Are Faculty Changing? How Reform Frameworks, Sampling Intensities, and Instrument Measures Impact Inferences about Student-Centered Teaching Practices

Gena C. Sbeglia,[†*] Justin A. Goodridge,[†] Lucy H. Gordon,[†] and Ross H. Nehm[†‡]
[†]Department of Ecology and Evolution and [‡]Program in Science Education, Stony Brook University, Stony Brook, NY 11794

ABSTRACT

Although recent studies have used the Classroom Observation Protocol for Undergraduate STEM (COPUS) to make claims about faculty reform, important questions remain: How should COPUS measures be situated within existing reform frameworks? Is there a universal sampling intensity that allows for valid inferences about the frequency of student-centered instruction within a semester or across semesters of a course? These questions were addressed using longitudinal COPUS observations (128 classes, three faculty, 4 years). COPUS behaviors were used to categorize classes into didactic, interactive lecture, or student-centered instructional styles. Sampling intensities (one to 11 classes) were simulated (1000 times) within a course and across semesters. The sampling intensities required for generating valid inferences about 1) the *presence* of student-centered instruction and 2) the *proportion* of instructional styles in a course and through time were calculated. Results indicated that the sampling intensity needed to characterize courses and instructors varied and was much higher than previously recommended for instructors with: 1) rare instances of student-centered classes, 2) variability in instructional style, and 3) longitudinal changes in instructional patterns. These conditions are common in early reform contexts. This study highlights the risks of broad, decontextualized sampling protocol recommendations and illustrates how reform frameworks, sampling intensities, and COPUS measures interact to impact inferences about faculty change.

INTRODUCTION

A decade ago, the American Association for the Advancement of Science (AAAS) published ambitious policy guidelines calling for evidence-based reform of the teaching, learning, and assessment of undergraduate biology (*Vision and Change*; AAAS, 2011). The report urged biology faculty to enact research-based instructional strategies (RBIS) and reduce didactic lecturing. A growing body of empirical work continues to bolster these policy recommendations. RBIS have been found to increase learning outcomes, enhance sense of belonging, and improve retention in science, technology, engineering, and mathematics (STEM; Booker, 2007; Freeman *et al.*, 2014; Salamone and Thomas, 2017). Moreover, RBIS have been found to disproportionately benefit students from backgrounds underrepresented in science (Theobald *et al.*, 2020). Given unambiguous research findings supporting the benefits of RBIS, biology education researchers have worked to develop and implement instruments capable of generating robust inferences about faculty teaching practices to determine whether the recommendations set forth by *Vision and Change* have been realized (e.g., Smith *et al.*, 2013; Stains *et al.*, 2018; National Science Foundation, 2020).

Tessa C. Andrews, *Monitoring Editor*
Submitted Nov 16, 2020; Revised May 11, 2021; Accepted May 20, 2021
CBE Life Sci Educ September 1, 2021 20:ar39
DOI:10.1187/cbe.20-11-0259
*Address correspondence to: Gena C. Sbeglia (gena.sbeglia@stonybrook.edu).
© 2021 G. C. Sbeglia *et al.* CBE—Life Sciences Education © 2021 The American Society for Cell Biology. This article is distributed by The American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0).
"ASCB®" and "The American Society for Cell Biology®" are registered trademarks of The American Society for Cell Biology.

One instrument that has figured prominently in efforts to empirically characterize faculty teaching practices is the Classroom Observation Protocol for Undergraduate STEM (COPUS). The COPUS is a published, validated, and widely used instrument designed to collect and categorize class-level observational data about student and instructor behaviors in undergraduate learning environments (Smith *et al.*, 2013). Trained raters (i.e., those able to reliably observe teaching behaviors) score specifically defined behaviors, many of which are characteristic of student-centered teaching. Since publication, the COPUS instrument has been used to study a range of educational topics, including: 1) quantitative comparisons of faculty teaching practices across different classroom contexts (e.g., flipped vs. traditional, large vs. small; e.g., Smith *et al.*, 2014; Maciejewski, 2016; Teasdale *et al.*, 2017; Stains *et al.*, 2018); 2) evaluation of the alignment between faculty perceptions and actual teaching practices (e.g., Smith *et al.*, 2013, 2014; Reisner *et al.*, 2020); 3) assessment of educational programs (e.g., Deligkaris and Chan Hilton, 2019); and 4) generation of inferences about the progress of reform (e.g., Lund *et al.*, 2015; Stains *et al.*, 2018).

This study focuses on the use of the COPUS to generate inferences about the enactment of and progress toward student-centered teaching (e.g., Lund *et al.*, 2015; Stains *et al.*, 2018). Stains *et al.* (2018) used COPUS measures to draw inferences about the magnitude of educational reform in undergraduate STEM instruction in the United States by studying more than 2000 classes taught by more than 500 instructors in 24 universities. The authors reported that these undergraduate STEM courses were dominated by teacher-centered pedagogies (e.g., lecturing) and argued that further efforts were needed to reform STEM education nationally. Although Smith *et al.* (2013) noted that the COPUS could be used to "[compare] practices longitudinally" (Smith *et al.*, 2013, p. 626) and Lund *et al.* (2015) claimed that the COPUS can "measure the extent of changes in instructional practices as a result of instructional reforms" (Lund *et al.*, 2015, p. 10), change in alignment with instructional reform does not appear to be the construct that the COPUS was originally designed to measure (The COPUS was designed to measure class-level frequencies of student and instructor behaviors). Therefore, important questions remain about how to use the COPUS instrument in a manner that permits valid inferences about 1) change in faculty behaviors through time and 2) how measures of change articulate with existing conceptual frameworks for educational reform. These questions are discussed in detail in this section.

First, it remains unclear how many classes must be sampled to generate valid inferences about learning environments both within a semester-long course (i.e., at the course level) and across semesters of the same course (i.e., at the faculty change level) or whether a universal sampling intensity recommendation is even possible to establish. For example, Stains *et al.* (2018) concluded that the characterization of instructional practices *for a given course* requires a sampling intensity of at least four classes (not three, as previously suggested by Lund and Stains, 2015). Yet less than 1% of the sample studied by Stains *et al.* (2018) included large numbers of classes (e.g., >10 classes) taught by a single instructor, raising questions about whether the source of evidence was sufficient to uphold this claim. Indeed, Stains *et al.* (2018) reported that as more classes were studied using the COPUS, the diversity of instructional styles increased (Stains *et al.*, 2018). Because this analysis appears to have lumped the sampling intensity of four classes with all higher sampling intensities, it is difficult to know whether four classes reached the threshold necessary to capture the actual diversity of instructional styles or proportions of each style. Goodridge *et al.* (2019) found that the sampling intensity needed to accurately and precisely characterize (i.e., with a 75% probability) individual COPUS behaviors for an instructor was often greater than four classes, especially for behaviors such as "working in groups" (WG), "lecturing" (Lec), and "asking a clicker question" (CQ). These findings suggest that 1) there is a relationship between sampling intensity and inferences about instructor behaviors and 2) prior data sets may not be well suited for testing claims about such relationships.

A second question in need of attention relates to the conceptual grounding of COPUS instrument scores. Instrument scores only have meaning (and thus can be used for measurement) when they are explicitly linked to a conceptual framework (Berger and Patchener, 1988; Nehm, 2019; King *et al.*, 1994; National Research Council [NRC], 2001; Leshem and Trafford, 2007; American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education [AERA *et al.*], 2014). The use of instruments that measure reform-based instructional practices require grounding in realistic and evidence-based models of how STEM faculty reform actually occurs (e.g., Lewis *et al.*, 2006; Henderson *et al.*, 2009, 2011). For example, the COPUS involves the characterization of behaviors from an individual class, whereas many reform frameworks conceptualize behavioral changes over time (e.g., Dancy *et al.*, 2007; Henderson *et al.*, 2011). At present, COPUS score interpretations have lacked 1) conceptualization at scales above the individual class (e.g., at the course level and the faculty change level), 2) grounding in conceptual reform frameworks that consider the nature of faculty change, and 3) delineation of the types of change that are considered meaningful vis-à-vis a conceptual framework. Indeed, the meaning of "faculty change" depends on the lens through which it is viewed (i.e., the conceptual framework); faculty change can occur in many different ways—quickly or slowly, linearly or nonlinearly, unidirectionally or with backtracking (Cuban, 1992; Cavallo, 2004; e.g., Silverthorn *et al.*, 2006; Hoyle and Wallace, 2007; Dancy and Henderson, 2008). These patterns may be linked to the affordances and constraints of the reform environment in which faculty work (Dancy and Henderson, 2005; Henderson *et al.*, 2009). The conceptual framing of a study therefore informs which of these changes should be considered meaningful, and researchers may therefore define meaningful change in different ways (e.g., faculty movement toward any evidence-based approaches, the majority of classes being student centered, alignment with institutional benchmarks). Explicit delineation of the reform target (e.g., frequency of instructor behaviors, patterns of instructor behavior change through time) must be made in order to select an appropriate sampling strategy. The sampling strategy will impact the measures that are generated and the inferences that are made using these measures. Figure 1 summarizes the essential features that

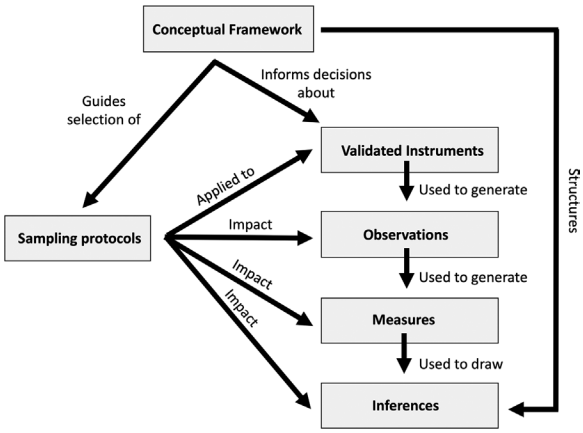

FIGURE 1. Relationship between conceptual frameworks, sampling protocols, instruments, instrument measures, and inferences.

need to be considered when embarking upon the measurement of faculty behaviors.

COPUS sampling approaches must be in alignment with a conceptual framework of reform for researchers to 1) delineate a sampling strategy appropriate for the reform context and 2) generate valid inferences about faculty change derived from instrument measures. For example, if a researcher's framework for reform indicates that slow, nonlinear patterns of faculty change are to be expected during early adoption stages, then capturing rare reform-based behaviors will be needed for generating evidence of reform-based progress. Addressing these limitations is needed to help researchers draw valid inferences about reform-based progress using COPUS scores and measures.

This study seeks to advance understanding of how the COPUS may be used to characterize undergraduate science education reform, particularly in early reform contexts. It does so by studying the impact that conceptual framing and sampling intensity have on inferences about faculty teaching practices and faculty change. In this study, meaningful change is defined as the movement, no matter how small, of faculty practices toward or away from evidence-based approaches. Specifically, when situated within a conceptual framework of reform (discussed below), we ask:

How many classes must be sampled in order to…

RQ1: generate valid inferences about the learning environment at the level of the *course*?
RQ1.1: document the *presence* of at least one student-centered class within a *course*?
RQ1.2: accurately and precisely estimate the *proportion* of different instructional styles within a *course*?

RQ2: generate valid inferences about *faculty change*?
RQ2.1: document *changes* in the *presence* of student-centered classes throughout *multiple semesters* of an instructor's course?
RQ2.2: accurately and precisely estimate *changes* in the *proportion* of different instructional styles throughout *multiple semesters* of an instructor's course?

CONCEPTUAL FRAMEWORKS, SAMPLING, AND MEASUREMENT

Situating a research question within a conceptual or theoretical framework is widely accepted as an essential feature of educational research (Institute of Education Sciences and National Science Foundation [IES and NSF], 2013; Creswell and Guetterman, 2019). Conceptual frameworks impact a researcher's sampling and measurement choices, which in turn impact inferences derived from those measures (cf. AERA *et al.*, 2014). The alignment (or lack thereof) among conceptual frameworks, measurement tools, and sampling approaches determines whether accurate or inaccurate inferences about the system being studied can be made. Observation protocols like the COPUS have great potential for facilitating the measurement of reform-based progress, but only when linked with an evidence-based conceptual framework of reform (e.g., Cuban, 1992, 1999; Dancy and Henderson, 2005; Henderson *et al.*, 2009). Many different alignments among conceptual frameworks, measurement tools, and sampling strategies are possible and acceptable, but they should be made explicit (AERA *et al.*, 2014).

One particularly influential conceptual framework for educational reform characterizes change in two ways: incremental change and fundamental change (Cuban, 1992, 1999). This framework specifically refers to the *nature* of change, not simply the *pace* of change. Incremental change assumes that the underlying structures of a system are appropriate and that improvement can be attained by incremental modifications that build upon the status quo (Cuban, 1992, 1999; Mathison, 2005). Incremental change tends to lack significant institutional barriers (Elmore, 1996; Henderson *et al.*, 2009) and is often expected to be linear and unidirectional in nature (Cavallo, 2004).

Fundamental change, on the other hand, assumes that the underlying system is problematic and that modifications that build on the existing system will not be able to accommodate the required improvements; rather, structural modifications are needed (Cuban, 1992; 1999). Fundamental change is likely to be slow and nonlinear, particularly in the early stages (Cuban, 1992; Cavallo, 2004; e.g., Silverthorn *et al.*, 2006; Hoyle and Wallace, 2007; Dancy and Henderson, 2008). This slow, nonlinear pattern reflects existing departmental or institutional structures that may act as barriers to change (Dancy and Henderson, 2005; Henderson *et al.*, 2009). Importantly, the teaching reforms outlined in *Vision and Change* (e.g., active learning and other student-centered practices; AAAS, 2011) often require fundamental (not incremental) change, particularly as related to the roles that instructors play in biology classrooms (Henderson *et al.*, 2009).

Other authors have proposed frameworks that make similar distinctions between types of educational change. For example, Kezar (2018) delineated between first- and second-order changes, which align rather closely with Cuban's incremental versus fundamental change. Specifically, first-order changes are described as those requiring minor improvement or adjustments to the system, whereas second-order changes (or deep changes) are described as those requiring engagement with "underlying values, assumptions, structures, processes, and culture" (Kezar, 2018, p. 49). Because the incremental versus fundamental change framework has been explicitly connected to the reform of STEM classrooms (e.g., Henderson *et al.*, 2009),

we will use Cuban's framing but emphasize that other frameworks are valuable and may also be appropriate.

The incremental change and fundamental change framework may be used to guide and inform sampling strategies for measuring reform-based progress within institutions through time. Because *incremental* change is expected to be linear, unidirectional, and less constrained by institutional barriers (Elmore, 1996; Cavallo, 2004; Henderson *et al.*, 2009), researchers who employ this framework could choose to collect data at the beginning and end of a study period (e.g., year 1 and 4 of a reform project) and expect to make valid inferences about progress. *Fundamental* change, in contrast, is expected to be nonlinear and to involve backtracking because of institutional barriers (Cuban, 1992; Cavallo, 2004; Dancy and Henderson, 2005; Henderson *et al.*, 2009). Therefore, a researcher who adopts the latter framework might choose to employ a sampling strategy in which data collection occurs at more frequent intervals so as to capture complex dynamics and better understand the barriers constraining reform. Furthermore, because fundamental change is expected to be slow (particularly in early reform contexts; Cuban, 1992; Cavallo, 2004), being able to detect *small* changes in faculty practice through time may be central to accurately and precisely characterizing movement forward (and backward). In this conceptual framework, the tools and sampling approaches used to measure reform-based progress must be capable of capturing small shifts toward or away from student-centered pedagogies. These hypothetical examples serve to illustrate the interconnections among conceptual frameworks and sampling.

The COPUS was originally designed to measure faculty teaching behaviors within an individual class (Smith *et al.*, 2013), and more recently it has been used to make inferences about the progress of reform (e.g., Lund *et al.*, 2015; Stains *et al.*, 2018). However, as noted earlier, the instrument has not been explicitly situated within a conceptual framework of reform. Because many COPUS behaviors are aligned with active-learning pedagogies (Smith *et al.*, 2013), and because COPUS measures have been used to classify a class as student centered (or not; e.g., Lund *et al.*, 2015; Stains *et al.*, 2018), one approach is to anchor the COPUS instrument within the *fundamental change* perspective.

For the COPUS to be used to measure faculty adoption of student-centered practices through time, the sampling approach (e.g., number or proportion of classes, number of semesters) must also align with this framework. Given the nature of fundamental change described earlier, *any* change in the adoption of student-centered practices may be viewed as central to the measurement of faculty change. Indeed, if one student-centered class in an otherwise didactic course were to be missed, then progress—however small—would not be detected (and conversely, backtracking or the abandonment of reform practices would also be missed). Therefore, according to the fundamental change framework, rare instances of student-centered practices within a course and through time must be able to be detected by the sampling approaches employed. In sum, using COPUS measures to draw inferences about faculty (and institutional) change requires grounding in a conceptual framework, which in turn guides the choice of sampling strategies (IES and NSF, 2013; Creswell and Guetterman, 2019).

MATERIALS AND METHODS

COPUS Measures

The COPUS was adapted from the Teaching Dimensions Observation Protocol (TDOP; Hora and Ferrare, 2010; Hora *et al.*, 2013) and was designed to characterize undergraduate STEM learning environments (Smith *et al.*, 2013). The COPUS facilitates documentation of 13 student and 12 instructor behaviors (see Supplemental Table 1) that collectively describe "the full range of normal classroom activities of students and instructors" (Smith *et al.*, 2013, p. 623). Trained raters document the presence or absence of these 25 behaviors at 2-minute intervals throughout the duration of a class session. Although several of these behaviors are known to be associated with student-centered classrooms, the COPUS does not require observers to make judgments about the quality of teaching or the alignment of instruction with student-centered pedagogies. Expert feedback regarding the extent to which the COPUS describes the full range of typical faculty and student classroom activities was used as validity evidence (Smith *et al.*, 2013). Scoring reliability evidence for the COPUS included high interrater reliability (Smith *et al.*, 2013).

Composite Measures Derived from the COPUS

Although the COPUS does not directly measure the extent to which classroom behaviors align with student-centered pedagogies, other authors have derived these measures. Specifically, Stains *et al.* (2018) conducted latent profile analysis on eight of the COPUS behaviors (four instructor, four student; see Supplemental Table 1) using a large sample ($n = 2008$) of classes. They identified seven clusters, each of which represented a distinct instructional profile. The clusters were based on the proportion of 2-minute intervals in which the particular behaviors occurred. The authors then compiled these seven clusters into three broad class-level instructional styles: didactic, interactive lecturer, and student centered (see Table 1). To categorize a *class* into one of these three class-level instructional styles, raw COPUS data can be uploaded into an online platform known as the COPUS Analyzer (copusprofiles.org; Stains *et al.*, 2018).

COPUS Measures of Instructional Patterns within and across Scales

The COPUS was designed to generate inferences about STEM faculty at the scale of an individual class. Classroom dynamics at other scales, such as the course level, faculty change level, departmental level, and institutional level, could also be studied using the COPUS (Smith *et al.*, 2013), but no evidence-based standards have been established for such applications. In this paper, we focus on the course-level and faculty change–level scales, which are some of the scales that have been used to characterize and enact reform (e.g., Pfund *et al.*, 2009; Dancy and Henderson, 2010; Henderson *et al.*, 2011; Prince *et al.*, 2013; McCourt *et al.*, 2017; Matz *et al.*, 2018; Dancy *et al.*, 2019; Table 1 and Figure 2).

The *course-level* scale addresses the measurement of learning environments *within a course* and takes into account the strategies used in multiple *classes*. Three course-level categories are outlined in this paper: didactic only, mixed course, student-centered only (described in Table 1). Higher sampling intensities are expected to be required for accurately and precisely measuring the learning environments of mixed courses (in which there

TABLE 1. Description of the scales and categories used to characterize COPUS data

Scale	Category	Description
Class level: Addresses the measurement of learning environments in an *individual class* and is the scale that the COPUS is most clearly designed to target.	Didactic[a]	>80% of class time consists of instructors lecturing (Lec); low frequency of active-learning behaviors (i.e., CQ, CG, 1o1, OG, WG, PQ, SAnQ; Stains *et al.*, 2018).
	Interactive lecturer[a]	<80% of class time consists of lecturing (Lec); higher frequency of active-learning behaviors (i.e., CQ, CG,1o1, OG, WG, PQ, SAnQ; Stains *et al.*, 2018).
	Student centered[a]	Replacement of most lecturing (Lec) with a variety of active learning strategies, particularly one on one assistance (1o1; Stains *et al.*, 2018).
Course level: Addresses the measurement of the learning environment *within a course* for a given instructor in a given semester.	Didactic only	Consistently didactic within a semester
	Mixed	Multiple class-level classifications within a semester
	Student-centered only	Consistently student centered within a semester
Faculty change level: Addresses the measurement of the learning environment across an instructor's semesters of the *same* course. This scale measures *faculty change*.	Consistent	No change in course-level *proportion or category of* instructional styles across semesters
	Dynamic	Change in *proportion of* course-level classification across semesters

[a]Categories and descriptions from Stains *et al.* (2018).

is variability in instructional styles) compared with didactic-only or student centered-only courses. The *faculty change* scale, on the other hand, addresses the measurement of the learning environment through time (i.e., *across semesters*) within the *same* course. Two faculty change–level categories include: consistent across semesters and dynamic across semesters (described in Table 1).

Sampling Strategy for Courses and Instructors

The strategies used to sample participants are fundamental considerations in educational research (IES and NSF, 2013; Creswell and Guetterman, 2019) and may impact the quality and nature of the inferences drawn from instrument measures (AERA *et al.*, 2014). There are various methods for sampling participants (and larger units) in educational and social science research. In probability sampling, researchers randomly select participants so that the sample is representative of the entire population, thus allowing generalizations to be made from the researcher's data to the population at large (Creswell and Guetterman, 2019). Probability sampling can be logistically challenging and requires 1) defining the entity that the sample is intended to represent (i.e., the target sample), 2) knowing the scope of the population from which to sample (i.e., the sampling frame), 3) identifying appropriate criteria for successful sampling to be achieved, and 4) targeting a sufficiently large sample size to capture the variation in the population (Nugent, 2019; Ramsey *et al.*, 2019). Because of the time-intensive nature of the COPUS (taking as long to generate as the length of the class being observed), a probabilistic sampling approach is not always feasible. As a result, purposeful sampling approaches may be more appropriate for some studies employing the COPUS. In purposeful sampling, researchers nonrandomly select participants who represent the full range of variation within the system so as to gain an understanding about a central phenomenon (Creswell and Guetterman, 2019).

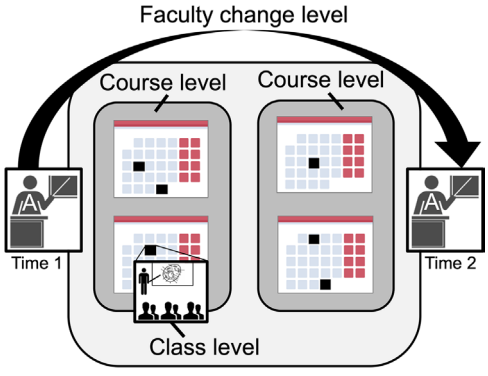

FIGURE 2. Scales (i.e., class level, course level, faculty change level) that may be used to characterize evidence-based teaching practices. The *class level* characterizes the measurement of learning environments in an individual class and is the scale that the COPUS is most clearly designed to target. The *course level* addresses the measurement of the learning environment within a course for a given instructor in a given semester. The *faculty change level* addresses the measurement of the learning environment across an instructor's semesters of the same course. This scale has the potential to measure faculty change. See Table 1 for additional details.

The sampling strategy selected for a study should align with the framework guiding the interpretation of the data (AERA *et al.*, 2014). The sampling strategy used to generate a COPUS data set that will be used to make inferences about reform should align with a researcher's conceptual framework for reform. At present, Stains *et al.* (2018) has generated the largest publicly available data set of COPUS measures. Unfortunately, this data set has several limitations that make it unsuitable for answering questions about how sampling intensity impacts inferences about reform-based progress. First, the Stains *et al.* (2018) data set is not situated within a single institutional context[1] and does not include many instances of an

[1]The data in Stains *et al.* (2018) came from 241 doctorate-granting universities and one primarily undergraduate institution. No further information about these institutions (e.g., R1 status, MSI status, geographic location) were available in the data set.

instructor over multiple semesters of the same course. As a result, it is poorly suited for making inferences about faculty change over time (i.e., the faculty change scale) within the context of an institution (i.e., the institutional change scale). Second, few faculty (~15%; Stains *et al.*, 2018) had COPUS observations for four or more classes within a course, even though most faculty teach more than four classes in a given course. Without some reasonable measure of the actual learning environment experienced by students over a time period reflecting an instructional sequence, the impact of varying levels of sampling intensity on inferences about that environment cannot be assessed. Third, it is not clear which sampling strategy was used to determine which faculty would be observed more or less frequently than others.

Given the limitations of the Stains *et al.* (2018) data set, a new data set suitable for answering questions about the impact of sampling intensity on inferences about reform-based progress within an institution was generated using an archive of previously-existing Echo recordings from several semesters of three introductory courses.[2] Specifically, high-intensity sampling at the level of a course (e.g., ≥10 *classes* per *course*) and at the level of the instructor (e.g., four *semesters* per instructor) was performed. Considering the time-consuming nature of COPUS observations, a form of purposeful sampling known as maximal variation sampling was employed. In maximal variation sampling, the researcher selects cases or individuals who represent the full range of variation that is thought to exist for some trait within the system of focus (Creswell and Guetterman, 2019). As described in more detail later, the system of focus for this study is a large public institution in the early stages of reform. Within this system, the alignment of the introductory courses to research-based instructional guidelines range from completely unaligned (e.g., 100% didactic with no active learning) to courses that include some research-based approaches. Therefore, this was the range of variation that our sampling approach sought to capture. The maximal variation sampling strategy employed in this study aligns with the fundamental change framework, because it enables the evaluation of how much sampling is required to generate enough sensitivity for detecting fundamental faculty change in an institution undergoing reform.

Institutional Reform Context and Faculty Studied

This study took place at a large, public, doctorate-granting university in the northeastern United States that is currently involved in a large reform initiative aimed at moving science faculty away from lecturing and toward RBIS (e.g., active learning). As such, this institution is in a period of transition, and its science courses have adopted student-centered practices to varying degrees (i.e., extreme nonadoption, intermediate adoption, and extreme adoption). Therefore, the choice of institution aligns with our sampling approach, because the faculty population varies in reform-based implementation. Within the context of this institution, we chose to focus sampling on gateway biology courses, which is where most reform efforts have been occurring. The institution selected for this study offers three large (>250 students) gateway biology courses on 1) ecology and evolution, 2) molecular and cellular biology, and 3) human physiology. These courses are required for the biology major and may be taken in any order. Sampling efforts were focused on the former two courses, because they are more typical of the introductory biology sequence at most institutions (i.e., most undergraduate programs require two introductory biology courses rather than three). In the semesters targeted for this study (Spring 2015, 2016, 2017, 2018), these two courses were consistently taught by a team of two faculty each (four faculty total). Other faculty also taught some sections of these courses in several semesters, but their participation in the course was not consistent through time. Because our framework conceptualizes the COPUS at multiple scales (including the faculty change level; Table 1 and Figure 2), it was necessary to limit the possible pool of faculty for generating COPUS data to only those whose archived Echo recordings could be observed consistently across semesters.

A pilot study involving classroom observations and self-reports suggested that the four faculty teaching these gateway courses represent various degrees of adoption of student-centered practices; Instructor 1 has been enacting student-centered practices for many years and has been involved in the reform of several courses at the institution (extreme adoption relative to other faculty at this institution). Instructor 2 had been a traditional lecture-based instructor for years, but recently began participating in reform efforts and instituting student-centered practices into the course (intermediate adoption relative to other faculty at this institution). Instructors 3 and 4 have both been traditional lecture-based instructors for decades, and they have not instituted reform-based practices into their courses (i.e., extreme nonadoption). They also taught the same course. Based on the maximal variation sampling approach used in this study (Creswell and Guetterman, 2019), only one of the two traditional instructors were sampled, given this redundancy. Therefore, three faculty (Instructors 1, 2, and 3) met our sampling criteria.

Although Instructors 1 and 2 taught different units of the same course, we followed the guidelines of Smith *et al.* (2014) and reported the results for each instructor as different courses, because they taught each class separately. The gateway courses taught by these instructors ran for 14 weeks and were taught two to three times a week for 80 or 50 minutes, respectively. Exams (two to three per course) were given during class time and several class periods were not sampled due to snow days or guest lectures.

Overall, COPUS data were collected from three faculty at varying degrees of reform-based progress in 10–11[3] classes (about half the semester) over four consecutive Spring semesters (2015–2018). This approach resulted in 40–44 classroom observations per instructor across four semesters for a total of 128 observations (154 hours total). COPUS data were gathered by three researchers who were certified to conduct COPUS observations by an expert evaluator (M. Smith, the developer of the COPUS). The COPUS raters received training and evaluation independent of the researchers (for more details, see

[2]It is standard practice at our university for large lecture classes to be Echo-recorded. The COPUS observations for this study were conducted using archived recordings from Spring 2015, 2016, 2017, and 2018.

[3]Some instructors taught more than 10 or 11 classes in some semesters, but we set the maximum number of observations for each instructor based on the semester with the fewest classes taught in order to make the semesters comparable.

Smith *et al.*, 2014). Using these procedures, all certified coders achieved an interrater reliability (IRR) score of >0.80. IRR was calculated by M. Smith using Cohen's kappa interrater scores. COPUS observations were conducted using previously-available Echo video recordings of each class. To examine for coding differences between observations conducted in person and using video recordings, we studied a subset of classes both in person *and* through Echo recordings in Spring 2018 and 2020. Comparisons of COPUS observations collected both in person and using video were found to have identical COPUS profiles in 100% of the cases ($n = 8$ classes). COPUS observations for each course were split among two or three COPUS observers and assigned such that observers alternated the classes they observed (e.g., Observer 1 was assigned to classes 1, 3, and 5, and Observer 2 was assigned to classes 2, 4, and 6). It is common practice in this institution for introductory courses to be team taught, with one instructor teaching about half of the classes in a given course. Therefore, our maximum within-course sampling intensity of no more than half a semester per instructor reflects typical instruction at this institution and is consistent with a possible structure of courses undergoing reform (~10 classes).

Actual and Sampled Measures

When determining the impact of sampling intensity on the characterization of the classroom environment at multiple scales, it was first necessary to determine the *actual* frequency and *actual* proportion of each instructional style using the 10–11 classes for which COPUS data were generated. Classes within each course were randomly subsampled at varying sampling intensities (ranging from one class per semester to the greatest number of classes per semester) to simulate varying numbers of classes for which an observer generated COPUS data. Unlike resampling, where the number of classes sampled would be that of the total sample, subsampling is typically done without replacement (Politis *et al.*, 1999). These random subsamples were repeated 1000 times for each sampling intensity. From these simulated data, two sampled measures were generated at each sampling intensity: 1) The proportion of the 1000 replicates in which the student-centered style was sampled *at least once*. 2) The proportion of *each instructional style* sampled for each of the 1000 replicates. We call this latter measure the "*sampled* proportions."

These sampled proportions may be similar to the actual proportions for some of the 1000 replicates but not others. A sampled proportion that closely approximates the actual proportion may be considered accurate (i.e., not biased). A sampling intensity for which a high percentage of replicates have accurate sampled proportions may be considered precise. Collectively, sampling bias (the systematic effects of sampling) and sampling precision (the random effects of sampling) can be used to estimate measurement uncertainty due to sampling (Ramsey *et al.*, 2019). Measurement uncertainty characterizes the range of values within which the true value of the population is expected to lie and "is the most important single parameter that describes the quality of measurements" (Ramsey *et al.*, 2019, p. 1).

Sampling bias was estimated for each replicate at each sampling intensity by calculating the difference between the sampled proportions at each sampling intensity and the actual proportion (those generated by the complete 10- or 11-class data set). The replicates characterized by sampled proportions that were "close enough" to the actual proportions for each sampling intensity were determined using the specific evidence-based measurement criterion described in the next section. Replicates that fit this criterion were considered to have low sampling bias and to accurately estimate the characteristics of the learning environment. The percentage of accurate replicates was calculated and plotted for each sampling intensity. The higher the percentage of accurate replicates, the higher the sampling precision of a given sampling intensity. A measure may have high sampling precision (and be considered precise) if it shows similar results (and in this case, similarly accurate results) with repeated measurement. Sampling intensities with a high probability of accurate estimation can be considered to be both unbiased *and* precise and to therefore have low measurement uncertainty.

The above steps were repeated across four semesters for each instructor. For example, Figure 3 illustrates the results of 1000 replicates at a sampling intensity of eight for three hypothetical courses ("didactic-only," "mixed," and "student centered–only" courses). For the didactic-only and student centered–only courses (Figure 3, A and C), the sampled proportions of each style for all 1000 replicates exactly matches the actual proportions. However, in the mixed course (Figure 3B), the sampled proportions for some of the 1000 replicates were closer to the actual proportion than others. The COPUS can only generate valid inferences about the course-level learning environment if the sampled proportions of each instructional style are "close enough" (see measurement criteria below) to the actual proportions.

Operationalizing the Measurement Criteria for Accurate (i.e., Unbiased) and Precise Estimation of the Learning Environment

An essential a priori consideration of sampling involves establishing the optimal sampling bias and sampling precision (i.e., measurement uncertainty) that is acceptable to allow valid inferences about the system of interest (Ramsey *et al.*, 2019). However, sampling considerations are frequently not made a priori, and errors in measurement that result from inappropriate sampling protocols can contribute to incorrect inferences (Loken and Gelman, 2017). This section describes the thresholds used to classify acceptable magnitudes of sampling bias and sampling precision for each sampling intensity.

The measurement criteria used to determine replicates in which *sampled* proportions were "close enough" to *actual* proportions (and thus had low sampling bias) were based on the fundamental change framework (i.e., Cuban, 1999; Dancy and Henderson, 2005) and operationalized by specifying a threshold within which the sampled proportions could acceptably differ from actual proportions. A conservative threshold of 9% (1/11 classes = 9%) was chosen to mandate the sampling of rare instructional styles and slow shifts (backward or forward) across time (as suggested by the fundamental change framework). Therefore, in this study, sampled proportions within a 9% threshold of the actual proportions were considered to have low sampling bias and to be accurate estimates of each instructional style and the learning environment more broadly. At this 9% threshold, if a single student-centered class was not sampled, then the resulting sampled proportion for this style would

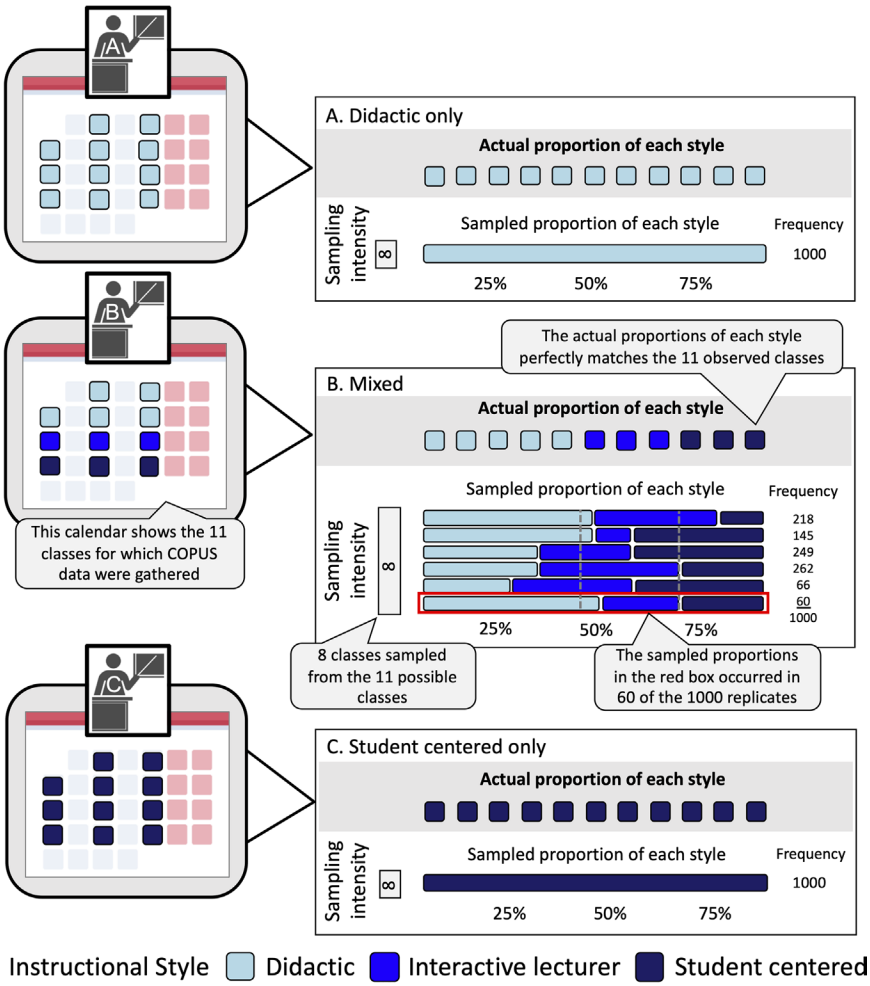

FIGURE 3. Overview of our methodological approach using hypothetical cases. Example data and results are shown for a (A) didactic-only, (B) mixed, and (C) student centered–only courses. Results of 1000 random samples at a sampling intensity of eight are shown for three hypothetical courses (A–C). For the didactic-only and student centered-only courses (i.e., courses A and C), all 1000 sampled proportions of each style *exactly* match the actual proportion of each style at all sampling intensities, including the one shown above. However, because the mixed course (i.e., B) has variation in its instructional styles, the 1000 random samples of eight classes can result in somewhat different patterns of sampled proportions of each instructional style (therefore B has more variation—or "rows"—of sampled data than A or C). Some of these replicates are closer to the actual proportions of each instructional style than others. The frequencies next to each sampled proportion represent the number of times (out of 1000) those particular proportions occurred. For example, for the mixed course, the first set of proportions occurred 218 times, the second set of proportions occurred 145 times, and so on.

not be considered to be an accurate estimate. Less conservative thresholds are also possible (see Supplemental Figures 1 and 2), but a 9% threshold aligns with the fundamental change framework, because it mandates that even small advances toward (or away from) student-centered instruction (e.g., one student-centered class added to or lost from a course) be sampled with a high probability.

The percentage of the 1000 replicates for each sampling intensity within this 9% threshold was then calculated and considered to represent the percentage of accurate (i.e., unbiased) estimates for each instructional style at each sampling intensity. Sampling intensities for which ≥75% of the 1000 replicates were classified as accurate estimates were considered to be both unbiased *and* precise. Therefore, this approach generated a value of the precision with which accurate sampling can be achieved at each sampling intensity.[4] In real terms, accurately classifying 75% of the 1000 replicates at a given sampling intensity means that a COPUS observer would have a 75% probability of accurately classifying the *course*-level learning environment if they sampled that number of classes. Stricter sampling precision criteria than the 75% cutoff described here could also be used, and the figures showing these results allow for interpretation based on more stringent sampling precision criteria. The COPUS (and its composite measures) has the potential to generate valid inferences about the course-level learning environment and reform-based progress when the actual proportion of each instructional style is accurately and precisely estimated. Therefore, establishing the sampling conditions under which the COPUS has a high probability of accurate and precise estimation is essential.

ANALYSES

To address RQ1.1 ("How many classes must be sampled in order to document the presence of at least one student-centered class within a course?"), we identified the semesters that included at least one student-centered class. For these classes, the percentage of the 1000 replicates at each sampling intensity for which the student-centered style was successfully sampled *at least once* was calculated (as described earlier). The sampling intensity needed to successfully sample at least one student-centered class with a ≥75% probability was reported.

To address RQ1.2 ("How many classes must be sampled in order to accurately and precisely estimate the proportion of different instructional styles within a course?"), for each of the 1000 replicates at each sampling intensity, we classified the sampled proportions as either accurate or inaccurate (according to the measurement criteria described earlier) and coded them as "1" or "0," respectively. This accuracy coding step was carried out for each instructional style individually (e.g., was the sampled proportion of the

[4]The standard error of the values of the 1000 replicates would be a more typical measure of sampling precision, but the approach used in this paper was chosen because it integrated accuracy and precision into one measure.

FIGURE 4. COPUS instructional styles for each instructor and course included in this study.

student-centered style for replicate 1 at a sampling intensity of six accurate or inaccurate?) and all instructional styles combined (e.g., were the sampled proportions of *all three* instructional styles for replicate 1 at a sampling intensity of six accurate or inaccurate?). Sampling intensities for which ≥ 75% of the 1000 replicates were classified as having accurate sampled proportions were considered to have high precision and thus a high probability of generating valid inferences of the learning environment.

To address RQ2.1 ("How many classes must be sampled in order to document changes in the presence of student-centered classes throughout multiple semesters of an instructor's course?"), we aligned the 1000 replicates at each sampling intensity across all four semesters to simulate longitudinally implemented (i.e., over four semesters) COPUS observations at various sampling intensities. This approach resulted in 1000 longitudinally aligned replicates (referred to as longitudinal replicates). The percentage of the 1000 longitudinal replicates for each sampling intensity that produced valid inferences about the presence of at least one student-centered class was calculated. The sampling intensity needed to successfully sample at least one student-centered class with a ≥75% probability in all semesters in which this style was present was calculated.

To address RQ2.2 ("How many classes must be sampled in order to accurately and precisely estimate changes in the proportion of different instructional styles throughout multiple semesters of an instructor's course?"), we generated accuracy codes (i.e., "0" or "1" described in RQ1.2) for each of the 1000 replicates at each sampling intensity across all four semesters to simulate longitudinally implemented (i.e., over four semesters) COPUS observations at various sampling intensities. For example, at a sampling intensity of four (i.e., four classes sampled), the first longitudinally aligned replicate could have the following accuracy pattern across semesters: didactic: 1, 1, 1, 1 (i.e., all four semesters were accurately estimated for this style); interactive lecturer: 0, 0, 0, 1 (i.e., only the last semester was accurately estimated for this style); student centered: 0, 0, 0, 0 (i.e., none of the semesters accurately estimated this style); all three styles: 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0 (i.e., replicate 1 did not accurately estimate *all* styles across *all* four semesters).

This approach was then used to ask: For how many of the 1000 longitudinally aligned replicates were there accurate estimates (i.e., 1's only) of the proportion of 1) didactic lecturing, 2) interactive lecturing, 3) student-centered instruction, and 4) all styles throughout *all* semesters? This approach assumes that faculty change *across semesters* can only be accurately measured if each composite semester is accurately measured, which aligns with our fundamental change framework, because the dynamics in *all* semesters are considered relevant to measuring change, given how slow and nonlinear it may progress in early reform contexts. Similarly to RQ1.2, sampling intensities for which ≥75% of the 1000 longitudinally aligned replicates were classified as having accurate sampled proportions were considered to have high precision and, thus, a high probability of generating valid inferences of the learning environment in that course over those four semesters.

RESULTS

The three instructors displayed various instructional styles (class level) in varying proportions in a given semester (course level) and through time (faculty change level; Figure 4). Didactic instruction was widespread in the sample and was the most common instructional style for all three instructors in most semesters. Instructor 3 used didactic lecturing (class level) for all classes (course level) in all four semesters (instructor level). This instructor could thus be classified as a *didactic-only* instructor who was consistent through time. Instructors 1 and 2 used interactive lecturing (class level) in at least one class (course level) in all semesters (instructor level) and student-centered practices in classes in some semesters, but most frequently in the last semester sampled. These data indicate that instructors 1 and 2 can be classified as mixed instructors who appeared to be transitioning toward RBIS.

TABLE 2. The proportion of the 1000 replicates at each sampling intensity in which a student-centered class was sampled in the semesters in which it was present

Instructor ID	Semester	Actual frequency of student- centered classes out of sample classes	Percent of 1000 replicates in which at least one student-centered class was sampled[a]				
			SI = 4 (36–40%)	SI = 5 (45–50%)	SI = 6 (54–60%)	SI = 8 (73–80%)	SI = 9 (82–90%)
Instructor 1	1	1	35.60%	45.60%	53.70%	71.80%	80.90%
	4	2	60.40%	74.20%	80.00%	93.80%	98.50%
Instructor 2	4	2	65.10%	76.60%	86.00%	96.90%	100.00%

[a]SI, sampling intensity. The values in parenthesis next to each SI indicate the percent of total classes that were sampled from a given instructor. Each SI is associated with a range of percentages, because the instructors taught a different number of total classes: Instructor 1 taught 11 classes, and Instructor 2 taught 10 classes.

RQ1.1

The sampling intensity required to sample at least one student-centered class within a course (in semesters in which it was present) with a ≥ 75% probability ranged from five to nine classes (50–82% of total classes; Table 2). At a sampling intensity of four classes (recommended by Stains *et al.*, 2018), at least one student-centered class was sampled in only 36–65% of the 1000 replicates. The probability of sampling at least one student-centered class depended on the actual frequency of student-centered classes in the course (Table 2); the rarer this style was in the course, the lower the probability of sampling it at any sampling intensity.

RQ1.2

The sampling intensity needed to accurately and precisely (i.e., with a ≥75% probability) estimate the proportion of instructional styles within a course ranged from one to 10 classes (10–91% of all classes taught; Figure 5, red dashed line) and varied depending on instructor characteristics. Therefore, the sampling intensity either had a very large impact or no impact on the accuracy and precision of the measures for each instructional style and the inferences about the course-level learning environment more broadly. This finding is largely explained by the variability of instructor behaviors within a course. Courses in our sample with variability in instructional styles among classes (i.e., "mixed" courses; Figure 3B) required higher sampling intensities to generate valid inferences about the learning environment than courses with no variability. Specifically, Instructors 1 and 2 used at least two instructional styles in every course for which COPUS data were gathered (Figure 4) and required a sampling intensity of at least seven classes to accurately and precisely estimate the proportions of each style (Figure 5). However, as the number of instances of the second style increased above one, the sampling intensity required also increased: semesters with more than one instance of a second style required a sampling intensity of at least nine classes to accurately and precisely estimate the proportions of each style (Figure 4). Unlike data for Instructors 1 and 2, the COPUS data gathered for Instructor 3 were classified as didactic for all classes, and therefore a sampling intensity of one class was sufficient to accurately and precisely estimate the proportions of each style and to make inferences about the learning environment in all four semesters (Figure 5). See Figure 6 for a summary of these results for RQ1.2. Overall, the number of classes

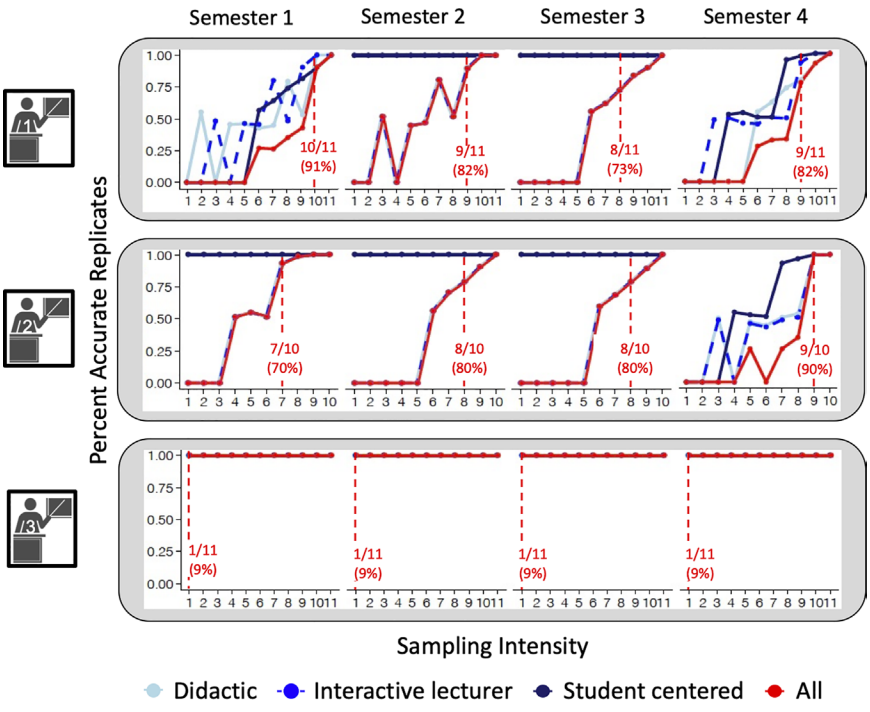

FIGURE 5. Percent of accurate replicates at each sampling intensity for each instructor and semester. The sampling intensity required to attain a ≥75% probability of accurately estimating all instructional styles in each semester is indicated by the vertical red dashed line and associated red text (e.g., red text that reads 10/11 means that 10 out of the 11 total classes (91% of classes) had to be sampled in order to attain a >75% probability of accurately estimating all instructional styles). Sampling intensities with accuracy probabilities at or above this 75% threshold can be considered capable of accurate and precise measurement for the instructors and courses in this sample. These results are based on COPUS data gathered from 128 full classes.

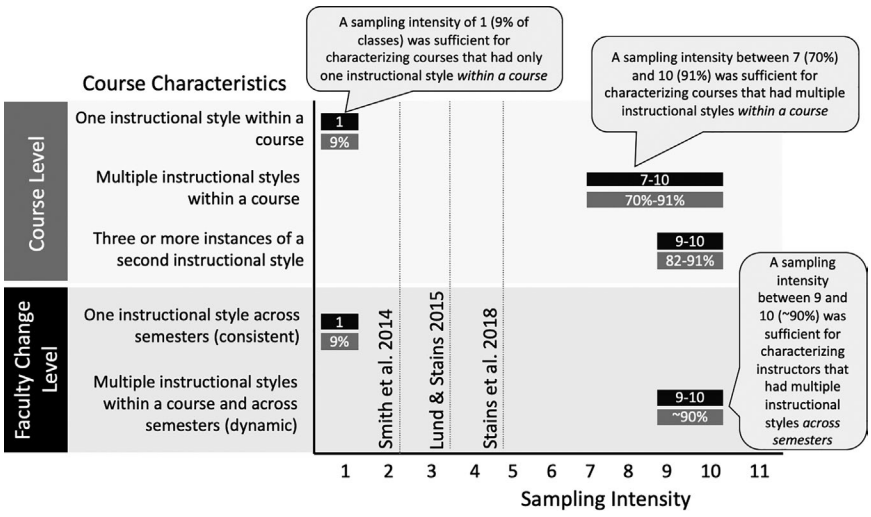

FIGURE 6. Summary of the sampling intensities required to accurately and precisely estimate the proportions of instructional styles within a course (course level) and through time (faculty change level) in this study. The black boxes indicate the *number* of classes and the gray boxes indicate the *percentage* of classes that had to be sampled for accurate and precise estimation in these courses. These values should *not* be taken as a gold standard for robust measurement using the COPUS. The gray vertical lines indicate the minimum sampling intensities recommended by Smith *et al.* (2014), Lund and Stains (2015), and Stains *et al.* (2018).

needed to characterize courses and instructors varied widely in our sample and depended on the specific course dynamics (one instructional style within a course, multiple styles within a course, etc.; Figure 6). It is also likely that the total number of classes taught by an instructor within a semester could impact the required sampling intensity. Preliminary work in which the current data sets were copied four times to simulate a course in which 40–44 classes occurred showed that the sampling intensity required to characterize course-level instructional patterns ranged from 14 to 28 classes (35–70% of classes; see Supplemental Figure 3). Although the exact numbers and percentages differed when more classes were added, the number of classes was still well above that recommended by prior work.

RQ 2.1

Two of the four semesters of COPUS data gathered for Instructor 1 included student-centered classes. However, when a sampling intensity of four classes was applied longitudinally to all of Instructor 1's semesters (i.e., four classes were sampled in 2015, four in 2016, four in 2017, four in 2018), at least one student-centered class was sampled in both of these semesters only 21% of the time, and not sampled *at all* 25% of the time (they were sampled in one of the two relevant semesters 54% of the time). Therefore, a researcher gathering COPUS data from Instructor 1's courses over four semesters would have only a 21% probability of making valid inferences about the presence of student-centered classes through time. A sampling intensity of nine classes was required to increase this probability above 75%.

One of the four semesters of COPUS data gathered for Instructor 2 included student-centered classes. However, when a sampling intensity of four was applied longitudinally to all of Instructor 2's semesters, at least one student-centered class was sampled only 65% of the time. Therefore, a researcher gathering COPUS data from Instructor 2's courses would have a 65% probability of making valid inferences about the presence of student-centered classes through time. A sampling intensity of five classes was required to increase this probability above 75%.

RQ2.2

Similar to the findings reported at the course level, the impact of sampling intensity on accurately and precisely estimating faculty change over time was either large or nonexistent based on faculty-specific patterns. For Instructor 3, whose courses were categorized as didactic-only, a sampling intensity of one class per semester was sufficient to accurately and precisely measure change (or lack thereof) through time (Figure 7C). In contrast, the courses taught by Instructors 1 and 2 included two instructional styles in some semesters and three in others, and the proportions shifted through time. Given the inconsistency of their instructional styles through time, higher sampling intensities (i.e., seven to eight classes) were required to accurately and precisely measure change (Figure 7, A and B). At a sampling intensity of four within a course (the sampling intensity recommended by Stains *et al.*, 2018), there was a 0% probability that an observer would have accurately and precisely measured change through time for all instructional styles (gray vertical lines in Figure 7, A and B). See Figure 6 for a summary of the results for RQ2.2. Therefore, as at the course level, the number of classes needed to characterize instructor change also varied widely and depended on the specific classroom context. The data set simulating a course with 40–44 classes per semester (as described in RQ1.2) was also analyzed over time. These analyses showed that the sampling intensity required to characterize faculty-change-level instructional patterns ranged from 14 to 28 classes (35–70% of classes; See Supplemental Figure 4). Although the exact numbers and percentages differed when more classes were added, the number of classes was still well above that recommended by prior work.

DISCUSSION

Using the COPUS to Measure Reform-Based Progress

Over the past several decades, many reform efforts have been implemented to help faculty move away from tradition-based didactic lecturing and toward RBIS (e.g., Stokstad, 2001; Wood and Gentile, 2003; Pfund *et al.*, 2009; AAAS, 2011; Deslauriers *et al.*, 2011; Haak *et al.*, 2011; Henderson *et al.*, 2011; Graham *et al.*, 2013; Matz *et al.*, 2018). Impact studies of these efforts have produced limited evidence of success (Waks, 2007; Henderson *et al.*, 2011; Kezar, 2018). Clearly, faculty and institutional change remains an ongoing challenge (Henderson *et al.*, 2011; Smith *et al.*, 2013; Kezar, 2018; Stains *et al.*, 2018). Measurement and sampling issues are foundational to documenting and evaluating the degree to which reform efforts are

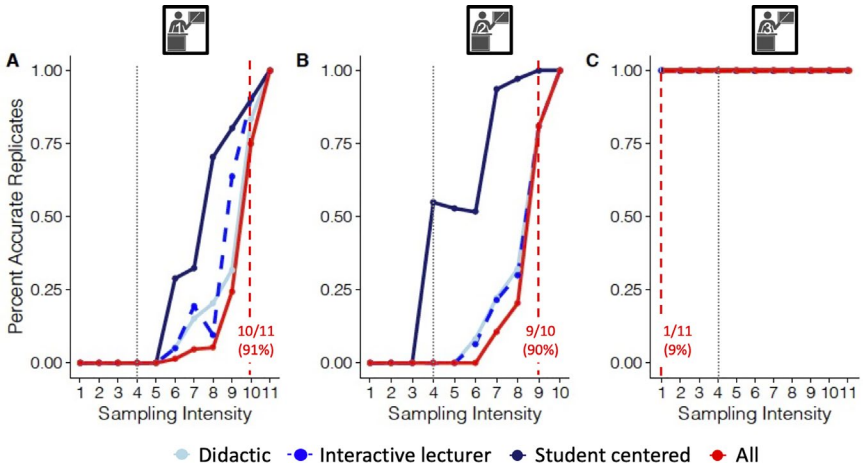

FIGURE 7. Percent of accurate longitudinally aligned replicates (i.e., all four semesters) at each sampling intensity for each instructor and semester. The sampling intensity required to attain a ≥75% probability of accurately estimating all instructional styles in every semester sampled for each instructor is indicated by the red dashed line and text. Sampling intensities with accuracy probabilities above this 75% threshold can be considered capable of accurate and precise measurement for the instructors and courses in this sample. The gray vertical line indicates the sampling intensity recommended by Stains *et al.* (2018).

succeeding, and yet they have received comparatively little attention in the reform literature.

The COPUS instrument has been proposed as a possible tool to aid in the measurement of faculty and institutional change (Lund *et al.*, 2015; Stains *et al.*, 2018). However, important questions remain: First, how should COPUS measures be situated within existing reform frameworks? Second, how many classes should be sampled to generate valid inferences about reform-based practices at multiple scales (i.e., course level and faculty change level)? These questions were addressed by generating a COPUS data set of three faculty teaching 10–11 classes in gateway biology courses over four consecutive semesters (128 classes and 154 observation hours) at an institution undergoing reform. Varying sampling intensities were simulated within a course and across semesters, and the findings were interpreted using a fundamental change framework of reform.

Reform-based progress can be measured at many scales, including the class, course, faculty change, departmental, and institutional levels (Smith *et al.*, 2013). Two of these scales—course and faculty—were the focus of this study, because they are commonly used to characterize reform-based progress (e.g., Pfund *et al.*, 2009; Dancy and Henderson, 2010; Henderson *et al.*, 2011; Prince *et al.*, 2013; McCourt *et al.*, 2017; Matz *et al.*, 2018; Dancy *et al.*, 2019). The faculty involved in this study were selected using a maximum variation sampling strategy (Creswell and Guetterman, 2019), and preliminary observations indicated that they represented the full range of variation along the continuum of reform-based progress at this institution (i.e., extreme nonadoption, a shift toward intermediate adoption, and a shift toward extreme adoption). The class-level COPUS measures presented in this study generally aligned with these preliminary observations, and the data set included a wide range of course and instructor characteristics that are likely to be found within institutions undergoing reform. Specifically, the sample included the following characteristics: 1) no variability in instructional styles within a semester; 2) the presence of rare instructional styles, two instructional styles, and three instructional styles; 3) the absence of change across semesters; and 4) the presence of change across semesters (both forward and backward).

The simulation studies indicated that the sampling intensity needed to make valid inferences about the presence of student-centered instruction at the course and faculty change level varied based on the characteristics of the course and the nature of change through time. Higher sampling intensities were required when the student-centered style was rare within a single semester or rare among multiple semesters of a course. Specifically, within the conceptual framework we used, a sampling intensity of nine classes (~82%) was needed to have a ≥75% probability of detecting the presence of student-centered instruction *within a course* when there was only one such class present in the course. When two such classes were present in the course, a lower sampling intensity was sufficient for detecting at least one of them. Furthermore, a sampling intensity of nine classes (~82%) was needed to have a ≥75% probability of consistently detecting the presence of student-centered instruction *across all semesters in which it occurred* when more than one semester included this style. When only *one* semester included this style, a lower sampling intensity was sufficient.

The sampling intensity needed to make valid inferences about the proportion of instructional styles at the course and faculty change level using the COPUS varied based on the characteristics of the courses in this study and the nature of their change through time. Higher sampling intensities were required for 1) courses with variability in instructional style (i.e., "mixed" courses) and 2) instructors demonstrating change in their instructional patterns through time (i.e., dynamic instructors; see Table 1). Specifically, in our sample, the following minimum sampling intensities were required to make valid inferences about the proportion of instructional styles within a course: For courses in our sample with no variability in instructional style within a semester, a sampling intensity of one class (~9%) was sufficient, but for courses in our sample with only one instance of a second style, seven classes (~70%) were required to accurately and precisely estimate their proportions. Therefore, in effect, a large amount of sampling was required to distinguish between a didactic-only course (like that of Instructor 3) and a course with a single nondidactic class. Courses in our sample with two or more instances of a second style required a sampling intensity of at least nine classes (~82%) to accurately and precisely estimate actual proportions.

Concerning faculty change, for instructors in our sample displaying no variability in their instructional style within or among the four semesters, sampling one class per semester

(~9%) was sufficient. For instructors in our sample with variability in their instructional styles within or among four semesters, a minimum of nine classes (~90%) was required to accurately and precisely estimate the proportions of each style.

Given the burden of such intensive COPUS observations, many authors have made sampling recommendations typically ranging from two (Smith *et al.*, 2014) to four (Stains *et al.*, 2018) classes per instructor. The detailed data set generated in this study allowed the drawing of inferences at multiple sampling intensities and the comparison of those inferences to the complete 11-class sample. The results of these comparisons suggest that the sampling intensity employed by COPUS observers are likely to impact the quality and nature of the inferences made about two of the instructors sampled. Specifically, when analyzing all observed classes (10–11 classes per instructor, per semester), Instructors 1 and 2 made measurable progress toward the adoption of student-centered approaches during the first few years of the institution's reform initiative. However, the sampling intensity recommended by Stains *et al.* (2018; i.e., four classes) frequently generated inaccurate inferences about the learning environment at the scales studied. Specifically, there was a < 65% probability that a COPUS observer would draw valid inferences about the *presence* of student-centered classes and a 0% probability they would accurately and precisely estimate the *proportions* of all instructional styles.

A significant implication of these findings is that the sampling intensities recommended (or used) by Smith *et al.* (2014; two classes), Lund and Stains (2015; three classes), Teasdale *et al.* (2017; one class), and Stains *et al.* (2018; four classes) were sufficient for measuring the course-level or faculty change–level scales in only one specific situation: the consistent use of one instructional style (sampling one class was sufficient). In the early stages of institutional reform and faculty adoption of evidence-based practices, a low frequency of adoption of student-centered practices is likely to be common. Our results show that rare styles are unlikely to be sampled at a sampling intensity of four classes and that higher sampling intensities are needed to make valid inferences about course-level learning environments and reform-based progress.

An early adoption framework for fundamental change requires detection of rare instructional styles—our simulations suggest that capturing these styles within a course will require a higher sampling intensity. Using this framework, the sampling intensities used and recommended in prior COPUS work are likely to inaccurately characterize progress in institutional and national enactment of student-centered instruction in many institutional settings. Our findings highlight the risks of employing sampling designs and measurement approaches untethered from conceptual frameworks about the nature of reform (cf. Lewis *et al.*, 2006; Dancy and Henderson, 2008; IES and NSF, 2013; AERA *et al.*, 2014).

Researchers would benefit from knowing if the early adoption of new and rare instructional styles is occurring for many reasons: reporting progress to funding agencies and university stakeholders (e.g., university deans, the NSF); providing professional support and scaffolding for early adopters (e.g., Pelletreau *et al.*, 2018); examining student learning disparities across evidence-based adoption intensities (e.g., Freeman *et al.*, 2014); and understanding the causes of RBIS advancement and retrenchment. Researchers would also benefit from having a valid measure of early reform-based progress that is distinct from student learning outcomes, because the low frequency of reform-based practices may not always be associated with improved student outcomes (e.g., Connell *et al.*, 2016; Theobald *et al.*, 2020), though they could still reflect meaningful progress toward student-centered teaching. More fine-grained evidence could further illuminate this relationship.

Overall, it is important to emphasize that the specific sampling intensity values reported in this section are *minimum* values (several courses required higher sampling intensities) and are specific to the dynamics of the courses sampled in this study. Given that the number of classes needed to characterize courses and instructors varied so widely and depended strongly on the dynamics within the classroom (one instructional style within a course, multiple styles within a course, etc.; see Figure 6), these values should *not* be taken as a "gold standard" for robust measurement using the COPUS. Contrary to prior work, the practice of recommending a general COPUS sampling intensity should be avoided without appropriate contextualization (Figure 1). The results presented here demonstrate this point nicely; the sampling intensity (both by number and percentage of classes) differed for the 10- to 11-class data set compared with the simulated 40- to 44-class data set, even though the latter was literally the same data copied four times. However, the general finding remained robust for both data sets; more varied and higher sampling intensities may be required to measure reform using the COPUS at the course or faculty change scale than has been previously reported.

An important takeaway from this study is that sampling protocols used in biology education research require evidence-based guidelines on sampling intensity that are informed by conceptual frameworks and institutional contexts. It is likely that this general conclusion may be relevant to other structured classroom observation protocols that are currently available to measure learning environments. Therefore, simulation studies investigating the impact of sampling strategies on inferences should be conducted on these protocols as well, particularly because some require observers to make judgments about alignment of teaching practices to some standard (e.g., Inside the Classroom: Observation and Analytic Protocol, Weiss *et al.*, 2003; Reformed Teaching Observation Protocol, Sawada *et al.*, 2002; Practical Observation Rubric to Assess Active Learning Classrooms, Eddy *et al.*, 2015) and others do not (e.g., TDOP; Hora *et al.*, 2013). Regardless of which protocol is used, future stimulation work should involve the measurement of outcomes in association with these observations.

Recommendations and Alternative Instruments for Sampling in Early Reform Contexts

Given the high, yet variable, sampling intensity needed for accurate and precise measurement in many instructional contexts, the COPUS might not be the most cost-effective tool to measure reform at the course level or faculty change level, particularly early in reform efforts or when little is known about the baseline instructional practices of faculty. The findings presented here indicate that the actual dynamics present within a course are very important to decisions about sampling strategy and that using the COPUS with no real knowledge of the amount of variability may be problematic. Therefore, implementing pre-observation assessments to estimate instructor

reform baselines could help to inform observation instrument and sampling protocol decisions. However, in early reform contexts, researchers may have access to less robust baseline information about instructors, and such information may be difficult to gather at the level of detail necessary to make informed decisions about sampling protocols. For example, early reform contexts are likely to include courses with didactic-only styles or rare instances of nondidactic styles, which is an important distinction in some conceptual frameworks. Unfortunately, pre-assessments may not reliably distinguish these functionally similar course characteristics a priori; thus, our results suggest that some variability should be assumed and high COPUS sampling intensities of courses are likely to be required (as shown in this study). Therefore, the findings in this study raise questions about whether the COPUS is best suited to the study of early reform contexts. Pairing detailed observation protocols like the COPUS with more cost-effective and time-efficient tools like the Decibel Analysis for Research in Teaching (DART; Owens *et al.*, 2017) could be one way forward. Unfortunately, the DART machine-learning algorithm did not generate accurate information about instructional practices using the audio-capture system at the institution we studied (Sbeglia, G. C., Goodridge, J. A., Gordon, L. H., Nehm, R. H., unpublished data). This failure may be a result of the recording method or the training data set. Despite the problems we encountered, it is possible that the DART will function in other settings. Overall, our work indicates that biology education researchers will be able to make most effective use of the COPUS when it is situated in a clear reform framework and aligned with anticipated instructor behaviors.

Limitations

This study used a maximal variation sampling approach (Creswell and Guetterman, 2019) to select faculty participants and courses, and as anticipated, the sample displayed a wide range of course and faculty characteristics that are likely to be found within institutions undergoing reform. Because the exact sampling intensity ranges reported in this study (Figure 6) are specific to the characteristics of the courses analyzed, they are not necessarily broadly generalizable to dynamics that were not observed in our sample (e.g., faculty with high variability in instructional styles in *all* semesters sampled). It would be a valuable contribution for researchers to investigate how other classroom and course dynamics may impact sampling intensity requirements. Regardless, a main conclusion in this study—that the currently recommended sampling intensity of four classes is likely too low to make valid inferences about most early reform course-level learning environments and faculty change—is likely generalizable to other samples.

Additionally, the sampling intensity ranges reported here (Figure 6) are also specific to the scale of reform targeted and the framework used. For example, other scales of reform (e.g., institutional level, national level) may align with different reform frameworks that warrant different sampling intensities. As a result, the framework employed by this study and the sampling intensity ranges that emerged from it should not be generalized to other scales without independent verification. However, it is likely that studies of instructional styles at large scales (e.g., at institutional or national scales) must still consider what sampling intensity is required within the courses that make up their sample. Furthermore, it is possible that there are frameworks for which a lower sampling intensity would be able to generate valid inferences about the status and progress of reform at the course or faculty change scale. For example, student-centered practices, such as active learning, may be most effective in high doses (e.g., 30% according to Theobald *et al.*, 2020), and researchers using outcome-based benchmarks may not require the accurate and precise measurement of rare styles or small shifts. Researchers or evaluators seeking to measure the general trajectory of change over a long period of time in an institutional context that has well-established reform initiatives may not require stringent sampling protocols. However, both of these examples are unlikely to apply to early reform contexts like the one that we studied. Despite these limitations, a main point of this study remains relevant for biology education research conducted at all scales: The sampling protocols used require evidence-based guidelines on sampling intensity that are informed by conceptual frameworks.

In this study, we used Stains *et al.*'s (2018) evaluative framework for how COPUS behaviors align with reform-based classifications of the learning environment (i.e., the three instructional styles). Measures generated from the COPUS can be used to make valid inferences about the progress of reform if they are supported by: 1) explicit validity evidence (i.e., evidence to support the claim that they measure what they are intended to measure), 2) reliability evidence (i.e., evidence that faculty use behaviors consistently across multiple classes), and 3) conceptual grounding (i.e., linking COPUS scores to a theoretical or conceptual framework; cf. AERA *et al.*, 2014; Nehm, 2019). At present, the COPUS instructional styles may not meet several of these standards for measuring faculty change. For example, Stains *et al.* (2018) have not provided robust validity evidence (e.g., AERA *et al.*, 2014; Campbell and Nehm, 2013) to support claims that their class-level measures (i.e., the three instructional styles) generate valid inferences, and recent empirical work has raised concerns over these three classifications (e.g., Denaro *et al.*, 2021). Furthermore, because the COPUS was not designed to measure the quality of instruction, it is possible that classroom learning environments could be characterized as student-centered by the Stains *et al.* (2018) framework but not by observers who are experts in reform-based instruction. In fact, it is not uncommon for student-centered instruction to be implemented in a manner that may appear on the surface to be evidence based, but that may actually represent an *inappropriate assimilation* of these ideas into prior instructional practices (Henderson *et al.*, 2009). Inappropriate assimilation of student-centered innovations may not achieve the desired outcomes and may lead faculty to conclude that these approaches are ineffective, possibly hindering reform progress (Henderson *et al.*, 2009). Regardless, although there is limited validity evidence for Stains *et al.*'s (2018) categorizations, this limitation is unlikely to affect our claims about the impact of sampling intensity, which is the focus of the current study.

CONCLUSION

Our work advances the measurement of educational reform by 1) aligning conceptual frameworks for reform with measurement goals as recommended by professional organizations and research policies (IES and NSF, 2013; AERA *et al.*, 2014) 2) identifying the measurement of RBIS at the course and

faculty change level as important scales to be measured, and 3) incorporating a probabilistic sampling approach to the measurement of course-level learning environments and faculty change within actual courses at an institution undergoing reform. Overall, the findings indicate that the sampling intensity needed to characterize courses and instructors varies widely and depends on classroom characteristics, indicating that recommendations of a universal COPUS sampling intensity should be avoided. Research designs in biology education require evidence-based guidelines on sampling intensity and must be informed by conceptual frameworks and institutional contexts for researchers and administrators to accurately and precisely measure the adoption of student-centered behaviors. This work is a small but important step in that direction.

ACKNOWLEDGMENTS

We thank the American Association of University Women and the Howard Hughes Medical Institute's Science Education Program (Inclusive Excellence Award) to Stony Brook University for support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of HHMI. We also thank two anonymous reviewers for insightful suggestions for improving the article.

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Berger, R. M., & Patchener, M. A. (1988). *Implementing the research plan: A guide for the helping professions*. (SAGE human services guide 51). London: SAGE Publications, Inc.

Booker, K. C. (2007). Perceptions of classroom belongingness among African American college students. *College Student Journal*, *41*(1), 178–186.

Campbell, C., & Nehm, R. H. (2013). A critical analysis of assessment quality in Genomics and Bioinformatics Education Research. *CBE—Life Sciences Education*, *12*(3), 530–541.

Cavallo, D. (2004). Models of growth—towards fundamental change in learning environments. *BT Technology Journal*, *22*(4), 96–112.

Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes about biology. *CBE—Life Sciences Education*, *15*(1), ar3. doi: 10.1187/cbe.15-03-0062

Creswell, J. W., & Guetterman, T. C. (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). Boston, MA: Pearson Education.

Cuban, L. (1992). Curriculum stability and change. In Jackson, P. W. (Ed.), *Handbook of research on curriculum* (pp. 216–247). New York: Macmillan.

Cuban, L. (1999). *How scholars trumped teachers: Change without reform in university curriculum, research, and teaching, 1890–1990*. New York: Teachers College Press.

Dancy, M., Brewe, E., & Henderson, C. (2007). Modeling success: Building community for reform. *AIP Conference Proceedings*, *951*, 77. doi: 10.1063/1.2820951

Dancy, M., & Henderson, C. (2008). *Barriers and promises in STEM reform* (Commissioned paper for National Academies of Science Workshop on Linking Evidence and Promising Practices in STEM Undergraduate Education). Washington, DC.

Dancy, M., & Henderson, C. (2010). Pedagogical practices and instructional change of physics faculty. *American Journal of Physics*, *78*(10), 1056–1063.

Dancy, M. H., & Henderson, C. (2005, September). Beyond the individual instructor: Systemic constraints in the implementation of research-informed practices. *AIP Conference Proceedings*, *790*, 113–116.

Dancy, M., Lau, A. C., Rundquist, A., & Henderson, C. (2019). Faculty online learning communities: A model for sustained teaching transformation. *Physical Review Physics Education Research*, *15*. doi: 10.1103/PhysRevPhysEducRes.15.020147

Deligkaris, C., & Chan Hilton, A. B. (2019, February 6). COPUS: A non-evaluative classroom observation instrument for assessment of instructional practices. In *3rd Celebration of Teaching & Learning Symposium*. Evansville, Indiana: University of Southern Indiana.

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. doi: 10.1187/cbe.20-04-0077

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, *862*, 862–864.

Eddy, S. M., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: Practical observation rubric to assess active learning classrooms. *CBE—Life Sciences Education*, *14*(2), ar23. doi: 10.1187/cbe.14-06-0095

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, *661*, 1–26.

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.

Goodridge, J., Gordon, L., Nehm, R., & Sbeglia, G. (2019, April 24). Measuring active learning in STEM classrooms: Quantifying the impact of sampling intensity on COPUS instrument scores. In *Undergraduate Research and Creative Activities Symposium*. Stony Brook, NY: Stony Brook University.

Graham, M. J., Frederick, J., Byars-Winston, A., Hunter, A. B., & Handelsman, J. (2013). Increasing persistence of college students in STEM. *Science*, *341*(6153), 1455–1456. doi: 10.1126/science.1240487

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*, 1213–1216.

Henderson, C., Beach, A., & Famiano, M. (2009). Promoting instructional change via co-teaching. *American Journal of Physics*, *77*(3), 274–283.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952–984.

Hora, M., & Ferrare, J. J. (2010). *The Teaching Dimensions Observation Protocol (TDOP)*. Madison: Wisconsin Center for Education Research. University of Wisconsin–Madison.

Hora, M. T., Oleson, A., & and Ferrare, J. J. (2013). *Teaching Dimensions Observation Protocol (TDOP) user's manual*, Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison. Retrieved August 7, 2013, from http://tdop.wceruw.org/Document/TDOP-Users-Guide.pdf.

Hoyle, E., & Wallace, M. (2007). Educational reform: An ironic perspective. *Educational Management Administration and Leadership*, *35*(1), 9–25.

Institute of Education Sciences and National Science Foundation. (2013). *Common guidelines for education research and development*. Washington, DC: U.S. Department of Education. Retrieved August, 1, 2020, from www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

Kezar, A. (2018). *How colleges change: Understanding, leading, and enacting change* (2nd ed.). New York: Routledge.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Leshem, S., & Trafford, V. (2007). Overlooking the conceptual framework. *Innovations in Education and Teaching International*, *44*(1), 93–105.

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? The case of lesson study. *Educational Researcher*, *35*(3), 3–14.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. doi: 10.1126/science.aal3618

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

Lund, T. J., & Stains, M. (2015). The importance of context: an exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, *2*(13). doi: 10.1186/s40594-015-0026-8

Maciejewski, W. (2016). Flipping the calculus classroom: An evaluative study. *Teaching Mathematics and Its Applications*, *35*(4), 187–201. doi: 10.1093/teamat/hrv019

Mathison, S. (2005). *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage. doi: 10.4135/9781412950558

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., … & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. *Science Advances*, *4*(10). doi: 10.1126/sciadv.aau0554

McCourt, J. S., Andrews, T. C., Knight, J. K., Merrill, J. E., Nehm, R. H., Pelletreau, K. N., … & Lemons, P. P. (2017). What motivates biology instructors to engage and persist in teaching professional development? *CBE—Life Sciences Education*, *16*(3), ar54. doi: 10.1187/cbe.16-08-0241

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment. In Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.), *Center for Education. Division on Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

National Science Foundation. (2020). *Vision and change in undergraduate biology education*. Retrieved November 13, 2020, from www.nsf.gov/funding/pgm_summ.jsp?pims_id=505859&more=Y

Nehm, R. H. (2019). Biology education research: building integrative frameworks for teaching and learning about living systems. *Disciplinary Interdisciplinary Science Education Research*, *1*, 15. doi: 10.1186/s43031-019-0017-6

Nugent, W. R. (2019). Probability and sampling. In Thyer, B. A. (Ed.), *The handbook of social work research methods* (2nd ed., pp. 37–50). London: Sage.

Owens, M. T., Seidel, S. B., Wong, M., Bejines, T. E., Lietz, S., Perez, J. R., … & Tanner, K. D. (2017). Classroom sound can be used to classify teaching practices in college science courses. *Proceedings of the National Academies of Sciences USA*, *114*(12), 3085–3090. doi: 10.1073/pnas.1618693114

Pelletreau, K. N., Knight, J. K., Lemons, P. P., McCourt, J. S., Merrill, J. E., Nehm, R. H., … & Smith, M. K. (2018). A faculty professional development model that improves student learning, encourages active-learning instructional practices, and works for faculty at multiple institutions. *CBE—Life Sciences Education*, *17*(2), es5. doi: 10.1187/cbe.17-12-0260

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. doi: 10.1126/science.1170015

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. New York: Springer.

Prince, M., Borrego, M., Henderson, C., Cutler, S., & Froyd, J. (2013). Use of research-based instructional strategies in core chemical engineering courses. *Chemical Engineering Education*, *47*(1), 27–37.

Ramsey, M. H., Ellison, S. L. R., & Rostron, P. (eds) (2019). *Eurachem/EUROLAB/ CITAC/Nordtest/AMC guide: Measurement uncertainty arising from sampling: A guide to methods and approaches* (2nd ed.). Eurachem. Retrieved February 1, 2021, from www.eurachem.org

Reisner, B. A., Pate, C. L., Kinkaid, M. M., Paunovic, D. M., Pratt, J. M., Stewart, J. L., … & Smith, S. R. (2020). I've been given COPUS (Classroom Observation Protocol for Undergraduate STEM) data on my chemistry class. Now what? *Journal of Chemical Education*, *97*(4), 1181–1189.

Salamone, M., & Thomas, K. (2017). Required peer-cooperative learning improves retention of STEM majors. *International Journal of STEM Education*, *4*. doi: 10.1186/s40594-017-0082-3

Silverthorn, D. U., Thorn, P. M., & Svinicki, M. D. (2006). It's difficult to change the way we teach: Lessons from the Integrative Themes in Physiology curriculum module project. *American Journal of Physiology—Advances in Physiology Education*, *30*(4), 204–214.

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*, 618–627.

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.

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 American universities: A snapshot from a large scale observation study. *Science*, *359*(6383), 1468–1470.

Stokstad, E. (2001). Reintroducing the intro course. *Science*, *293*(5535), 1608–1610.

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*, 245–253.

Teasdale, R., Viskupic, K., Bartley, J. K., McConnell, D., Manduca, C., Bruckner, M., …& Iverson, E. (2017). A multidimensional assessment of reformed teaching practice in geoscience classrooms. *Geosphere*, *13*(2), 608–627.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., … & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences USA*, *117*(12), 6476–6483.

Waks, L. J. (2007). The concept of fundamental educational change. *Educational Theory*, *57*(3), 277–295.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K–12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research.

Wood, W. B., & Gentile, J. M. (2003). Teaching in a research context. *Science*, *302*(5650), 1510.