Exploring the Complementarity of Measures of Instructional Practices

Lu Shi,[1] Maia Popova,[2] Robert M. Erdmann,[3] Anthony Pellegrini,[4] Victoria Johnson,[5] Binh Le,[6] Trina Popple,[7] Zachary Nelson,[8] Molly Undersander Gaston,[9] and Marilyne Stains[1]*

[1]Department of Chemistry, University of Virginia, Charlottesville, VA 22904; [2]Department of Chemistry and Biochemistry, University of North Carolina at Greensboro, Greensboro, NC 27412; [3]Campus Learning Data and Technology, University of Minnesota Rochester, Rochester, MN 55904; [4]Applied Research Associates, Panama City, FL 32401; [5]Illinois College of Optometry, Chicago, IL 60616; [6]ProScribe, Omaha, NE 68007; [7]Gretna High School, Gretna, NE 68028; [8]Lincoln North Star High School, Lincoln, NE 68504; [9]Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588

ABSTRACT

The assessment of instructional quality has been and continues to be a desirable, yet difficult endeavor in higher education. The development of new teaching evaluation frameworks along with instruments to measure various aspects of teaching practices holds promise. The challenge rests in the implementation of these frameworks and measures in authentic settings. Part of this challenge is for instructors, researchers, and administrators to parse through and select a meaningful set of tools from the plethora of existing instruments. In this study, we aim to start clarifying the landscape of measures of instructional practice by exploring the complementarity of two existing instruments: the Classroom Observation Protocol for Undergraduate STEM (COPUS) and the Learner-Centered Teaching Rubrics (LCTR). We collected classroom observations and course artifacts from 28 science instructors from research-intensive institutions across the United States. Results show the need to use both instruments to capture nuanced and comprehensive description of a faculty member's instructional practice. This study highlights the messiness of measuring instructional quality and the need to explore the implementation of teaching evaluation frameworks and measures of instructional practices in authentic settings.

INTRODUCTION

The development and implementation of measures that provide reliable and valid evidence for the quality of teaching are critical for the improvement of the learning experiences provided to students enrolled in science, technology, engineering, and mathematics (STEM) courses. Such evidence is essential for instructors to assess the effectiveness of their teaching and monitor changes as they work to improve students' learning outcomes in their courses, and for them to be objectively recognized during evaluation and promotion proceedings. Moreover, institutions could leverage validated measures to inform the implementation of targeted support structures for instructors and employ them to demonstrate to accreditation agencies and other stakeholders the quality of the learning environments they provide to their students. Unfortunately, the measures currently employed at most institutions are inadequate. An analysis of the promotion and tenure policies of 51 research universities revealed that the most common measures of teaching effectiveness were student evaluations and peer classroom observations (Dennin *et al.*, 2017). Indeed, at most universities, the primary metric is student course evaluations (Shao, Anderson, and Newsome, 2007; Henderson *et al.*, 2014) despite extensive evidence of their inappropriateness. For example, it is well established that student evaluations are influenced by various factors not related to instructional quality, including the instructor's identities, student

Abdi Warfa, *Monitoring Editor*

Submitted Mar 21, 2022; Revised Sep 27, 2022; Accepted Nov 7, 2022

CBE Life Sci Educ March 1, 2023 22:ar1

DOI:10.1187/cbe.22-03-0047

*Address correspondence to: Marilyne Stains (mstains@virginia.edu)

demographics, and subject area (e.g., Hornstein, 2017; Fan *et al.*, 2019; Heffernan, 2022). Peer classroom observations are also problematic, because they are often conducted without guidelines or observation protocols aligned with effective teaching practices, and the peers observing may not be well versed in these practices. A survey of more than 1000 instructional staff across eight institutions supported by the Association of American Universities STEM Education Initiative indicated that only 13% of respondents described the quality of the evidence collected to measure effective teaching as high (Dennin *et al.*, 2017). There is thus an incredible need to develop measures and frameworks for the assessment of teaching that enable instructors' growth toward targeted effective instructional practices (Bradforth *et al.*, 2015; Wieman, 2015; Dennin *et al.*, 2017).

Discipline-based education researchers have had a keen interest in measuring instructional practices to enhance teaching evaluation processes and to monitor changes resulting from the implementation of instructional reform efforts. Consequently, they have developed tools such as surveys, observation protocols, and rubrics that aim to provide valid and/or reliable characterizations of an instructor's teaching practices.

Surveys can be collected at scale with minimal time required for input and analysis, while capturing a range of practices such as in-class behaviors, assessment practices, and out-of-class activities. Some prominent examples include the Teaching Practices Inventory (Smith *et al.*, 2014; Wieman and Gilbert, 2014), the Postsecondary Instructional Practices Survey (Walter *et al.*, 2016), and the Measurement Instrument for Scientific Teaching (Durham *et al.*, 2017, 2018). While these surveys have shown alignment with other measures of instructional practices (Smith *et al.*, 2014; Durham *et al.*, 2018), they can also be prone to validity threats. For example, Ebert-May *et al.* (2011) demonstrated the discrepancy between self-report surveys and classroom observations of instructional practices within the context of the evaluation of a pedagogical workshop.

Observation protocols have been considered a robust alternative to surveys, because biases from the instructors themselves are removed. Most of the observation protocols can be grouped in two categories (American Association for the Advancement of Science, 2013): 1) holistic observation protocols that require the observers to make judgments at the end of the class session by answering survey-like questions or writing descriptive narratives based on the field notes taken from the class session (e.g., reformed teaching observation protocol [RTOP]; Sawada *et al.*, 2002); and 2) segmented observational protocols that require the observers to capture elements of the classroom instruction within a certain time frame (e.g., Classroom Observation Protocol for Undergraduate STEM [COPUS]; Smith *et al.*, 2013). For example, COPUS provides information about how instructors and students spend their time in the classroom. It requires observers to code every 2 minutes of class time for 13 student behaviors (listening, answering questions, etc.) and 12 instructor behaviors (lecturing, posing questions, etc.). COPUS does not require observers to make judgments of teaching quality, only the frequency of particular behaviors is recorded. The drawback of these observation protocols is that they solely focus on in-class practices and require extensive resources (observers) and time (for training and analysis).

Rubrics can address some of the weaknesses of the surveys and observation protocols by removing the instructors' biases while also capturing more holistically the experience provided to the students in a course rather than just in the classroom. An example of such a rubric is the Learner-Centered Teaching Rubrics (LCTR; Blumberg, 2008). As described by Maryellen Weimer (2002), this set of rubrics assesses instructors' use of learner-centered teaching. In a learner-centered teaching environment, the focus is on learning "what the student is learning, how the student is learning, the conditions under which the student is learning, whether the student is retaining and applying the learning, and how current learning positions the student for future learning" (Weimer, 2002, p. xvi). In this environment, the instructor guides, facilitates, and designs the learning experiences and is no longer simply transmitting information. The five LCTR align with the five dimensions that Weimer advocates need to change to achieve learner-centered teaching: the Function of Content (i.e., students develop disciplinary skills along with in-depth conceptual understanding and understand the relevance of these acquired skills and knowledge), the Role of the Instructor (i.e., the instructor is a facilitator as opposed to a conveyor of knowledge), the Responsibility for Learning (i.e., the instructor fosters students' responsibility for learning), the Purposes and Processes of Assessment (i.e., assessments are ongoing and promote reflection and learning), and the Balance of Power (i.e., the students have some control over the learning process). The LCTR were originally designed as a tool to help faculty in one-on-one consultations with a pedagogical expert to reflect on their instructional practices and identify areas for improvement (Blumberg, 2008, 2016). These consultations included analysis of classroom artifacts (syllabus, exams, lecture notes, etc.), which constitute authentic evidence of students' experiences in the course.

Teaching evaluation frameworks advocate for the triangulation of teaching data across different measures (Association of American Universities, 2019; Simonson *et al.*, 2022). For example, one framework popular among academic institutions and organizations is the Benchmarks for Teaching Effectiveness, developed by the Center for Teaching Excellence at the University of Kansas (KU Center for Teaching Excellence, 2021a). This framework aims to measure teaching by evaluating various facets of instruction and leveraging a variety of sources of evidence. A multidimensional rubric was developed as part of this framework to guide academic departments in their approach to teaching evaluation (Follmer Greenhoot *et al.*, 2020). The rubric includes seven dimensions: Goals, Content, and Alignment; Teaching Practices; Achievement of Learning Outcomes; Classroom Climate; Reflection and Iterative Growth; Mentoring and Advising; and Involvement in Teaching Service, Scholarship, or Community. The framework provides guidelines for the types of evidence that can be collected to support the evaluation of each of these dimensions (KU Center for Teaching Excellence, 2021b). For example, the following pieces of evidence are suggested to help assess the Teaching Practices dimension: syllabus, a sample of course materials, class observations supported by an observation protocol, dialogue with the instructor, and students' ratings and comments. The amount and breadth of evidence that should be collected and the need to identify tools to evaluate this evidence can be overwhelming for instructors, departments, and institutions and may contribute to

TABLE 1. Descriptive demographics for the participants

Demographic variables		NFW instructors, *n*	STEM instructors, *n*	Total
Gender[a]	Female	10	7	17
	Male	5	6	11
Course level taught	Undergraduate	7	10	17
	Graduate	8	3	11
Teaching experience	Less than 3 years	13	1	14
	At least 3 years	2	12	14
Total		15	13	28

[a]Participants self-identified on a survey that asked them to provide their demographic information.

limited uptake of these teaching evaluation frameworks. One strategy to mitigate this potential "overload" is to characterize the complementarity of existing instruments in order to assist with the educated selection of instruments.

In this study, we pursue this strategy and explore the relationship between COPUS and the LCTR. COPUS was chosen because it is an observation protocol that has been adopted extensively by both researchers (e.g., TRESTLE, 2017) and instructors (e.g., Arts and Sciences Support of Education Through Technology, 2022) due to its ease of use and objective output (i.e., capturing the behaviors of instructors and students occurring every 2 minutes). However, COPUS only captures the in-class learning experience and does not explicitly measure its quality. The set of LCTR, on the other hand, provides a holistic characterization of students' learning experiences in a course and aligns with research on effective practices to support learning (Blumberg, 2008). We chose this tool, rather than a survey, because teaching evaluation frameworks often request the collection of course artifacts (KU Center for Teaching Excellence, 2021b). The LCTR represent one of the few tools that we are aware of that provide a systematic, evidence-based approach to analyzing these course artifacts. The LCTR provide comprehensive evidence for two of the dimensions on the Benchmarks for Teaching Effectiveness rubric (Goals, Content, and Alignment and Teaching Practices). However, the LCTR are time and resource intensive, because there is a need for extensive analyst training, as well as collection and analysis of a large amount of data (course artifacts along with classroom observations). While it is reasonable to hypothesize that the way an instructor engages students in the classroom is a good predictor of the students' experience in the course overall, this hypothesis has yet to be fully explored. If the hypothesis holds, then it would be more effective to only use one instrument (presumably the one being least resource intensive) while retaining confidence in the ability to capture the instructional practices enacted in the course overall. This would leave resources to collect and analyze other sources of evidence, such as student learning outcomes or reflection statements from instructors. We were thus interested in understanding the extent to which in-class behaviors (captured with COPUS) relate to the way an instructor implements a course (captured with the LCTR). The overall research question explored in this study is thus: To what extent do in-class instructional behaviors, as measured using COPUS, relate to how an instructor approaches the teaching of their course, as determined using the LCTR?

METHODS

Participants

The participants in this study were recruited from two different professional development workshops. Participants were recruited after enrolling in the workshops. The first set of participants (n = 15) are new chemistry assistant professors (within the first 3 years of their appointment) who attended the New Faculty Workshop (NFW; Stains *et al.*, 2015). The NFW participants taught a variety of chemistry courses ranging from general chemistry to upper-level graduate courses. The second set of participants (n = 13) are STEM instructors who attended a pedagogical workshop provided at one research-intensive institution. The STEM instructors represented different STEM departments (entomology, earth and atmospheric sciences, astronomy, etc.) and taught courses at either the undergraduate or graduate level. The two sets of participants bring the total sample size to 28 college instructors. Demographic information for each set of participants is listed in Table 1. The study was approved by the University of Nebraska–Lincoln and the University of Virginia Institutional Review Boards.

Data Collection

Participants' classroom video observations and course artifacts were collected during the semester following their participation in the workshop. Observations and course artifacts were associated with a particular unit/topic/chapter taught by each participant. Each participant was allowed to select the unit/topic/chapter to be analyzed for this study. The number of classroom video observations ranged from two to eight per participant with a mode of four. The course artifacts included the syllabus, any course material used to teach the selected unit/topic/chapter (slides, class notes, etc.), and any assessment tools used to assess the selected unit/topic/chapter (homework, quiz, mid-term/final exam, etc.).

Analysis of In-Class Learning Experience Using COPUS

The classroom videos were analyzed with COPUS (Smith *et al.*, 2013). This protocol requires observers to select student and instructor behaviors that occurred within each 2-minute time interval of a class. Examples of student behaviors include Listening, Clicker Question Discussion, and Answer Instructor Question. Examples of instructor behaviors include Lecturing, Pose Questions, and Moving through the Classroom. The researchers who used COPUS to code classroom videos in this study (LS, VJ, BL, TP, MUG) were trained as a cohort, with the training process led by other researchers (RME, ZN) who had

conducted prior studies using COPUS. The training and coding processes are summarized in the following steps:

1. Three sets of training videos were selected based on the degree of difficulty in coding them in prior coding sessions ("easy," "moderate," and "hard"). For the easy videos, most of the classroom time consists of the instructor lecturing and students listening. The hard videos include more varied forms of interactions between instructors and students, necessitating the use of a wide spectrum of COPUS codes.
2. During the training process, the researchers watched and coded the easy videos independently. Researchers entered the results of their coding into a joint spreadsheet. This spreadsheet was used to facilitate the discussions of coding disagreements with the entire cohort of coders ($N = 7$). The coding spreadsheet was used to calculate Fleiss's kappa, which allows for the determination of interrater agreement between more than two raters (Nichols *et al.*, 2010). Fleiss' kappa was calculated using the irr package in R (Gamer *et al.*, 2012). If Fleiss' kappa values for a classroom observation reached a threshold of above 0.8, the group shifted to training on the next harder set of videos. However, if the value was below 0.8, the process was repeated on a video of similar difficulty following the discussion of coding disagreements.
3. After reaching the desired level of agreement for all training videos (above 0.8), 106 classroom observations collected from the 28 participants were distributed across the seven trained coders. Each coded 10 to 20 videos.
4. Once all videos were coded, the spreadsheet containing the coding was submitted to the COPUS Analyzer (Stains *et al.*, 2018). This tool leveraged the results of a latent profile analysis on a large set of classroom observations (Stains *et al.*, 2018) to classify each observation into three instructional styles: didactic (i.e., the instructor lectures for, on average, more than 80% of class time), interactive lecture (i.e., the lecture is supplemented with some student-centered strategies such as group work, asking clicker questions), and student centered (i.e., instructors incorporate student-centered strategies into a large portion of their class time, such that, on average, only about 50% of class time is spent lecturing).

For each instructor, we calculated the proportion of videos that fell into each of the three COPUS instructional styles.

Analysis of In- and Out-of-Class Learning Experiences Using the LCTR

The LCTR (Blumberg, 2008, 2016) were used to more broadly characterize how an instructor taught a course. Course artifacts and two classroom observations for each participating instructor were used to assign scores on the rubrics. The five rubrics (the Function of Content, the Role of the Instructor, the Responsibility for Learning, the Purposes and Processes of Student Assessment, and the Balance of Power) each contain several components. Each component is measured on a four-point scale with 1 representing Instructor-centered Approach; 2 and 3 representing Lower and Higher Level of Transitioning, respectively; and 4 representing Learner-centered Approach. The original rubrics (Blumberg, 2008) required some modifications for this study, as some of the components were not measurable with the obtained classroom observations and course artifacts (Popova *et al.*, 2020). One example is that, based on classroom observations collected and the course artifacts, which did not contain student answers, it was not possible to identify whether the students had opportunities to justify their answers when they did not agree with those of the instructor. This eliminated one component within the rubric Purposes and Processes of Assessment. The modified LCTR employed in this study contained 14 components across the five rubrics (Supplementary Table S1).

Three members of the research team (LS, MP, AP) were involved in the coding process. First, the researchers coded the selected classroom observations and course artifacts independently for three to four instructors. Then, the coders discussed their assigned scores for each component of each of the LCTR to resolve any disagreements. After reaching a 100% agreement, the researchers coded the rest of the instructors independently. However, to ensure adherence to the rubrics over time, two researchers coded the same instructor after they had coded three or four other instructors independently. In total, 10 instructors were coded by two researchers and 18 by a single researcher.

For each instructor, the scores on the components within a rubric were averaged to obtain a rubric-level score. The rubric-level scores were then averaged to obtain what we will refer to as the "LCTR score." The LCTR scores ranged from 1 to 4, with 1 representing instructors using Instructor-centered Approach; 4 representing instructors using Learner-centered Approach; and 2 and 3 identified as Lower and Higher Level of Transitioning, respectively. The interpretation of each of this score for each component of the LCTR is provided in Supplementary Table S2.

RESULTS

We share here the distribution of instructional styles across the 28 instructors as described by COPUS and the LCTR. We then explore the relationship in the characterization of instructional styles between COPUS and the LCTR. Finally, we report on the influence of observation intensity (i.e., the number of classroom observations conducted per participant) on this relationship.

Participants' Instructional Styles According to COPUS and the LCTR

Analysis of the COPUS data (Supplementary Table S3) shows that half of the instructors ($n = 14$) were classified within the same instructional style across all video recordings of their course. More than half of these instructors ($n = 8$) were classified as Didactic, five as Interactive Lecture, and one as Student-centered. The other half of the instructors demonstrated different instructional styles across the video recordings. Seven of them taught didactically in at least one of the video and in a more engaging way in the others. Four instructors had a mix of Interactive Lecture and Student-centered styles.

Based on the LCTR data (Supplementary Table S3), two-thirds of the instructor had an average LCTR score between 2 and 3, indicating that most instructors were between a Lower and Higher Level Of Transitioning with respect to teaching their courses. A little more than a third ($n = 7$) had scores below 2, indicating a dominating Instructor-centered Approach. Two instructors had scores slightly above 3. None had scores above 3.2, indicating their courses were instructed in a learner-centered style.

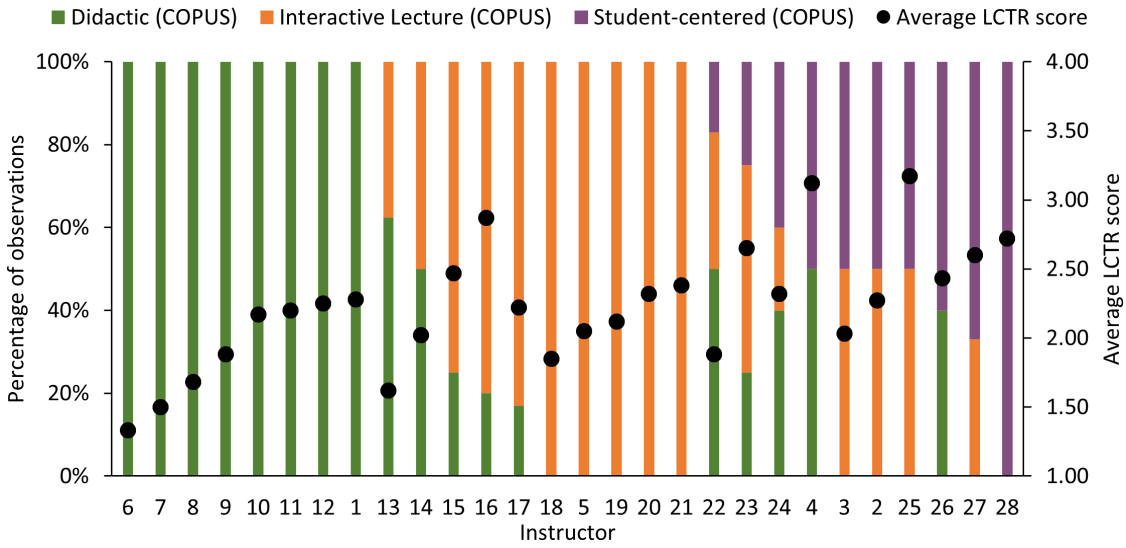

FIGURE 1. Proportion of COPUS instructional styles observed in the set of videos collected from each instructor and average score across all five LCTR for each instructor. Data are organized from highest proportion of Didactic videos per participant to highest proportion of Student-centered videos per participant.

Relationships between the Instructional Styles Described by COPUS and the LCTR

The relationship between the instructional styles identified according to COPUS and average LCTR score is presented in Figure 1. Instructors in Figure 1 are organized based on the proportion of COPUS instruction styles within the set of videos collected from the instructor, from sets of videos that are solely classified as Didactic to sets of videos that are solely classified as Student-centered. This organization allows us to analyze the trend in the LCTR average score as the proportion of COPUS instructional styles involving student engagement increases. Figure 1 shows a positive trend between the two measures: an increase in the proportion of videos that include student engagement in the classroom (as measured by COPUS) is related to an increase in the average LCTR score. This indicates that the two measures provide similar overall characterization of the level of student-centeredness in a course. However, there are some notable discrepancies that illuminate the nuances about an instructor's practice that both tools provided. Four of these discrepancies are highlighted below.

First, we describe two instructors who had similar LCTR scores but different COPUS instructional styles. Instructor 1 was teaching an undergraduate engineering course and had been teaching for 2 years at the time of data collection. The COPUS analysis of the three videos that were collected indicated that they spent, on average, 93.1 ± 6.5% of class time lecturing. No group work was observed in these three videos, and all three videos were classified as Didactic. Instructor 1 had a 2.28 average LCTR score based on the analyses of course artifacts and videos. The level of learner-centered instruction in this instructor's course can thus be described as Lower Level of Transitioning. Instructor 2 was teaching an introductory meteorology course for the first time at this institution when data were collected. The analysis of course artifacts and the two videos collected resulted in an average LCTR score almost identical to Instructor 1, 2.27. Both instructors had scores ranging in the Higher Level of Transitioning for two of the five rubrics, Function of Content and Role of the Instructor. For example, Instructor 1 facilitated learning by using a variety of instructional strategies, including minute papers (captured on video and described in the syllabus), group work (evidence from syllabus), and opportunities for students to critique one another's work (evidence from syllabus). Instructor 2 also employed various instructional strategies, including in-class discussion, group work, and clicker questions (evidence from videos and syllabus). Both instructors also had low- and high-level learning objectives listed on the syllabus, and Instructor 1 was observed talking to students about some of these learning objectives during class. All these items represented a high level of transitioning within the Role of Instructor rubric. Both instructors had lower scores for the other three LCTR. Although Instructor 2 had a similar LCTR score than Instructor 1, Instructor 2's overall instructional style based on COPUS was more learner-centered, with one video classified as Didactic (94.4% of class time spent lecturing) and one video classified as Student-centered (59.5% of class time spent lecturing). The analysis of these two instructors shows that, in this case, the COPUS instrument provided some evidence for the level of use of group work. Indeed, while both instructors indicated using group work during class time in their syllabi, group work was only observed in the videos collected from Instructor 2. This potentially indicates that group work is employed less frequently in Instructor 1's than Instructor 2's course, although more videos from each instructor would need to be analyzed via COPUS to validate this claim or a follow-up interview would need to be conducted to characterize frequency of use of group work.

Second, we describe two instructors who showcase contrasting assessments of instructional practices when both instruments are used. Instructor 3 was teaching an undergraduate astronomy course and had been teaching for 3 years at this institution at the time of data collection. The LCTR analysis of their course artifacts and four videos provided an average LCTR score of 2.03, with a score less than 2.50 on each of the LCTR. While the LCTR analysis characterized Instructor 3's practices

as Lower Level of Transitioning, the COPUS analysis classified two of the four videos collected from them as Interactive Lecturing and two as Student-centered. Across the four videos, they lectured for an average of 79.7 ± 15.4% of class time, and group work took place in every session observed with an average of 17.8 ± 16.4% of class time. During one observation, they lectured for only 58% of the class, and students worked in groups on a worksheet for 42% of the class time. This instructor was thus characterized as minimally learner-centered with the LCTR but as student-centered with COPUS. Instructor 4 was the opposite. Instructor 4 was teaching an upper-level biology course and had been teaching for about 10 years at this institution by the time of data collection. The LCTR analysis of their four videos and course artifacts resulted in an average LCTR score of 3.12, which corresponds to Higher Level of Transitioning toward learner-centered practices. However, while two of the videos were classified by COPUS as Student-centered ($M_{Lecturing}$ = 44.5 ± 2.3% lecturing), the other two were classified as Didactic ($M_{Lecturing}$ = 78.6 ± 11.7% lecturing). The analysis of these two instructors demonstrates that each instrument provides nuances about instructional practices.

Upon further analysis of the relationships between the COPUS data and each of the five LCTR, we noticed that, although positive trends are observed for each rubric (Figure 2), extensive discrepancies can be seen, especially for instructors classified with COPUS as Didactic. For example, Instructor 6 and Instructor 12 both were classified as Didactic on each of the videos collected from them (four and three, respectively) but Instructor 6 received a 1.50 score on the Role of Instructor rubric while Instructor 12 received a 3.25 (Figure 2b). Similar large discrepancies in LCTR scores among solely Didactic instructors can be found for the Function of Content (Figure 2a) and the Balance of Power (Figure 2e) rubrics.

Taken together, these results demonstrate that the outputs from both instruments are necessary to provide a more comprehensive and nuanced characterization of an instructor's practice. The information collected from both instruments better informs both the instructor and other stakeholders (e.g., consultant from a teaching and learning center, researcher characterizing instructional practices of STEM faculty, or colleagues evaluating a peer's teaching portfolio as part of a formal evaluation process) about the learning environments students experience in an instructor's course.

Comparing the Relationship between COPUS and LCTR Outputs by Observation Intensity

Several studies, including ours, have remarked that the level of accuracy in the description of an instructor's in-class practice relies on coding several class sessions, although the minimum number of sessions to be observed and their timing is still being determined (Lund *et al.*, 2015; Stains *et al.*, 2018; Denaro *et al.*, 2021; McConnell *et al.*, 2021; Sbeglia *et al.*, 2021; Weston *et al.*, 2021). We thus intended to explore whether the positive trend between the LCTR and COPUS reported in the previous section was related to the number of classroom observations coded with COPUS. Across the 28 instructors, the number of classroom observations ranged from two to eight per instructor, with a mode of four (Supplementary Table S3). Participants were classified into two groups based on the number of observations collected, with one group having a number of observations below the mode (two to three observations, n = 14) and the other group having a number of observations at or above the mode (four to eight observations, n = 14). We replicated Figure 1 for each group (Figure 3). While the figure for each group shows a positive trend between the average LCTR score and the level of student engagement as measured by COPUS, the trend is stronger for the group with two to three observations (Figure 3a). Notably, the variety of instructional styles measured by COPUS that instructors exhibited in their videos varied a lot less for the group with less observations than for the group with more observations. The large discrepancies in average LCTR score described earlier among the Didactic instructors are still observed, regardless of the number of observations collected from the instructors.

These results reinforce the results from the previous section. Each instrument provides nuances in an instructor's practice that the other instrument does not capture. However, the data indicate that numerous videos should be analyzed via COPUS in order to increase the likelihood of capturing the various practices an instructor exhibits in the classroom. The adequate number of videos necessary is still an area open to empirical investigation.

Limitations

The small sample size limits the generalizability of the findings. At the same time, the diversity of instructors and courses represented alleviates some of the generalizability concerns. This study and associated findings should be considered exploratory; more extensive studies need to be conducted to reach generalizable results.

DISCUSSION AND IMPLICATIONS

Teaching evaluation frameworks have advocated for the use of multiple sources of evidence to characterize instructional practices (Association of American Universities, 2019; KU Center for Teaching Excellence, 2021a; Simonson *et al.*, 2022). Thanks to extensive efforts within the education research community, we now have access to a plethora of instruments that measure various aspects of teaching. Each comes with its own strengths and weaknesses concerning the validity and reliability of the data collected and the level of resources required to collect and analyze the data. To assist instructors, researchers, and institutions in their selection of instruments, it is important to explore and report on the complementarity, or lack thereof, of instruments. In this study, we set out to identify the complementarity between a popular observation protocol focused on instructor and student behaviors in the classroom: COPUS and the LCTR, which measure the level of learner-centeredness of a course.

Our results show that the outcomes from COPUS and the LCTR are related to each other but that each instrument provides nuanced information about an instructor's practice that is not captured by the other instrument. For example, COPUS provides information about the frequency of use of group work, while the LCTR capture the purpose of assessment and the nature of learning objectives for the course. While COPUS has gained popularity as a measure of instructional practice, it is critical to limit the implication of the output data to what COPUS is measuring: classroom behaviors within the sample of classroom sessions observed, rather than use them to infer about the level of student-centeredness exhibited in an instructor's course.

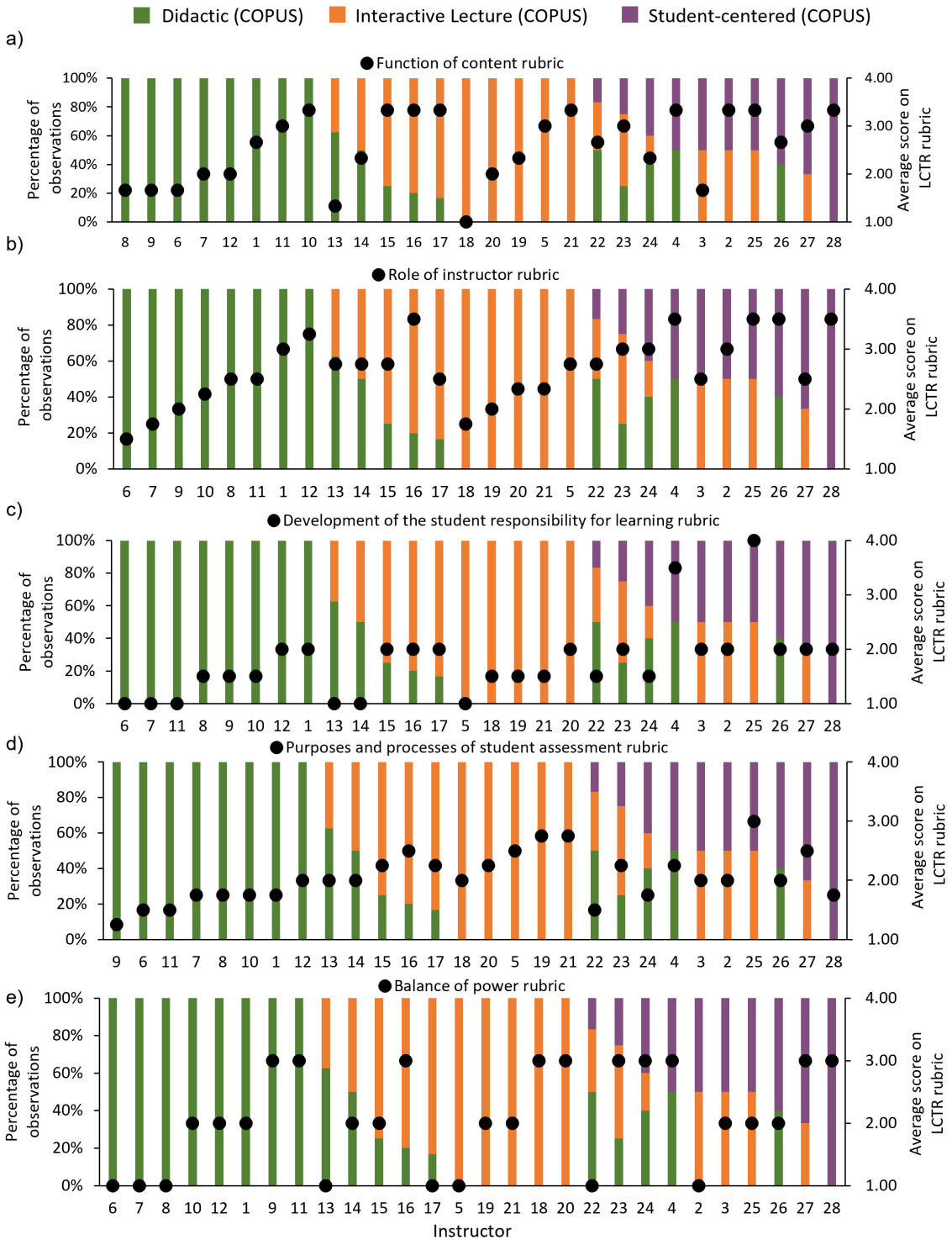

FIGURE 2. Proportion of COPUS instructional styles observed in the set of videos collected from each instructor and LCTR average score for each instructor on (a) the Function of Content rubric, (b) the Role of Instructor rubric, (c) the Development of Student Responsibility for Learning rubric, (d) the Purposes and Processes of Student Assessment rubric, and (e) the Balance of Power rubric. Data are organized from highest proportion of Didactic videos per participant to highest proportion of Student-centered videos per participant.

Each instrument should be selected to align with the teaching dimension that it is intended to capture, and claims should be limited to that dimension. For example, researchers interested in capturing the impact of a new professional development initiative focused on in-class group work would benefit from using COPUS and not LCTR. Similarly, if one of the main criteria in evaluating teaching effectiveness for an institution is the implementation of group work in the classroom, then COPUS would

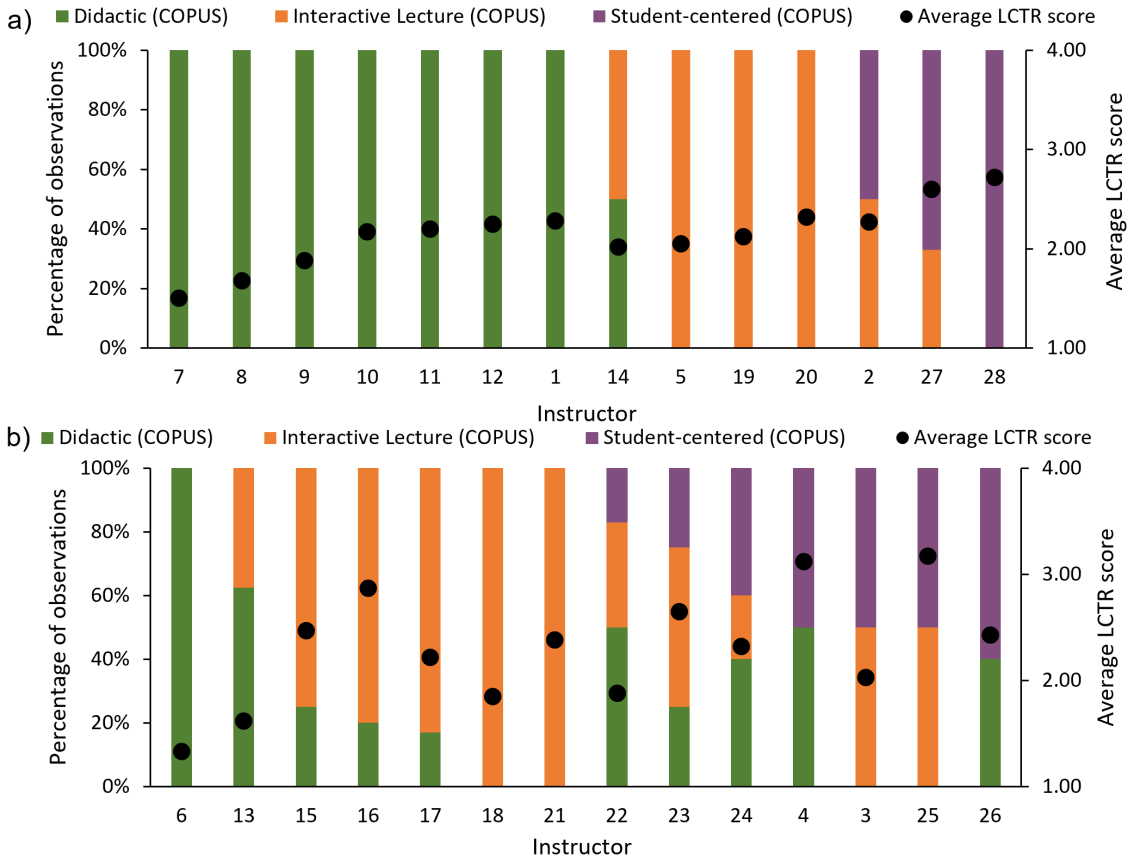

FIGURE 3. Proportion of COPUS instructional styles observed in the set of videos collected from each instructor and average score for each instructor across all five LCTR: (a) instructors for whom we collected two to three videos and (b) instructors for whom we collected four to eight videos. Data are organized from highest proportion of Didactic videos per participant to highest proportion of Student-centered videos per participant.

be an appropriate measure. However, if one operationalizes teaching effectiveness as student-centered practices in various aspects of a course, then both the LCTR and COPUS should be employed.

While the results and their implications relate to the use of these two instruments for the purpose of summative evaluation, these instruments can also serve a formative purpose. Indeed, formative evaluation can be used to promote growth among instructors, and each of these tools could be valuable depending on the nature of the growth expected. COPUS can be a great entry point when working with faculty with limited exposure to and understanding of learner-centered teaching. It could be used, for example, in the formative years of new professors to help them become comfortable and develop an appreciation for the benefits of collaborative learning. For more seasoned instructors, the LCTR can help them reflect on their instructional practices and provide them with benchmarks to ensure a holistic, learner-centered approach to their courses.

Importantly, this study displays the complexity of measuring instructional effectiveness and supports teaching evaluation frameworks' recommendation that several sources of evidence and analytical tools should be used to describe an instructor's teaching practices (Association of American Universities, 2019; KU Center for Teaching Excellence, 2021a; Simonson *et al.*, 2022). Yet it is still unclear how to combine the multiple sources of evidence to provide an overall description and assessment of teaching quality for an instructor. In this study, we saw variations in the level of implementation of student/learner-centered practices across the two measures and within measures. For example, Instructor 5 scored at the Higher Level of Transitioning on the Function of Content rubric but at the Instructor-centered level on the Development of the Student Responsibility for Learning rubric (Supplementary Table S3). Moreover, Instructor 3 showed great variability in their use of group work in their classroom, with only 8% of class time spent on group work in one session versus 42% in another session. It is quite challenging to evaluate the effectiveness of these two instructors with the evidence collected here, which is quite extensive and robust compared with typical measures of teaching evaluations (Dennin *et al.*, 2017). Notably absent from the evidence collected for this study are measures of student learning. While the addition of this type of evidence could bring some clarity, providing too much weight to such evidence could have its own pitfalls, especially if the measures are inappropriate (such as student evaluations). The development of rubrics such as those developed as part of the Benchmarks for Teaching Effectiveness framework (KU Center for Teaching Excellence, 2021a) addresses some of these challenges. However, more research is needed to explore how instructors, committee members, and administrators actually apply these rubrics.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant nos. DUE-1256003, DUE-1347814, and CAREER 1552448/2021491.

REFERENCES

American Association for the Advancement of Science (AAAS). (2013). *Describing and measuring undergraduate STEM teaching practice: A report from a national meeting on the measurement of undergraduate science, technology, engineering and mathematics (STEM) teaching*. Retrieved January 12, 2022, from https://cns.utexas.edu/images/CNS/Deans_Office/ResearchFacilities/strategic_initiatives/measuring-stem-teaching-practices.pdf

Arts & Sciences Support of Education Through Technology. (2022). COPUS Tool Details. Retrieved March 18, 2022, from www.colorado.edu/assett/programs/vips/copus

Association of American Universities. (2019). *AAU Undergraduate STEM Education Initiative: Matrix of summative evaluation of teaching strategies*. Retrieved March 7, 2022, from www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/P&T-Matrix.pdf

Blumberg, P. (2008). *Developing learner-centered teaching: A practical guide for faculty*. Hoboken, NJ: Wiley.

Blumberg, P. (2016). Assessing implementation of learner-centered teaching while providing faculty development. *College Teaching*, *64*(4), 194–203. doi: 10.1080/87567555.2016.1200528

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., … & Smith, T. L. (2015). University learning: Improve undergraduate science education. *Nature*, *523*(7560), 282–284.

Denaro, K., Sato, B., Harlow, A., Aebersold, A., & Verma, M. (2021). Comparison of cluster analysis methodologies for characterization of Classroom Observation Protocol for Undergraduate STEM (COPUS) data. *CBE—Life Sciences Education*, *20*(1), ar3.

Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., … O'Dowd, D. K. (2017). Aligning practice to policies: Changing the culture to recognize and reward teaching at research universities. *CBE—Life Sciences Education*, *16*(4), es5.

Durham, M. F., Knight, J. K., Bremers, E. K., DeFreece, J. D., Paine, A. R., & Couch, B. A. (2018). Student, instructor, and observer agreement regarding frequencies of scientific teaching practices using the Measurement Instrument for Scientific Teaching–Observable (MISTO). *International Journal of STEM Education*, *5*(1), 1–15.

Durham, M. F., Knight, J. K., & Couch, B. A. (2017). Measurement Instrument for Scientific Teaching (MIST): A tool to measure the frequencies of research-based teaching practices in undergraduate science courses. *CBE—Life Sciences Education*, *16*(4), ar67.

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 programs. *BioScience*, *61*(7), 550–558. doi: 10.1525/bio.2011.61.7.9

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE*, *14*(2), e0209749.

Follmer Greenhoot, A., Ward, D., Bernstein, D., Patterson, M. M., & Colyott, K. (2020). *Benchmarks for teaching effectiveness (revised 2020)*. Retrieved from https://cte.ku.edu/sites/cte.ku.edu/files/docs/KU%20Benchmarks%20Framework%202020update.pdf

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr'. *Various coefficients of interrater reliability and agreement in R. R package version 0.84.1*. Retrieved December 2, 2022, from https://rdocumentation.org/packages/irr/versions/0.84.1

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144–154.

Henderson, C., Turpen, C., Dancy, M., & Chapman, T. (2014). Assessment of teaching effectiveness: Lack of alignment between instructors, institutions, and research recommendations. *Physical Review Special Topics—Physics Education Research*, *10*(1). doi: 10.1103/PhysRevSTPER.10.010106

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1304016.

KU Center for Teaching Excellence. (2021a). *Benchmarks for teaching effectiveness*. Retrieved March 7, 2022, from https://cte.ku.edu/benchmarks-teaching-effectiveness-project

KU Center for Teaching Excellence. (2021b). *KU benchmarks rubric: Evidence matrix, by teaching dimension and source*. Retrieved March 7, 2022, from https://cte.ku.edu/benchmarks-teaching-effectiveness-project

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE—Life Sciences Education*, *14*(2), ar18. doi: 10.1187/cbe.14-10-0168

McConnell, M., Boyer, J., Montplaisir, L. M., Arneson, J. B., Harding, R. L., Farlow, B., & Offerdahl, E. G. (2021). Interpret with caution: COPUS instructional styles may not differ in terms of practices that support student learning. *CBE—Life Sciences Education*, *20*(2), ar26.

Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. *Quality Assurance Journal*, *13*(3–4), 57–61. doi: 10.1002/qaj.481

Popova, M., Shi, L., Harshman, J., Kraft, A., & Stains, M. (2020). Untangling a complex relationship: Teaching beliefs and instructional practices of assistant chemistry faculty at research-intensive institutions. *Chemistry Education Research and Practice*, *21*(2), 513–527. doi: 10.1039/c9rp00217k

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.

Sbeglia, G. C., Goodridge, J. A., Gordon, L. H., & Nehm, R. H. (2021). Are faculty changing? How reform frameworks, sampling intensities, and instrument measures impact inferences about student-centered teaching practices. *CBE—Life Sciences Education*, *20*(3), ar39.

Shao, L. P., Anderson, L. P., & Newsome, M. (2007). Evaluating teaching effectiveness: Where we are and where we should be. *Assessment & Evaluation in Higher Education*, *32*(3), 355–371. doi:10.1080/02602930600801886

Simonson, S. R., Earl, B., & Frary, M. (2022). Establishing a framework for assessing teaching effectiveness. *College Teaching*, *70*(2), 1–18.

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*(4), 618–627. doi: 10.1187/cbe.13-08-0154

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE—Life Sciences Education*, *13*(4), 624–635. doi: 10.1187/cbe.14-06-0108

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., … & Laski, F. A. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470.

Stains, M., Pilarz, M., & Chakraverty, D. (2015). Short and long-term impacts of the Cottrell Scholars Collaborative New Faculty Workshop. *Journal of Chemical Education*, *92*(9), 1466–1476. doi: 10.1021/acs.jchemed.5b00324

TRESTLE. (2017). *COPUS observations resources*. Retrieved March 18, 2022, from https://trestlenetwork.ku.edu/copus-observation-resources

Walter, E. M., Henderson, C. R., Beach, A. L., & Williams, C. T. (2016). Introducing the Postsecondary Instructional Practices Survey (PIPS): A concise, interdisciplinary, and easy-to-score survey. *CBE—Life Sciences Education*, *15*(4), ar53. doi: 10.1187/cbe.15-09-0193

Weimer, M. (2002). *Learner-centered teaching: Five key changes to practice*. Hoboken, NJ: Wiley.

Weston, T. J., Hayward, C. N., & Laursen, S. L. (2021). When seeing is believing: Generalizability and decision studies for observational data in evaluation and research on teaching. *American Journal of Evaluation*, *42*(3), 377–398.

Wieman, C. (2015). A better way to evaluate undergraduate teaching. *Change: The Magazine of Higher Learning*, *47*(1), 6–15.

Wieman, C., & Gilbert, S. (2014). 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. doi: 10.1187/cbe.14-02-0023