middle schools (see Osthoff, Clune, Ferrare, Kretchmar, & White, 2009). The original protocol was substantively revised and adapted to specifically fit postsecondary classroom practices.[4] In particular, the adaptation emphasized systems-of-practice theory from distributed leadership studies in K–12 schools (e.g., Halverson, 2003). The TDOP comprises five categories of teacher and/or student behaviors (i.e., teaching methods, pedagogical strategies, student-teacher interactions, available cognitive engagement, and instructional technology). These categories are not intended as latent variables but instead are higher-order descriptive categories of teaching behaviors under which more detailed codes are located. In addition to collecting data for these dimensions of practice, analysts take notes about the class content and other features of the class that are of interest to the observer.

Use of the TDOP requires extensive training and testing interrater reliability (IRR) is of critical importance, and Cohen’s kappa scores for pairs of analysts using the TDOP have varied between 0.652 and 0.926 depending on the category being tested, with the high-inference category of available cognitive engagement being the most problematic. Published studies using TDOP data include analyses of discipline-specific teaching practices in undergraduate courses (Hora & Ferrare, 2012), the relationship of instructors’ prior experiences on their classroom behaviors (Oleson & Hora, 2012), and instructors’ cultural models of teaching and learning (Ferrare & Hora, 2012). The TDOP has also been digitized into a web-based platform so that all data collection, IRR testing, and data management are automated.

Classroom observation rubric. Similarly, Turpen and Finkelstein (2009) developed a descriptive protocol to examine interactive teaching methods in undergraduate physics classrooms. As part of this research, Turpen and Finkelstein (2009) developed an observation protocol that hones in on a single aspect of instruction: student-teacher dialog in the context of clicker use in undergraduate courses. The protocol is based on an activity theory framework that posits that classroom practice consists of the norms governing the behaviors (i.e., use of tools

[3] The Culture, Cognition, and Evaluation of STEM Higher Education Reform project (Grant No. DRL-0814724).

[4] For a description of the initial development of the TDOP, see Hora & Ferrare, 2012.

such as clickers) of both students and teachers. The protocol focuses on the types of questions posed with clickers (e.g., content or logistic), answer options, distribution of student responses, professor wait time, actions during this wait time, and what the researchers call “dialogic interactions.” Published studies using data from this protocol include Turpen and Finkelstein (2009) and Turpen and Finkelstein (2010).

Methodological Considerations

As with any measurement instrument, classroom observations should be designed, tested, and used with careful attention to the methodological quality of the instrument. Reliable instruments will help ensure that trained observers use a given protocol consistently across different teachers and contexts. Of particular importance is ensuring that all sources of measurement error are minimized to the greatest extent possible. Porter (2011) is critical of the overall lack of attention to instrument reliability and validity in higher education, particularly with student surveys such as the National Survey of Student Engagement. This critique of student surveys applies equally to classroom observations, where information about psychometric quality of instruments is sometimes not readily available or questions persist about the reliability and validity of resulting data. While many of the protocols reviewed above have provided documentation of the development and psychometric qualities of the instruments, there has been relatively little outside scrutiny and testing of these instruments. One exception is the in-depth analysis of the RTOP and Inside the Classroom Observation and Analytic Protocol (ITC) instruments conducted by Henry et al. (2007, 2009). In this section we briefly review some of the most important psychometric concerns that should be applied to observation protocols.

Validity

Consideration about an instrument’s validity refers to whether or not the protocol measures what it is supposed to measure or test. Validity theory has been debated and discussed among psychometricians for many decades, with different approaches including criterion validity, content validity, and construct validity.[5] Criterion-based validity refers to how well a score predicts or estimates a measure that is external to the test. This approach requires the identification of a suitable criterion that is operationalized in a clear and concise manner. Concepts related to criterion-validity include predictive and concurrent validity. Predictive validity refers to how well a measure can predict an outcome in the future (e.g., persistence in the major), while concurrent validity refers to how well a measure is correlated with another measure observed at the same time. A key idea underlying criterion-validity is that the criterion must be unambiguous, pertaining to a clearly specified domain (Martinez & Raudenbush, 2009). In critiquing criterion-referenced protocols such as the RTOP, Wainwright, Morrell, and Flick (2003) note that “there is no agreed upon set of practices that represent the mathematics and science standards” (pp. 24–25) and that the expected outcomes are “open to wide interpretation” (p. 25).

[5] For a summary of validity issues in relation to educational measurement see Martinez and Raudenbush (2009).

In contrast, content-based validity refers to how well a measure adequately captures or measures the domain of interest (e.g., teaching, student learning.). The estimation of content validity is often done by experts who vouch for the fact that an instrument does in fact measure the domain under consideration. This approach is common in educational testing where instruments are designed to measure knowledge of a topic, such that experts (e.g., a physics faculty member) can determine if an item does measure competency in that domain. Problems with content-based validity arise when experts have different interpretations of the validity of a measure, a prospect that has given rise to critiques that content-based validity is overly subjective (Kane, 2006). Another model of validity is that of construct validity, which is based on the notion that educational and psychological constructs are latent variables, and which researchers must define and then measure these unobserved constructs in a convincing manner (Cronbach & Meehl, 1955). According to this view, the theoretical underpinning of a construct must be considered in conjunction with test results. In the case of observation protocols, construct validity is the most common type of validity addressed by developers, who generally use factor analytic techniques to ascertain the validity of latent variables built into protocols.

Reliability

Reliability refers to the consistency with which a particular instrument measures the same thing each time data are collected, regardless of who is doing the data collection. Different types of reliability estimates are used to assess the quality of research instruments, including IRR that is particularly important in the case of classroom observations. IRR is used to determine if multiple observers will use the protocol in a similar fashion while scoring the same class period. Researchers use a variety of techniques to assess IRR. The specific approach used to assess IRR is contingent upon the measurement properties of the data. For example, during the initial development of the RTOP—which uses a Likert-type scale—researchers used best-fit linear regression to assess IRR. For instruments such as the TDOP—consisting of dichotomous observations—researchers typically use Cohen’s Kappa and/or one of many measures of similarity for dichotomous data (see Gower, 1985).

Challenges and Opportunities in Classroom Observations

Based on the review of widely used classroom protocols and their main characteristics, coupled with a consideration of methodological and practical issues, we can begin to identify strengths and weaknesses of this technique for measuring postsecondary teaching.

Challenges and Limitations

1. Questions about the reliability of observational data. Perhaps the most significant limitation is the demonstrable difficulty in obtaining reliable data across multiple raters. This limitation applies to all types of observation protocols, but especially to evaluative instruments that place additional burden on analysts to judge the quality and/or efficacy of instruction in real time. Based on a review of the Bill and Melinda Gates Foundation’s “Measures of Effective Teaching” Project that used evaluative protocols, Guarino and Stacy (2012) found that “classroom observation measures of teacher performance are as variable and imprecise as value-added measures based on student test scores and should be considered equally controversial" (p. 9). The researchers found high variability in how different protocols measure a teacher and how a single teacher is rated from analyst to analyst (even with the same protocol). Additionally, pre-existing beliefs about teaching and learning can influence how raters rate teachers they observe (Cash et al., 2012).

2. Conducting observations can be time consuming and disruptive. While many instruments used in postsecondary settings are free of charge, most require a significant investment in time to train analysts, collect data, and then analyze and interpret findings. Furthermore, observations can be disruptive and some faculty resent having their classes interrupted or having an analyst's presence disrupt classroom dynamics (Eison, 1988; Shearer, 2012).

3. Tradeoffs of structured and unstructured protocols. There are benefits unique to using either structured or unstructured protocols, but there also exist challenges related to each type of protocol. Structured protocols, in providing predetermined options and categories for the analyst to observe, necessarily lose some of the richness and depth of classroom dynamics and class content (Millis, 1992). Conversely, unstructured protocols result in highly idiosyncratic data that cannot be compared across individuals or institutions.

Opportunities

1. Observations can provide rich and detailed accounts of teaching. Classroom observations can capture nuances and details of practice and classroom dynamics not otherwise available through other techniques. These accounts also capture aspects of the local organizational and classroom context that are difficult to measure with other techniques. It is only through the observation of action that researchers can obtain richly contextualized accounts of what people do and when.

2. Observation data can be used for multiple purposes. Data obtained from classroom observations are unique in that they can be used productively for a variety of purposes, including professional development, assessment, and research. While survey and interview data are versatile and can be put to many uses, faculty developers typically rely on observation data when coaching and mentoring faculty, largely due to the fact that observation data result in detailed and actionable knowledge that faculty can immediately relate to and build upon in their own practice.

3. Observation-based data can be used to identify aspects of effective teaching. Using observation data to identify high-quality or effective instructional practices is challenging, yet there are opportunities to use these data for these purposes. First, research has linked certain dimensions of teaching practice, such as clarity (i.e., of speech, of class purpose), organization (i.e., of the lesson, of materials), and enthusiasm, to student achievement and learning-related outcomes (e.g., Murray, 1997). Second, Guarino and Stacy (2012) suggest that observations are particularly good at capturing the conveyance or modeling of non-cognitive skills like collaboration or self-expression that have been shown to be important for long-term student success (e.g., Heckman & Rubinstein, 2001). Thus, if used carefully, reliably collected observation data could capture aspects of effective teaching.

Recommendations for Research

As observation protocols for higher education are further developed, it is important to reflect on the role of these instruments in educational reform. As we noted, the use of observation protocols and other "accountability measures" in higher education lags far behind that of K–12 education (see Pianta & Hamre, 2009). However, we suggest that this puts the field of higher education at a distinct advantage because a critical examination of K–12 reform efforts can provide more thoughtful direction to similar efforts in higher education. For example, some researchers have argued that current efforts to reform teaching quality result in a misguided prescriptive approach that "de-skills" teaching and narrows the curriculum (e.g., Darling-Hammond, 2010; Au, 2007). The prescriptive approach to instructional reform views teaching as a set of practices that can be universally adopted across contexts. This recipe-based view, which is not totally unheard of in higher education, ignores factors that educators must regularly negotiate and which may make the implementation of particular practices difficult, such as a lack of adequate training, insufficient resources and support, and student background and learning preferences, among others. Thus, as the use of observation instruments to assess quality becomes more widespread in higher education, it is essential to give careful consideration to the potential impact these efforts will have on the working conditions of instructors as well as the learning experiences of students.

Keeping these concerns in mind, we offer the following suggestions and recommendations for future research.

- **Focus on protocol development and testing.** Any research instrument should be scrutinized in terms of its origins, psychometric properties, and issues related to its use in the field. More work needs to be done on testing the validity and reliability of many observation protocols, especially in relation to IRR.

- **Use evaluative protocols with caution.** With growing evidence that asking observers to both describe and evaluate the quality of teaching leads to unreliable results, the use of these protocols and the interpretation of research findings should be done with caution and skepticism.

- **Support the use of observations for professional development.** While classroom observations can be used for assessment and research applications, perhaps the most effective and defensible use of these data is to support professional development activities.

- **Conduct large-scale observational studies.** Despite limitations, observation-based research provides a robust accounting of classroom practices that are unavailable through other methods. Yet most studies using observations have been conducted at single institutions or classrooms, and large-scale studies would provide policymakers and researchers with a detailed accounting of what is going on in the nation's undergraduate STEM classrooms.

- **Integrate observations with other types of data.** Observation data are best interpreted in conjunction with insights into instructors' goals and plans for the observed class and accounts of student experiences in the classroom and their learning outcomes.

- **Observe multiple venues of instruction.** Classrooms are not the only locations where teaching and learning take place, but laboratories, recitations, and online venues should also be carefully observed and studied. In assessing the educational experience of undergraduate students, accounting for each of these venues would result in a more accurate portrayal of the nature and quality of undergraduate STEM education.

References

Adamson, S. L., Bank, D., Burtch, M., Cox, F. III, Judson, E., Turley, J. B., Benford, R., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching, 40*(10), 939–957.

Amidon, E. J., & Flanders, N. A. (1967). *The role of the teacher in the classroom: A manual for understanding and improving teachers' classroom behavior.* (Rev. ed.) Minneapolis, MN: Association for Productive Teaching.

Au, W. (2007). High-stakes testing and curricular control: A qualitative meta-synthesis. *Educational Researcher, 36*(5), 258–267.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work*. San Francisco, CA: Jossey-Bass Publishers.

Brent, R., & Felder, R. M. (2004). A protocol for peer review of teaching. Proceedings of the American Society for Engineering Education Annual Conference. Washington, DC: American Society for Engineering Education.

Calvin College (2012). Teaching development evaluation tools. Teaching Behaviors Inventory. Retrieved from http://www.calvin.edu/admin/provost/teaching/instructional/tools/behaviors.htm

Cash, A. H., Hamre, B. K., Pianta, R. C., & Meyers, S. S. (2012). Rater calibration when observational assessment occurs at large scales: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529–542.

Cashin, W. E. (1989). *Defining and evaluating college teaching* (IDEA Paper No. 21). Manhattan, KS: Center for Faculty Evaluation and Development. Kansas State University.

Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement*. San Francisco, CA: Jossey-Bass.

Chism, N. (2007). *Peer review of teaching: A sourcebook* (2nd ed.). San Francisco, CA: Jossey-Bass.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement*. Consortium for Policy Research in Education Rep. No. RR-43. Philadelphia: University of Pennsylvania, Graduate School of Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

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

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future.* New York, NY: Teachers College Press.

Eison, J. (1988). Designing effective peer observation programs. *Journal of Staff, Program and Organization Development*, *6*(2), 51–59.

Ferrare, J. J., & Hora, M. T. (2012). *Cultural models of teaching and learning: Challenges and opportunities for undergraduate math and science education* (WCER Working Paper No. 2012-8). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Gilbert, M. B., & Haley, A. (2010). Faculty evaluations: An alternative approach based on classroom observations. *Faculty Focus*. Retrieved from http://www.facultyfocus.com/articles/faculty-evaluation/faculty-evaluations-an-alternative-approach-based-on-classroom-observations/

Good, T., & Brophy, J. (2000). *Looking in classrooms*. (8th Ed). New York, NY: Longman.

Gower, J. C. (1985). Measures of similarity, dissimilarity, and distance. *Encyclopedia of statistical sciences (volume 5)* (pp. 397–405). New York, NY: Wiley.

Guarino, C., & Stacy, B. (2012). *Review of gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. National Educational Policy Center. Boulder, CO.

Halverson, R. (2003). Systems of practice: how leaders use artifacts to create professional community in schools. *Educational Policy Analysis Archives, 11*(37), 1–35.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *The American Economic Review, 91*(2), 145–149.

Henry, M. A., Murray, K. S., & Phillips, K. A. (2007). *Meeting the challenge of STEM classroom observation in evaluating teacher development projects: A comparison of two widely used instruments.* M.A. Henry Consulting, LLC. St. Louis, MO.

Henry, M. A., Murray, K. S., Hogrebe, M., & Daab, M. (2009). *Quantitative analysis of indicators on the RTOP and ITC observation instruments.* M.A. Henry Consulting, LLC. St. Louis, MO.

Hora, M. T. & Ferrare, J. J. (2012). Instructional systems of practice: A multi-dimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences*. doi: 10.1080/10508406.2012.729767

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Martinez, A., & Raudenbush, S. W. (2009). Validity studies involving measures of classroom quality. W.T. Grant Foundation. Retrieved from http://www.wtgrantfoundation.org/resources/measures-of-social-settings

Millis, B. J. (1992). Conducting effective peer classroom observations. *To Improve the Academy*. The Professional and Organizational Development Network, volume 11.

Murray, H. G. (1997). Effective teaching behavior in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171–204). New York, NY: Agathon Press.

Neumann, A. (2012). *Presidential address.* Unpublished paper presented at the 2012 annual meeting of the Association for the Study of Higher Education. Las Vegas, NV.

Oleson, A., & Hora, M. T. (2012). *Teaching the way they were taught? Revisiting the sources of teaching knowledge and the role of prior experience in shaping faculty teaching practices* (WCER Working Paper No. 2012-9). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Osthoff, E., Clune, W., Ferrare, J., Kretchmar, K., & White, P. (2009). *Implementing immersion: Design, professional development, classroom enactment and learning effects of an extended science inquiry unit in an urban district.* Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Educational Research.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom Assessment Scoring System—CLASS*. Baltimore, MD: Brookes.

Piburn, M., Sawada, D., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). *Reformed teaching observation protocol (RTOP)*. Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education, 35*(1), 45–76.

Roehrig, G. H., & Kruse, R. A. (2005). The role of teachers' beliefs and knowledge in the adoption of a reform-based curriculum. *School Science and Mathematics, 105*(8), 412–422.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175–214.

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.

Sawada, D., Piburn, M., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). Reformed Teaching Observation Protocol: Technical Report No. IN00-1. Tempe, AZ: Arizona State University.

Shearer, T. M. (2012, March 6). A pleasing observation. *The Chronicle of Higher Education.* Retrieved from http://chronicle.com/article/A-Pleasing-Observation/131074/

Shulman, L. S. (1987). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan*, 39–44.

Sorcinelli, M. D., & Sorcinelli, G. (1988). *Effective use of time in the classroom: An instructor's guide.* Chicago, IL: Northwestern University.

Turpen, C., & Finkelstein, N. D. (2010). The construction of different classroom norms during Peer Instruction: Students perceive differences. *Physical Review Special Topics–Physics Education Research 6*(2), 020123.

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: Variation in physics professors' implementation of peer instruction. *Physical Review Special Topics–Physics Education Research, 5,* 020101.

University of Washington. (2012). Center for Instructional Development and Research (CIDR). Retrieved from http://depts.washington.edu/cidrweb/

Wainwright, C., Morrell, P. D., Flick, L., & Shepige, A. (2004). Observation of reform teaching in undergraduate level mathematics and science courses. *School Science and Mathematics, 104*(7), 322–335.

Wainwright, C. L., Flick, L. B., & Morrell, P. D. (2003). Development of instruments for assessment of instructional practices in standards-based teaching. *Journal of Mathematics and Science: Collaborative Explorations, 6*(1), 21–46.

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2011). Development of the UTeach Observation Protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach Preparation Program (UTeach Technical Report 2011-01). UTeach Natural Sciences, University of Texas at Austin.