Interpret with Caution: COPUS Instructional Styles May Not Differ in Terms of Practices That Support Student Learning

Melody McConnell,[†] Jeffrey Boyer,[‡] Lisa M. Montplaisir,[§] Jessie B. Arneson,[¶] Rachel L.S. Harding,[§] Brian Farlow,[‖] and Erika G. Offerdahl,[¶*]

[†]Division of Science and Mathematics, Mayville State University, Mayville, ND 58257; [‡]Office of the Provost and [§]Department of Biological Sciences, North Dakota State University, Fargo, ND 58105; [¶]Washington State University, School of Molecular Biosciences, Pullman, WA 99164; [‖]Minnesota State Community and Technical College, Moorhead, MN 56563

ABSTRACT

There is a growing need for valid and reliable measures to monitor the efficacy of undergraduate science, technology, engineering, and mathematics (STEM) reform initiatives. The Classroom Observation Protocol for Undergraduate STEM (COPUS) is a widely used tool originally designed to measure the presence of overt instructor and student behaviors. It has subsequently been used to characterize instruction along a continuum from didactic to student centered, and more recently to categorize instruction into one of three styles. Initiatives focused on professional development often support instructors' progression from didactic to student-centered styles. There is a need to examine COPUS instructional styles in terms of behaviors that research has shown to improve student learning. Formative assessment is a research-based practice that involves behaviors accounted for by the COPUS (e.g., posing a question). We qualitatively compared the formative assessment behaviors in 16 biology class sessions categorized into each of the three COPUS styles. We were unable to detect differences in formative assessment behaviors between the COPUS styles. Caution should be taken when interpreting COPUS data to make inferences about the effects of reform efforts. This study underscores the need for additional measures to monitor national reform initiatives in undergraduate STEM.

INTRODUCTION

National reform initiatives strive to transform undergraduate science, technology, engineering, and mathematics (STEM) education by supporting individual instructors (e.g., Pfund *et al.*, 2009; Emery *et al.*, 2020), academic departments (Brancaccio-Taras *et al.*, 2016; Reinholz *et al.*, 2019), and institutions of higher education (e.g., Network of STEM Education Centers, https://serc.carleton.edu/StemEdCenters/index.html) with the ultimate goal of improving student learning. Measuring the efficacy of these efforts requires establishing indicators to determine the degree to which instructors enact evidence-based instructional practices (Rosenberg *et al.*, 2018). In undergraduate STEM education, as in K–12 systems (Weisberg *et al.*, 2009; Kraft and Gilmour, 2017), there is a need for valid and reliable measures that can document progress in the adoption of practices that support improved student learning.

In a landmark study, Stains and colleagues (2018) completed a first step in documenting progress by characterizing the landscape of undergraduate STEM instructional practices in North America. They used the Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith *et al.*, 2013), a widely accepted instrument that documents the presence of overt instructor and student behaviors, to observe more than 2000 undergraduate STEM classes. The authors conducted a latent profile

Tessa C. Andrews, *Monitoring Editor*
Submitted Sep 17, 2020; Revised Mar 10, 2021; Accepted Mar 15, 2021
CBE Life Sci Educ June 1, 2021 20:ar26
DOI:10.1187/cbe.20-09-0218

*Address correspondence to: Erika G. Offerdahl (erika.offerdahl@wsu.edu).

analysis using the COPUS observation data and determined that undergraduate STEM instruction can be described by seven instructional profiles, each of which can be further categorized into one of three instructional styles (didactic, interactive lecture, and student centered). Each of the instructional styles exhibits different frequencies of student-centered instructional behaviors, with didactic having the lowest frequency. Stains and colleagues determined that didactic instructional practices were still the most prevalent in the undergraduate STEM courses sampled for their study, despite substantial evidence that didactic instruction is less effective than active, student-centered practices (Freeman *et al.*, 2014; Stains *et al.*, 2018).

The extensive research demonstrating the efficacy of active, student-centered instruction has sent a clear message to the community that didactic-only instruction must be replaced (Wieman, 2014). The COPUS was originally recommended for use by faculty because it could provide the necessary "information to report about their use of active-learning strategies" (Smith *et al.*, 2013, p. 626). Further, it has been implied that efforts to improve undergraduate STEM instruction will facilitate a progression from didactic instructional profiles to student-centered profiles, with interactive lecture profiles as intermediate steps in the progression (Lund *et al.*, 2015; Holland *et al.*, 2018; Stains *et al.*, 2018; Weir *et al.*, 2019). As the Stains and colleagues' study points out, this progression has the potential to "inform incremental and diverse paths toward student-centered teaching" (Stains *et al.*, 2018, p. 1470).

Formative assessment and associated feedback have been identified as an integral component of active, student-centered classrooms that positively affects student learning (Committee on STEM Education of the National Science and Technology Council, 2018; Offerdahl *et al.*, 2018; Rosenberg *et al.*, 2018). Formative assessment is a cycle through which evidence of student thinking is elicited using an instructional prompt, student progress toward desired learning outcomes is diagnosed, and feedback is provided to support student learning (Nicol and Macfarlane-Dick, 2006; Nicol, 2010; Offerdahl *et al.*, 2018). The efficacy of instructor-generated feedback on student learning has been documented in a variety of classroom settings, with recent metasyntheses suggesting effect sizes as high as 0.70 and 0.79 (Hattie and Timperley, 2007; Hattie, 2009; Hattie and Zierer, 2019). Random effects models suggest a more modest, medium effect of feedback ($d = 0.48$) on student learning (Wisniewski *et al.*, 2020). It has been suggested that the observed variation in effect sizes is influenced by the type of feedback provided and anticipated outcome (e.g., cognitive skills, motivational outcomes; Hattie and Timperley, 2007; Hattie, 2009; Hattie and Zierer, 2019). The implication of this collective work is that formative assessment and feedback positively affect learning, but variations in instructional practices mediate their effectiveness (Offerdahl *et al.*, 2018).

Formative assessment is a student-centered practice, so it is unsurprising that the COPUS explicitly accounts for some instructor behaviors that would occur during a formative assessment cycle (e.g., posing a question, following up on a question). The COPUS was not intended to provide information about the quality of behaviors or the content covered. Therefore, while the COPUS may be expected to provide insight into the frequency with which formative assessment and feedback occur in an undergraduate STEM course, it is unlikely it will be able to distinguish between the types of formative assessments or feedback generated by an instructor that may have the greatest effect on student learning. Thus, caution is warranted in the meaning we ascribe to COPUS styles until we know whether and how they correspond to differences in high-impact instructional practices such as formative assessment and feedback. To that end, the primary goal of this study was to determine whether the three COPUS instructional styles reflect distinct differences in formative assessment and feedback practices in a sample of introductory biology class sessions. This work was guided by the research question: How do formative assessment practices compare between the three COPUS instructional styles?

Formative Assessment: A Student-Centered Practice That Improves Student Learning

Formative assessment is an integral part of active, student-centered learning environments, because it provides an opportunity for students to work actively with the course material and receive feedback to support their learning (Hattie and Timperley, 2007; Evans, 2013; Offerdahl *et al.*, 2018). Typically, the formative assessment process begins with the instructor giving the student or class some type of prompt, which allows students to respond and thereby reveal evidence of their understanding. Alternatively, students may demonstrate their understanding by asking a question. Finally, the instructor may respond to the evidence of student understanding in some way, including feedback or asking an additional prompt (Ruiz-Primo and Furtak, 2006).

Formative assessment allows students to gauge where they are in relation to the desired learning objectives and discern strategies for how to reach those objectives (Nicol and Macfarlane-Dick, 2006; Hattie and Timperley, 2007). While there is some consensus in the literature about the critical components of formative assessment (Offerdahl *et al.*, 2018), instructors are likely to tailor their formative assessment practices according to their personal preferences or to fit within the constraints of the physical classroom environment and course content (Gess-Newsome *et al.*, 2003). This freedom to adapt allows instructors to maintain student engagement by employing a variety of assessments while at the same time collecting and responding to student understanding about the content (Carless, 2019). Conversely, this freedom may lead to instructors adapting the critical components of formative assessment in a way that reduces its effectiveness in supporting student learning (Briggs *et al.*, 2012; Stains and Vickrey, 2017; Offerdahl *et al.*, 2018).

There are a number of ways in which the efficacy of formative assessment might be reduced. One critical component of the formative assessment cycle is instructor-generated feedback (Hattie and Timperley, 2007; Evans, 2013; Offerdahl *et al.*, 2018). The quality of feedback that can be generated depends on the evidence of student thinking revealed by a formative assessment prompt (Esterhazy and Damsa, 2019). Evidence of student thinking is produced in a number of forms ranging from a single word called out by a single student to worksheets, diagrams drawn on whiteboards by groups, and clicker response distributions. Call and response–style verbal prompts that produce single-word responses do not provide enough information about student thinking to adequately diagnose progress toward

the desired learning outcome (Nicol, 2010; Offerdahl and Montplaisir, 2014). Relatedly, prompts that elicit simple recall of facts as opposed to revealing students' connection-making between ideas will provide insight into lower cognitive level learning outcomes that would be insufficient if the desired learning outcome is to create and evaluate complex solutions to problems (Crowe *et al.*, 2008). The type of formative assessment prompt affects the quality of instructor-generated feedback (Offerdahl and Montplaisir, 2014; Offerdahl *et al.*, 2018), and therefore the potential for supporting student learning.

Once an instructor solicits evidence of student learning, how the instructor uses the evidence matters. Many of the potential benefits of formative assessment are predicated on a robust dialogue between instructor and student, which creates an iterative feedback loop (Duschl and Gitomer, 1997; Nicol, 2010; Offerdahl *et al.*, 2018). Effective formative assessment diagnoses how close students are to the learning objectives and provides the opportunity to give feedback that will support students' progression toward those objectives (Nicol and Macfarlane-Dick, 2006; Hattie and Timperley, 2007). Optimal feedback also facilitates student reflection and metacognition about the current performance or state of understanding, the target objectives, and how a gap between the two can be bridged (Hattie and Timperley, 2007; Evans, 2013; Offerdahl *et al.*, 2018). Feedback that simply transmits evaluative information about the correctness of an answer may constrain students more than it helps them (Hattie and Timperley, 2007). Students should be given an opportunity to actively work to revise their thinking, regulate their own learning, and move closer to the learning goals (Boud and Molloy, 2013; Metcalfe, 2017; Carless, 2019). Ultimately, with appropriate feedback, guidance, and continued questioning, students will be expected to develop the metacognition to self-assess their understanding of the content and their own work (Tai *et al.*, 2018; Carless, 2019).

The Classroom Observation Protocol for Undergraduate STEM (COPUS)

The COPUS is a widely used protocol developed for documenting the presence of 13 student and 12 instructor behaviors in 2-minute time intervals (Smith *et al.*, 2013). COPUS allows for the quantification of student-centered instructor behaviors such as posing questions (PQ), including clicker questions (CQ). Such behaviors represent the first step of a formative assessment cycle, that is, eliciting evidence of student understanding. Intervals during which instructors respond to evidence of student understanding or follow up on prompts they have given to students are often coded as FUp. Instructor interactions with individual students (1o1) or groups of students (MG) are opportunities to guide the students and collect evidence of student understanding.

The COPUS has subsequently been used to further characterize the degree to which student-centered instruction is applied in and across university STEM settings (Lund *et al.*, 2015; Stains *et al.*, 2018). Lund *et al.* (2015) conducted a cluster analysis using eight informative and nonredundant COPUS codes (AnQ-S, SQ, GW, CQ, FUp, Lec, RtW, MG), resulting in 10 COPUS profiles that aligned with other observation protocols that document a continuum of reformed teaching. Stains and colleagues (2018) used eight of the COPUS codes in their analysis of more than 2000 classroom observations. They chose the following codes due to their heterogeneity, noncorrelation, and theoretical association with active learning: four instructor codes (Lec, PQ, CQ, 1o1) and four student codes (CG, WG, OG, SQ) (Stains *et al.*, 2018, p. 1469). The latent profile analysis allowed them to identify seven instructional profiles that fell into one of three distinct COPUS instructional styles: didactic, interactive lecture, and student centered. Didactic refers to class periods during which 80% or more of 2-minute time intervals include lecturing. The interactive lecture style consists of greater than 50% of time intervals with lecture, but also incorporates some student-centered activities and group work. The student-centered instructional style incorporates large amounts of student-centered activities and group work into each class period (Stains *et al.*, 2018). As part of the analysis, it was determined that a minimum of four classroom observations are required to accurately capture an instructor's teaching style. Researchers can submit their observation data to the COPUS analyzer (http://copusprofiles.org), an online tool that categorizes an observation into one of the seven instructional profiles, which are further categorized into the three instructional styles.

The COPUS can provide insight into the frequency of 2-minute time blocks within which formative assessment prompts occur but was not designed to provide information about the nature or quality of assessment prompts or feedback. For example, assessment prompts require varying degrees of cognitive investment ranging from recall of facts to synthesis of complex solutions to problems (Krathwohl, 2002; Crowe *et al.*, 2008). The COPUS codes CQ and PQ are used to record that a prompt has been provided, but do not differentiate between levels of cognitive demand. Similarly, instructor behaviors that are coded as follow up by COPUS (FUp) can include a variety of instructor moves that vary in terms of student learning (e.g., providing a correct answer with little additional explanation, providing feedback to shape future student behavior). The nature and quality of formative assessment prompts and instructor-generated feedback have important implications for student learning (Evans, 2013; Offerdahl and Montplaisir, 2014; Offerdahl *et al.*, 2018; Carless, 2019).

METHODS

We observed four sections of a two-semester introductory biology course for majors over three academic semesters (two instructors for two academic semesters each; Figure 1). Consistent with the recommendation of Stains and colleagues (2018), we observed four class sessions (meetings) for each course section. Observations were equally spaced across each semester. The two instructors had attended multiple professional development opportunities (e.g., HHMI Summer Institutes, www.summerinstitutes.org) and were knowledgeable about promising practices in formative assessment. They also enjoyed ample departmental support to incorporate active-learning practices and were embedded within a smaller collaborative group that had been working to create an assessment-rich environment for students in introductory biology (McConnell *et al.*, 2019). The courses and instructors were selected using a purposive sampling strategy as part of a larger study because they were anticipated to include many formative assessment and feedback behaviors in an introductory biology classroom (Creswell and Poth, 2016).

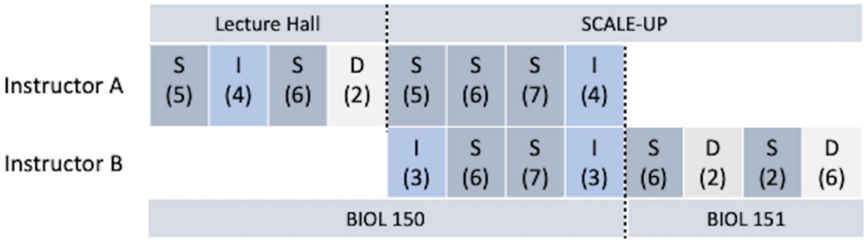

FIGURE 1. Two instructors were observed teaching four times per semester either in a fixed-seating lecture hall or a SCALE-UP classroom. Observations were conducted over three academic semesters; for one semester, instructors taught separate sections of the same course. Each instructional day was categorized as didactic (D), interactive lecture (I), or student centered (S). Numbers in parentheses indicate the COPUS cluster into which the class session was categorized. One semester, the instructors taught different sections of the same course (BIOL 150 in the SCALE-UP classroom).

Two of the sections were the same course taught by the same instructor but in consecutive academic years (Figure 1, Instructor A). In the first year, the instructor taught in a fixed-seating lecture hall and in the second year taught the same course in a SCALE-UP classroom (Beichner, 2008). The other two sections were taught by a different instructor exclusively in a SCALE-UP classroom in back-to-back semesters (Figure 1, Instructor B). We used the COPUS to document instructor behaviors and a separate analytical protocol to document formative assessment and feedback (see *Formative Assessment Practices*).

Study Context

Observations were conducted in the introductory biology majors' sequence at a doctoral-granting research-intensive land grant university in the upper Midwest. The courses were large-enrollment (100+ students per section) and served a variety of majors, primarily biology, pre-health, pre-pharmacy, and agriculture. Most students were first- or second-year students, although juniors and seniors were also enrolled in the courses.

The instruction observed for this study occurred either in a large lecture hall (300+ seats) with rows of seats facing the instructor station and slide viewing in the front of the classroom only (Figure 1, one semester of data) or a SCALE-UP classroom (Figure 1, three semesters of data). In the SCALE-UP setting, there were 15 round tables that seated nine students each (135 seats total). Each table was equipped with connections to an individual monitor, and access to whiteboards was provided on all walls of the room. Six large screens as well as the individual monitors were situated around the perimeter of the room, with the instructor station near the middle. Undergraduate learning assistants (Otero *et al.*, 2010) were employed in both learning environments to increase instructional interaction with groups.

Observations

Class sessions were 75 minutes long and were observed and video-recorded (with the camera focused on the instructor station) four times each semester, the minimum number suggested by Stains and colleagues (2018). Observation days were chosen to be as evenly spaced as possible throughout the semester and avoided days with quizzes or exams. Instructors consented for observations to happen at any point during the semester and did not know ahead of time which days they would be observed.

Transcripts were generated from the video-recorded observations. The transcripts included all recorded utterances of the instructor and students that were directed or audible to the entire class. These transcripts were used for coding formative assessment prompts and responses, with reference to the video when needed to clear up any ambiguity in the transcript.

COPUS Instructional Styles

Trained observers used the COPUS to document instructor and student behaviors in real time. We submitted the COPUS data from each classroom observation to the COPUS analyzer (http://copusprofiles.org), which categorized each observation into a COPUS profile and instructional style. Observations fell into all three COPUS instructional styles and six of the seven profile clusters (2–7). Each instructor exhibited at least four different COPUS profiles and all three styles during the two semesters they were observed (Figure 1). Additionally, all three styles were observed in both the lecture hall and SCALE-UP settings.

Formative Assessment Practices

The practice of formative assessment involves eliciting evidence of student reasoning, diagnosing students' progress toward desired learning outcomes, and providing feedback that ultimately shapes student learning (Nicol and Macfarlane-Dick, 2006; Boud and Molloy, 2013; Offerdahl *et al.*, 2018). We characterized formative assessment practices in terms of: 1) the frequency with which student reasoning was elicited (i.e., frequency of formative assessment prompts), 2) cognitive level of the formative assessment prompts, 3) the type and frequency of feedback provided by the instructor, and 4) the extent to which students were asked to expand on their answers or explain their reasoning. We documented only the formative assessment and feedback that was provided to the entire class. We did not record instructors' interactions while moving through groups or with individual students.

Formative Assessment Prompts. We defined a formative assessment prompt as any nonrhetorical instructional move that elicited evidence of student thinking, requiring students to respond from the cognitive domain. Formative assessment prompts most often included clicker questions, verbal questions, and worksheets. Prompts that followed students' responses to an initial question often asked them to expand on their response (EXP) which provided additional, and sometimes deeper, student thinking. Questions such as "Everybody with me?" or "Any questions on that?" were excluded from analysis, because they did not directly assess the cognitive domain.

The total number of prompts for each class session was counted (n = 458), and cognitive level was determined. Two coders who were experienced in assigning Bloom's levels (E.G.O. and J.B.A.) independently reviewed each prompt and determined the cognitive level using Bloom's taxonomy (Krathwohl, 2002) with an initial agreement of 95.6% (linear

TABLE 1. Instructional feedback and responses that were coded after a student response to a formative assessment prompt or after a student question

Code	Description
PR	Praising or providing encouragement
DD/V	Displaying or verbally describing a distribution of student responses or themes within student responses
EV	Providing an evaluation of the correctness of the student response
C	Clarifying or providing further information
SH	Shaping student behavior such as study skills
EXP	Prompting students to explore or elaborate on ideas (also coded as a new prompt)

weighted kappa = 0.94). When coders disagreed, the assigned codes differed by one level in most cases (e.g., knowledge vs. comprehension). On prompts for which agreement differed by two levels, the coders agreed on the type of skill assessed but initially disagreed on the depth of conceptual understanding required to answer the prompt. For example, analysis questions assess similar skills to comprehension-level tasks but require more contextualization from novice learners (Anderson *et al.*, 2001; Arneson and Offerdahl, 2018). All instances of disagreement were negotiated to reach consensus.

Feedback. When student thinking is revealed by a formative assessment prompt, an instructor has the opportunity to diagnose students' in-progress learning and provide feedback. Therefore, we examined the instructor moves that followed student responses to formative assessment prompts to identify and characterize instructor-generated feedback (Ruiz-Primo and Furtak, 2006; Furtak *et al.*, 2017). We also attended to instructor responses to student questions as a source of instructor-generated feedback, because student questions provide information about and an opportunity to change a student's trajectory toward the desired learning outcome.

For the purposes of this analysis, we used Shute's (2008) definition of formative feedback: "information communicated to the learner that is intended to modify the learner's thinking or behavior for the purpose of improving learning" (p. 1). We first went through transcripts using Shute's definition to identify all instances of feedback. Three coders (MM, EO, JB) then worked independently to read the instances of feedback and identify themes. Through an iterative process of identifying categories, discussing, and re-examining the data to refine categories, we arrived at a final coding scheme consisting of five types of instructor-generated feedback (Table 1): praise (PR); displaying distributions of student responses (DD/V); evaluations of the correctness of student responses (EV); clarifications or further information about the content (C); and shaping student behavior, such as using metacognitive strategies (SH). Additionally, we coded one type of instructional response that does not fit the definition of feedback (EXP) but often occurs in formative assessment cycles. The EXP code is not feedback but describes instances in which an instructor responded to evidence of student learning by soliciting additional, and potentially deeper, student thinking. As such, it was coded both as a new prompt and as EXP. Three raters (MM, EO, JB) coded for instructional feedback and EXP and a Fleiss's kappa (Fleiss, 1971) was calculated ($\kappa = 0.883$). All instances of disagreement were negotiated until a consensus was reached.

Comparison of Formative Assessment Practices by COPUS Instructional Style

We employed a descriptive approach to compare the formative assessment practices within and across the three COPUS instructional styles as observed over four academic semesters. Our intent here is not to provide a formal statistical comparison to generalize to a broader population of instructors; moreover, our sample size precludes that. Rather, we adopt a descriptive approach to support the claim that COPUS instructional styles may not reflect distinct patterns in formative assessment practices.

RESULTS

Throughout this section, our comparison of COPUS instructional styles will focus on two aspects of formative assessment practice: the types of instructor-generated feedback and responses given to student in-class work; and the frequency and cognitive level of formative assessment prompts given by the instructor.

We examined the types of instructor-generated feedback provided in class sessions that were categorized into each of the three COPUS styles (Figure 2). As might be expected, the types of feedback varied from one class session to another. Yet, with the exception of shaping feedback (SH), there were no discernible patterns in the types and frequency of feedback observed in each of the three COPUS instructional styles. Shaping feedback (SH) was only observed in class sessions categorized into the student-centered and interactive lecture styles, but not all student-centered or interactive lecture class sessions included shaping feedback. For all instructional styles, evaluative (EV) and clarifying (C) feedback were the most prevalent. Similarly, all instructional styles had instances of the provision of normative feedback in the form of a distribution of student answers (DD/V) to a voting "clicker-style" type question. Further, the frequency with which instructors asked students to expand on their thinking (EXP) varied between class sessions, but with no clear distinction between the three COPUS instructional styles (Figure 3).

Instructor-generated feedback depends on the nature of the student thinking produced by a formative assessment prompt (Offerdahl and Montplaisir, 2014). Formative assessment prompts that assess different cognitive levels provide distinct insights into student thinking and therefore affect the quality of feedback. We therefore examined the cognitive level of formative assessment prompts in each of the COPUS styles (Figure 4). As might be expected, the proportion of prompts at any of the cognitive levels varied between class sessions. Knowledge-, comprehension-, and application-level prompts predominated across all class periods. Synthesis-level prompts

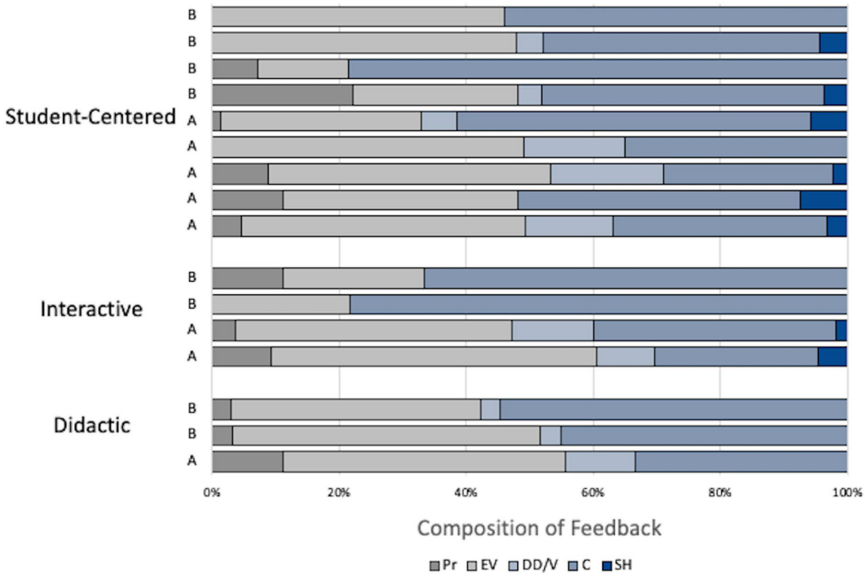

FIGURE 2. Percent of total instances of instructor-generated feedback provided in each class session for each of the five types of feedback: praise (Pr), evaluation (EV), displaying class vote (DD/V), clarifying (C), and shaping (S). Class sessions for both instructors (A and B) are organized by COPUS instructional style (didactic, three sessions; interactive, four sessions; student centered, nine sessions).

were only observed in class sessions categorized as interactive lecture or student centered. With the exception of synthesis-level prompts, the relative proportions of prompts observed in one COPUS instructional style can also be observed in one or more of the other styles (Figure 4 and Supplementary Figure S1).

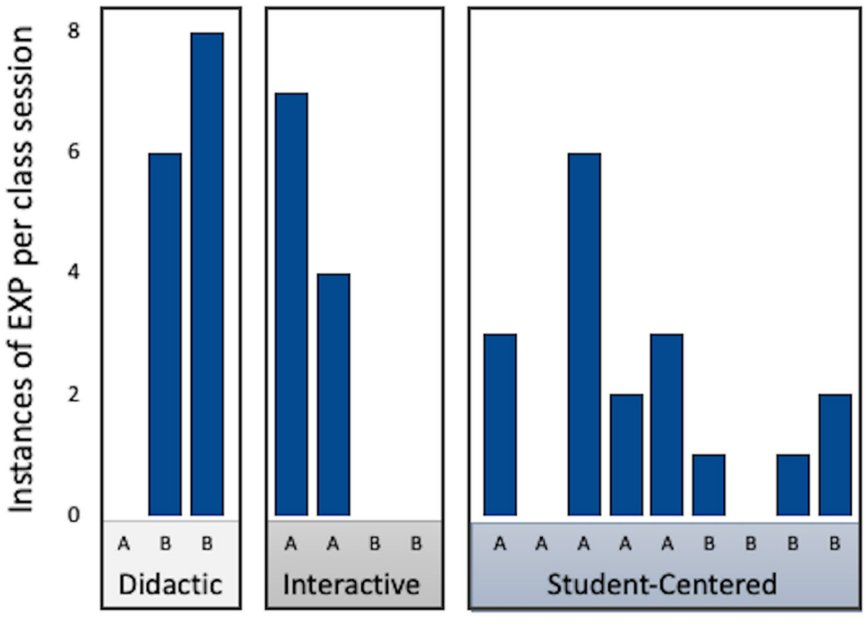

FIGURE 3. Frequency of instructor EXP moves within class sessions organized by COPUS instructional style.

To further illustrate the lack of alignment of COPUS styles with formative assessment practices, we identified three pairs of class sessions that were taught in two sections of biology in the same semester, covered the same content, and were categorized into the same COPUS styles (Figure 5a, blue box). Within each pair of class sessions, the proportions of types of feedback were qualitatively different both across sections and within instructors (Figure 5b, top panel). In terms of the cognitive level of the formative assessment prompts, the three class sessions taught by Instructor A were not easily distinguishable from one another (Figure 5b, bottom panel), yet they were categorized into two different instructional styles (student centered [S] and interactive lecture [I]). The three class sessions taught by Instructor B were more readily distinguishable from one another in terms of cognitive level but were only categorized into two distinct instructional styles.

DISCUSSION

The COPUS has been used widely to quantify classroom practice and how class time is spent in undergraduate STEM courses (e.g., Smith *et al.*, 2014; Stains *et al.*, 2018). COPUS data have also been used to categorize instructional styles ranging from didactic to student centered (e.g., Lund *et al.*, 2015; Stains *et al.*, 2018). We examined classroom observations categorized into each of the three COPUS instructional styles (didactic, interactive lecture, and student centered) to investigate implicit assumptions that student-centered styles are higher quality by characterizing formative assessment and feedback moves, which have been associated with increases in student learning (e.g., Hattie and Timperley, 2007; Hattie, 2009; Hattie and Zierer, 2019; Offerdahl *et al.*, 2018).

Formative assessment and feedback were detected in all class sessions. Not surprisingly, the formative assessment practices varied from one observation to another. Yet these differences did not correspond with the three COPUS instructional styles. Class sessions with nearly identical patterns in formative assessment practices were categorized into different instructional styles. Conversely, class sessions within the same instructional style were observed to have very different patterns of formative assessment. Although not designed to measure formative assessment or feedback, the COPUS does account for instructor behaviors that are part of the formative assessment process (e.g., posing a question, following up).

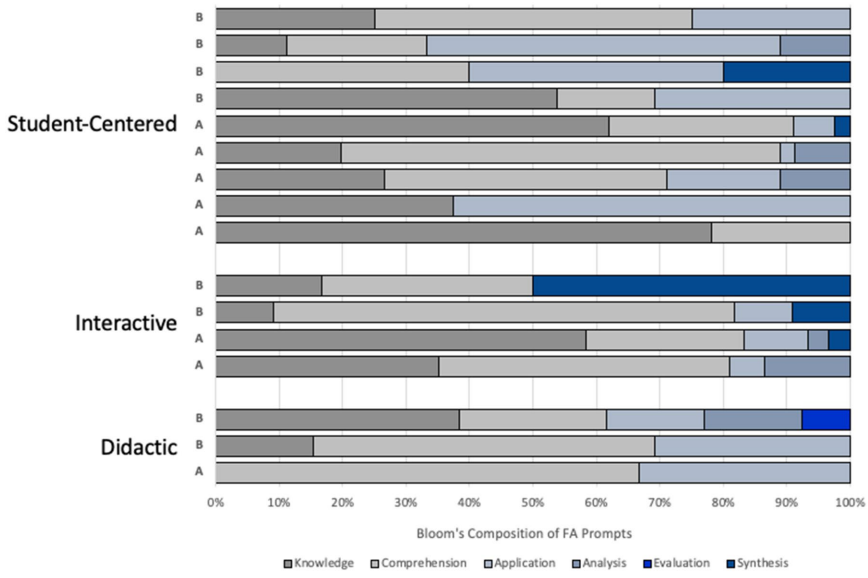

FIGURE 4. Proportion of formative assessment (FA) prompts at each of the observed cognitive levels (knowledge, comprehension, application, analysis, and synthesis) for each class session taught by both instructors (A and B) organized by COPUS style.

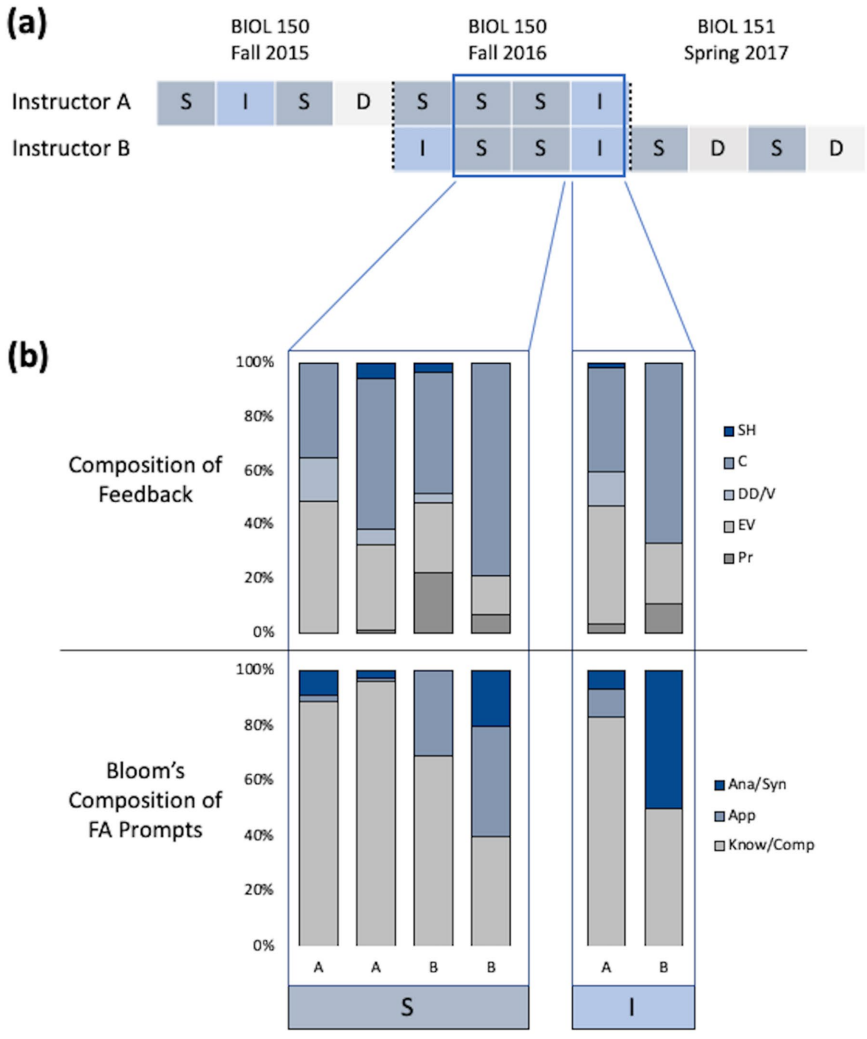

FIGURE 5. (a) Three pairs of class sessions covering the same content and categorized into the same instructional style were identified. These pairs of class sessions were further examined in terms of the frequency of types of feedback observed in each class session (b, top) and cognitive level (b, bottom). See Table 1 for codes.

Our data indicate that COPUS styles may not reliably distinguish differences in instructors' formative assessment and feedback practices, particularly between the interactive lecture and student-centered profiles. These results are important, particularly given the tacit assumption that moving from didactic to interactive lecture to student-centered instruction is a desirable and productive trajectory. Our results underscore the need for caution when interpreting COPUS data, and COPUS styles in particular, and the need for additional measures to more fully characterize the degree to which high-impact instructional practices are occurring.

We followed the recommendation of Stains and colleagues (2018) to conduct four classroom observations per semester. All three COPUS instructional styles were observed in this study, and all but one of the seven profiles (Figure 1). Yet no single instructor exhibited a single instructional style. Two or more instructional styles were observed for each instructor. Our results suggest that a sampling intensity of greater than four observations is needed to validly characterize instructional practices for a single instructor. Emerging work supports our finding and recommends at least double this sampling size (Goodridge *et al.*, 2020). These data demonstrate the importance of considering multiple time points in a semester when investigating instructor practices; the more observations that are collected, the more complete the picture of evidence-based practices like assessment and feedback will be (Reichenbach, 2017).

Finally, our results suggest potential areas for growth in instructor practice of formative assessment in active-learning classrooms. Two productive aspects of formative assessment that were less frequently observed in our study were the use of higher-order formative assessment prompts and the prompting of students to expand on their thinking (EXP). Both of these instructional moves probe student thinking at a deeper level, helping students

recognize and move toward reaching learning objectives (Hattie and Timperley, 2007; Evans, 2013). Although evidence of student thinking was prompted in all semesters, prompts primarily elicited lower-order cognitive skills (i.e., knowledge and comprehension). Additionally, students were much more likely to receive evaluative feedback (e.g., "that is correct") or clarifying feedback (e.g., "hydrogen bonds are just one type of hydrophobic interaction") than to be encouraged to expand their thinking. Both higher-order prompts and eliciting further evidence of student thinking tend to take up more class time. Thus, it is difficult to prescribe the proportion of prompts that "should" be higher order or the amount of student thinking that "should" be further explored with additional prompting. Yet higher-order prompts and asking students to expand on their reasoning were infrequent even within the observed classrooms, selected because they were likely to be assessment rich. These data suggest that further attention needs to be paid to the formative assessment practices in undergraduate courses even if taught by instructors in supportive teaching environments who are motivated and have taken advantage of multiple opportunities for professional development.

LIMITATIONS

This study was limited in scope (i.e., one university, a single course series, two instructors over four academic semesters) and therefore prohibits generalization to other instructors in different contexts. Further, while we took care to observe class sessions throughout each semester, we were only able to collect detailed formative assessment and feedback data on four equally spaced class sessions per semester. We were not able to capture feedback or prompts given to individual students or groups in this context, but only at the whole-class level. Therefore, formative assessment and feedback given by the learning assistants (Thompson *et al.*, 2020) or the instructor as they interacted with individual students or groups was not captured.

CONCLUSION

Over the last couple decades, evidence has accumulated about the types of instructional practices likely to positively impact student learning. Lecture-based pedagogies have been repeatedly demonstrated to be less effective than student-centered, active-learning classrooms (Freeman *et al.*, 2014). Formative assessment and feedback are not only critical for student learning but are by their very design a hallmark of active-learning classrooms (Rosenberg *et al.*, 2018). As national initiatives to transform undergraduate STEM education maintain momentum, there is a continuing need for measures that can document the efficacy of these initiatives. The COPUS is a straightforward and reliable tool that documents many observed instructor and student behaviors (Smith *et al.*, 2013) and can be used to categorize class sessions into three instructional styles along a continuum from didactic to student centered (Stains *et al.*, 2018). The goal of this study was to understand the degree to which COPUS styles represent meaningful differences in formative assessment practices. To this end, we purposively studied an undergraduate biology department that was supportive of instructional innovation and instructors who had received extensive professional development in evidence-based instructional practices (including formative assessment). Our sample, though small, adhered to current recommendations for minimum number of observations. Yet we were unable to detect distinct differences in the formative assessment practices enacted in class sessions that were categorized into each of the three styles. These results underscore the need for a cautious interpretation of COPUS styles, particularly the implicit assumption that interactive lecture and student-centered profiles differ from each other in terms of practices that support student learning. The work presented here is one example of how using additional data streams to provide a more complete picture of undergraduate teaching can help reform initiatives concentrate on the most important aspects and continue to have a positive impact on undergraduate STEM education.

ACKNOWLEDGMENTS

The authors acknowledge the instructors who participated in this study, as well as Sam Pazcini, Rebecca Reichenbach, and Lisa Wiltbank for fruitful conversations about this work. The authors also thank two anonymous reviewers for their critical feedback and assistance framing this work. This material is based on work supported by the National Science Foundation under grant no. 1431891.

REFERENCES

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Arneson, J. B., & Offerdahl, E. G. (2018). Visual literacy in Bloom: Using Bloom's taxonomy to support visual learning skills. *CBE—Life Sciences Education*, *17*(1), ar7.

Beichner, R. J. (2008). *The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project*. Retrieved December 13, 2020, from www.per-central.org/items/detail.cfm?ID=4517

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Brancaccio-Taras, L., Pape-Lindstrom, P., Peteroy-Kelly, M., Aguirre, K., Awong-Taylor, J., Balser, T., ... & Zhao, J. (2016). The PULSE Vision and Change Rubrics, Version 1.0: A valid and equitable tool to measure transformation of life sciences departments at all institution types. *CBE—Life Sciences Education*, *15*(4), ar60. https://doi.org/10.1187/cbe.15-12-0260

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*(4), 13–17. https://doi.org/10.1111/j.1745-3992.2012.00251.x

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment and Evaluation in Higher Education*, *44*(5), 705–714. https://doi.org/10.1080/02602938.2018.1531108

Committee on STEM Education of the National Science and Technology Council. (2018). *Charting a Course for Success: America's Strategy for STEM Education*. Washington, DC: Executive Office of the President of the United States.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, *7*(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, *4*(1), 37–73.

Emery, N. C., Maher, J. M., & Ebert-May, D. (2020). Early-career faculty practice learner-centered teaching up to 9 years after postdoctoral professional development. *Science Advances*, *6*(25), eaba2091. https://doi.org/10.1126/sciadv.aba2091

Esterhazy, R., & Damsa, C. (2019). Unpacking the feedback process: An analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, *44*(2), 260–274. https://doi.org/10.1080/03075079.2017.1359249

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences USA*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Furtak, E. M., Ruiz-Primo, M. A., & Bakeman, R. (2017). Exploring the utility of sequential analysis in studying informal formative assessment practices. *Educational Measurement: Issues and Practice*, *36*(1), 28–38. https://doi.org/10.1111/emip.12143

Gess-Newsome, J., Southerland, S. A., Johnston, A., & Woodbury, S. (2003). Educational reform, personal practical theories, and dissatisfaction: The anatomy of change in college science teaching. *American Educational Research Journal*, *40*(3), 731–767.

Goodridge, J. A., Gordon, L., Nehm, R., & Sbeglia, G. C. (2020). Faculty adoption of evidence-based teaching practices: The role of observation sampling intensity on measures of change held in 24 July 2020, Minneapolis, MN. Society for the Advancement of Biology Education Research.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112.

Hattie, J., & Zierer, K. (2019). *Visible learning insights*. New York, NY: Routledge.

Hattie, J. A. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. New York, NY: Routledge.

Holland, T., Sherman, S. B., & Harris, S. (2018). Paired teaching: A professional development model for adopting evidence-based practices. *College Teaching*, *66*(3), 148–157.

Kraft, M., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

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.

McConnell, M., Montplaisir, L., & Offerdahl, E. (2019). Meeting the conditions for diffusion of teaching innovations in a university STEM department. *Journal for STEM Education Research*, *3*, 43–68. https://doi.org/10.1007/s41979-019-00023-w

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment and Evaluation in Higher Education*, *35*(5), 501–517. https://doi.org/10.1080/02602931003786559

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Offerdahl, E. G., McConnell, M., & Boyer, J. (2018). Can I have your recipe? Using a fidelity of implementation (FOI) framework to identify the key ingredients of formative assessment for learning. *CBE—Life Sciences Education*, *17*(4), es16. https://doi.org/10.1187/cbe.18-02-0029

Offerdahl, E. G., & Montplaisir, L. (2014). Student-generated reading questions: Diagnosing student thinking with diverse formative assessments. *Biochemistry and Molecular Biology Education*, *42*(1), 29–38. https://doi.org/10.1002/bmb.20757

Otero, V., Pollock, S., & Finkelstein, N. (2010). A physics department's role in preparing physics teachers: The Colorado Learning Assistant Model. *American Journal of Physics*, *78*(11), 1218–1224. https://doi.org/10.1119/1.3471291

Pfund, C., Miller, S., Brenner, K., Bruns, P., Chang, A., Ebert-May, D., ... & Handelsman, J. (2009). Summer Institute to improve university science teaching. *Science*, *324*(5926), 470–471. https://doi.org/10.1126/science.1170015

Reichenbach, R. S. D. (2017). Surfing the semester: A *study of the flow of active learning implementation* (Master's thesis). North Dakota State University, Fargo. Retrieved November 10, 2020, from https://library.ndsu.edu/ir/handle/10365/28420

Reinholz, D. L., Pilgrim, M. E., Corbo, J. C., & Finkelstein, N. (2019). Transforming undergraduate education from the middle out with departmental action teams. *Change: The Magazine of Higher Learning*, *51*(5), 64–70. https://doi.org/10.1080/00091383.2019.1652078

Rosenberg, M. B., Hilton, M. L., & Dibner, K. A. (2018). *Indicators for monitoring undergraduate STEM education (Consensus study report)*. Washington, DC: National Academies Press.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning *Educational Assessment*, *11*(3–4), 237–263.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Smith, M. K., Jones, F. H. M., 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. https://doi.org/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. https://doi.org/10.1187/cbe.14-06-0108

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16*(1), rm1. https://doi.org/10.1187/cbe.16-03-0113

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Thompson, A. N., Talbot, R. M., Doughty, L., Huvard, H., Le, P., Hartley, L., & Boyer, J. (2020). Development and application of the Action Taxonomy for Learning Assistants (ATLAs). *International Journal of STEM Education*, *7*(1), 1. https://doi.org/10.1186/s40594-019-0200-5

Weir, L. K., Barker, M. K., McDonnell, L. M., Schimpf, N. G., Rodela, T. M., & Schulte, P. M. (2019). Small changes, big gains: A curriculum-wide study of teaching practices and student learning in undergraduate biology. *PLoS ONE*, *14*(8), e0220900.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, J., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York: New Teacher Project.

Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. *Proceedings of the National Academy of Sciences USA*, *111*(23), 8319–8320.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087.