Measurement Instrument for Scientific Teaching (MIST): A Tool to Measure the Frequencies of Research-Based Teaching Practices in Undergraduate Science Courses

Mary F. Durham,[†] Jennifer K. Knight,[‡] and Brian A. Couch[†]*
[†]School of Biological Sciences, University of Nebraska, Lincoln, NE 68588; [‡]Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309

ABSTRACT

The Scientific Teaching (ST) pedagogical framework provides various approaches for science instructors to teach in a way that more closely emulates how science is practiced by actively and inclusively engaging students in their own learning and by making instructional decisions based on student performance data. Fully understanding the impact of ST requires having mechanisms to quantify its implementation. While many useful instruments exist to document teaching practices, these instruments only partially align with the range of practices specified by ST, as described in a recently published taxonomy. Here, we describe the development, validation, and implementation of the Measurement Instrument for Scientific Teaching (MIST), a survey derived from the ST taxonomy and designed to gauge the frequencies of ST practices in undergraduate science courses. MIST showed acceptable validity and reliability based on results from 7767 students in 87 courses at nine institutions. We used factor analyses to identify eight subcategories of ST practices and used these categories to develop a short version of the instrument amenable to joint administration with other research instruments. We further discuss how MIST can be used by instructors, departments, researchers, and professional development programs to quantify and track changes in ST practices.

INTRODUCTION

National calls over the past several decades recommend that science programs alter their undergraduate teaching to optimize student learning and achievement (National Research Council [NRC], 1999, 2003a,b; American Association for the Advancement of Science [AAAS], 2011; President's Council of Advisors on Science and Technology [PCAST)], 2012). These reports propose a wide range of changes based on research-based models of how students learn and the types of expertise and skills that will best serve students in their future careers. They also emphasize the use of teaching strategies that consider the experiences of all students and alleviate historic achievement and representation gaps for particular demographic groups. As a result of these calls, many educators and researchers have made efforts to implement new teaching practices, generate improved curricula, train instructors in research-based instructional strategies, and conduct research on the impacts of pedagogical transformation.

Among the many recent educational movements, a pedagogical approach called "Scientific Teaching" (ST) has gained prominence, particularly in biology disciplines (Handelsman *et al.,* 2004, 2007; AAAS, 2011). Consistent with recommendations in national reports, ST aims to make the teaching of science more closely resemble how science is practiced by infusing courses with the nature and rigor of the scientific process and by incorporating teaching strategies supported by empirical evidence.

Sarah L. Eddy, *Monitoring Editor*
Submitted February 13, 2017; Revised July 14, 2017; Accepted August 10, 2017
CBE Life Sci Educ December 1, 2017 16:ar67
DOI:10.1187/cbe.17-02-0033

*Address correspondence to: Brian A. Couch (bcouch2@unl.edu).

Building on previous ST descriptions, we developed a taxonomy of ST practices to provide a framework for future investigations (Couch *et al.*, 2015). This taxonomy defines the core pedagogical goals of ST and articulates a general approach and specific practices that fulfill each goal. In this manner, the taxonomy translates ST into a list of behaviors, artifacts, and conditions that can be observed and documented in a course.

With respect to its scope, ST promotes the active engagement of students in the learning process through group activities and formative assessments (Frederick, 1987; Prince, 2004) and recommends that instructors use a backward design process to align their learning objectives, course activities, and assessments (Wiggins and McTighe, 2005). ST also highlights the importance of cognitive processes critical for the practice of science and learning, including connecting science with society (Sadler *et al.*, 2004; Zeidler *et al.*, 2005; Chamany *et al.*, 2008; Labov and Huddleston, 2008; Pierret and Friedrichsen, 2009), using science process skills (Hanauer *et al.*, 2006; Bao *et al.*, 2009; Coil *et al.*, 2010; Wei and Woodin, 2011; Goldey *et al.*, 2012), incorporating concepts from across different disciplines (Bialek and Botstein, 2004; Labov *et al.*, 2010; Tra and Evans, 2010), and developing metacognitive reflection (Ertmer and Newby, 1996; Pintrich, 2002; Schraw *et al.*, 2006; Tanner, 2012). Finally, ST further emphasizes inclusive teaching practices that reduce unconscious biases and affirm students with diverse backgrounds as members of the scientific community (Steele, 1997; Seymour, 2000; Dasgupta and Greenwald, 2001; Uhlmann and Cohen, 2005; Tanner and Allen, 2007).

Over the past two decades, a number of faculty development programs have been created to promote the use of research-based instructional practices, including those associated with ST. In particular, the Summer Institutes on Scientific Teaching (SI)[1] has trained more than 1600 instructors in ST strategies between 2004 and 2016 (Pfund *et al.*, 2009). The SI is a 4- to 5-day workshop in which participants learn about ST and work in groups to develop an ST-based teaching module. Participants are then encouraged to implement ST practices in their courses and share this pedagogical approach with peers at their home institutions. The practices associated with ST are also used in a variety of other teacher development workshops, such as the On the Cutting Edge program in geosciences (Manduca *et al.*, 2010), the Cottrell Scholars program in chemistry (Baker *et al.*, 2014), the Workshop for New Physics and Astronomy Faculty (Henderson, 2008), and the FIRST IV (Faculty Institutes for Reforming Science Teaching IV) workshop in biology (Ebert-May *et al.*, 2015). While initial reports from the SI have detected promising changes in instructional practices among SI alums (Pfund *et al.*, 2009; Aragón *et al.*, 2016), many questions still remain regarding the degree to which instructors trained at the SI or through other programs implement ST practices in their courses, how successfully participants disseminate the ST approach within and across departments, and whether changes in teaching practices lead to corresponding changes in student outcomes. In the longer term, addressing these questions requires the development of instruments to quantify the use of ST practices in courses.

Many different instruments have been used by researchers to characterize teaching in undergraduate science courses (AAAS, 2013). In addition to differences in their underlying development frameworks, these instruments also vary in who completes the evaluation. Some instruments ask students to answer survey questions based on their experiences in a course. For example, the Student Evaluation of Educational Quality (SEEQ) asks students to evaluate the quality of various course components, such as overall learning, instructor enthusiasm, course organization, group interactions, instructor rapport, topical breadth, exams, and assignments (Marsh, 1982). With other instruments, instructors report on the strategies used in their own courses. The Teaching Practices Inventory (TPI) and the Postsecondary Instructional Practices Survey both ask instructors about the extent to which they implement various research-based teaching practices and include questions related to how students engage with course content, whether students interact with their peers, and how the instructor gauges and provides feedback on student learning (Wieman and Gilbert, 2014; Williams *et al.*, 2015; Walter *et al.*, 2016). Finally, a number of instruments rely on an external observer to evaluate or document classroom dynamics. For example, the Reformed Teaching Observation Protocol (RTOP) evaluates whether a course incorporates certain reformed teaching strategies that create a student-centered learning environment and includes questions regarding lesson design, propositional knowledge, procedural knowledge, student–instructor interactions, and student–student interactions (Sawada *et al.*, 2002). The Teaching Dimensions Observation Protocol (TDOP) and Classroom Observation Protocol for Undergraduate STEM (COPUS) describe teaching practices by recording whether certain behaviors occur during 2-minute intervals throughout a class period (Hora *et al.*, 2013; Smith *et al.*, 2013). The Practical Observation Rubric to Assess Active Learning (PORTAAL) is used by observers to document the frequency and duration of class activities that employ specific active-learning techniques documented to improve student learning (Eddy *et al.*, 2015). These instruments also vary in the extent to which they use human judgment to evaluate the quality of teaching or solely describe practices with no judgment of teaching quality or efficacy (Hora, 2013). For example, the COPUS is strictly descriptive, whereas the SEEQ is largely evaluative.

While the existing instruments can measure various teaching practices within undergraduate science courses, none of them is explicitly aligned with the ST framework and, therefore, they do not account for the full spectrum of ST practices. For example, none of the abovementioned instruments measures alignment of formative or summative assessments with learning goals or how often examples and analogies highlight diverse groups or perspectives. Several other ST practices are measured by only one instrument from this group. The RTOP is the only instrument that measures the use of interdisciplinary content, how often students design or evaluate experimental strategies, how often the instructor mentions contributions from diverse people or perspectives, and the level of instructor sensitivity (Sawada *et al.*, 2002). The TPI is the only instrument that determines how often students are asked to read or evaluate scientific articles, how often instructors describe the historical context of breakthrough discoveries, or whether summative assessment items use a variety of question formats

[1] The Summer Institutes on Scientific Teaching was previously called the National Academies Summer Institute for Undergraduate Education in Biology.

(Wieman and Gilbert, 2014). The TDOP is the only instrument that measures how often students are asked to reconcile conflicting data, use scientific judgment to address challenges, or use appropriate statistical methods (Hora *et al.*, 2013). Finally, PORTAAL is the only instrument that explores how often an instructor uses strategies to promote individual accountability within group exercises (Eddy *et al.*, 2015).

While all of the instruments described were rigorously developed and serve their designed purposes well, no single existing instrument fully accounts for the current breadth of recommended ST practices. We developed the Measurement Instrument for Scientific Teaching (MIST) to fill this role. Here, we describe the development of MIST, including how we established instrument validity during the item development process, and we report factor analyses and reliability statistics from a large-scale administration of the instrument with undergraduate students. We further demonstrate how results from this instrument can be used for the documentation and ongoing improvement of teaching practices in science courses.

METHODS

Item Development and Revision

We began the instrument development process by translating supporting practices from the ST taxonomy into survey questions (Couch *et al.*, 2015). To the extent possible, questions focused on activities, opportunities, and structures provided by an instructor to students, and items were worded in objective terms using limited educational jargon to ensure they could be interpreted and answered by any person affiliated with the course (e.g., student, instructor, observer, teaching assistant, or administrator). In some cases, definitions and examples were provided to help survey respondents better understand the range of activities satisfying a given question. Similar to other existing instruments, item response scales varied based on the type of question being asked (Brawner *et al.*, 2002; Wieman and Gilbert, 2014). In total, MIST uses 49 items to capture the 37 supporting practices of ST delineated in the ST taxonomy, because some taxonomy practices, such as course alignment with learning goals, require more than one MIST item to adequately capture the extent of their use in the course. Most MIST items used one of four answer choice scales: a seven-point Likert-style frequency scale, a six-point Likert-style agree–disagree scale, a 0–100% slider-bar scale, or a no/yes answer.

We used interviews to optimize the clarity of the individual items and improve the face and content validity of the instrument (Reeves and Marbach-Ad, 2016). Interviewees completed an online version of MIST while participating in a think-aloud session in which they shared the thought process they used to answer each question (Anders and Simon, 1980). This helped identify issues with question interpretability and answer choices. Question revisions proceeded in an iterative cycle in which we conducted two to five interviews between each round of item editing. In total, we conducted 54 interviews with undergraduate students at the University of Nebraska–Lincoln (UNL), instructors from multiple institutions, and other individuals involved in educational efforts (e.g., program evaluators, professional society representatives). In addition, a draft version of MIST was piloted to 29 students in a 2015 summer session course to test software and participation logistics. The full MIST instrument is provided in Supplemental Material 1.

MIST Structure and Administration

MIST items were composed in the third-person tense so that any person with access to a course could complete the instrument. In this article, we have focused on responses from the student perspective. We purposefully recruited student participants from courses using a wide range of teaching practices. Table 1 presents a complete description of institution, course, and student demographics. Institutions spanned a range of sizes and were primarily classified as research institutions. Courses represented a balance of enrollment sizes and course levels and were largely from biology disciplines. Participating students had gender and race/ethnicity distributions roughly reflective of the

TABLE 1. MIST 2015–2016 administration demographics

	n	Percent of sample
Institutions		
Carnegie classification		
Highest research activity (R1)	5	56
Higher research activity (R2)	3	33
Primarily undergraduate institution	1	11
Undergraduate enrollment		
Small (<10,000)	2	22
Medium (10,000–20,000)	1	11
Large (20,000–30,000)	3	33
Very large (>30,000)	3	33
Courses	87	
Discipline		
Biology	79	91
Other STEM	8	9
Enrollment		
Small (<25 students)	18	21
Medium (26–100 students)	28	32
Large (>100 students)	41	47
Course level		
Lower division (100–200 level)	46	53
Upper division (300–400 level)	41	47
Students	7767	
Class year		
First year	1542	20
Sophomore	2080	27
Junior	2233	29
Senior	1694	22
Other	218	3
Gender		
Female	4788	62
Male	2873	37
Other	18	0.2
Not specified	88	1
Ethnicity		
Underrepresented minority (URM)	1224	16
Non-URM	6543	84

broader student populations at each institution and ranged from first-years to seniors.

We administered the final version of MIST containing 49 total items[2] to students enrolled in 87 courses at nine different institutions during the 2015–2016 academic year. Students completed MIST between weeks 13 and 14 of a 15-week semester. The instrument was administered online, outside class through Qualtrics, and was followed by a demographics questionnaire. To streamline the participant experience, the survey included conditional questions that appeared only if certain teaching strategies were reported in prior questions. We asked instructors to give students a small amount of course credit for participating in the survey. Of the 9960 students enrolled in participating courses, 8006 accessed the online survey. After removal of incomplete responses and responses from nonconsenting students and students under 18 years of age, the final data set contained 7767 complete student survey responses, representing 78% of the total enrollment in participating courses.

Data Processing

Survey responses were converted to numerical codes for data analysis. Responses were assigned a value of 0–6 for Likert-style scales,[3] 1–6 for agree–disagree scales,[4] 0–10 for slider-bar scales, and 0/1 for no/yes questions. Conditional response questions that were not displayed were scored as zero, indicating that the practice did not occur.

Thirteen participating instructors taught duplicate course sections in the same semester. We collected data separately for each of these sections and examined the correlation in student question-level responses between sections. In all cases, paired section responses had a Pearson's correlation greater than 0.80, so the data were combined and treated as a single course for that instructor. There were also 10 team-taught courses in which separate instructors taught different portions of the course. For these courses, students were randomly assigned to complete MIST based on the teaching of one instructor or the other, and these responses were treated as separate courses.

To calculate survey durations, we tabulated individual page dwell times. For any cases in which a student stayed on the same page for longer than 20 minutes, we replaced this dwell time with the average dwell time for that page for the student's course section. We then used the sum of page dwell times to calculate total survey completion time. Students completed MIST in an average of 11.2 minutes, with 80% of students completing the instrument in less than 15 minutes.

Analysis of MIST as a Single Scale

To analyze MIST responses as a single scale, we calculated the internal reliability across all survey items using scale reliability analyses in SPSS. While we did not expect that instructors who implemented one practice would necessarily implement all the other practices, we did suspect that each of the ST practices in MIST would be more likely to be implemented by more transformed instructors compared with more traditional instructors. Output from the reliability procedure included Cronbach's alpha coefficient and item-total correlations, which are both measures of the internal consistency of survey items (Netemeyer *et al.*, 2003; Hanauer and Dolan, 2014). The alpha coefficient reflects the degree of covariance between survey items and ranges from 0 to 1, with values above 0.7 considered acceptable. Item-total correlations indicate the degree to which responses for each item are consistent with responses on the entire instrument and range from −1 to 1, with correlations above 0.3 considered representative of the overall scale (Pallant, 2010).

We also conducted confirmatory factor analysis (CFA), using the lavaan package in R to determine whether response patterns were consistent with a single underlying factor (Rosseel, 2012). CFA model goodness of fit was evaluated following established recommendations (Hu and Bentler, 1999). The comparative fit index (CFI) and Tucker-Lee index (TLI) are comparative fit indices that compare the fit of the specified model with the fit of a baseline model in which covariances between items are set to zero (Brown, 2015). The root-mean-square error of approximation (RMSEA) is a population-based parsimony measure that estimates the extent to which the model fits the data, taking sample size into account. The standard root-mean-square residual (SRMR) estimates absolute fit of the model by measuring the difference between observed and model-predicted item correlations (Brown, 2015). We calculated factor loadings to determine the extent to which each item can be explained by the underlying factor, and nearly all MIST items saliently loaded above 0.3 (Fabrigar *et al.*, 1999; Costello and Osborne, 2005; Field, 2014).

Identification of MIST Subcategories

In developing MIST, we recognized that certain groups of practices were related, in that they reflected a more general teaching approach. For example, we might expect an instructor committed to active learning to score high on items related to in- and out-of-class activity, group work, peer feedback, and polling methods. Similarly, an instructor wishing to help students develop fluency with data analysis and interpretation might have students apply statistical approaches, construct graphs, interpret different data representations, and use models. We used a combination of factor analyses and theoretical grounding to identify groups of related practices and ensure that each group aligned with the underlying ST framework (Woolley *et al.*, 2004; Brown, 2015; Harshman and Stains, 2017).

We began by using exploratory factor analysis (EFA) to determine whether we could detect underlying factors that explained the variance in student responses to particular groups of questions (Thompson, 2004; Hanauer and Dolan, 2014). The underlying factors identified by EFA thus reflected groups of teaching practices that tended to be implemented together by instructors (Fabrigar *et al.*, 1999; Field, 2014). This analysis was initially conducted on 63 courses from Fall 2015. Several criteria indicated that the data set was suitable for this analysis: many correlation coefficients in the correlation matrix were above a 0.3 threshold; the Kaiser-Meyer-Olkin measure of

[2] One question (Q19) inquiring about how students were grouped is asked of instructors only.

[3] One question referring to learning goal dissemination had a select-all answer format with seven possible answers to select. For this question, a single code of 0–6 was assigned corresponding to the highest frequency at which learning goals were provided to students.

[4] Three of the four agree/disagree items had a "not applicable" answer choice (e.g., "This course did not include whole-class discussions"). The n/a responses were assigned a zero score. All remaining disagree-agree responses were scored as 1–6.

sampling adequacy was 0.929, which exceeded the 0.6 recommended threshold (Kaiser, 1970; Kaiser and Rice, 1974); and a significance of $p < 0.001$ was reached with the Bartlett's test of sphericity (Bartlett, 1950).

We completed EFA procedures in SPSS using maximum-likelihood extraction and direct oblimin rotation. We considered three types of criteria to determine the number of factors to accept (Fabrigar *et al.*, 1999; Brown, 2015). The Kaiser-Guttman rule recommends including all factors with eigenvalues above 1.0 in the correlation matrix (Guttman, 1954; Kaiser, 1960). The scree test recommends including all factors with eigenvalues that are substantially lower than the previous factor, as inferred by the inflection point on a "scree plot" (Catell, 1966). Parallel analysis compares eigenvalues of the sample with eigenvalues of random numbers to determine the number of factors to include in the EFA (Horn, 1965). We used the initial EFA output to determine the Kaiser-Guttman rule and the scree test results, and we completed parallel analysis using a syntax for SPSS (O'Connor, 2000).

Based on an initial EFA with no a priori number of factors, the Kaiser-Guttman rule specified seven factors, the scree test indicated between five and eight factors, and parallel analysis indicated the presence of 10 factors. We explored each of these models by running separate EFAs with five, six, seven, eight, and 10 factors. Preliminary EFA analyses revealed that two items (one question about exam frequency and one question about incorporating the historical context of scientific breakthroughs) did not factor consistently into any category, so these items were excluded from this and all subsequent subcategory analyses. We eliminated the eight- and 10-factor EFA models, because they resulted in one or more factors with less than three items (Fabrigar *et al.*, 1999; Costello and Osborne, 2005). We concluded that the seven-factor model, which explained 47.6% of the variance in the data, was the best fit to the data. We intended to use EFA solely as an initial guide to identify subcategories from a data-driven perspective, so we assigned items to the factors in which they had the highest factor loadings and did not set a rigid cutoff. Nonetheless, most MIST items loaded on their respective factors above 0.30 (Fabrigar *et al.*, 1999; Costello and Osborne, 2005; Field, 2014), and no items were cross-loaded above 0.30 (Supplemental Material 2).

While EFA procedures represent a rigorous approach to obtaining empirically derived factor structures, these structures are highly contingent on the particular sample. In fact, a recent investigation of the widely used Approaches to Teaching Inventory (ATI) indicated that at least 23 different plausible EFA structures have been used to categorize ATI items in 39 different studies (Harshman and Stains, 2017). Furthermore, we recognized that unrelated items may factor together for other reasons, such as their co-occurrence in professional development programs. To address these limitations and ensure that the MIST subcategories would be meaningful to users in broader contexts, we made theoretically grounded adjustments to the EFA structure to bring the groups into alignment with the ST framework (Woolley *et al.*, 2004; Hughes *et al.*, 2006; Harshman and Stains, 2017).

The three questions related to polling methods initially appeared as a separate factor in EFA; however, polling methods could also be viewed as a specific active-learning modality. Thus, we removed this factor and reassigned the polling method questions to the factors related to active learning and learning goal alignment. In light of the groupings and resultant factor loadings, we retained five of the six remaining factors that resonated with the ST framework. The sixth factor appeared to be combining subsets of ST practices related to different cognitive processes, so we split this factor into three subcategories. We then confirmed these eight subcategories by performing CFA and calculating coefficient alphas for each factor separately on the Fall 2015 sample, the Spring 2016 sample, and the full sample of 87 courses from both semesters (Fabrigar *et al.*, 1999; Netemeyer *et al.*, 2003; Hanauer and Dolan, 2014).

To verify that the revised groupings reflected sets of related teaching practices and that we had adequately defined the approach underlying each group, we solicited feedback from 10 faculty with expertise in ST. We asked experts to indicate whether or not they agreed that each MIST item fit with the other items in its assigned category. In the case of disagreement, experts were asked to explain their reasoning and indicate an alternative category. The expert panel generally agreed with our MIST subcategory groupings, and no concerns were raised with respect to the categories that were modified from the original EFA structure. Forty-one of the 46 MIST items included in the subcategory model had 90–100% expert agreement with their assigned categories. The remaining five items had 50–70% expert agreement. Three of these items referred to the instructor providing feedback to students on formative or summative assessments, one referred to students stating interests and asking original questions during whole-class discussions, and one referred to incorporating real-life examples. In each of these cases, expert concerns were related to the MIST subcategory titles being inclusive of the items contained within the subcategory, which we addressed by adding appropriate descriptions to the titles. The final MIST subcategory model consists of eight subcategories of ST practices: Active-Learning Strategies, Learning Goal Use and Feedback, Inclusivity, Responsiveness to Students, Experimental Design and Communication, Data Analysis and Interpretation, Cognitive Skills, and Course and Self-Reflection.

Development of a MIST Short Version

We developed a short version of MIST (MIST-Short) for users with survey time constraints, such as instructors or researchers who want to pair MIST with other instruments. MIST-Short was developed to retain representation of each subcategory. Thus, two or three items within each subcategory were selected based on several criteria, including high factor loadings, high response variation across courses, low variation within courses, and centrality to the ST framework. To analyze MIST-Short as a single scale, we calculated coefficient alphas and conducted CFA with a single-factor solution using data extracted from the full version of MIST. We also calculated Pearson's correlations between MIST-Short subcategory scores and corresponding subcategory scores from the full MIST instrument. On the basis of the timing per question, we estimate that students can complete MIST-Short in approximately 5 minutes.

Scoring System

To determine MIST scores, we calculated the mean response for students in a given course for each question, and this value was normalized to the maximum scale value for that question, using the equation

$$x_j = \overline{s}_j \,/\, a_{\text{max}j}$$

where x_j is the normalized response for question j, $\overline{s}_j$ is the mean student response for question j, and $a_{\text{max}j}$ is the maximum scale value possible for question j (i.e., 10 for slider-bar questions and 6 for Likert-style questions). Item 15, a no/yes question on group work, was not included in score calculations, because group-work information is included in subsequent questions.

For determination of total scores for MIST as a single scale, the eight MIST subcategories and MIST-Short, normalized mean responses from relevant MIST items were summed and divided by the number of contributing questions, using the equation

$$\text{MIST}_{\text{scale score}} = \left[\left(X_{Q1} + X_{Q2} + \ldots + X_{Qn}\right) / \, n\right]$$

where $X_{Q1} \ldots X_{Qn}$ are the normalized mean responses for each question contributing to the specified scale and n is the number of questions included in the scale calculation. Total scores were normalized to a 0–100 scale by multiplying by 100. Note that nonnormalized MIST subcategory scores will not add up to the full MIST score because two MIST items were not included in any subcategories and because each MIST subcategory score is drawn from a different number of MIST items.

This project was classified as exempt from institutional review board review at UNL (project ID 15016) and all other participating institutions.

RESULTS

MIST Can Provide an Overall Estimate of ST

We conducted several analyses to determine the extent to which student responses to MIST items aligned as a single scale. MIST had high internal reliability, with an overall alpha of 0.93. Nearly all the MIST items had item-total correlations above 0.30; however, some items pertaining to inclusivity or exam alignment showed weaker correlations with the overall MIST scale (Table 2). In addition, the exam frequency item did not correlate with the overall scale. A variety of minimum factor-loading cutoffs are recommended in the literature to indicate salient loadings, including 0.4 (Matsunaga, 2010), 0.32 (Tabachnick and Fidell, 2001; Costello and Osborne, 2005), and 0.30 (Costello and Osborne, 2005; Field, 2014). Aside from the inclusivity and exam items, all MIST items saliently loaded at 0.30 or above.

MIST Contains Discernible Subcategories of Teaching Practices

Student MIST responses were also used to develop a scoring system that provides information on discrete aspects of ST. To examine and identify the underlying structure of MIST, we performed an iterative series of factor analyses aimed at identifying the number of subcategories present within the instrument and determining which items aligned with each subcategory. On the basis of results of these analyses and theoretical groundings in the ST framework, we arrived at a final "subcategory model" that specified eight latent variables with three to 13 individual items loading on each factor (Table 3). Both semester samples and the full sample produced similar model fit characteristics (full sample: CFI = 0.73, TLI = 0.71, RMSEA = 0.082, SRMR = 0.079). All factors loaded saliently onto their respective subcategories at 0.4, except one item that loaded at 0.316. Furthermore, each subcategory showed evidence of acceptable internal reliability with alphas of 0.69–0.86 (Table 3).

MIST Shows a Wide Range of Responses at Different Levels of Resolution

To determine the range of teaching practices used across the sample courses, we visualized the distribution of MIST results at the level of overall scores, subcategory scores, and individual teaching practices. Overall MIST scores ranged from 24 to 71 on a scale of 0–100, with a relatively normal distribution and higher scores representing higher levels of ST implementation (Figure 1). Based on the structure of the survey, it was unlikely that MIST scores would have fallen in the extreme ranges of the scale (i.e., outside 15–85).

MIST subcategory scores showed varying degrees of implementation across the courses sampled (Figure 2). Three subcategories generally had score distributions closer to the upper end of the scale (learning goal use and feedback, inclusivity, and responsiveness to students), four subcategories had moderate implementation levels (active-learning strategies, experimental design and communication, data analysis and interpretation, and cognitive skills), and one subcategory was noticeably lower than the others (course and self-reflection).

Individual items showed the broadest range of response distributions (Table 2 and Supplemental Material 3). Items with the highest normalized responses included instructor sensitivity to socially controversial issues (Q26), students stating interests and asking questions in class (Q29), and exam alignment with learning goals (Q13). Items with the lowest implementation levels were out-of-class group work (Q18), group participation strategies (Q20), and scientific communication in formal written papers or oral presentations (Q41).

With respect to global measures of in-class activity, students reported engaging in nonlecture activities for an average of 48% of class time (Q1) and working in groups for an average of 42% of class time (Q16). On average, three polling questions were asked each week, and students completed in-class activities about once per week.

SI Participants Show Higher MIST Scores

We examined associations between MIST results and instructor and course characteristics for the given sample (Figure 3). Students in courses taught by instructors who had attended an SI reported significantly higher perceptions of ST practices than students in courses taught by instructors who had not attended an SI (Figure 3A, SI participants, mean = 53.8 ± 1.6 SE; non-SI participants, mean = 47.0 ± 1.3 SE; $t = 3.07$, $df = 84$, $p = 0.003$, effect size as Cohen's $d = 0.71$). We found no difference in overall MIST scores between lower-division (100–200 level) and upper-division (300–400 level) courses (Figure 3B, lower division, mean = 49.5 ± 1.4 SE; upper division, mean = 48.6 ± 1.8 SE; $t = 0.39$, $df = 85$, $p = 0.70$). The sample also showed no trend in overall MIST scores based on course size (Figure 3C, $r = -0.05$, $p = 0.65$).

MIST Provides Feedback for Individual Instructors

Instructor MIST profiles showed different instructional strengths and weaknesses. We highlight results from three

TABLE 2. Item-total correlations, factor loadings, and descriptive statistics for individual MIST items on full MIST scale: ST single-scale model (alpha = 0.93)

Question no.	Item description	Item-total correlation	Full MIST factor loading	Max scale value	Mean normalized score[a]	SD[a]	Mean course SD[a]
1	Percent active	0.52	0.561	10	0.45	0.27	0.20
2	Learning goal maximum frequency	0.35	0.345	6	0.64	0.27	0.25
3	Polling method: frequency	0.41	0.396	6	0.60	0.39	0.19
4	Polling method: % alignment	0.50	0.459	10	0.55	0.39	0.27
5	Polling method: % peer learning	0.47	0.389	10	0.52	0.39	0.25
6	In-class: frequency	0.59	0.622	6	0.42	0.29	0.19
7	In-class: % alignment	0.57	0.399	10	0.59	0.39	0.30
8	In-class: % feedback	0.60	0.500	10	0.49	0.37	0.31
9	Out-of-class: frequency	0.34	0.350	6	0.49	0.25	0.16
10	Out-of-class: % alignment	0.43	0.365	10	0.67	0.35	0.28
11	Out-of-class: % feedback	0.46	0.429	10	0.45	0.36	0.32
12	Exams: frequency	0.03	−0.009	6	0.65	0.20	0.16
13	Exams: % alignment	0.26	0.185	10	0.79	0.27	0.24
14	Exams: % feedback	0.40	0.372	10	0.53	0.35	0.31
15	Group work: y/n[b]	0.53	0.589	1	0.37	0.35	0.20
16	Group work: % of class time	0.53	0.623	10	0.42	0.37	0.20
17	Group work: in-class frequency	0.59	0.678	6	0.18	0.27	0.21
18	Group work: out-of-class frequency[c]	0.40	0.545	6	0.18	0.28	0.24
20	Group work: group participation strategy	0.46	0.611	6	0.40	0.38	0.22
21	Group work: share results with whole class	0.56	0.649	6	0.27	0.33	0.27
22	Peer feedback	0.56	0.615	6	0.54	0.36	0.28
23	Students respond to each other	0.54	0.584	6	0.58	0.35	0.31
24	Diverse examples and analogies	0.29	0.300	6	0.63	0.33	0.29
25	Diverse scientist/researcher contributions	0.28	0.284	6	0.78	0.19	0.18
26	Instructor sensitivity	0.23	0.195	6	0.29	0.29	0.24
27	Students provide feedback on activities/content	0.44	0.492	6	0.38	0.38	0.33
28	Make adjustment from student feedback	0.46	0.507	2	0.77	0.18	0.16
29	Student state interests and ask original questions	0.37	0.357	6	0.66	0.27	0.23
30	Instructor aware of student nonunderstanding	0.45	0.424	6	0.64	0.30	0.27
31	Follow-up activities provided if not understood	0.48	0.467	6	0.50	0.30	0.25
32	Make hypotheses/predictions	0.62	0.687	6	0.37	0.31	0.26
33	Critique hypotheses and experimental strategies	0.58	0.670	6	0.29	0.30	0.25
34	Design experiments	0.57	0.647	6	0.40	0.31	0.25
35	Summarize, interpret, analyze data with math	0.53	0.598	6	0.28	0.29	0.22
36	Make graphs or tables	0.51	0.587	6	0.50	0.28	0.22
37	Analyze/interpret data graphs/tables	0.54	0.586	6	0.45	0.30	0.25
38	Use data to make decisions/defend conclusions	0.60	0.661	6	0.48	0.29	0.26
39	Use models	0.51	0.549	6	0.28	0.29	0.23
40	Scientific literature or media articles	0.40	0.463	6	0.16	0.25	0.22
41	Science communication: written papers/oral pres.	0.33	0.408	6	0.59	0.27	0.24
42	Course concepts applicable to life	0.37	0.355	6	0.49	0.28	0.24
43	Historical context	0.32	0.316	6	0.46	0.32	0.29
44	Use nonwritten formats	0.42	0.449	6	0.42	0.32	0.29
45	Interdisciplinary	0.48	0.511	6	0.66	0.27	0.23
46	Higher-level thought processes	0.45	0.446	6	0.47	0.32	0.26
47	Open-ended exercises/case studies	0.58	0.633	6	0.35	0.28	0.27
48	Reflection: effective study habits	0.49	0.533	6	0.37	0.30	0.27
49	Reflection: problem-solving strategies	0.55	0.600	6	0.45	0.27	0.20

[a] Mean normalized score and SD are calculated from all individual student responses. Mean course SD is the mean of SDs from each course.

[b] Question 15 was included in initial scale analyses, but was not included in MIST scores because it was accounted for in questions 16–21.

[c] Question 19 was asked only of instructor participants.

TABLE 3. MIST subcategory model, subcategory reliabilities, and factor loadings of MIST items

Item	Item description	Factor loading
Active-Learning Strategies (alpha = 0.86)		
Q1	Percent active	0.598
Q3	Polling method: frequency	0.405
Q5	Polling method: % peer learning	0.513
Q6	In-class: frequency	0.645
Q9	Out-of-class: frequency	0.356
Q15	Group work: y/n[a]	0.840
Q16	Group work: % of class time	0.806
Q17	Group work: in-class frequency	0.894
Q18	Group work: out-of-class frequency	0.636
Q20	Group work: group participation strategy	0.680
Q21	Group work: share results with whole class	0.838
Q22	Peer feedback	0.600
Q23	Students respond to each other	0.510
Learning Goal Use and Feedback (alpha = 0.79)		
Q2	Learning goal maximum frequency	0.418
Q4	Polling method: % alignment	0.536
Q7	In-class: % alignment	0.783
Q8	In-class: % feedback	0.773
Q10	Out-of-class: % alignment	0.549
Q11	Out-of-class: % feedback	0.523
Q13	Exams: % alignment	0.429
Q14	Exams: % feedback	0.475
Inclusivity (alpha = 0.69)		
Q24	Diverse examples and analogies	0.835
Q25	Diverse scientist/researcher contributions	0.854
Q26	Instructor sensitivity	0.316
Responsiveness to Students (alpha = 0.73)		
Q29	Student state interests and ask original questions	0.555
Q30	Instructor aware of student nonunderstanding	0.820
Q31	Follow-up activities provided if not understood	0.803
Q42	Course concepts applicable to life	0.431
Experimental Design and Communication (alpha = 0.83)		
Q32	Make hypotheses/predictions	0.743
Q33	Critique hypotheses and experimental strategies	0.848
Q34	Design experiments	0.777
Q40	Scientific literature or media articles	0.588
Q41	Science communication: written papers/ oral pres.	0.525
Data Analysis and Interpretation (alpha = 0.85)		
Q35	Summarize, interpret, analyze data with math	0.714
Q36	Make graphs or tables	0.656
Q37	Analyze/interpret data graphs/tables	0.798
Q38	Use data to make decisions/defend conclusions	0.845
Q39	Use models	0.663
Cognitive Skills (alpha = 0.72)		
Q44	Use nonwritten formats	0.584
Q45	Interdisciplinary	0.640
Q46	Higher-level thought processes	0.591
Q47	Open-ended exercises/case studies	0.689
Course and Self-Reflection (alpha = 0.77)		
Q27	Students provide feedback on activities/ content	0.503
Q28	Make adjustment from student feedback	0.488
Q48	Reflection: effective study habits	0.853
Q49	Reflection: problem-solving strategies	0.903

[a] Question 15 was included in factor analyses but was not included in the subcategory score because it was accounted for in questions 16–21.

instructor participants to demonstrate how individuals could derive information from their MIST reports to guide instructional decisions (Figure 4). Instructors A and B had a high degree of ST implementation, evidenced by overall MIST scores in the 70th and 85th percentiles, respectively, but these instructors showed different strengths in MIST subcategories. Instructor A showed higher levels of inclusivity and experimental design and communication, while Instructor B had higher rankings in responsiveness to students, cognitive skills, and course and self-reflection. Conversely, Instructor C's overall ST implementation levels were much lower, but this instructor showed relative strength in having students consider aspects of experimental design and communication, with responses in this MIST subcategory reaching the 84th percentile.

MIST-Short Approximates Scores from the Full Version

Because we did not administer MIST-Short by itself, we estimated characteristics of the shortened instrument by analyzing student responses to the selected subset of items from the full version of MIST. From this subset of student data, estimated

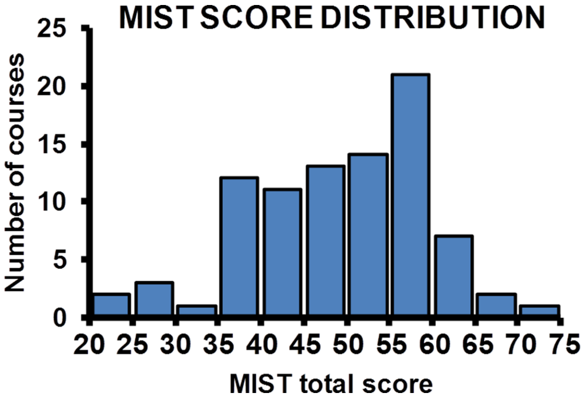


FIGURE 1. Frequency distribution of overall MIST scores. Bars represent the number of courses within each score bin. For example, the rightmost bin contains MIST scores greater than 70 and less than or equal to 75. *n* = 87 courses.

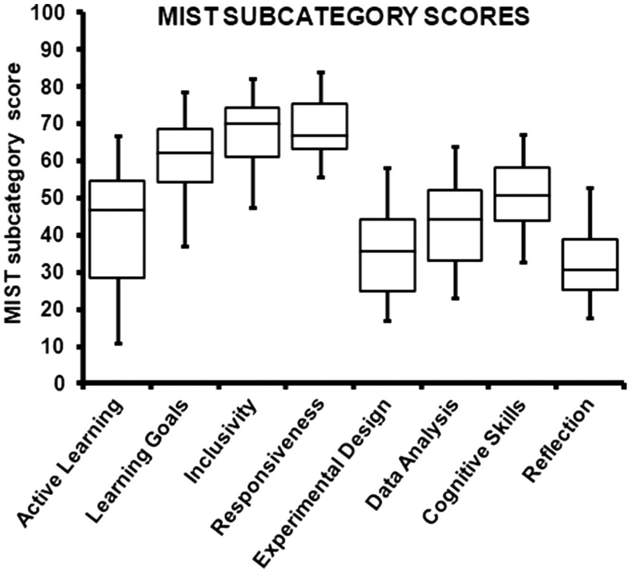


FIGURE 2. Score distributions for the eight MIST subcategories. Central bars represent subcategory median scores, boxes represent inner quartiles, and whiskers represent the 5th and 95th percentile values. $n = 87$ courses.

results for MIST-Short showed good internal reliability (alpha = 0.85), and each item on the short version saliently loaded at 0.30 or above with a single factor specified in CFA (Table 4). Simulated MIST-Short total scores correlated very closely with MIST full-version total scores (r = 0.97). Each simulated short-version subcategory also showed a strong correlation (r range: 0.87–0.98) with the corresponding subcategory score from the full version (Table 5).

DISCUSSION

The ST pedagogical framework and its supporting instructional practices have been emphasized in many national calls to improve undergraduate science, technology, engineering, and mathematics (STEM) education (AAAS, 1990, 2011; NRC, 2003a). To further understanding of how ST practices influence student success, we recognized a need for a descriptive instrument to gauge the extent of ST implementation in undergraduate courses. While existing instruments capture some aspects of ST, we developed MIST to specifically align with the full range of potential ST practices. The 49 items on MIST represent nearly all the supporting practices identified in the ST taxonomy (Couch *et al.*, 2015). Furthermore, the development process and results presented here provide evidence for the validity and reliability of the full scale, eight subcategories, and individual items corresponding to the frequency or extent of specific teaching practices. MIST-Short also demonstrated a capacity to approximate scores from the full version.

Integrating Response Patterns with Underlying Theory

Building on the framework specified in the ST taxonomy, our response modeling process revealed both expected and unexpected aspects of how instructors implement ST. In developing a subcategory model, we discovered that student responses empirically grouped into seven factors. This implies a degree of correspondence in the implementation and perception of certain groups of teaching practices. For example, the items in the active-learning strategies factor address the extent to which students were actively engaged, answering questions, and working together during a course. To ensure that all item groupings had practical significance for survey users, we also made theoretically grounded decisions to adjust some factors to arrive at a final subcategory model. We recognize that these categorical groupings reflect the current state of implementation patterns and may change over time, so different factoring models should be considered in the future as the state of transformed teaching advances. Fit statistics for the final model were on par with those of a recently published instrument measuring instructional practices (Walter *et al.*, 2016). Thus, the final eight subcategories represent an integration of response patterns with underlying theory and provide an additional level at which to consider ST implementation.

Our modeling process also revealed that current perceived implementation patterns of some practices are not tightly aligned with the overall ST framework. When considered as a single scale, two items related to exam frequency and alignment did not correlate strongly with the full scale. The misalignment of the exam frequency question was not surprising, because ST does not have explicit directives on an ideal exam frequency. Having exams that align with underlying learning

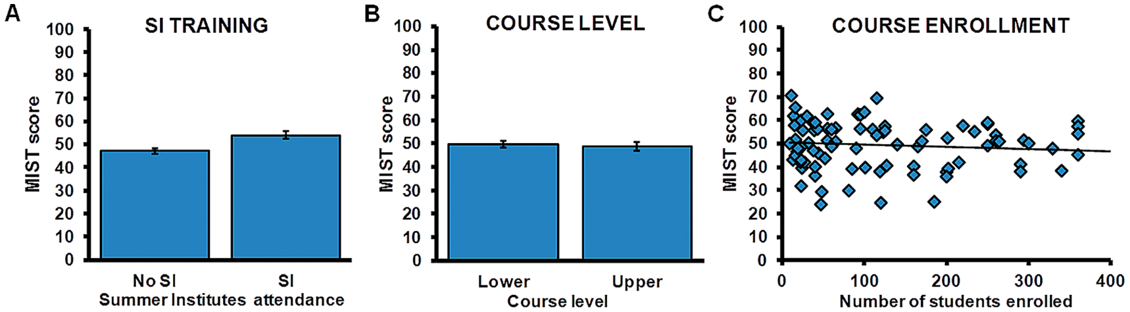
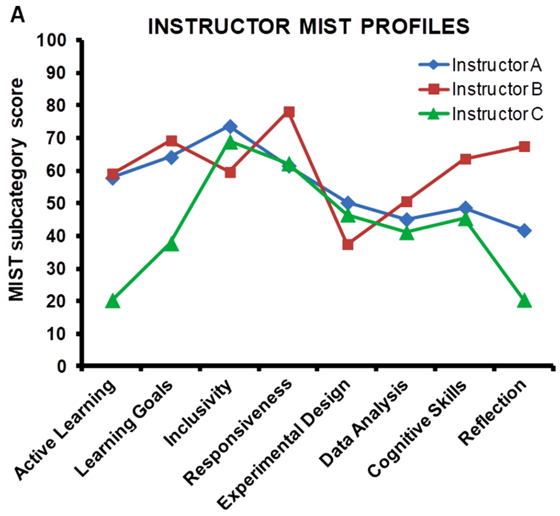


FIGURE 3. MIST scores based on (A) SI participation status, (B) course level, and (C) course enrollment. Bars represent mean ± SE for courses in each group. Diamonds correspond to MIST scores for each individual course of the indicated enrollment size. The solid line represents the regression line. $n = 58$ non-SI participants, 28 SI participants; $n = 48$ lower-division, 39 upper-division courses; $n = 87$ total courses.



B

Scale	Instructor A		Instructor B		Instructor C	
	Score	Percentile	Score	Percentile	Score	Percentile
Full MIST	56.6	70th	59.1	85th	38.3	15th
MIST Subcategory:						
Active Learning	57.8	80th	59.1	85th	20.3	16th
Learning Goals	64.2	58th	69.3	75th	37.8	8th
Inclusivity	73.7	70th	59.6	23rd	68.9	47th
Responsiveness	61.5	17th	78.2	84th	62.2	22nd
Experimental Design	50.2	90th	37.5	52nd	46.6	84th
Data Analysis	45.1	53rd	50.6	71st	41.2	43rd
Cognitive Skills	48.7	42nd	63.6	92nd	45.5	34th
Reflection	41.7	77th	67.5	99th	20.4	8th

FIGURE 4. MIST profiles for three instructors across MIST subcategories. (A) Points represent MIST subcategory scores for Instructors A, B, and C based on mean student responses in each course. (B) Full MIST score, MIST subcategory scores, and percentile rankings in the full sample are displayed for each instructor.

objectives, however, is an explicit part of the ST taxonomy. Because responses to the item on exam alignment were very high, we suspect that students may have considered the content of exams to be synonymous with course objectives, limiting the ability of this question to discern between courses with high and low exam alignment. We also found that the items pertaining to inclusivity did not align well with the full scale. While inclusivity represents a central part of ST, these results suggest that the degree to which instructors implement certain inclusive teaching practices is partly decoupled from their broader implementation of other ST practices. Lower variance in responses to inclusivity items may also have contributed to the lower degree of alignment with the larger scale.

MIST Reveals Factors That Influence ST Implementation Levels

While our initial efforts focused on instrument development, our data also provide insights into potential factors correlated with the extent of ST implementation. Among the courses sampled, student responses indicated that courses taught by individuals who had attended an SI workshop had higher overall ST implementation scores than nonattending counterparts (Figure 3A). This finding agrees with previous self-reported data suggesting that attending an SI facilitates instructor adoption of ST practices (Pfund *et al.*, 2009; Aragón *et al.*, 2016). Importantly, the results presented here relied on student observations of instructional practices and, therefore, avoided the potential issue of instructors inflating their self-reported teaching practices (Ebert-May *et al.*, 2011). However, work

TABLE 4. MIST-Short single-factor model item loadings: MIST-Short model (alpha = 0.85)

Item	Item description	Factor loading
Q2	Learning goal maximum frequency	0.355
Q3	Polling method: frequency	0.403
Q4	Polling method: % alignment	0.472
Q6	In-class: frequency	0.584
Q7	In-class: % alignment	0.542
Q17	Group work: in-class frequency	0.577
Q24	Diverse examples and analogies	0.314
Q25	Diverse scientist/researcher contributions	0.301
Q27	Students provide feedback on activities/content	0.474
Q30	Instructor aware of student nonunderstanding	0.464
Q31	Follow-up activities provided if not understood	0.509
Q32	Make hypotheses/predictions	0.710
Q34	Design experiments	0.615
Q37	Analyze/interpret data graphs/tables	0.612
Q38	Use data to make decisions/defend conclusions	0.681
Q46	Higher-level thought processes	0.501
Q47	Open-ended exercises/case studies	0.650
Q48	Reflection: effective study habits	0.519

TABLE 5. Correlations of total scores and subcategories between the MIST-Short and the MIST full version

MIST scale/subcategory title	No. of questions	*r* with full instrument
Overall MIST-Short	18	0.97
Active-Learning Strategies	3	0.95
Learning Goal Use and Feedback	3	0.87
Inclusivity	2	0.98
Responsiveness to Students	2	0.95
Experimental Design and Communication	2	0.89
Data Analysis and Interpretation	2	0.93
Cognitive Skills	2	0.95
Course and Self-Reflection	2	0.96

from student course evaluations suggests that student ratings may also reflect various sources of bias, including course grading policies, required student workload, student skill level, course entertainment value, and instructor demographics (Becker and Watts, 1999; Spooren *et al.*, 2013; Braga *et al.*, 2014). Given that the structure of MIST questions differs from standard student course evaluations, further investigation is needed to understand the extent to which MIST scores are susceptible to student biases. In addition, investigations including pre–post SI surveys and more directed sampling strategies will be needed to determine whether instructors with already high ST implementation levels are more likely to attend an SI or whether the SI itself enables instructors to increase their use of ST.

We found that course level and enrollment were not correlated with MIST overall scores, suggesting that ST can be implemented in varying course environments (Figure 3, B and C). These results agree with several other studies demonstrating that course transformation can be achieved despite practical constraints associated with large courses (Hake, 1998; Crouch and Mazur, 2001; Allen and Tanner, 2005; Knight and Wood, 2005; Freeman *et al.*, 2007; Derting and Ebert-May, 2010; Smith *et al.*, 2014). Furthermore, these findings suggest that MIST does not have an implicit bias toward detecting ST practices in a particular course context.

MIST Enables Investigation of Particular Research Questions

In addition to tracking changes in teaching practices over time or after professional development workshops, MIST can also be used to investigate specific research questions. For example, while many studies have linked active learning to improved course performance and decreased failure rates (Freeman *et al.*, 2014), comparatively fewer studies have investigated whether and how other recommended teaching practices influence student outcomes. The MIST subcategories provide a means to empirically decipher the contributions of a particular factor to a set of student outcomes, such as engagement, conceptual learning, skills development, science identity, and persistence (Graham *et al.*, 2013). Thus, MIST can help support more nuanced studies of teaching practices (Freeman *et al.*, 2014; Wieman, 2014; Hora, 2015).

Recent reports have begun to investigate differences in how instructors, students, and observers document course practices. This issue remains critical to advancement in the education field, because many studies on the efficacy of professional development programs and the impact of teaching practices hinge on having accurate measures of instructional practice (AAAS, 2013; Smith *et al.*, 2013; Wieman and Gilbert, 2014). Some data suggest that instructors may systematically overestimate the adoption of transformed practices in self-report surveys, particularly after professional development programs (Ebert-May *et al.*, 2011). Conversely, other work indicates an association between instructor self-reports and course observations that may be attributed to the low-stakes nature of the instructor survey and questions that target very specific teaching practices (Smith *et al.*, 2014; Wieman and Gilbert, 2014). More recently, researchers compared instructor and student reports of teaching practices in a lab course (Beck and Blumer, 2016). While this study found significant correlations between student and instructor survey responses, these relationships only accounted for a moderate amount of variance. In this case, course observations were not available to determine which veiwpoint agreed more closely with an observation-based perspective. The syntax of MIST items enables them to be interpretable to any course affiliates (i.e., instructors, students, or observers). Thus, MIST will lay a groundwork for future studies to understand differences in these modes of documenting course practices.

Given the limitations of any one mode of course documentation, we chose to use student reports for initial MIST studies for several reasons. First, most national reports focus on the implementation of student-centered instruction, which places student perceptions, behaviors, and learning at the center of instructional design. Accordingly, transformed instructional practices should have detectable effects on student course experiences. Second, we wished to develop a mechanism for documenting course practices that circumvents the possibility that instructors who participate in professional development programs, such as the SI, could inflate their scores. Third, despite potential biases and limitations, student surveys and evaluations of teaching (e.g., the National Survey of Student Engagement) have long been used by instructors and institutions as a common benchmark. Finally, students represent a universally available resource: every course has students who can provide insight on teaching practices, and these students attend class for the entire semester. By comparison, few courses have resources for or access to trained observers, and it becomes increasingly cost-prohibitive to employ multiple observers or too time-consuming to have faculty observers attend more than a few class sessions.

While the student viewpoint represents an accepted and pragmatic way for instructors to document their teaching practices, future studies are needed to understand additional affordances and limitations of relying on the student perspective. For example, MIST items ask students to make judgments with respect to their perception of events that occurred over a full semester time span. While we do not expect each student to report on practices with exact precision, the frequency response scales on most questions were designed to indicate rough approximations of monthly, weekly, or daily frequencies. During validation interviews, students expressed comfort with their ability to identify the appropriate frequency at these levels, but student perceptions of the frequencies varied among students and could have been influenced by a host of variables, including individual student characteristics and the activities implemented in recent class sessions.

Interpreting and Using MIST Results

As with any educational instrument, MIST results must be interpreted and used in a manner consistent with the overall goals of a course or academic program. We developed frequency scales of individual MIST items to capture the full extent of potential variation of practices, ranging from completely absent to very frequent. It is unlikely that every ST practice will be implemented at the highest level in an individual course. Thus, when interpreting MIST scores, instructors should focus on questions that align with their own goals. Subcategory percentile ranks can be helpful for determining how a given course compares with other courses, but this sample does not constitute a representative cross-section of the national course population. Furthermore, while MIST subcategories reflect national educational priorities, further research is needed to understand how scores in these subcategories relate to various student outcomes. Although we predict that higher MIST subcategory scores will correspond with positive student outcomes, we do not propose an "ideal" profile of MIST subcategory scores, in part because the optimal MIST profile may vary for different course levels, institutions, types of students, and instructors. While the full MIST score is beneficial for a variety of research purposes that inform broadscale research questions, such as how the level of ST changes with different course sizes or at different types of institutions, we recommend a focus on MIST subcategories, because they reflect more discrete teaching approaches.

Throughout the development process, we envisioned a wide range of potential uses for MIST. At the individual level, instructors can use MIST results to learn about different teaching practices, decide whether they are satisfied with their perceived implementation of these practices, and recognize pathways for improvement. In the example case, the three instructors showed noticeably different implementation levels of ST and each subcategory (Figure 4). While Instructors A and B had relatively high overall ST implementation, Instructor A may wish to take steps to improve his/her awareness of student thinking by collecting samples of student work and using student responses to modify his/her teaching. Likewise, Instructor B might consider ways to incorporate inclusivity or experimental design and communication practices in the course. Conversely, Instructor C had a lower level of ST implementation but had a clear emphasis on experimental design and communication. This instructor might consider whether the focus in this area could be complemented by additional growth in the subcategories of data analysis and interpretation or cognitive skills. Instructors can further use individual MIST items within each section to gain specific ideas on how to grow in these areas.

Instructors can also use MIST across multiple semesters to track changes in their teaching practices and document growth for promotion and tenure. At a higher level, departments can characterize their teaching practices and compare them with a broader sample as a means to identify areas of strength and ideas for growth. For example, departments may implement active-learning strategies at high levels as a result of focused efforts to transform their undergraduate curricula, but they may identify growth in the areas of cognitive skills and course and self-reflection as a next step for advancement. MIST data can be also be used by leaders of professional development workshops to characterize the teaching practices of incoming participants and tailor their programs accordingly. Furthermore, instructor profiles can be used to form mentoring or peer observation pairs that are mutually beneficial. In these ways, MIST provides many different avenues to advance calls for increased use of transformed teaching practices for individuals, departments, institutions, researchers, and the broader education community.

Availability of MIST

The MIST survey is designed to be administered in an electronic form that can accommodate conditional responses. We have included two qsf files of MIST and MIST-Short for use with Qualtrics, and two Excel files with embedded formulas for calculating overall scores and subcategory scores for each instrument (Supplemental Materials 4, 5, 7, and 8). A one-page front-and-back handout summary of MIST questions, suitable for distribution at workshops or among colleagues, is available in Supplemental Material 6.

ACKNOWLEDGMENTS

This work was supported by a National Science Foundation TUES 3 award (DUE- 1323019) to B.A.C., J.K.K., and others. We are grateful to the team (Oriana Aragón, Meghan Bathgate, Aiyana Bobrownicki, Jane Buckley, Andrew Cavanagh, Xinnian Chen, Jennifer Frederick, Mark Graham, Monica Hargraves, Claire Hebbard, William Trochim, and Jonathan Waterhouse) for critical support and feedback on instrument development, survey logistics, and data analyses. We also thank the institutional site coordinators, instructor participants, student participants, and interviewees. We appreciate the insights given by David Hanauer on instrument design and Jordan Harshman on the article. We thank Tanya Brown, Kati Brazeal, Joey Hubbard, and the UNL Discipline-Based Education Research community for critical discussions.

REFERENCES

Allen, D., & Tanner, K. (2005). Infusing active learning into the large-enrollment biology class: Seven strategies, from the simple to complex. *Cell Biology Education*, *4,* 262–268.

American Association for the Advancement of Science (AAAS). (1990). *The liberal art of science*. Washington, DC.

AAAS. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

AAAS. (2013). *Describing and measuring undergraduate STEM teaching practices*. Washington, DC.

Anders, K., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87,* 215–251.

Aragón, O. R., Dovidio, J. F., & Graham, M. J. (2016). Colorblind and multicultural ideologies are associated with faculty adoption of inclusive teaching practices. *Journal of Diversity in Higher Education*, *10*(3), 201–215.

Baker, L. A., Chakraverty, D., Columbus, L., Feig, A. L., Jenks, W. S., Pilarz, M., … Wesemann, J. L. (2014). Cottrell scholars collaborative new faculty workshop: Professional development for new chemistry faculty and initial assessment of its efficacy. *Journal of Chemical Education*, *91,* 1874–1881.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., … Wu, N. (2009). Learning and scientific reasoning. *Science*, *323,* 586–587.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3,* 77–85.

Beck, C. W., & Blumer, L. S. (2016). Alternative realities: Faculty and student perceptions of instructional practices in laboratory courses. *CBE—Life Sciences Education*, *15,* ar52.

Becker, W. E., & Watts, M. (1999). How departments of economics evaluate teaching. *American Economic Review*, *89,* 344–349.

Bialek, W., & Botstein, D. (2004). Introductory science and mathematics education for 21st-century biologists. *Science*, *303,* 788–790.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41,* 71–88.

Brawner, C. E., Felder, R. M., Allen, R., & Brent, R. (2002). A survey of faculty teaching practices and involvement in faculty development activities. *Journal of Engineering Education*, *91,* 393.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford.

Catell, R. (1966). The scree test for the number of factors. *Multivariate Behaviour Research*, *1,* 245–276.

Chamany, K., Allen, D., & Tanner, K. (2008). Making biology learning relevant to students: Integrating people, history, and context into college biology teaching. *CBE—Life Sciences Education*, *7,* 267–278.

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE—Life Sciences Education*, *9,* 524–535.

Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research and Evaluation*, *10,* 1–9.

Couch, B. A., Brown, T. L., Schelpat, T. J., Graham, M. J., & Knight, J. K. (2015). Scientific teaching: Defining a taxonomy of observable practices. *CBE—Life Sciences Education*, *14,* ar9.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69,* 970–977.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81,* 800–814.

Derting, T. L., & Ebert-May, D. (2010). Learner-centered inquiry in undergraduate biology: Positive relationships with long-term student achievement. *CBE—Life Sciences Education*, *9,* 462–472.

Ebert-May, D., Derting, T. L., Henkel, T. P., Maher, J. M., Momsen, J. L., Arnold, B., & Passmore, H. A. (2015). Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE—Life Sciences Education*, *14,* ar22.

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

Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE—Life Sciences Education*, *14,* ar23.

Ertmer, P. A., & Newby, T. J. (1996). The expert learner: Strategic, self-regulated, and reflective. *Instructional Science*, *24,* 1–24.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4,* 272–299.

Field, A. (2014). *Discovering statistics using SPSS*. London: Sage.

Frederick, P. J. (1987). Student involvement: Active learning in large classes. *New Directions for Teaching and Learning*. *1987,* 45–56.

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,* 8410–8415.

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education*, *6,* 132–139.

Goldey, E. S., Abercrombie, C. L., Ivy, T. M., Kusher, D. I., Moeller, J. F., Rayner, D. A., … Spivey, N. W. (2012). Biological inquiry: A new course and assessment plan in response to the call to transform undergraduate biology. *CBE—Life Sciences Education*, *11,* 353–363.

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

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19,* 149–161.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66,* 64–74.

Hanauer, D. I., & Dolan, E. L. (2014). The Project Ownership Survey: Measuring differences in scientific inquiry experiences. *CBE—Life Sciences Education*, *13,* 149–158.

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., & Hatfull, G. F. (2006). Teaching scientific inquiry. *Science*, *314,* 1880–1881.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., … Wood, W. B. (2004). Scientific teaching. *Science*, *304,* 521–522.

Handelsman, J., Miller, S., & Pfund, C. (2007). *Scientific teaching*. New York: Freeman.

Harshman, J., & Stains, M. (2017). A review and evaluation of the internal structure and consistency of the Approaches to Teaching Inventory. *International Journal of Science Education*, *39,* 918–936.

Henderson, C. (2008). Promoting instructional change in new faculty: An evaluation of the physics and astronomy new faculty workshop. *American Journal of Physics*, *76,* 179–187.

Hora, M. T. (2013). A Review of Classroom Observation Techniques in Postsecondary Settings (WCER Working Paper No. 2013-01), Madison: Wisconsin Center for Education Research School of Education, University of Wisconsin–Madison.

Hora, M. T. (2015). Toward a descriptive science of teaching: How the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms. *Science Education*, *99,* 783–818.

Hora, M. T., Oleson, A., & Ferrare, J. J. (2013). Teaching Dimensions Observation Protocol (TDOP) User's Manual. Madison: Wisconsin Center for Education Research School of Education, University of Wisconsin–Madison.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30,* 179–185.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6,* 1–55.

Hughes, S. O., Anderson, C. B., Power, T. G., Micheli, N., Jaramillo, S., & Nicklas, T. A. (2006). Measuring feeding in low-income African-American and Hispanic parents. *Appetite*, *46,* 215–223.

Kaiser, H. F. (1960). An index of factorial simplicity. *Psychometrika*, *39,* 31–36.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika*, *35,* 401–415.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34,* 111–117.

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4,* 298–310.

Labov, J. B., & Huddleston, N. F. (2008). Integrating policy and decision making into undergraduate science education. *CBE—Life Sciences Education*, *7,* 347–352.

Labov, J. B., Reid, A. H., & Yamamoto, K. R. (2010). Integrated biology and undergraduate science education: A new biology education for the twenty-first century? *CBE—Life Sciences Education*, *9,* 10–16.

Manduca, C. A., Mogk, D. W., Tewksbury, B., Macdonald, R. H., Fox, S. P., Iverson, E. R., … Bruckner, M. (2010). On the cutting edge: Teaching help for geoscience faculty. *Science*, *327,* 1095–1096.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52,* 77–95.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3,* 97–110.

National Research Council (NRC). (1999). *Transforming undergraduate education in science, mathematics, engineering, and technology*. Washington, DC: National Academies Press.

NRC. (2003a). *BIO2010: Transforming undergraduate education for future research biologists*. Washington, DC: National Academies Press.

NRC. (2003b). *Evaluating and improving undergraduate teaching in science, technology, engineering, and mathematics*. Washington, DC: National Academies Press.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32,* 396–402.

Pallant, J. (2010). *SPSS survival manual: A step by step guide to data analysis using SPSS*. Berkshire, England: Open University Press/McGraw-Hill.

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,* 470–471.

Pierret, C., & Friedrichsen, P. (2009). Stem cells and society: An undergraduate course exploring the intersections among science, religion, and law. *CBE—Life Sciences Education*, *8,* 79–87.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory and Practice*, *41,* 219–225.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93,* 223–231.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers.. *CBE—Life Sciences Education*, *15,* rm1.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48,* 1–36.

Sadler, T. D., Chambers, F. W., & Zeidler, D. L. (2004). Student conceptualizations of the nature of science in response to a socioscientific issue. *International Journal of Science Education*, *26,* 387–409.

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.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36,* 111–139.

Seymour, E. (2000). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview.

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,* 624–635.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83,* 598–642.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist*, *52,* 613–629.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn and Bacon.

Tanner, K., & Allen, D. (2007). Cultural competence in the college biology classroom. *CBE—Life Sciences Education*, *6,* 251–258.

Tanner, K. D. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, *11,* 113–120.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Tra, Y. V., & Evans, I. M. (2010). Enhancing interdisciplinary mathematics and biology education: A microarray data analysis course bridging these disciplines. *CBE—Life Sciences Education*, *9,* 217–226.

Uhlmann, E., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16,* 474–480.

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

Wei, C. A., & Woodin, T. (2011). Undergraduate research experiences in biology: Alternatives to the apprenticeship model. *CBE—Life Sciences Education*, *10,* 123–131.

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

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE—Life Sciences Education*, *13,* 552–569.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Williams, C. T., Walter, E. M., Henderson, C., & Beach, A. L. (2015). Describing undergraduate STEM teaching practices: A comparison of instructor self-report instruments. *International Journal of STEM Education*, *2,* 18.

Woolley, S. L., Benjamin, W.-J. J., & Woolley, A. W. (2004). Construct validity of a self-report measure of teacher beliefs related to constructivist and traditional approaches to teaching and learning. *Educational and Psychological Measurement*, *64,* 319–331.

Zeidler, D. L., Sadler, T. D., Simmons, M. L., & Howes, E. V. (2005). Beyond STS: A research-based framework for socioscientific issues education. *Science Education*, *89,* 357–377.