Guides to Advance Teaching Evaluation (GATEs): A Resource for STEM Departments Planning Robust and Equitable Evaluation Practices

Sandhya Krishnan,[†] Jessica Gehrtz,[‡] Paula P. Lemons,[§] Erin L. Dolan,[§] Peggy Brickman,[‖] and Tessa C. Andrews[¶*]

[†]Department of Mathematics, Science, and Social Studies Education; [§]Department of Biochemistry and Molecular Biology; [‖]Department of Plant Biology, and [¶]Department of Genetics, University of Georgia, Athens, GA 30602; [‡]Department of Mathematics, University of Texas at San Antonio, San Antonio, TX 78249

ABSTRACT

Most science, technology, engineering, and mathematics (STEM) departments inadequately evaluate teaching, which means they are not equipped to recognize or reward effective teaching. As part of a project at one institution, we observed that departmental chairs needed help recognizing the decisions they would need to make to improve teaching evaluation practices. To meet this need, we developed the Guides to Advance Teaching Evaluation (GATEs), using an iterative development process. The GATEs are designed to be a planning tool that outlines concrete goals to guide reform in teaching evaluation practices in STEM departments at research-intensive institutions. The GATEs are grounded in the available scholarly literature and guided by existing reform efforts and have been vetted with STEM departmental chairs. The GATEs steer departments to draw on three voices to evaluate teaching: trained peers, students, and the instructor. This research-based resource includes three components for each voice: 1) a list of departmental target practices to serve as goals; 2) a characterization of common starting places to prompt reflection; and 3) ideas for getting started. We provide anecdotal examples of potential uses of the GATEs for reform efforts in STEM departments and as a research tool to document departmental practices at different time points.

INTRODUCTION

Slow uptake of evidence-based teaching by college science, technology, engineering, and mathematics (STEM) faculty has brought increased attention to the systems in which faculty work. In particular, widespread and effective implementation of evidence-based teaching may depend on the systems in place to recognize and reward effective teaching and instructors' efforts to continuously improve (e.g., Dennin *et al.*, 2017; Stains *et al.*, 2018; Laursen *et al.*, 2019). To incentivize evidence-based teaching, we must have systems capable of robustly evaluating teaching quality. Yet many higher education institutions and their departments lack such systems (e.g., Berk, 2005; Dennin *et al.*, 2017).

Recent work undertaken by prominent national organizations underscores the desire for better teaching evaluation systems in STEM higher education. The Association of American Universities has emerged as a leader in teaching evaluation reform, repeatedly gathering stakeholders and providing financial support to member institutions to shift culture and practices surrounding teaching evaluation (e.g., Bradforth *et al.*, 2015; Dennin *et al.*, 2017). The National Academies of Science, Engineering, and Medicine has convened university administrators, change agents, and researchers to share and learn about teaching evaluation reform (NASEM, 2020). Researchers and

Cynthia Bauerle, *Monitoring Editor*

Submitted Aug 6, 2021; Revised Mar 18, 2022; Accepted Apr 26, 2022

CBE Life Sci Educ September 1, 2022 21:ar42

DOI:10.1187/cbe.21-08-0198

Author contributions: S.K. and T.C.A. conceived and designed the experiments; S.K., P.P.L., E.L.D., P.B., and T.C.A. performed the experiments; S.K., J.G., and T.C.A. analyzed the data; S.K. and T.C.A. drafted the article; S.K. and T.C.A. prepared the digital images.

*Address correspondence to: Tessa C. Andrews (tandrews@uga.edu).

change agents have developed, tested, and promoted better models of teaching evaluation (e.g., Andrews *et al.*, 2020; Finkelstein *et al.*, 2020; Weaver *et al.*, 2020; Simonson *et al.*, 2021; TEval, 2019).

These efforts have coalesced around the principle that teaching evaluation should rely on multiple perspectives, including the perspectives of students, peers, and the instructor (Finkelstein *et al.*, 2020; Weaver *et al.*, 2020). The Teaching Quality Framework, a project that is part of TEval, refers to these as three "voices" that contribute evidence of teaching quality (Andrews *et al.*, 2020; TEval, 2019). Relying on three voices for teaching evaluation recognizes that these different perspectives can illuminate specific aspects of teaching (Reinholz *et al.*, 2019). Additionally, because any form of evidence is subject to bias, relying on multiple sources of evidence is more robust and equitable. Students are best positioned to provide information about what occurs regularly in class and the accessibility of the instructor. For example, students are uniquely able to comment on the climate created in a course because they experience it over an entire semester and do so as full participants rather than outside observers. Trained peer observers are well positioned to evaluate the alignment of course content and skills with the discipline and to gauge the effectiveness of teaching strategies in promoting equitable learning opportunities (Thomas *et al.*, 2014). Additionally, college faculty value constructive feedback from their peers and report being more likely to implement changes based on peer feedback than student feedback (Brickman *et al.*, 2016). The instructor's own voice is also essential. Instructors alone can speak to their goals, intentions, and efforts to learn and improve. Instructors also have the most comprehensive view of their courses, students, and disciplines, as well as the changes they have made in their teaching over time, allowing them to contextualize evidence regarding teaching effectiveness.

Currently, most STEM departments lack consistent teaching evaluation practices that draw on multiple voices and therefore are not equipped to recognize nor reward effective teaching (Dennin *et al.*, 2017; NASEM, 2020). Importantly, inconsistent and ad hoc teaching evaluation practices can result in inequities among faculty as they are being reviewed for promotions, tenure, and salary raises. Therefore, departmental teaching evaluation practices not only need to produce robust evidence, they must also aim to treat faculty equitably. With these challenges in mind and as part of one institutional transformation effort, we developed the Guides to Advance Teaching Evaluation (GATEs; Appendix A in the Supplemental Material). Our primary objective was to develop a resource that could help steer departments toward robust and equitable teaching evaluation practices. Our secondary objective was to develop a tool that researchers could use to systematically document teaching evaluation practices in STEM departments. This article has several complementary goals, and these do not align neatly with the conventional research paper format. Therefore, hereafter we use informative headings and granular subheadings to guide readers. This article:

1. articulates the need in our local context that led us to develop the GATEs (see *Local Context and Need*);
2. describes the iterative process of developing the GATEs and vetting it with departmental chairs (see *Iterative Development Process*);
3. presents the GATEs, including the reasoning and evidence behind each component (see *GATEs: Description and Evidence*); and
4. provides anecdotal examples of potential uses of the GATEs for departmental change and for research (see *Examples of Potential GATEs Uses*).

LOCAL CONTEXT AND NEED

Inspired by the ongoing reform efforts described in the Introduction, we aimed to help local STEM departments reconsider and reform teaching evaluation practices so that they could better recognize and reward high-quality teaching and ultimately improve student outcomes. We undertook this work through a National Science Foundation (NSF)-funded project at the University of Georgia called DeLTA,[1] which pursues transformative change in undergraduate STEM education. This project convened 12 STEM departmental chairs who gathered for facilitated meetings four to six times per year to reconsider and reform their departmental practices. We designed these meetings to provide departmental chairs with opportunities to reflect on current departmental practices, recognize their own underlying assumptions about teaching evaluation, critically consider alternative practices, and set goals for action (Andrews *et al.*, 2021). As a result of leading these meetings, the project leadership gained valuable insights about the current practices and thinking of departmental chairs. We leveraged these insights to develop and refine resources to meet the needs of STEM departments.

At the start of our transformation project, collaborating departments engaged minimally in teaching evaluation, as is common in the United States (Dennin *et al.*, 2017; NASEM, 2020). Based on many interactions with departmental chairs, we knew that departments relied primarily on mandatory end-of-course student surveys to evaluate teaching. We learned that departmental chairs were almost all dissatisfied with their current teaching evaluations, expressing concern that student evaluation results did not provide useful evidence of teaching effectiveness. At the same time, our work with departmental chairs in meetings revealed that most could not articulate concrete ways to improve teaching evaluation and were not taking action to reform their departments' practices. Changing departmental practices requires a lot of decision making, because practices encompass what occurs, how and when it occurs, and who completes this work. The departmental chairs in our local project did not have time to become teaching evaluation experts or to discover the various departmental practices they might need to develop.

Based on these observations, we concluded that these STEM departmental chairs needed help to recognize the various decisions they would need to make to improve teaching evaluation and also the chance to consider examples and criteria for making those decisions. We wanted to help departmental chairs capitalize on what had already been discovered about teaching evaluation, including resources and processes developed in other initiatives and institutions. Finally, we wanted to help

[1] DeLTA = Departmental and Leadership Teams for Action. To avoid overwhelming readers with acronyms, we refer to DeLTA as the "project" or "transformation project" hereafter. We include the formal name so that readers can draw connections across publications about this project, such as Andrews *et al.* (2021).

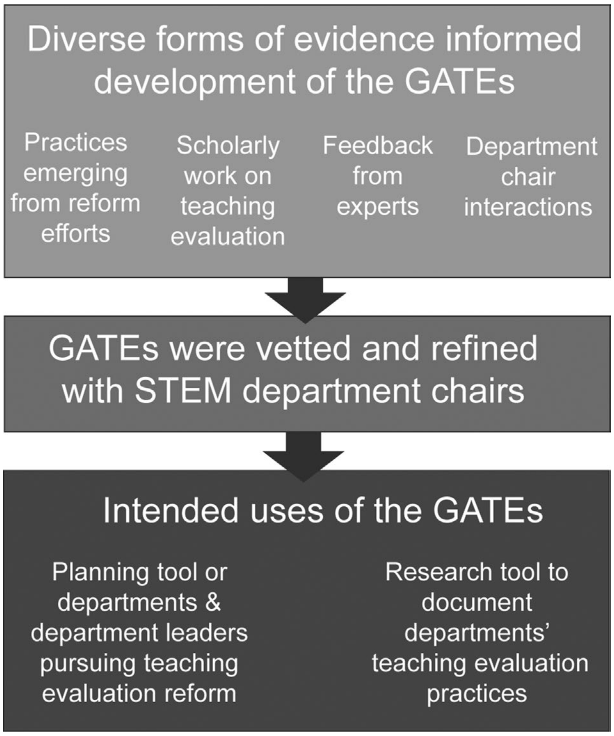

Diverse forms of evidence informed development of the GATEs
Practices emerging from reform efforts
Scholarly work on teaching evaluation
Feedback from experts
Department chair interactions
GATEs were vetted and refined with STEM department chairs
Intended uses of the GATEs
Planning tool or departments & department leaders pursuing teaching evaluation reform
Research tool to document departments' teaching evaluation practices

FIGURE 1. **Overview of the development and potential uses of the GATEs. We developed the GATEs based on diverse forms of evidence and vetted the GATEs with departmental chairs to generate a research-based resource with two potential purposes.**

departmental chairs see teaching evaluation reform as a surmountable challenge.

ITERATIVE DEVELOPMENT PROCESS

To aid readers, we first briefly describe the final product of the iterative development process: the Guides to Advance Teaching Evaluation (GATEs). There is one GATE each for the three voices that contribute to evaluating teaching: trained peers, students, and self. Each GATE has three components: 1) a list of research-based target practices to serve as goals for departments; 2) a characterization of common starting places departments may be when they begin considering teaching evaluation reform; and 3) ideas for getting started, enacting multiple target practices at once, and learning more. We discuss the development of each component of the GATEs in the following subsections; the overall development process and intended uses are summarized in Figure 1. The iterative development and vetting process occurred across 2 years. All research was determined to be exempt by the University of Georgia Institutional Review Board under protocol ID no. STUDY00006754.

Development of Target Practices

The central component of the GATE for each voice is a list of target practices that can serve as long-term goals for departments. We used several approaches to identify target practices that can contribute to robust and equitable teaching evaluation using each voice. Robust evaluation produces trustworthy and useful evidence of effectiveness, and equitable evaluation ensures that faculty are treated fairly and that steps are taken to mitigate biases.

Departmental teaching evaluation practices have rarely been the subject of scholarly inquiry, so we had to consider diverse forms of evidence that a target practice was important, rather than just peer-reviewed literature. We describe the various sources of evidence here, with some examples. We particularly valued examples of practices that were emerging from teaching evaluation reform efforts in STEM departments. We assumed that a teaching evaluation practice that had been pursued repeatedly had proven practically important for departments, so we looked for practices that were similar across multiple departments or institutions. For example, reform efforts at University of Oregon, University of Southern California, and the University of Colorado–Boulder recommend that faculty involved in peer observations take part in some form of training (Appendix B in the Supplemental Material). For some target practices, we could rely on a body of scholarly work, such as with practices related to acknowledging and accounting for potential bias in mandatory student evaluations (Appendix B in the Supplemental Material). We also drew on the core commitments of our transformation project to inform target practices. The project's core commitments, which are goals that many departments and projects hold (e.g., Corbo *et al.*, 2016), include basing education decisions on evidence, fostering continuous teaching improvement, and promoting inclusion and diversity (Andrews *et al.*, 2021). Grounded by the project's commitment to fostering continuous teaching improvement and similar emphases in other teaching evaluation reform efforts, several target practices guide departments to compare evidence of teaching effectiveness at multiple time points. Longitudinal data are necessary to recognize and reward teaching improvements.

In addition to synthesizing these sources to develop target practices, we relied on the perspectives of experts, including researchers studying teaching evaluation and change agents working to shift teaching evaluation practices. We shared the target practices in one-on-one or small-group meetings and invited expert feedback on the relevance and necessity of each practice, as well as the clarity of their organization and description. Specific feedback was solicited through questions like: "We would like to make the practices described in each level more realistic. What stands out to you as unrealistic for a STEM department?" We gathered feedback at multiple time points in the development process, which helped us hone the set of target practices for each voice. As one example, experts provided critical feedback about the need for an organizing framework for the target practices and suggested organizing characteristics, which were then refined further through vetting with departmental chairs.

The product of synthesizing diverse forms of evidence was the target practices in the GATEs, and an organizing structure for the target practices that is used for all three voices. Appendix B in the Supplemental Material describes the synthesized supporting evidence and rationale for each practice for each voice.

Identification of Starting Places

Most STEM departments in our project did not originally have any target practices in place. We wanted the resource that we developed to help departmental chairs quickly see the ways in which their current practices did not align with the target practices. We also wanted to normalize the reality that most STEM departments at research-intensive institutions currently fall far short of robust teaching evaluation practices (e.g., Dennin *et al.*, 2017). Therefore, in addition to the target practices, we also developed a description of common starting places where departments may find themselves when they begin the process of teaching evaluation reform.

We developed the characterizations of starting places using data about the practices in our collaborating departments. We collected these data through one-on-one interviews in Summer 2019 with departmental chairs and other department members. We anticipated that there could be differences between how teaching evaluation was intended to occur and how it actually occurred, so we wanted to interview at least one faculty member in addition to the departmental chair. Aiming to recruit faculty who were likely to be familiar with recent teaching evaluation in each department, we focused on faculty who had experienced being reviewed for tenure and/or promotion and who were reported by their colleagues to be influential regarding undergraduate education (as indicated on an anonymous survey). In total, we interviewed 12 departmental chairs and 13 other faculty from across the departments involved in our project. These semistructured interviews lasted about 60 minutes, and the interview protocol asked questions about departmental policies and practices related to teaching evaluation. The questions that provided data about teaching evaluation practices are included in Appendix C in the Supplemental Material.

We systematically analyzed interview transcripts to determine the current practices in each department. Two researchers (S.K. and J.G.) independently read the interview transcripts and identified and documented specific teaching evaluation practices described in each of the interviews. The researchers then met to compare their analyses and dealt with any differences in findings by returning to the interviews and confirming the presence, absence, or specific details of mentioned practices. Thus, the researchers reached consensus about specific actions that local departments took around teaching evaluation and organized these practices by the three voices. Finally, they organized the teaching evaluation practices for each voice into three starting places that represented the variation among collaborating departments. Departments' starting places ranged from not using a specific voice for teaching evaluation to practices that reflected some deliberate action to improve practices.

Refinement and Further Development of the GATEs

At this point in the development process, the pilot GATEs consisted of two components for each voice: a list of target practices and a characterization of common starting places. We examined how intended users interpreted and responded to this pilot version of the GATEs. Specifically, we sought to understand whether departmental chairs understood the target practices as we intended and whether they recognized their own departments' practices in the characterization of starting places. We were also interested in emotional responses, because we worried that a negative emotional response would prevent the GATEs from supporting change. Therefore, we aimed to minimize negative emotional reactions to the GATEs when it was possible to do so without compromising content.

We first gathered evidence of responses to the GATEs through observation of departmental chairs working in groups to review the pilot version. Twelve departmental chairs participated in a meeting in which they reviewed the target practices for one voice and placed their department within a starting place for that voice. This provided initial evidence about how the GATEs were interpreted by departmental chairs. We include a brief description of this meeting and observations about how departmental chairs responded in *Examples of Potential GATEs Uses.*

We also conducted one-on-one think-aloud interviews with six departmental chairs. We asked interviewees to read through the GATE for one voice and to share aloud everything they were thinking. We interviewed a broad range of departmental chairs, including two who had no prior experience with the GATEs and four who had worked with the project for more than a year and had previously interacted with the GATEs. We selected the voice for each interviewee to ensure that we had two interviews for each voice and that each participant was seeing the chosen voice for the first time. We also asked participants how they would use the GATEs in their departments.

These opportunities to vet the GATEs with departmental chairs resulted in multiple revisions to make the GATEs more user-friendly. We made changes to the wording used in the target practices and the starting places. As an example, we changed the name of one of the starting places from "consistency lacking" to "closer to cohesion," because departmental chairs misread "consistency" as "consistently" and commented that the focus on deficiency through the use of "lacking" came across as judgmental. Another change to wording that resulted from evidence of the responses of departmental chairs was replacing the word "trustworthy" with "reliable" to describe one of the three organizing characteristics of the target practices. We initially used the term "trustworthy" as a lay description for practices that give confidence that the evidence collected can be trusted to accurately represent someone's teaching. However, in our investigation of departmental chairs' engagement with the resource, some STEM departmental chairs prickled at the term "trustworthy" and responded more positively to the word "reliable." Given that a key goal of the GATEs is to serve as a resource to STEM faculty and departments, we opted to use a word that was both understandable and less likely to prompt a negative emotional reaction.

We also altered the formatting based on evidence of how departmental chairs interacted with the GATEs. For example, the component of the GATEs that characterizes starting places was originally formatted as a table. However, this led departmental chairs to interpret the starting places as stages to pass through on the way to target practices. We revised the formatting to clarify that departments at any of the starting places can move directly to target practices. As another example, we added the self-assessment formatting of the target practices following observations that users wanted to treat the target practice list as a checklist. Departmental chairs wanted to use a self-assessment themselves and envisioned providing it to faculty as part of conversations about reform. We deliberately kept this to one page, because users anticipated printing it to physically mark their progress and sharing it at in-person meetings.

Finally, we augmented the GATEs with one additional component. We observed that some departmental chairs felt overwhelmed by the number of distinct target practices for each voice and floundered as they considered where to start. In contrast, other chairs noticed starting places that seemed feasible as well as ways to work toward multiple target practices at once. Departmental chairs also quickly recognized that, although the GATEs provided a big-picture and long-term plan, they needed more guidance. Therefore, the last component of the GATEs provides: 1) suggestions about target practices that can be fruitful initial achievements, 2) groups of target practices that could be efficiently accomplished together (i.e., "bundles"), and 3) direct links to a Google document with additional reading and tools.

GATES: DESCRIPTION AND EVIDENCE

The goal of this section of the paper is to describe the GATEs that resulted from the iterative development process, as well as highlighting evidence supporting this research-based tool. The GATEs include a one-page overview and three components for each voice (see Appendix A in the Supplemental Material): 1) a list of target practices, organized as a departmental self-assessment; 2) characterizations of common starting places, titled "Where Is Your Department Starting?"; 3) and ideas and resources to support movement toward the target practices, titled "Starting Strong and Engaging Efficiently."

Target Practices

The GATEs aims to provide departments with clearly articulated, long-term goals for robust and equitable teaching evaluation practices. We refer to these as "target practices" to emphasize that they describe departmental-level decisions and actions (i.e., practices) that are likely to be aspirational for many departments (i.e., targets). The target practices address the breadth and specifics of the decisions, standards, and expectations that are important for robust and equitable teaching evaluation for each voice.

Through the iterative development process, three characteristics emerged as a useful organizing framework for target practices, and we discuss specific target practices using this framework. Robust and equitable teaching evaluation is 1) structured, 2) reliable, and 3) longitudinal. Target practices that lend structure to teaching evaluation help to minimize bias, create more consistency, and thereby result in more equitable evaluation experiences across faculty, much like structure fosters equity in other contexts in higher education (e.g., Haak *et al.*, 2011; Eddy and Hogan, 2014; O'Meara *et al.*, 2019; Laursen and Austin, 2020). Evaluation that is reliable is informed by multiple sources of evidence, making it less subject to bias and more trustworthy. Evaluation that is longitudinal is able to document improvement over time and provide feedback to faculty about strengths and room for improvement.

The target practices are organized as a self-assessment (Tables 1–3) that invites users to record their departments' current status for each practice as "fully in place," "working on it," "want to work on it," and "not right now." These options acknowledge a few realities that we observed in interactions with departmental chairs. First, a department may engage in a year or more of activity that is fruitful but falls short of having practices fully in place (i.e., "working on it"). We included "want to work on it," because departments may aspire to a target practice, but may not yet have taken any action. Finally, we worded the lowest level of commitment (i.e., "not right now") to leave room for making progress in the future.

Depending on their goals, readers will appreciate different levels of detail about the diverse forms of evidence supporting the target practices. Readers who are primarily interested in the target practices and the GATEs as a tool should focus their attention on Tables 1–3, potentially only skimming this section. Readers who want to know more about the underlying rationale and research may appreciate the detail in this section of the paper, which summarizes supporting evidence, including relevant research literature. For the sake of brevity, we primarily describe peer voice target practices. Change agents may want even more detail and resources. Appendix B in the Supplemental Material describes the rationale and supporting evidence for every target practice for all three voices. As more scholarship about teaching evaluation and more reform efforts are undertaken, we will learn more about which teaching evaluation practices are most important. Therefore, we present the GATEs as a valuable resource to guide reform now and in the future and also as a living resource that can be updated as our collective knowledge grows.

The next three subsections describe target practices and supporting evidence and are ordered to follow the GATEs (Tables 1–3): structured, reliable, and longitudinal.

Structured. Evaluation that is structured involves formalized processes, expectations, training, and support for faculty. As Table 1 shows, structured use of peer voice includes eight target practices. One target practice calls for a formal observation form to influence what is observed and which other data are collected. Reform efforts across multiple institutions have mandated or supported the development of peer observation forms, because they help to standardize what observers pay attention to and externalize a department's expectations for effective teaching (Appendix B in the Supplemental Material). Transparent expectations help create equity among faculty, because everyone has access to the same information about what is expected of them. Importantly, target practices do not dictate the particular standards that a department should adopt, because each discipline, institution, and department has unique needs and contingencies to consider. Rather, the target practices outline key decisions that departments will need to make to bring structure to collecting and analyzing evidence of teaching effectiveness from peer voice. Working toward structured target practices requires moving away from teaching evaluation that is inconsistent across faculty or guided primarily by historical precedent.

Structured teaching evaluation requires the development, refinement, and maintenance of standards and expectations. To support this work, the target practices for achieving structure address the need for human resources, collective decision making, and training for faculty. For example, peer voice target practice 5 recognizes that one or more faculty will need to organize peer observation (Table 1). The departments that have made the most progress reforming their teaching evaluation practices in our project have appointed committees or identified faculty to lead the development and implementation of new practices.

This service work is not an insignificant time commitment, and guided by scholarly literature about inequities in

TABLE 1. Peer voice target practices, organized by three characteristics of robust and equitable teaching evaluation, and formatted as a self-assessment

Peer voice target practices: What is your status and what actions will you take?			Not right now	Want to work on it	Working on it	Fully in place
Structured	1	Department uses a formal observation form to guide what is observed and which other data are collected (e.g., class materials, assessments, pre-observation meeting). Forms may be adopted or adapted from other departments.	□	□	□	□
	2	Department has a formal template for writing a report based on peer review, potentially distinguishing between formative and summative review.	□	□	□	□
	3	Department uses formal processes or criteria to select peer observer(s) for all instructors.	□	□	□	□
	4	Department enacts policy about the number of peer observations and observers during a review period and/or across review periods.	□	□	□	□
	5	Department designates a coordinator, leader, or committee to carry out and refine peer observation practices.	□	□	□	□
	6	Department has a process for allocating and recognizing workload related to coordinating and conducting observations.	□	□	□	□
	7	Department periodically discusses and improves peer evaluation practices to maximize utility to instructors and the department.	□	□	□	□
	8	Department provides or arranges formal training about the departmental peer review process for peer observers.	□	□	□	□
Reliable	9	Department relies on multiple observations for all instructors, such as using multiple observers, observing multiple lessons, and/or observing multiple courses.	□	□	□	□
	10	Department specifies which class materials (e.g., syllabi, exams, homework, slides, handouts) are collected and evaluated as part of peer observation.	□	□	□	□
	11	Department expects observers to talk with instructors to properly contextualize observations and review of materials. This might include discussing course goals, lesson goals, class structure, and students.	□	□	□	□
Longitudinal	12	Department conducts peer observation over multiple time points in a review period for all instructors to document teaching improvements.	□	□	□	□
	13	Department ensures that the peer observation process provides feedback to instructors via follow-up discussion that covers strengths and areas for improvement.	□	□	□	□

faculty work, peer voice target practice 6 calls for recognizing faculty work associated with organizing and providing peer observations (Table 1). Recognizing these service contributions is a crucial equity concern, because there is growing recognition of inequities in faculty work by gender and race (e.g., Baez, 2000; Griffin and Reddick, 2011; Guarino and Borden, 2017; O'Meara *et al.*, 2017a, b; Misra *et al.*, 2021). Though service work related to teaching evaluation has not been specifically investigated to determine whether this workload is distributed equitably, investigations of other teaching and service work provide cautionary tales. Thus, departments pursuing peer evaluation should explicitly recognize this service, for example, by accounting for it as departmental service akin to other departmental committee work.

Developing formal processes and securing buy-in from faculty likely necessitates some degree of discussion in departments, which is recognized in peer voice target practice 7 (Table 1). Departments will approach this in different ways. Some departments develop new processes and policies through vigorous discussion among the entire faculty, whereas other departments pursue change by strategically building support for new policies among committees or informal subsets of faculty. Target practices for each voice highlight the importance of discussion among faculty and also recognize that faculty and departments will determine the best way to approach consensus building in their local contexts.

STEM faculty within our institution were largely unfamiliar with collecting, analyzing, and using evidence to evaluate their own and others' teaching. Thus, it makes sense that research institutions that have reformed peer observation often require or recommend training for peer observers (see Appendix B in the Supplemental Material). Peer voice target practice 8 focuses on departments arranging for or providing training. Training for peer observers supports the structure that departments build through other target practices. For example, a standard peer observation form is likely to be interpreted and used differently by observers. Training can help peer observers come to consensus about what is important to observe, thereby resulting in more consistent and fairer peer observation across faculty.

Though training for peer observation was advocated by other reform efforts (Appendix B in the Supplemental Material), we did not find examples of institutions or departments training faculty about appropriately using student voice or engaging in systematic teaching self-reflection. Yet we observed just as much need for support among faculty in these areas within our collaborating departments. Faculty who are not experienced with teaching evaluation deserve training and

TABLE 2. Student voice target practices, organized by three characteristics of robust and equitable teaching evaluation, and formatted as a self-assessment

Student voice target practices: What is your status and what actions will you take?			Not right now	Want to work on it	Working on it	Fully in place
Structured	1	Department has formal standards for how and when instructors collect, analyze, and report student data (e.g., response rate expectation, standard quantitative and qualitative analysis).	□	□	□	□
	2	Department makes appropriate distinctions in their expectations about student data for different review periods (e.g., annual review, third-year review, promotions) and different levels of teaching experience with a given course.	□	□	□	□
	3	Department periodically discusses and improves expectations for collecting and analyzing data from students to maximize utility to instructors and the department.	□	□	□	□
	4	Department provides or arranges formal training, or other support, for instructors about collecting and analyzing student data, including achieving high response rates, analyzing quantitative and qualitative data systematically and appropriately, gathering data beyond mandatory evaluations, and making comparisons across time.	□	□	□	□
Reliable	5	Department expects instructors to do everything they can to achieve high response rates on mandatory student evaluations (e.g., course credit offered, class time set aside).	□	□	□	□
	6	Department recognizes known biases, such as bias against women, minoritized groups, and large class size, and limits comparisons of mandatory student evaluations between instructors.	□	□	□	□
	7	Department specifies that quantitative questions on mandatory student evaluations be analyzed as distributions of scores, rather than averages. Because quantitative questions often use an ordinal rating scale (excellent, very good, good, poor), average scores and standard deviations are inappropriate. We cannot assume the points on ordinal scales are equidistant.	□	□	□	□
	8	Department specifies which set of quantitative student evaluation questions are used for each review period (e.g., annual, promotion).	□	□	□	□
	9	Department specifies that student comments on mandatory evaluations be systematically examined to determine teaching strengths and room for improvement.	□	□	□	□
	10	Department expects instructors to collect, analyze, and interpret some data beyond mandatory student evaluations.	□	□	□	□
Longitudinal	11	Department expects instructors to document change (or consistently exemplary results) by comparing data from students across multiple time points.	□	□	□	□

support so that they can meaningfully participate, and thus the GATEs include a target practice related to training for each voice (Tables 1–3).

Reliable. Target practices that address reliability help departments trust the conclusions drawn from evidence of teaching effectiveness. These practices involve drawing on multiple sources of evidence, considering potential sources of bias, and relying on intentional and appropriate analysis of collected evidence. As a reminder, we used the term “reliable” rather than alternatives because it elicited more favorable responses from STEM departmental chairs. We use “reliable” as a lay term that means “you can rely on this evidence.” As Table 1 shows, reliable use of peer voice includes three target practices.

These target practices focus on broadening the information used for peer review, because inferences drawn from multiple sources of data are likely to be more informative and trustworthy. Target practice 9 calls for using multiple observations of a class, rather than just one, which is recommended across institutions (Appendix B in the Supplemental Material). Target practice 10 describes the review of class materials, rather than just classroom observations. This is important, because students’ learning experiences extend well beyond the classroom, and thus examining class materials provides a more robust view of a course. For example, exams and projects influence students’ grades and even their approaches to learning (Stanger-Hall, 2012), but class periods focused on these are generally purposely avoided for peer observation because they differ from typical instruction. Additionally, class materials like the syllabus are important tools for equitably communicating key information to students and can set the tone for a welcoming class climate (e.g., Gin *et al.*, 2021). Finally, target practice 11 recognizes that peer observation, even repeated over a few lessons, offers a limited vantage point. It calls for discussions between peer observers and the instructor to place observations within the context of the course goals, course structure, and student body (Table 1). Without such conversations, peer observers may not be able to appreciate how instructors’ decisions reflect

TABLE 3. Self voice target practices, organized by three characteristics of robust and equitable teaching evaluation, and formatted as a self-assessment

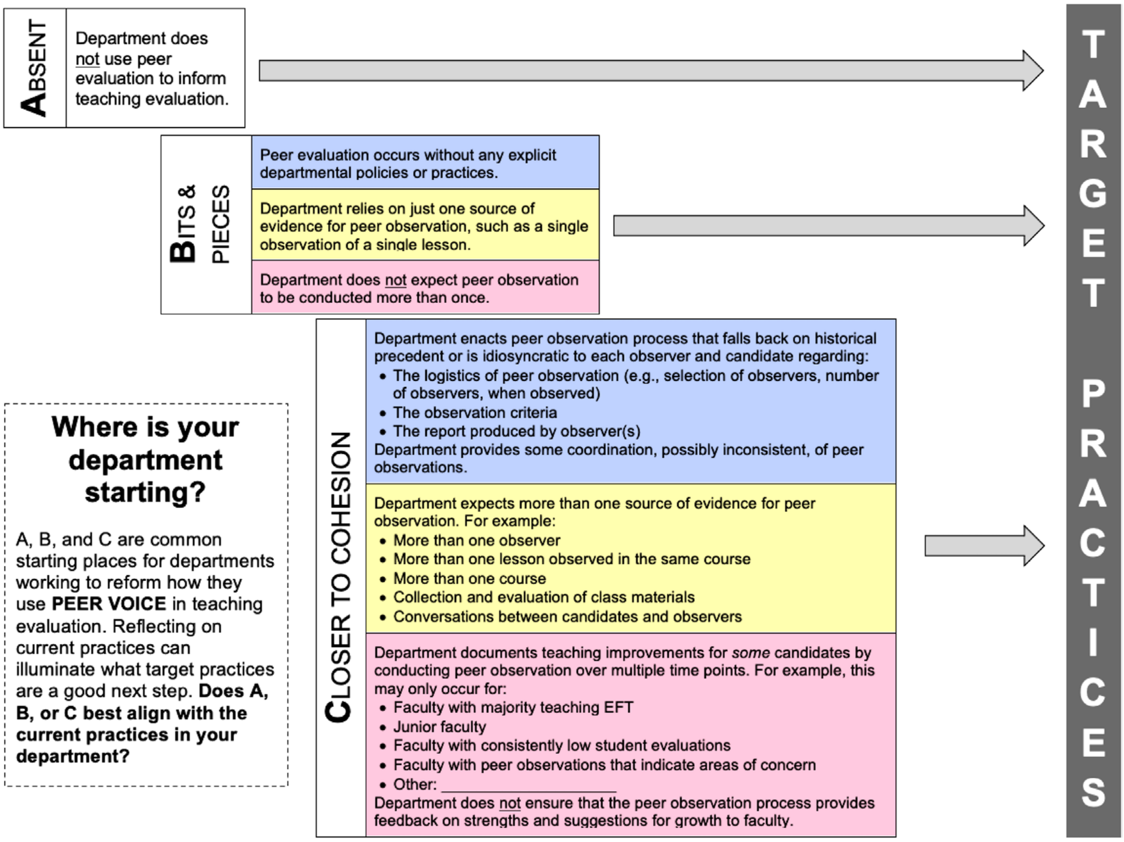
Self voice target practices: What is your status and what actions will you take?			Not right now	Want to work on it	Working on it	Fully in place
Structured	1	Department uses a formal self-reflection form to guide the scope and content of written self-reflection narratives, including standards for what constitutes evidence-based self-reflection. Forms may be adopted or adapted from other departments.	□	□	□	□
	2	Department periodically discusses and improves standards for written teaching reflections to maximize utility to instructors and the department.	□	□	□	□
	3	Department provides or arranges formal training or other support for instructors concerning the self-reflection process and to help instructors meet departmental expectations for documenting self-reflection.	□	□	□	□
Reliable	4	Department expects instructors to engage in a self-reflection process and provide written documentation thereof that is focused on tackling teaching challenges (e.g., concerns raised in student evaluations or peer observation, student learning difficulties, lack of engagement).	□	□	□	□
	5	Department expects the self-reflection process and written documentation thereof to rely on the systematic analysis of evidence about student learning and experiences.	□	□	□	□
	6	Departmental expectations for self-reflection consider the experience level of instructors. For example, instructors new to a course or teaching may primarily rely on informal sources of data (e.g., notes, brief written feedback from students), whereas more experienced instructors rely on formal sources of data (e.g., assessment data) and systematic observation (e.g., feedback from trained peers).	□	□	□	□
Longitudinal	7	Department expects that written reflections discuss how instructors have built on prior self-reflections, including the outcomes of planned improvements and innovations.	□	□	□	□
	8	Department expects that written reflections discuss efforts to grow and learn as educators. This can include learning from both successes and failures.	□	□	□	□

attention to their particular students, learning objectives, and course. Multiple institutions recommend these target practices as part of peer review of teaching (Appendix B in the Supplemental Material).

We also highlight a few student voice target practices that deal with considering the potential for bias and appropriately analyzing data. Biases in the data collected can make teaching evaluation both unfair and uninformative. Biases have been most thoroughly documented in student evaluation, and student voice target practice 6 specifies that departments recognize known biases and act accordingly, including limiting comparisons of these data between instructors (Table 2). Comparisons among instructors often will not be trustworthy, because the ratings can depend on irrelevant instructor characteristics. A variety of factors unrelated to effective teaching can be associated with scores on mandatory evaluations, including instructor's gender (e.g., Boring, 2017; Fan *et al.*, 2019; Adams *et al.*, 2022), instructor's race and ethnicity (e.g., Anderson and Smith, 2005; Smith and Hawkins, 2011), instructor's native language (Fan *et al.*, 2019), and class size (Bedard and Kuhn, 2008).

Another potential source of bias is introduced for student evaluations when response rates are low. Student voice target practice 5 stipulates that departments set expectations for high response rates for these surveys (Table 2). There are several practical solutions to achieving high response rates, such as allowing time in class to complete evaluations or offering a small incentive to students for completing the evaluation (e.g., Berk, 2012; see details in the external resources linked in Starting Strong and Engaging Efficiently with Student Voice in Appendix A in the Supplemental Material). Departments can encourage faculty to use these simple strategies to increase response rates by setting an expectation for the outcome (e.g., response rate of at least 85%) or the process (e.g., offer nominal extra credit if 85% of the class completes the survey).

Reliable teaching evaluation also requires appropriate and intentional analysis of the data collected. Student voice target practice 7 calls for analyzing quantitative results of mandatory student course evaluations as distributions rather than means (Table 2). This is important, because these questions often use an ordinal rating scale (excellent, very good, good, poor), and thus it cannot be assumed that the points on the scales are interpreted as equidistant (Bishop and Herron, 2015). For example, students may interpret the distance between good and very good as small compared with the distance between very good and excellent. If the points on the scale are not interpreted as equidistant by respondents, then means and standard deviations are not meaningful ways to summarize these data. Similarly, student comments from mandatory evaluations must be analyzed reliably. A common practice in our collaborating departments was selecting a sample of student comments for promotion and tenure dossiers that most positively portrayed the instructor and course. Sometimes referred to as "cherry-picking," this practice makes the written data from students an entirely unreliable source of information. Departments using student voice target practice 9 expect that faculty undertake a systematic approach to analyzing student comments for



ABSENT

Department does not use peer evaluation to inform teaching evaluation.

BITS & PIECES

Peer evaluation occurs without any explicit departmental policies or practices.

Department relies on just one source of evidence for peer observation, such as a single observation of a single lesson.

Department does not expect peer observation to be conducted more than once.

CLOSER TO COHESION

Department enacts peer observation process that falls back on historical precedent or is idiosyncratic to each observer and candidate regarding:

- The logistics of peer observation (e.g., selection of observers, number of observers, when observed)
- The observation criteria
- The report produced by observer(s)

Department provides some coordination, possibly inconsistent, of peer observations.

Department expects more than one source of evidence for peer observation. For example:

- More than one observer
- More than one lesson observed in the same course
- More than one course
- Collection and evaluation of class materials
- Conversations between candidates and observers

Department documents teaching improvements for *some* candidates by conducting peer observation over multiple time points. For example, this may only occur for:

- Faculty with majority teaching EFT
- Junior faculty
- Faculty with consistently low student evaluations
- Faculty with peer observations that indicate areas of concern
- Other: ____________________

Department does not ensure that the peer observation process provides feedback on strengths and suggestions for growth to faculty.

TARGET PRACTICES

Where is your department starting?

A, B, and C are common starting places for departments working to reform how they use **PEER VOICE** in teaching evaluation. Reflecting on current practices can illuminate what target practices are a good next step. **Does A, B, or C best align with the current practices in your department?**

FIGURE 2. The Where Is Your Starting Place? component for each voice can help departments recognize their current practices and the ways in which those practices fall short of the target practices. Departments at each starting place can proceed directly to developing target practices. Shading is the same as the target practices: blue = structured, yellow = reliable, pink = longitudinal.

evidence of teaching strengths and areas for improvement (Table 2).

Longitudinal. Evaluation that is longitudinal is able to document change over time and provide feedback to faculty about teaching strengths and areas for improvement. These target practices allow departments to value instructors' efforts to continuously improve, rather than just valuing teaching achievements. This is realistic for a few reasons. First, not all faculty will have had equivalent opportunities to develop their teaching skills and expertise. Second, not all faculty will aim to be exceptional college teachers. Third, it is likely that not all faculty in a department will be equally effective. Yet all faculty can improve their teaching. Thus, teaching evaluation that recognizes efforts toward continuous improvement is likely to be more equitable and more effective at supporting the diversity of goals and skills among faculty in a department. Accordingly, multiple research institutions have created policies and recommendations for teaching evaluation that value improvement over time (see Appendix B in the Supplemental Material).

As Table 1 shows, longitudinal use of peer voice includes two target practices focused on documenting and supporting continuous improvement. Peer voice target practice 12 involves conducting peer observation at multiple time points in a review period with the goal of documenting teaching improvements over time (Table 1). Repeated evaluation of teaching, using peer, student, or self voice, offers both accountability for faculty to pursue improvements and the opportunity for teaching improvements to be recognized and rewarded. Target practice 13 calls for the peer observation process to provide actionable feedback for instructors about both strengths and areas for improvement (Table 1). Our collaborating departmental chairs desired better approaches for providing instructors with feedback to inform teaching improvements, and feedback from peers can foster more reflective teaching and learning about teaching among faculty (e.g., Dillon *et al.*, 2020). Additionally, faculty desire constructive feedback from respected peers because they expect such feedback to help them improve (e.g., Brickman *et al.*, 2016).

Where Is Your Department Starting?

In addition to the list of target practices, the GATEs include a component titled "Where Is Your Department Starting?" that describes three common starting places for departments for each voice: 1) Absent, 2) Bits & Pieces, and 3) Closer to Cohesion (Figure 2). These three categorizations emerged from our interview data with departments at the start of our project. "Absent" recognizes that a department may not currently use a particular voice to evaluate teaching. "Bits & Pieces" applies when a department uses a voice to evaluate teaching but has few (or no) formalized processes or expectations. For example, a department might conduct peer evaluation without any

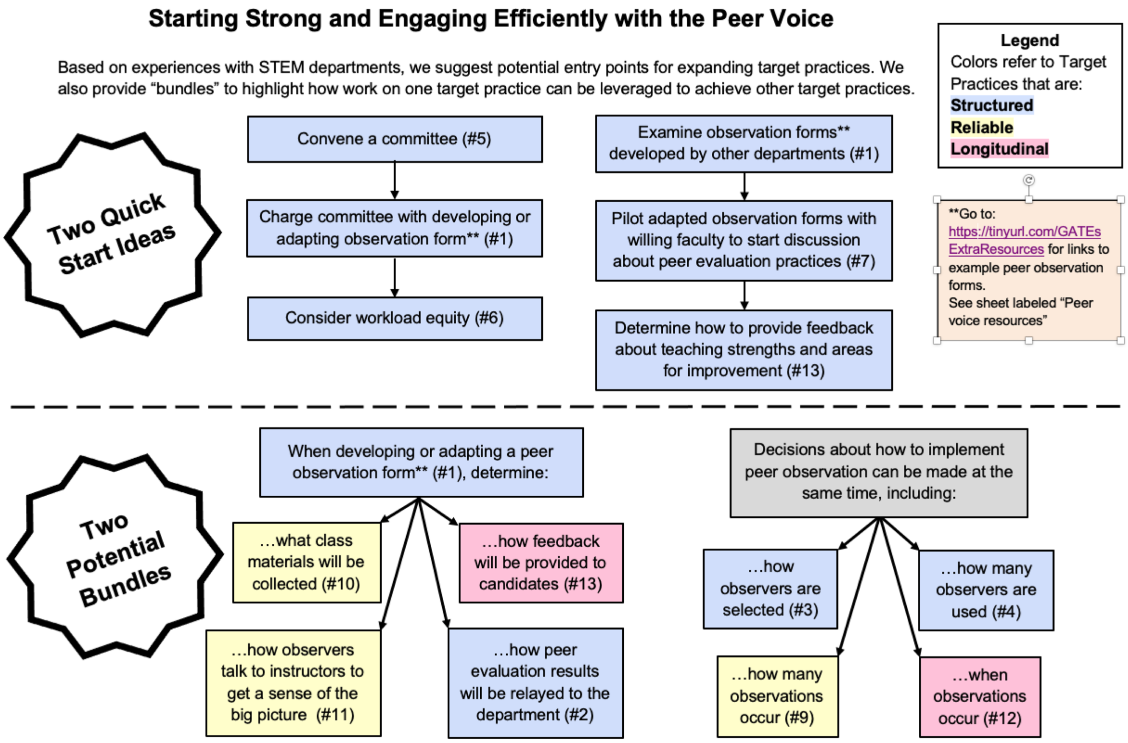

Starting Strong and Engaging Efficiently with the Peer Voice
Based on experiences with STEM departments, we suggest potential entry points for expanding target practices. We also provide "bundles" to highlight how work on one target practice can be leveraged to achieve other target practices.
Legend
Colors refer to Target Practices that are:
Structured
Reliable
Longitudinal
Two Quick Start Ideas
Convene a committee (#5)
Charge committee with developing or adapting observation form** (#1)
Consider workload equity (#6)
Examine observation forms** developed by other departments (#1)
Pilot adapted observation forms with willing faculty to start discussion about peer evaluation practices (#7)
Determine how to provide feedback about teaching strengths and areas for improvement (#13)
**Go to: https://tinyurl.com/GATEsExtraResources for links to example peer observation forms.
See sheet labeled "Peer voice resources"
Two Potential Bundles
When developing or adapting a peer observation form** (#1), determine:
…what class materials will be collected (#10)
…how feedback will be provided to candidates (#13)
…how observers talk to instructors to get a sense of the big picture (#11)
…how peer evaluation results will be relayed to the department (#2)
Decisions about how to implement peer observation can be made at the same time, including:
…how observers are selected (#3)
…how many observers are used (#4)
…how many observations occur (#9)
…when observations occur (#12)

FIGURE 3. The Starting Strong and Engaging Efficiently component for each voice offers more direction for those who want it. This component has three parts: ideas about which target practices to tackle first, bundles of target practices that can be efficiently developed together, and links to outside resources directly related to the target practices. The outside resources are stored in a Google sheet that can be accessed by anyone with the link. Shading is the same as for the target practices: blue = structured, yellow = reliable, pink = longitudinal.

explicit policies or practices. This is likely to result in inconsistent, and therefore inequitable, peer observation across faculty. "Closer to Cohesion" describes departments that have established some specific departmental practices for using a voice in teaching evaluation. This starting place is important to recognize, because the practices described here likely represent deliberate efforts by a department or leader to improve teaching evaluation. Yet these practices fall short of robust and equitable teaching evaluation in important ways. For example, the evidence produced by the practices described in Closer to Cohesion for peer voice may vary considerably across observers and faculty, making it both less trustworthy and inequitable (Figure 2). As for the target practices, this component of the GATEs makes distinctions between structured, reliable, and longitudinal.

The main purpose of the Where Is Your Department Starting? component is to help departments recognize their starting places and the ways in which their current practices fall short of the target practices. In our experience, that is achieved relatively quickly. When we used this component in a meeting with departmental chairs, we aimed for these leaders to read the descriptions of starting places for one voice, recognize their existing practices and a need for change, and move on to carefully consider the target practices within about 5 minutes.

Starting Strong and Engaging Efficiently

The final component of the GATE for each voice aims to help users envision how they could immediately and efficiently work toward target practices. As described in *Refinement and Further Development of the GATEs*, we observed that some departmental chairs felt overwhelmed when they read the list of target practices and reflected on the fact that none were currently in place in their own departments. One department leader explained that the "activation energy" needed to get started felt too high. Others wanted advice about how to get started in their own departments. Therefore, we developed this component for departmental chairs who wanted more direction. This offers three things: ideas about which target practices to tackle first, bundles of target practices that can be efficiently developed together, and links to outside resources directly related to the target practices.

Using peer voice as an example (Figure 3), this component suggests two "quick start ideas." One way that departments could get started building peer evaluation practices is by convening a committee (target practice 5). A committee can then take responsibility for developing or adapting a peer observation form (target practice 1). As noted earlier, departments will need to explicitly recognize the workload of committee members as they develop and deploy new peer observation practices, so this is highlighted as an important early step (target practice 6).

Another starting place that appealed to some departments was considering existing peer observation forms used in other STEM departments. A productive next step could be piloting an adapted observation form with a subset of willing faculty. Our collaborating departmental chairs could often name faculty who would see immediate value in participating in peer observation, such as new faculty eager for teaching feedback and faculty looking toward promotion who wanted peer observation feedback as part of their dossiers. Departmental chairs anticipated engaging these faculty in a pilot enactment of peer observation and then creating opportunities for these faculty to share their experiences with others, generating conversations that conveyed the benefits that faculty had experienced as a result of peer observation (target practice 7). This approach would also lead a department to prioritize the development of teaching evaluation practices that provide constructive feedback about strengths and areas for improvement (target practice 13).

This component of the GATEs also suggests bundles of practices that could be efficiently accomplished together. We expect this approach to appeal to taxed departmental chairs and maybe especially to certain professional identities who prioritize maximizing efficiencies, such as engineers. For peer voice, the first bundle includes five target practices and essentially encompasses the decisions that departments will need to make about the breadth of information that observers will rely on and how observers will communicate their evaluations (Figure 3). The second suggested bundle similarly groups a set of decisions that departments can make at one time, in this case about the logistics of implementing peer observation. These are examples of how departments could achieve multiple related target practices through one coherent action.

The last offering of this component of the GATEs is a link to curated resources directly related to the target practices highlighted on the page (Figure 3). For peer voice, this includes links to peer observation forms and related resources at nine research-intensive institutions.

EXAMPLES OF POTENTIAL GATES USES

In this section, we provide anecdotal examples of how we have used the GATEs to help the reader imagine possible uses. We propose that the GATEs can be useful in two contexts: 1) as a planning tool that provides concrete goals to guide the reform of departmental teaching evaluation practices in research-intensive institutions and 2) as a research tool to document departmental practices at different time points.

Example of Using the GATEs as a Resource for Departmental Change

We describe how our project used the GATEs in a facilitated meeting of STEM departmental chairs, and a few observations we made about how chairs interacted with the GATEs. We facilitated a meeting in which departmental chairs considered two components of the GATEs: the target practices and Where Is Your Starting Place? Our meeting goals were for departmental chairs to 1) recognize how their departmental practices aligned (or not) with target practices for at least one voice, 2) recognize the types of practices their departments may need that they currently lack, and 3) identify one or more target practices that they wanted to pursue in their departments.

The meeting consisted of a short presentation, discussions in breakout rooms, and goal setting. The lead facilitator (P.P.L.) reminded departmental chairs of the three-voice framework and explained that the meeting would focus on considering target practices aligned with these voices. She emphasized that the GATEs were based on practices from other research-intensive institutions and national reform efforts, as well as what was occurring within local STEM departments. P.P.L. oriented chairs to the GATEs, explaining that structured, reliable, and longitudinal are key characteristics of robust evaluation.

Next, we assigned departmental chairs to a breakout room to discuss the GATE for one voice, based on interests they had expressed in prior meetings. Once in breakout rooms in groups comprising two to four departmental chairs and a facilitator, chairs read through the target practices and starting places. Facilitators prompted chairs to reflect on which practices jumped out at them and why. After several minutes of reading and reflection, facilitators prompted chairs to place their departments in a starting place for structured, reliable, and longitudinal and to note whether their departments had achieved any target practices. After discussing their assessments of their current practices, the facilitators prompted departmental chairs to consider what target practices they would work on during the next year, with attention to progress already made, practices that would resonate with members of their departments, and the availability of human resources.

Based on this experience of using the GATEs, we offer a few early insights that may be useful to change agents. First, the GATEs helped make facilitated conversations about teaching evaluation concrete and productive. Departmental chairs were able to quickly make sense of target practices and to consider how their departments' approaches to teaching evaluation differed from the target practices. Chairs with varying levels of knowledge about teaching evaluation could engage in discussions about specific departmental practices that they may never have considered previously. Reflecting on the target practices also helped departmental chairs decide or confirm what voice and target practices they wanted to prioritize for further action and consideration.

Second, departmental chairs' judgments of their current practices, including both starting places and target practice status, aligned with the research team's judgments, suggesting that departmental chairs may be able to accurately self-assess departmental teaching evaluation practices. Chairs recognized and felt comfortable sharing their starting places, even when they placed themselves in the Absent or Bits & Pieces category. The self-assessment process also prompted departmental chairs to consider which target practices seemed more and less palatable their colleagues and to honestly recognize progress yet to be made. Recognizing a need for change is often a key component of motivation or readiness to change (Armenakis and Harris, 2002; Rogers, 2010; Andrews and Lemons, 2015), and engaging with the GATEs may help departments recognize ways in which their current practices need improvement.

Third, departmental chairs envisioned different ways to use the GATEs in their departments, suggesting this resource can serve different purposes. A few described using the GATEs to set goals for themselves as leaders and to think about which departmental colleagues could help them work toward those goals. Departmental chairs saw the GATEs as a conversation starter and resource within a department. One departmental chair described how the GATEs could work in conjunction with related tools (e.g., peer observation forms) to serve as a comprehensive resource and to help convince faculty of the need to change teaching evaluation practices. Another departmental chair planned to use the GATEs to form a "charge" for a committee and to help the committee develop a long-term "map" of the change needed.

TABLE 4. Number of departments with teaching practices aligned with three different "starting places" for 12 STEM departments at the beginning of the project, divided by voice and characteristic of evaluation[a]

	Absent	Bits & Pieces	Closer to Cohesion
Peer voice			
Structured	1	4	6
Reliable	1	5	6
Longitudinal	1	7	4
Student voice			
Structured	0	7	5
Reliable	0	12	0
Longitudinal	0	12	0
Self voice			
Structured	8	4	0
Reliable	8	4	0
Longitudinal	8	4	0

[a]The criteria for each starting place for each voice are found in Appendix A in the Supplemental Material, and peer voice is also provided in Figure 2.

Example of Using the GATEs in Research

In addition to serving as a source of long-term goals for robust and equitable teaching evaluation practices, parts of the GATEs may be useful to researchers. We conducted modest pilot tests using different sources of data to characterize a department's current teaching evaluation practices. We used recordings of meetings of the departmental chairs, departmental chair goal-setting notes, and one-on-one interviews. We found that the most robust data for this assessment came from one-on-one interviews with two or more faculty who were directly involved with teaching evaluation in the department, including the departmental chair. In our study, these data came from interviews that directly asked about departmental teaching evaluation practices (see Appendix C in the Supplemental Material). The interview asked direct questions about both peer evaluation and student evaluation practices, and the data regarding self-reflection practices came from more general questions about how teaching was evaluated annually and for promotion and tenure.

We determined the interrater reliability for categorizing departmental teaching evaluation practices using the GATEs. We calculated interrater reliability using a weighted Cohen's kappa. This calculation accounts for the ordered nature of categories, weighting disagreements that are further apart more than disagreements that are closer. We characterized departments' teaching evaluation practices at the start of the project, when not many departments had target practices in place. Therefore, we rated departments' practices as best aligned with one of four categories: Absent, Bits & Pieces, Closer to Cohesion, or some target practices in place. We treated these categories as ordinal. We made these judgments for each voice and for the three characteristics of robust and equitable teaching evaluation (e.g., structured, reliable, longitudinal). Therefore, raters made nine judgments for each department. One rater (S.K.) was very familiar with these departments' practices, because she had conducted the interviews and attended all project meetings with departmental chairs. The other rater was new to the project, reading the interview transcripts for the first time. These two raters achieved high interrater reliability (weighted Cohen's kappa = 0.925; Fleiss and Cohen, 1973) and discussed all disagreements to reach consensus.

Across 12 departments, only one department had any target practices in place at the start of our institutional transformation project. The two raters agreed that this department exhibited two peer voice target practices (2 and 7), which are both related to structure. Table 4 shows the number of departments at each starting place, by voice and characteristic of teaching evaluation, at the start of their involvement in the project. On average, our local STEM departments had more advanced starting places for peer voice than for student voice, and most commonly lacked practices for using the instructor's own perspective (i.e., self voice) for teaching evaluation. Bits & Pieces as a starting place was most common across voices, meaning that departments used that voice in teaching evaluation, but lacked any standards or formalized processes. Overall, these data emphasize the considerable dearth of robust and equitable teaching practices among STEM departments at one research institution before intervention.

DISCUSSION

This paper describes the development and vetting of a novel resource to support STEM departments in building robust and equitable teaching evaluation practices. Given the documented problems with student course evaluations (e.g., Bedard and Kuhn, 2008; Smith and Hawkins, 2011; Boring, 2017; Fan *et al.*, 2019) and widespread dissatisfaction with these data among faculty (Brickman *et al.*, 2016), departments need to advance beyond sole reliance on student evaluations. The GATEs can help departments leverage the distinct and important perspectives of trained peers, students, and the instructors themselves. We drew on the best available evidence to develop the GATEs, but the scholarly literature about departmental teaching evaluation practices is sparse. We encourage users to view the GATEs as a useful resource for right now and also as a resource subject to change as we learn more from teaching evaluation scholarship and reform efforts over time.

The GATEs were designed to strike a balance between being prescriptive and flexible, recognizing that departments will need to develop practices suited to their context while staying true to principles of robust and equitable teaching evaluation. Formalizing expectations by writing them down in forms and policies, and then consistently using them, is crucial to ensuring that evaluation is equitable across faculty. Teaching evaluation that is optional or unstructured may communicate to faculty that the intellectual work of teaching and their efforts to continuously improve are not valuable or measurable. Therefore, the GATEs call on departments to formalize expectations for teaching, which is prescriptive. The GATEs are also fundamentally flexible because each department determines the expectations they have for teaching and continuous teaching improvement. Research expectations for promotion, an analogous reality familiar to STEM departments, are both prescriptive and flexible. It is common for departments to expect faculty to publish their work and garner funding, but the exact number of publications or external funding amounts are not stipulated to allow for differences among research areas and faculty. The GATEs direct departments to create standards for teaching evaluation but do not specify the content of those standards to allow for differences across departmental contexts.

We observed that departmental chairs differed in the level of prescriptiveness they preferred, and thus they responded to the GATEs differently. For example, peer voice target practice 1 and self voice target practice 1 relate to the use of standard forms for peer evaluation and written self-reflection, respectively (Tables 1 and 3). These target practices intentionally do not specify the content of these forms, because departments will need to discuss the subject of observation and reflection. Some departmental chairs objected to having any standard forms, due to concerns that faculty would resist anything prescriptive and that a departmental form could curtail faculty freedom of expression. On the other hand, some departmental chairs considered the target practices insufficiently prescriptive. They worried that having to develop forms to suit their departments was too burdensome for faculty, and they desired examples (e.g., peer observation forms, rubrics to assess self-reflections, etc.) that could be used as provided or tweaked to suit their departments. Luckily, departments do not have to start this work from scratch. They can rely on extensive prior research on effective teaching and work done by multiple groups to define effective teaching and build tools to evaluate teaching effectiveness (e.g., Simonson *et al.*, 2021; Weaver *et al.*, 2020). As described earlier, we added the Starting Strong and Engaging Efficiently component to meet the needs of users who desired more specific guidance and examples they could adapt to their settings. Change agents should anticipate that some colleagues may object to creating standards for teaching evaluation.

Limitations

Though the GATEs fill an important gap, this resource does not address every shift that departments may need to make to incentivize effective teaching. Most critically, the GATEs do not specify how departments should use judgments about teaching effectiveness to inform high-stakes decisions about merit raises, promotion, or tenure (Dennin *et al.*, 2017). Establishing how judgments of teaching effectiveness will be used is a necessary step in achieving the ultimate goal of improving students' experiences in undergraduate STEM classrooms. Robust and equitable teaching evaluation practices may have little effect on faculty and students if the results of these evaluations are not seriously considered in decisions about salaries, appointments, and promotions. Our institutional transformation project is guided by the philosophy that it is not fair to faculty to immediately consider teaching effectiveness in high-stakes decisions if it has largely been overlooked in the past, nor is it fair to ask faculty to invest time in robust evaluation practices that have no actual consequences. Therefore, departments should consider developing a plan for how to transition to a system in which robust evidence of teaching effectiveness meaningfully informs decisions.

Another limitation of the GATEs is that it was developed to meet the needs of STEM departments in one institutional change project. This is limited in both scope and time. We relied on both expert feedback and evidence of teaching evaluation practices emerging from other reform efforts, which broadens the relevance of the GATEs well beyond one institution. Nonetheless, extrainstitutional, institutional, departmental, and cultural factors may make some target practices ill-suited to some contexts. For example, faculty unions or institutional policies may dictate some teaching evaluation practices, such as the number of peer observations allowed in a given time period. Therefore, a departmental practice would need to align with external requirements. Additionally, this work does not establish the utility of the GATEs in other institution types and non-STEM disciplines. Furthermore, this work does not allow us to draw conclusions about the long-term impacts of the GATEs on departmental teaching evaluation practices.

Researchers studying departments outside their own institutions or departments with which they have not interacted with previously will likely need to interview more faculty to gather sufficiently detailed and contextualized data about current teaching evaluation practices. We had access to detailed information about departmental practices and often insider knowledge of such practices, which allowed us to make reliable judgements about target practices for collaborating departments. We collected data to determine the status of target practices using one-on-one interviews with multiple faculty from each department. The project team also includes members of multiple collaborating departments, providing additional insider knowledge. We have worked with the collaborating departments for more than two years, providing multiple opportunities to confirm the details of teaching evaluation practices, or lack thereof. We also did not thoroughly test other methods of data collection, such as surveys or focus groups. However, we have concerns about data collection methods that would not allow for follow-up questioning because informants may not have thought much about teaching evaluation in the past, and thus may need repeated prompting to provide sufficiently detailed information.

Second, the data that we analyzed about current departmental teaching evaluation practices came from departments with few or no target practices in place. Therefore, we often were limited to categorizing starting places rather than a department's status for each target practice. In a context wherein departments had been working to adopt target practices, a research team will likely need to clearly define the distinction between a target practice being "fully in place" versus "working on it" in order to reliably judge target practice status. We encourage researchers to disseminate the distinctions that they make so that others can benefit from this work.

Key Areas for Future Research

Future work should investigate how the GATEs, other resources, and specific interventions influence departmental teaching evaluation practices, and how those teaching evaluation practices influence instructional practices. The current research literature is insufficient to know what is necessary to support meaningful teaching evaluation reform in STEM departments. Each effort toward teaching evaluation reform is essentially a case study, and it is only by looking for patterns across cases that we can grow our collective knowledge. There is also a complete lack of research about how departmental teaching evaluation practices ultimately influence instructional practices of individuals, and whether this differs for faculty at different career stages and in different position types.

We have endeavored to gather the best evidence currently available about what other research-intensive institutions have found productive and feasible as they have pursued teaching evaluation reform, but the existing evidence base is limited. As reform efforts expand across more departments and institutions, researchers can study which target practices are most essential to shifting how teaching is perceived, recognized, and rewarded. This may involve studying departments, departmental leadership, and promotion and tenure discussions and decisions, as well as how faculty perceive and respond to teaching evaluation practices.

It will also be important to study which target practices promote continuous teaching improvement among faculty. Expectations for ongoing, evidence-informed teaching self-reflection, including self voice target practices (Table 3), could foster continuous improvement. Additionally, the development and implementation of a peer-review process, especially one that includes faculty discussions and training (Table 1), may result in faculty expecting that teaching will be seriously and rigorously considered by their colleagues for promotion and tenure decisions. Future work may be able to investigate the influence of specific target practices on faculty perceptions of departmental climate and expectations.

Future research should also consider what supports the sustainability of robust and equitable teaching evaluation. The guide for each voice includes a target practice related to periodically discussing and improving evaluations practices for that voice. Each voice also includes a target practice related to training faculty, which will help build capacity for and expertise about teaching evaluation in the department. Yet the role of ongoing discussions and training in maintaining robust and equitable teaching evaluation practices has not been investigated. Sustaining teaching evaluation practices is essential to shifting the culture of departments and thereby impacting students, but the current research literature has little to offer in this area.

ACKNOWLEDGMENTS

We are very grateful to the DeLTA team and our collaborating departmental chairs for their commitment to improving undergraduate education and support for faculty. We are also grateful to many colleagues for their guidance, feedback, and contributions to this work, including Sarah Andrews, Susan Shadle, Emily Miller, Hannah Ericson, and the Biology Education Research Group at the University of Georgia. The ideas for this work emerged from the Scientists Engaged in Education Research (SEER) center, which is part of the Owens Institute for Behavioral Research. Support for this work was provided by the NSF's Improving Undergraduate STEM (IUSE) program under award 1821023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We thank three anonymous reviewers and the monitoring editor for generous feedback that helped us streamline and improve this article.

REFERENCES

Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L. J., Slavich, E., & Waters, D. (2022). Gender bias in student evaluations of teaching: "Punish [ing] those who fail to do their gender right." *Higher Education*, *83*(4), 787–807.

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184–201. https://doi.org/10.1177/0739986304273707

Andrews, S. E., Keating, J., Corbo, J. C., Gammon, M., Reinholz, D. L., & Finkelstein, N. (2020). Transforming teaching evaluation in disciplines: A model and case study of departmental change. In White K., Beach A., Finkelstein N., Henderson C., Simkins S., Slakey L., Stains M., Weaver G., & Whitehead L. (Eds.), *Transforming institutions: Accelerating systemic change in higher education*. Pressbooks. Retrieved May 27, 2022, from http://openbooks.library.umass.edu/ascnti2020/

Andrews, T. C., Brickman, P., Dolan, E. L., & Lemons, P. P. (2021). Every tool in the toolbox: Pursuing multilevel institutional change in the DeLTA Project. *Change: The Magazine of Higher Learning*, *53*(2), 25–32. https://doi.org/10.1080/00091383.2021.1883974

Andrews, T. C., & Lemons, P. P. (2015). It's personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. *CBE—Life Sciences Education*, *14*(1), ar7. https://doi.org/10.1187/cbe.14-05-0084

Armenakis, A. A., & Harris, S. G. (2002). Crafting a change message to create transformational readiness. *Journal of Organizational Change Management*, *15*(2), 169–183. https://doi.org/10.1108/09534810210423080

Baez, B. (2000). Race-related service and faculty of color: Conceptualizing critical agency in academe. *Higher Education*, *39*(3), 363–391.

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*(3), 253–265. https://doi.org/10.1016/j.econedurev.2006.08.007

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48–62.

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching & Learning*, *8*(2), 98–107.

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International Journal of Exercise Science*, *8*(3), 297–302.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., … & Smith, T. L. (2015). University learning: Improve undergraduate science education. *Nature News*, *523*(7560), 282. https://doi.org/10.1038/523282a

Brickman, P., Gormally, C., & Martella, A. M. (2016). Making the grade: Using instructional feedback and evaluation to inspire evidence-based teaching. *CBE—Life Sciences Education*, *15*(4), ar75. https://doi.org/10.1187/cbe.15-12-0249

Corbo, J. C., Reinholz, D. L., Dancy, M. H., Deetz, S., & Finkelstein, N. (2016). Framework for transforming departmental culture to support educational innovation. *Physical Review Physics Education Research*, *12*(1), 010113. https://doi.org/10.1103/PhysRevPhysEducRes.12.010113

Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., … & Miller, E. R. (2017). Aligning practice to policies: Changing the culture to recognize and reward teaching at research universities. *CBE—Life Sciences Education*, *16*(4), es5. https://doi.org/10.1187/cbe.17-02-0032

Dillon, H., James, C., Prestholdt, T., Peterson, V., Salomone, S., & Anctil, E. (2020). Development of a formative peer observation protocol for STEM faculty reflection. *Assessment & Evaluation in Higher Education*, *45*(3), 387–400. https://doi.org/10.1080/02602938.2019.1645091

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*(3), 453–468. https://doi.org/10.1187/cbe.14-03-0050

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE*, *14*(2), e0209749. https://doi.org/10.1371/journal.pone.0209749

Finkelstein, N., Greenhoot, A. F., Weaver, G., & Austin, A. E. (2020). A department-level cultural change project: Transforming evaluation of teaching. In White K., Beach A., Finkelstein N., Henderson C., Simkins S., Slakey L., Stains M., Weaver G., & Whitehead L. (Eds.), *Transforming institutions: Accelerating systemic change in higher education*. Pressbooks. Retrieved May 27, 2022, from http://openbooks.library.umass.edu/ascnti2020

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Gin, L. E., Scott, R. A., Pfeiffer, L. D., Zheng, Y., Cooper, K. M., & Brownell, S. E. (2021). It's in the syllabus … or is it? How biology syllabi can serve as communication tools for creating inclusive classrooms at a large-enrollment research institution. *Advances in Physiology Education*, *45*(2), 224–240. https://doi.org/10.1152/advan.00119.2020

Griffin, K. A., & Reddick, R. J. (2011). Surveillance and sacrifice: Gender differences in the mentoring patterns of Black professors at predominantly White research universities. *American Educational Research Journal*, *48*(5), 1032–1057. https://doi.org/10.3102/0002831211405025

Guarino, C. M., & Borden, V. M. (2017). Faculty service loads and gender: Are women taking care of the academic family? *Research in Higher Education*, *58*(6), 672–694. https://doi.org/10.1007/s11162-017-9454-2

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216. 10.1126/science.1204820

Laursen, S., Andrews, T., Stains, M., Finelli, C. J., Borrego, M., McConnell, D., … & Malcom, S. (2019). *Levers for change: An assessment of progress on changing STEM instruction*. Washington, DC: American Association for the Advancement of Science. Retrieved May 27, 2022, from www.aaas.org/sites/default/files/2019-07/levers-for-change-WEB100_2019.pdf

Laursen, S., & Austin, A. E. (2020). *Building gender equity in the academy: Institutional strategies for change*. Baltimore, MD: Johns Hopkins University Press.

Misra, J., Kuvaeva, A., O'Meara, K., Culpepper, D. K., & Jaeger, A. (2021). Gendered and racialized perceptions of faculty workloads. *Gender & Society*, *35*(3), 358–394. https://doi.org/10.1177/08912432211001387

National Academies of Sciences, Engineering, and Medicine. (2020). *Recognizing and evaluating science teaching in higher education: Proceedings of a workshop—in brief*. Washington, DC: National Academies Press. https://doi.org/10.17226/25685

O'Meara, K., Kuvaeva, A., & Nyunt, G. (2017a). Constrained choices: A view of campus service inequality from annual faculty reports. *Journal of Higher Education*, *88*(5), 672–700. https://doi.org/10.1080/00221546.2016.1257312

O'Meara, K., Kuvaeva, A., Nyunt, G., Waugaman, C., & Jackson, R. (2017b). Asked more often: Gender differences in faculty workload in research universities and the work interactions that shape them. *American Educational Research Journal*, *54*(6), 1154–1186. https://doi.org/10.3102/0002831217716767

O'Meara, K., Lennartz, C. J., Kuvaeva, A., Jaeger, A., & Misra, J. (2019). Department conditions and practices associated with faculty workload satisfaction and perceptions of equity. *Journal of Higher Education*, *90*(5), 744–772. https://doi.org/10.1080/00221546.2019.1584025

Reinholz, D., Finkelstein, N., Corbo, J., & Bernstein, D. (2019). Evaluating scholarly teaching: A model and call for an evidence-based approach," In Lester J., Klein C., Johri A., & Rungwala H. (Eds.), *Learning analytics in higher education: Current innovations, future potential, and practical applications*. New York, NY: Routledge.

Rogers, E. M. (2010). *Diffusion of innovations* (4th ed.). New York, NY: Simon and Schuster.

Simonson, S. R., Earl, B., & Frary, M. (2021). Establishing a framework for assessing teaching effectiveness. *College Teaching*, *70*(2), 1–18. https://doi.org/10.1080/87567555.2021.1909528

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of Black college faculty: Does race matter? *Journal of Negro Education*, *80*(2), 149–162.

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

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

TEval. (2019). *Transforming higher education—Multidimensional evaluation of teaching*. Retrieved May 27, 2022, from https://teval.net

Thomas, S., Chie, Q. T., Abraham, M., Jalarajan Raj, S., & Beh, L. S. (2014). A qualitative review of literature on peer review of teaching in higher education: An application of the SWOT framework. *Review of Educational Research*, *84*(1), 112–159. https://doi.org/10.3102/0034654313499617

Weaver, G. C., Austin, A. E., Greenhoot, A. F., & Finkelstein, N. D. (2020). Establishing a better approach for evaluating teaching: The TEval project. *Change: The Magazine of Higher Learning*, *52*(3), 25–31. https://doi.org/10.1080/00091383.2020.1745575