How Do We Teach Thee? Let Me Count the Ways
A Syllabus Rubric with Practical Promise for Characterizing Mathematics Teaching

Sandra Laursen
University of Colorado Boulder

Tim Archie
IDEA Center

Good methods to characterize teaching are needed to describe both current status and changes in teaching practice, and to link student outcomes to particular instructional practices. Such methods are understudied and thus the relative merits of different methods are not well understood. As part of a study examining multiple methods for characterizing teaching in college mathematics, we analyzed syllabi using three rubrics. Syllabi are authentic course artifacts that reflect course design and instructor's intentions; they are readily available from instructors. One of these rubrics, an evaluative rubric known as Measuring the Promise (MtP), proved useful in distinguishing courses taught by a sample of seven early-career instructors and a comparison sample of experienced active learning practitioners. Good correlation of MtP scores with observation scores using the well-established Reformed Teaching Observation Protocol suggest that the MtP may be a useful alternative to costly and time-consuming observations.

Keywords: syllabi, observations, measurement of teaching

For many studies of higher education, it is important to characterize teaching: to describe teaching practice within an institution or across a discipline, to relate student outcomes to teaching practices, or to measure change in teaching practice over time. However, we do not yet have a good understanding of what can be learned from different approaches to describing teaching, nor the strengths and limitations of these approaches (AAAS, 2013). Yet descriptions of teaching practice form the foundation for claims about the effectiveness of various instructional practices, and thus also the basis of many current efforts to change such practices.

Our study is motivated in particular by a need for good methods to measure change in teaching after professional development of college instructors (CIPD). Funders and institutions need good evidence to determine whether, how, and in what forms, professional development may be a good investment. While better student learning is the ultimate goal, measuring student outcomes directly is not always possible, and faculty may need time to gain skill in the new techniques before student learning is measurably improved.

An alternative is to measure the degree to which faculty implement evidence-based teaching practices, such as those introduced in CIPD programs, together with a "golden spike" approach that links these practices to prior studies that demonstrate how these teaching practices influence student outcomes (Brown Urban & Trochim, 2009). Because teaching is a complex activity, this measurement, too, is challenging. Observational studies are viewed as most objective but are complicated and costly, while well-validated instructor surveys are not yet available (Felder, Brent & Prince, 2011). Course artifacts offer a wealth of teaching-related material that is readily accessible and authentic in representing the instructor's actual work, rather than her later representation of it—but it is less clear how to make inferences from these materials about the instructor's classroom instruction, teaching decisions or philosophy. To understand whether and how CIPD improves STEM teaching, we need valid and reliable measures of teaching practice that can be used to learn whether and how instructors' practices and choices change after CIPD.

This methods development study was exploratory by design, involving close examination of teaching practices in seven undergraduate mathematics courses taught by early-career instructors. We compared insights gained from a full suite of teaching measures: student and instructor surveys, observations, and coding of course syllabi and assessment items. The study focused on the potential of these measures to detect change in instruction, such as the changes that might result from professional development, but we did not directly study change. Thus the study is a close look at a small sample. We sought to identify methods that are both informative and practical for measuring teaching practice, and to make judgments about when and how these methods may be useful—alone or combined—in characterizing teaching. This work thus offers advice to researchers and evaluators to make intelligent choices for their own studies.

Our broad research questions were:

1. What are the affordances and limitations of behavior-oriented and outcome-oriented observations, faculty self-reports, student reports and classroom artifacts as methods for characterizing teaching in undergraduate mathematics classrooms?
2. What are the differences among characterizations of teaching in undergraduate mathematics classrooms that are based on these distinct types of measures?

Our study explored both descriptive and evaluative measures of teaching. Evaluative or "outcome-oriented" measures examine the aims and effects of instruction rather than the choesn activities, rating instruction against a specific standard for "good teaching," thus differing from strictly descriptive or "behavior-oriented" measures. Here we focus on two evaluative measures, a widely used observation protocol and a rubric for analyzing course syllabi. These two methods, observation and syllabus analysis, represent extremes of simplicity and complication in the logistics and invasiveness of data gathering and the demands of data analysis, so it is interesting to compare their potential as measures for characterizing mathematics teaching.

Study Sample

We recruited instructors from MAA Project NExT (PN), New Experiences in Teaching, a professional development program for early-career mathematics instructors. Working with early-career instructors whose teaching methods were still developing, we were able to observe a range of teaching behaviors and skill levels that are likely comparable to those encountered in studying professional development outcomes for other instructors of undergraduate mathematics. We also sought to gather data from courses for varied student audiences and at varied curricular levels.

We solicited study participants through selected PN listservs, inviting respondents to read an online description of the study, review the consent form, and complete a pre-screening questionnaire about their courses and academic calendars. Ultimately, seven instructors took part.

This sample included variety across instructors, courses, departments and institutions. The *instructors* included 4 women and 3 men. Six were white and one was multi-racial; none were Hispanic. Five held tenure-track positions, one a long-term instructorship and one a visiting position. Their teaching experience (including TA work) was 3 to 10 years. Their *courses* spanned the early (3), middle (2) and late (2) undergraduate curriculum and a range of mathematics, STEM, non-STEM, and pre-service teaching audiences. Class size was small, 8-28 students. The six semester-based and one quarter-based courses met for 35-56 hours each. The courses were situated in *departments* that granted bachelors (3), masters (2) or doctoral (2) degrees as the highest mathematics degree. The *institutions* were diverse in geography, institution type and student enrollment; two had high minority student populations.

Study Methods

Working with each instructor, we selected a single target course from which we collected all data. With this small sample we could not generalize about any particular instructional setting, but we could test the applicability of these methods in varied settings. Each instructor contributed data to support six study components:

a) Video observations of ten class periods, coded with descriptive and evaluative protocols
b) Instructor survey, end of course, self-reporting teaching practices
c) Instructor interview
d) Student surveys, end of course, including both descriptive and evaluative items
e) Course syllabus, coded with descriptive and evaluative schemes
f) Course-specific subset of assessments identified from the syllabus.

In this report, we focus on methods (a) and (e), using observation data as a benchmark to evaluate a syllabus rubric as a possible tool to characterize teaching with relatively low effort. The rubric also offers good potential as a tool for formative evaluation or for providing feedback to instructors as professional development. Elsewhere we discuss results from other methods.

Why Study Syllabi?

As Eberly, Newton and Wiggins (2001) point out, the syllabus is both "the initial communication tool that students receive" and "the most formal mechanism for sharing information with students" (p. 1) about a course. Ideally, it is "a learning-focused document that communicates clearly and compellingly what students will gain from the course, what they will do to achieve the promise it lays out, how they will know whether they are getting there, and how to best go about studying" (Palmer, Bach & Streifer, 2014b). If, as these authors argue, the syllabus serves as a "framework for designing meaningful learning environments," then it follows that we may be able to diagnose the presence of such intentional and student-focused design from syllabi. Syllabus analysis offers advantages for data collection too: the syllabus is already written and widely available, so collecting it requires low instructor effort; it is brief, thus rapid to analyze; and it is public, so gathering it should not require special IRB permission.

Analysis of Syllabi

We tested three syllabus analysis tools found in the literature. The SPROUT-S protocol is a descriptive protocol developed at UC Irvine to study the relationship of student academic outcomes to the use of "promising instructional practices" in undergraduate STEM courses (Reimer et al., 2016). The Penn State Engineering Education protocol (Zappe et al., 2015, 2016) is also descriptive, a list of 47 research-based practices in engineering education that is drawn from synthetic work by Hattie (2008) and Chi (2009) examining factors related to student achievement and student learning. While in principle a descriptive approach could assist in analyzing the presence or prevalence of certain instructional methods or philosophies, we found neither of these descriptive tools useful for our study, as we will describe in our presentation.

Instead, the analysis presented here focuses on an evaluative rubric, Measuring the Promise (MtP), a validated rubric from faculty developers at the University of Virginia (Palmer, Bach & Streifer, 2014a). As a rubric, it defines a coherent set of criteria and describes different levels of performance quality on the criteria (Brookhart, 2013). It could thus be used in formative evaluation—to guide professional development on course design—as well as a summative assessment tool. It is designed for use in any discipline.

The holistic and evaluative rubric is strongly literature-based (Palmer, Bach & Streifer, 2014b). The full rubric uses 16 items grouped into five categories: learning goals and objectives, assessment, schedule, classroom learning environment, and learning activities. Each item is rated gold, silver or bronze to indicate its relative importance in the scoring rubric—which is in turn based on its expected influence on student outcomes—and the rater classifies the strength of evidence about each as strong, moderate, or low. The authors specify their assumptions about the rater's background knowledge and provide examples of the kinds of evidence used to assess each criterion. Raters must understand Fink's (2013) significant learning goals, distinguish learning goals and objectives, and assess alignment of goals, objectives, activities and assessments in course design. With this background, modest training is required to achieve interrater reliability.

To emphasize the presence and quality of essential features, the scoring system weights both the features (3,2,1) and the evidence for them (2,1,0) (CTE, 2017). The maximum score is 58: a 'learning-focused' syllabus will score in the range of 41 and higher; a 'content-focused' syllabus at 18 or below. Syllabi in between are called 'transitional.'

For the early-career instructors, syllabi from seven courses taught in 2015-16 were coded using the MtP, plus two more syllabi representing earlier versions of the same course. For comparison, 12 syllabi were coded for courses taught in 2016-17 by 'expert' instructors known to use strongly student-centered teaching approaches.

Observation Coding with RTOP

The Reformed Teaching Observation Protocol (Sawada et al., 2002) was developed to evaluate the degree of "reform" toward student-centered teaching in science. RTOP's 25 items assess the degree to which the classroom is learner-centered in five categories: lesson design and implementation, propositional knowledge, procedural knowledge, communicative interactions, and student-teacher relationships. RTOP scores in K12 classrooms have been correlated with student achievement and used to assess change as a result of CIPD. It requires significant training and nuanced judgment against externally defined criteria for effective teaching.

We collected observation data for 8-10 class sessions taught by the early-career instructors using a portable video camera shipped to instructors and mounted behind students, facing forward. We followed RTOP data analysis methods outlined by Ebert-May et al. (2011). Five items forming five subscales are scored on a Likert scale 0-4. After initial training, six videos were randomly selected and coded by three raters. We tested interrater reliability by computing intraclass correlation coefficients (ICC) and achieved an acceptable ICC level (>0.80) for overall RTOP scores and for each subscale.

One rater then coded five randomly chosen class sessions for each of the seven courses. We computed RTOP and subscale scores for each class and calculated means of the five observations for each course. Total scores of 0-100 are classified into five categories using score breakpoints of 30, 45, 60, and 75, where scores ≤30 are interpreted as "straight lecture" and scores >75 as "active involvement in open-ended inquiry," with intermediate scores placed along a spectrum of interaction and inquiry. Because the classifications are framed in language common in discussing inquiry-based science, such as carrying out experiments, we re-interpreted the classifications for college mathematics, considering inquiry-oriented processes such as preparing, explaining and critiquing proofs or problem solutions, and explicitly considering alternative solutions.

Results

Figures 1 and 2 show MtP scores for syllabi from courses taught by early-career (designated PN for Project NExT alumni) and experienced instructors (EX). As a group, total scores for courses of experienced instructors (mean 32 ± 12) were not statistically distinguishable from those of early-career instructors (mean 30 ± 13). However, 3 of 7 courses by early-career instructors, vs. only 1 of 9 courses of experienced instructors, were rated as content-centered. In both figures, arrows link pairs of syllabi for a single course that represent a particular instructor's historical and current practices; these pairs are discussed further below.

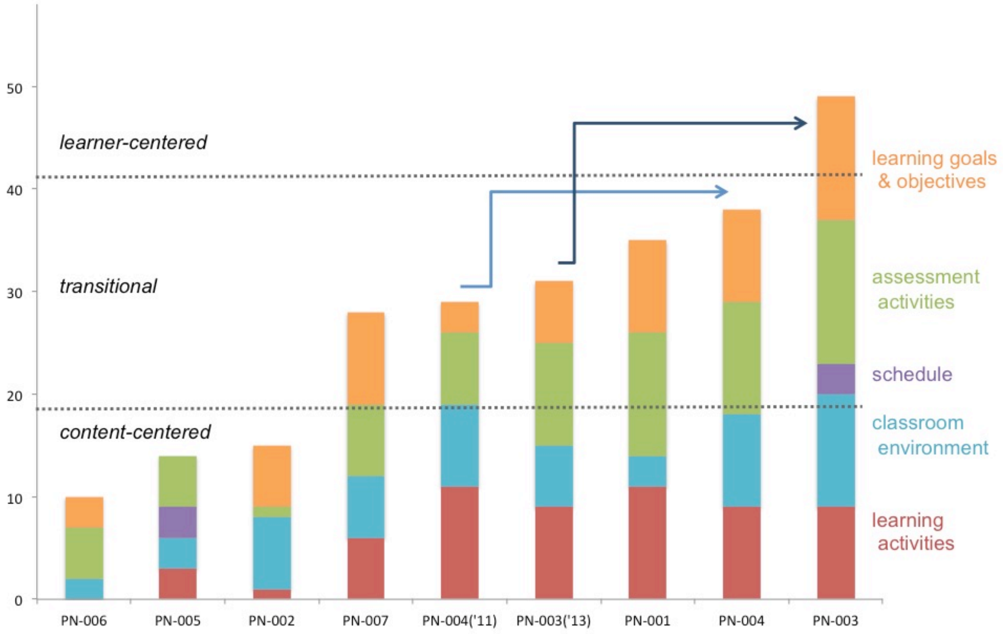

Figure 1: MtP syllabus ratings for early-career instructors, by item category.

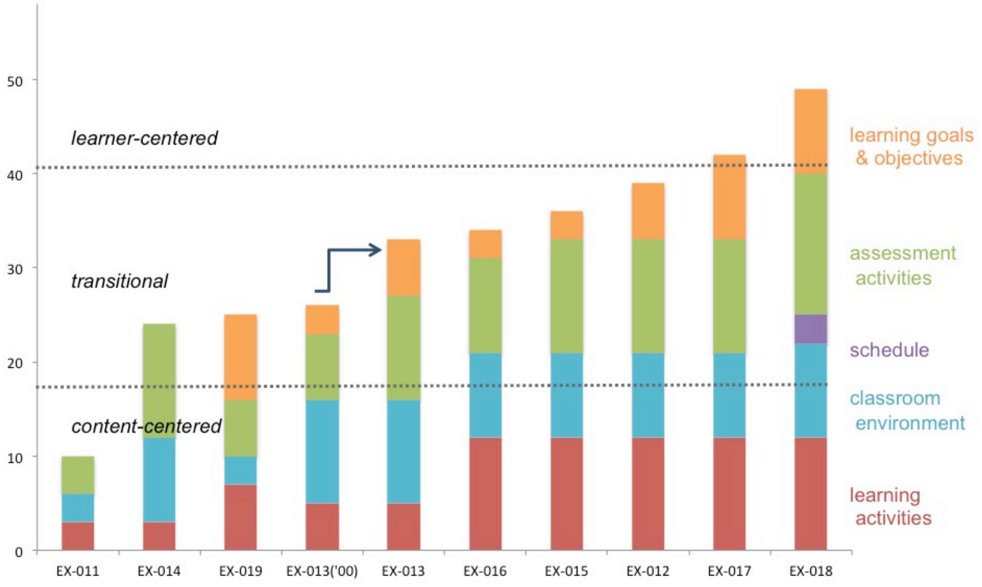

Figure 2: MtP syllabus ratings for experienced instructors, by item category.

Comparing subscores in detail reveals some more distinctive patterns of difference between early-career (PN sample) and experienced (EX sample) instructors. Scores on *learning goals and objectives* were slightly higher among early-career instructors (mean of 6.3 for PN vs 4.8 for EX, of 12 points maximum). Scores on *assessment activities* were fairly high for both groups (8.0 PN, 10.1 EX, of 12). In mathematics, homework is frequently assigned and used to give formative feedback. Scores on the *schedule* were low across the board (0.7 PN, 0.3 EX, of 6). Scores for the *classroom environment* were higher among experienced instructors (8.3 EX, 6.1 PN, of 12). Scores for *learning activities* were moderate to high among both groups, but somewhat more consistent among experienced instructors (6.6 PN, 8.3 EX, of 12).

Comparison of current/historic pairs of syllabi (marked by arrows in Figures 1 and 2) for individual instructors shows positive change over time for the three cases available that reflect the start of a teaching career as compared to now. Other data (not shown) suggests that very experienced instructors find a teaching groove and stick to it; their MtP score does not change.

These data suggest that the MtP has good discrimination on aspects of course planning that may differ between instructors of differing experience and/or skill. To relate the syllabus score to a measure based on actual classroom practice, we compared the MtP scores to mean scores on the RTOP (Figure 3) for the PN sample, for which we had both data types.

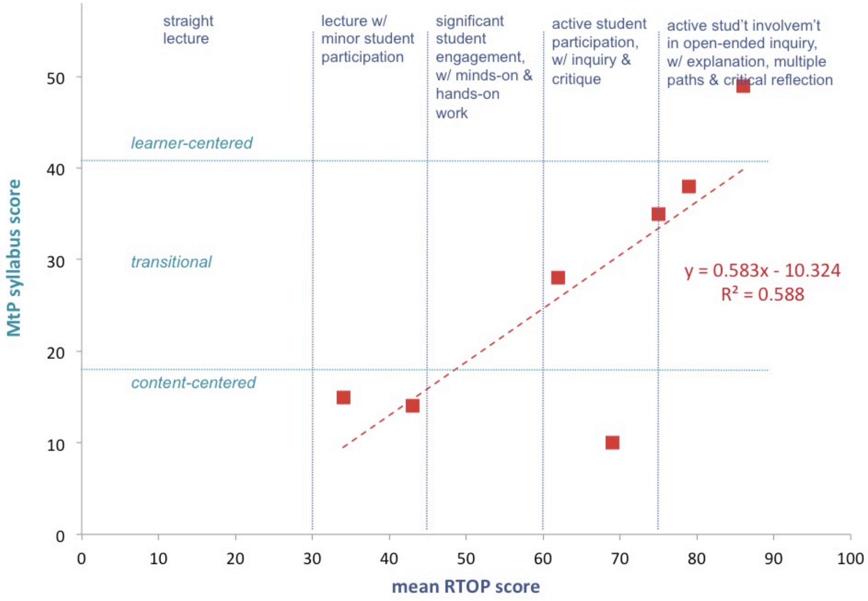

Figure 3: Correlation of MtP syllabus scores with RTOP scores, including score classifications

Five of seven courses score in the active learning ranges of the RTOP scale, while two scores are described as interactive lecture. Moreover, MtP syllabus scores correlate well (R=0.59) with mean RTOP scores. In general, low MtP scores reflect underdeveloped or incomplete syllabi; actual classroom practice may be more interactive and inquiry-driven than is shown in the document. For example, the outlying point in Figure 3 represents a course where we observed inquiry activities, peer to peer collaboration, and extensive use of multiple representations of mathematical ideas, yielding a medium-high RTOP score, but the syllabus was disorganized, overly rule-oriented, and uninviting to the learner as an entrée into the discipline.

Discussion

In general experienced active-learning instructors scored high on the MtP items for *classroom environment, assessment* (especially formative assessment), and *learning activities*, thus the rubric does show evidence of their student-centered orientations. Some of the early-career instructors were also IBL users, and their syllabi reflect aspirations toward the same student-centered practices. The reverse trend for *learning goals and objectives*, where early-career instructors scored higher, may reflect greater exposure of early-career instructors to learning goal-setting through professional development or exposure to RUME work. In addition, interview data revealed that some departments had set common learning objectives for particular courses; thus the learning objectives may be inherited rather than originated by the instructor.

Low scores on *schedule* arose because information on the choice and sequence of topics was commonly missing in syllabi from courses using inquiry-based learning (IBL), which reduced the score on items related to the intellectual organization or conceptual flow of the course and its pacing. It is also possible that college mathematics instructors take course content as canonical, whether or not they use IBL. For instance, with high consensus about what goes into a Calculus 1 course, and with many students required to take it, instructors may not think to justify to students their choices about the selection and sequencing of big ideas. Content sequencing may also be seen as given if it is decided departmentally and used by all who teach the same course.

In the cases where we could compare two versions of a course, the observed changes in MtP score suggest that the rubric is sensitive to change over time in instructors' practice.

The strength of correlation between the MtP and the RTOP is somewhat surprising, given that the MtP is based solely on the written plan for the course, and the RTOP rates instruction as implemented in class. However, both are holistic measures that focus on instructional design and set standards for 'good teaching' that are literature-based and thus aligned in many respects. Coding of a separate observation sample with RTOP will tell us if this correlation is robust.

Both our study groups, early-career and experienced, were more learning-focused than a general university instructor population (CTE, 2017). Mean MtP scores exceeded the median pre-test score for faculty who enrolled in a week-long institute on course design—but were lower than the post-test scores for those faculty. However, our study samples do not represent college math instructors nationally; they were volunteers already participating in educator communities.

Conclusions and Implications

The MtP emphasizes instructors' design choices, as reflected in their syllabus, and focuses on clarity and alignment of the course design. It does not attempt to judge how well a course is executed but does capture elements of how instructors view students, teaching, and their subject. The rubric offers high face validity, due to its grounding in instructional design literature, and good discrimination, due to the weighted scoring system. Moreover, syllabus analysis with the MtP offers advantages for both gathering and analyzing data. In these ways, we find the MtP rubric a tool with significant potential to be useful in studies of teaching or change in teaching.

The correlation of MtP with RTOP in this small data set is particularly intriguing, because it suggests the potential of MtP to substitute, in some studies, for time-consuming and costly observations. Like the RTOP, MtP does require specialized expertise to apply, but it is well supported with coder training materials. Analysis of a limited data set suggests that the MtP has promise in detecting change in individuals' practice over time, but may be less useful in characterizing entire groups of instructors, due to the variability within groups.

References

American Association for the Advancement of Science (AAAS) (2013). Describing & measuring undergraduate STEM teaching practices: A report from a national meeting on the measurement of undergraduate science, technology, engineering and mathematics (STEM) teaching, December 17-19, 2012. Washington, DC: AAAS. Accessed 8/15/17 from http://ccliconference.org/files/2013/11/Measuring-STEM-Teaching-Practices.pdf

Brookhart, S. (2013). What are rubrics and why are they important? Chapter 1 in *How to create and use rubrics for formative assessment and grading*. ASCD. http://www.ascd.org/publications/books/112001/chapters/What-Are-Rubrics-and-Why-Are-They-Important%C2%A2.aspx

Brown Urban, J., & Trochim, W. (2009). The role of evaluation in research-practice integration Working toward the ''golden spike.' *American Journal of Evaluation*, 30(4), 538-553.

BYU Idaho (2012). Dee Fink's Taxonomy of Significant Learning. https://www.byui.edu/outcomes-and-assessment-old/the-basics/step-1-articulate-outcomes/dee-finks-taxonomy-of-significant-learning

Center for Teaching Excellence (CTE) (2017). Syllabus Rubric. University of Virginia. http://cte.virginia.edu/resources/syllabus-rubric/

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73-105.

Eberly, M. B., Newton, S. E., & Wiggins, R. A. (2001). The syllabus as a tool for student-centered learning. *The Journal of General Education*, *50*(1), 56-74.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development. *BioScience*, 61 (7), 550-558.

Felder, R. M., Brent, R., & Prince, M. J. (2011). Engineering instructional development: Programs, best practices, and recommendations. *Journal of Engineering Education*, 100(1), 89-122.

Fink, L. D. (2013). *Creating significant learning experiences: an integrated approach to designing college courses* (2nd ed.). San-Francisco: Jossey-Bass.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research*, *1*(1), 7-34.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hoover, M. A., & Pelaez, N. J. (2008). Blood circulation laboratory investigations with video are less investigative than instructional blood circulation laboratories with live organisms. *Advances in Physiology Education*, *32*(1), 55-60.

Palmer, M., Bach, D., & Streifer, A. (2014a). *Measuring the Promise: A Valid and Reliable Syllabus Rubric. Guide to Assessing the Focus of Syllabi*. University of Virginia Teaching Resource Center.

Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014b). Measuring the promise: A learning-focused syllabus rubric. *To Improve the Academy*, *33*, 14-36. doi:10.1002/tia2.20004

Reimer, L. C., Schenke, K., Nguyen, T., O'Dowd, D. K., Domina, T., & Warschauer, M. (2016). Evaluating promising practices in undergraduate STEM lecture courses. *Russell Sage Foundation Journal of the Social Sciences*, *21*(2), 212-233.

Sawada, D., et al. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, *102*(6), 245-253.

Zappe, S. E., Hochstedt, K. S., & Litzinger, T. A. (2015). Exploration of course syllabi as a potential source of data on the use of evidence-based instructional practices in engineering courses. Research in Engineering Education Symposium (REES2015), Dublin, Ireland, July 13-15.

Zappe, S. E., Hochstedt, K. S., Merson, D., Schrott, L., & Litzinger, T. A. (2016). Development and implementation of quantitative methods to study instructional practices in engineering programs. *International Journal of Engineering Education*, *32*(5A), 1942-1959.