The Effects of Practice-Based Training on Graduate Teaching Assistants' Classroom Practices

Erin A. Becker,[†‡*] Erin J. Easlon,[§] Sarah C. Potter,[†] Alberto Guzman-Alvarez,[†] Jensen M. Spear,[†] Marc T. Facciotti,[‖¶] Michele M. Igo,[§] Mitchell Singer,[§] and Christopher Pagliarulo[†]

[†]Center for Educational Effectiveness, Department of Undergraduate Education, [§]Department of Microbiology and Molecular Genetics, [‖]Department of Biomedical Engineering, and [¶]Genome Center, University of California, Davis, Davis, CA 95616; [‡]Data Carpentry, datacarpentry.org

ABSTRACT

Evidence-based teaching is a highly complex skill, requiring repeated cycles of deliberate practice and feedback to master. Despite existing well-characterized frameworks for practice-based training in K–12 teacher education, the major principles of these frameworks have not yet been transferred to instructor development in higher educational contexts, including training of graduate teaching assistants (GTAs). We sought to determine whether a practice-based training program could help GTAs learn and use evidence-based teaching methods in their classrooms. We implemented a weekly training program for introductory biology GTAs that included structured drills of techniques selected to enhance student practice, logic development, and accountability and reduce apprehension. These elements were selected based on their previous characterization as dimensions of active learning. GTAs received regular performance feedback based on classroom observations. To quantify use of target techniques and levels of student participation, we collected and coded 160 h of video footage. We investigated the relationship between frequency of GTA implementation of target techniques and student exam scores; however, we observed no significant relationship. Although GTAs adopted and used many of the target techniques with high frequency, techniques that enforced student participation were not stably adopted, and their use was unresponsive to formal feedback. We also found that techniques discussed in training, but not practiced, were not used at quantifiable frequencies, further supporting the importance of practice-based training for influencing instructional practices.

INTRODUCTION

Introductory science, technology, engineering, and mathematics courses at research-intensive universities tend to employ didactic, lecture-based instruction (Lund *et al.*, 2015). This type of learning environment, on its own, offers few opportunities to engage in the kind of iterative practice and feedback necessary for students to test and improve their current state of knowledge. To provide opportunities for structured practice, many of these courses include an associated discussion or laboratory section, often taught by graduate teaching assistants (GTAs). Analogously, existing pedagogy training for GTAs is often structured as didactic workshops, peer discussions, and/or literature readings (Prieto and Scheel, 2008) and often does not facilitate the type of iterative practice and feedback needed to learn and master such skills.

Here, we focus on the development and implementation of a practice-based teacher-training program for GTAs. Practice-based teaching frameworks are widely used in K–12 teacher training (Zeichner, 2012), with evidence showing that such "hands-on" work is an important element in the success of professional development in impacting teachers' classroom practices (Garet *et al.*, 2001). Such training engages novice teachers in deliberate practice (Ericsson *et al.*, 1993) of well-specified instructional activities

Marilyne Stains, *Monitoring Editor*
Submitted June 8, 2016; Revised June 27, 2017; Accepted July 27, 2017
CBE Life Sci Educ December 1, 2017 16:ar58
DOI:10.1187/cbe.16-05-0162

*Address correspondence to: Erin A. Becker (ebecker@datacarpentry.org).

Supplemental Material can be found at:
http://www.lifescied.org/content/suppl/2017/11/16/16.4.ar58.DC1

and incorporates targeted feedback and coaching (Lampert *et al.*, 2010). Training focuses on a core set of evidence-based high-impact teaching practices that occur with high frequency and can be used across many different environments and that instructors can begin to master early in their teaching careers (Grossman *et al.*, 2009). Complex teaching practices are first broken down into learnable component skills. Novices are then provided with repeated and progressive practice of these skills to develop automaticity (Darling-Hammond and Bransford, 2007). Focus on a limited number of component skills allows novices to build efficacy with demanding practices in small steps; this occurs through repeated cycles of practice in increasingly complex situations (Lampert *et al.*, 2013). During practice, component skills are first modeled by trainers and then drilled in a simulated, simplified situation with peers. Finally, component skills are implemented in an authentic classroom environment, under close observation, with coaching.

One mechanism that has been proposed to explain improvements in student learning outcomes observed in active-learning environments is that increased student participation in the learning process requires students to take responsibility for their own knowledge level (Eddy *et al.*, 2015). When participation is enforced, students should be more likely to hold themselves and each other accountable, leading to an increase in focused time on task. Based on this proposed mechanism, GTAs were instructed to call on every student during each class period. This constituted a "task standard" for performance. This type of "cold-calling" to increase and distribute classroom participation has been reported to increase student voluntary participation and comfort levels in class (Dallimore *et al.*, 2006, 2012).

Weekly drills constituted the "task repetition" element of deliberate practice. In addition, we used a comprehensive feedback program, consistent with current recommendations for best practices (Gormally *et al.*, 2014). This program included timely and repeated feedback sessions (Rezler and Anderson, 1971; Fedor and Buckley, 1987; O'Reilly and Renzaglia, 1994) focused on an explicit task standard for performance (Hattie and Timperley 2007) and provided measurable and specific guidelines for improvement (Englert and Sugai, 1983; Liden and Mitchell, 1985; O'Reilly and Renzaglia, 1994). Together, the task repetition and feedback elements of our training program fulfill the demonstrated requirements for effective deliberate practice (Ericsson *et al.*, 1993) and closely align with practice-based training guidelines developed for K–12 teaching training.

Our focus was on helping GTAs create safe and supportive classroom environments where students actively participated in practice and were routinely given feedback to help improve their performance. To this end, our study addressed five research questions:

1. To what extent did a practice-based training program prompt GTAs to implement evidence-based instructional practices?
2. To what extent did GTAs implement the practices that led to the highest levels of student participation?
3. How did GTA practices and student participation levels change throughout the training program?
4. Were changes in GTA practice associated with the nature of feedback provided (appreciation, coaching, or none)?
5. Which, if any, of these practices were associated with changes in student learning outcomes?

Traditionally, the impact of GTA-training programs has been assessed using indirect measures, such as self-efficacy surveys (Hardré, 2003; Komarraju, 2008; Young and Bippus, 2008), student evaluations (Davis and Kring, 2001; Marbach-Ad *et al.*, 2012; Pentecost *et al.*, 2012; Schussler *et al.*, 2015), and surveys of perceived student learning (Hardré, 2003). Use of such indirect measures can be valuable. However, by themselves, they do not provide detailed information about the actual use of instructional practices and therefore cannot address the question of how training efforts translate into concrete changes in teaching behavior. Thus, it is crucial to measure GTA classroom practices directly, both to enable evaluation of the effectiveness of the training program and to provide accurate feedback to trainees.

To assess the success of our training program, we evaluated two outcome variables that align with two of Kirkpatrick's four levels of evaluation (Kirkpatrick, 1994). This framework has previously been used to evaluate professional development in higher education contexts (Steinert *et al.*, 2006; Wyse *et al.*, 2014). Our outcome variables were 1) demonstration of GTAs' ability to apply target instructional practices in the classroom and 2) impact of training on student learning. These two outcome variables correspond to two of the types of outcome variables in Reeves's framework for GTA professional development evaluation and research: 1) GTA teaching and 2) undergraduate student outcomes (Reeves *et al.*, 2016).

Our study is unique in implementing a GTA-training program strongly aligned with practice-based frameworks found in K–12 teaching training and in directly assessing both the change in GTA classroom practices longitudinally throughout the training process and associations between GTA instructional practices and student learning.

METHODS

Course Structure

This study was conducted at a 4-year, residential R1 research university in the western United States. All 15 GTAs involved in the study led discussion sections for the same course, taught by the same lecture instructor in Fall 2014. Involvement in the training program was mandatory. The course was a large-enrollment (~1000 students/quarter) first course in a three-quarter introductory general biology series. Course content focused on molecular and cellular biology, including: biochemistry, energetics, metabolism, cellular structure, information flow, and regulation. The course met four times per week (three 1-hour lecture periods and one 2-hour GTA-lead discussion period) for 10 weeks. Student attendance at discussion was mandatory, with discussion scores comprising 20% of the final course grade. Content covered in each weekly discussion was standardized across all 45 discussion sections. A summary of topics covered each week is shown in Supplemental Table 1.

Before attending each weekly discussion, students completed graded prediscussion quizzes (preparatory assignments) through Carnegie Mellon University's Open Learning Initiative (OLI) Introduction to Biology online course (https://oli.cmu.edu), modified to fit specific course content. These quizzes were accompanied by substantial reading material and ungraded practice problems that included immediate, automated feedback. Student preparatory assignment responses provided feedback to GTAs about their students' level of understanding and were used by each GTA to design individualized 45- to

60-minute question-and-answer warm-up sessions held at the beginning of each discussion. The remainder of class time (40–55 minutes) was spent on POGIL (process-oriented, guided-inquiry learning)-like problem sets focusing on case studies and guided-inquiry activities (Farrell *et al.*, 1999) that were graded and served as weekly review assignments. Students worked in groups of four on problem sets and during group-based warm-up questions. During the warm-up question-and-answer session, GTAs were free to use any instructional method they felt would best elicit student participation; however, they were strongly encouraged to use the engagement techniques covered in training (see *Training Program*). During the remainder of class time, GTAs were directed to move through the classroom and interact with student groups to gauge their level of understanding and to supply individualized feedback to students (detailed in *Circulate and Check for Understanding*).

Information on GTA demographics and previous teaching experience was compiled from GTA self-reported data and university records (Table 1). Forty percent of GTAs had previously served as GTAs for the study course, and 66% had acted as GTAs for some course before the study term. One-third of GTAs had other (non-GTA) teaching experience before the study term, either at the university level or in other educational settings.

TABLE 1. GTA demographics[a]

	Value	% responding
% Female	33	100
% Domestic	87	100
% In PhD program	73	100
Year in current program (mean ± SD)	3.5 ± 1.8	87
% Nonuniversity teaching experience	20	93
% Other university teaching experience (non-GTA)	33	93
% Prior GTA experience	66	73
of which terms GTA (mean ± SD)	4.1 ± 4.2	100
% Prior Bis2A GTA experience	40	100
of which terms Bis2A GTA (mean ± SD)	3.3 ± 2.3	86

[a]All teaching experience is before and not including study term. Mean and SD for length of teaching experience do not include respondents who reported no experience in that category. $n = 15$ GTAs.

Training Program

Each of the 15 GTAs taught three 2-hour discussion sections per week of approximately 24 students each. Before the first discussion section each week, GTAs participated in a 2-hour practice-based training session. Trainings were designed to quickly ramp up GTAs' ability to effectively and consistently implement instructional techniques that emphasized student accountability, logic development, and practice of problem-solving skills. The training period was divided into two 1-hour sessions. The first hour focused on course-specific content review. Content review was conducted as a "mock warm-up," with training leaders modeling target techniques and GTAs acting as students. During the first 5 weeks, the second hour of training was dedicated to practice of a specific instructional technique (see *Target Techniques and Drills* for more details). Before GTA practice ("drill"), the theory behind the target technique was explained and step-by-step implementation guidelines were provided, followed by a demonstration by the training leaders. For most drills, GTAs were split into two groups of approximately seven GTAs plus a training leader, with each GTA practicing the target technique under the training leader's guidance. In week 6, the structure of the second hour of training changed, and GTAs used this time to collaboratively design interactive, question-based warm-up sessions for their discussion sections using the techniques learned in the first 5 weeks. The schedule of the training program is shown in Table 2. In addition to weekly training, GTAs also attended a 1-hour meeting the week before the start of term that covered course organization, grading policies, and an overview of the goals of the training program. Undergraduate learning assistants (ULAs) were present during GTA-training sessions, but for the most part participated in drills as mock students, not as mock instructors, due to time constraints. ULAs served a supplementary instructional role during discussion sessions. However, GTAs did occasionally arrange for a ULA to lead a class session. ULA technique use was included in analysis of student learning outcomes but not in analysis of GTA teaching practices.

TABLE 2. Training schedule: focus areas for first and second hours of 2-hour weekly GTA-training meetings

Week	First hour	Second hour
0	Administration/buy-in and pedagogy discussion	N/A
1	Content review	Drill (cold call)
2	Content review	Drill (stretch it: explain logic)
3	Content review	Drill (right is right)
4	Content review	Drill (stretch it: follow-up)
5	Content review	Drill (circulate/check for understanding)
6	Content review	Warm-up development
7	Content review/ Drill (circulate/check for understanding)	Warm-up development
8	No meeting	No meeting
9	Content review	Warm-up development
10	Content review	Warm-up development

Target Techniques and Drills

Target techniques were selected that facilitated our overall goal of creating a highly engaged classroom characterized by a high degree of student participation in iterative practice and feedback within a safe and supportive classroom environment. At the beginning of the training program, GTAs were provided with literature describing the target techniques, including the theory behind each technique, implementation tips and selected references to relevant research literature (Supplemental Document 1). Target techniques were (in alphabetical order): circulate/check for understanding, cold call, debrief, no apology, normalize error, praise effort, praise improvement, right-is-right, stretch it: explain logic, and stretch it: follow-up (Lemov, 2010). Descriptions and selected references are provided in Table 3. Due to time constraints, only five of the 10 target techniques were explicitly drilled in training. Drilled techniques were selected from target techniques based on the trainers' perception of the most urgent needs. All 10 techniques were discussed in training sessions and/or one-on-one feedback meetings between the GTA and the training leaders.

TABLE 3. Target techniques[a]

Technique	Description	Category	References
Circulate	Moving through classroom and engaging with students to monitor understanding	Practice	Lemov, 2010
Check for understanding*		Practice	
Cold call*	Calling on nonvolunteering students by name to answer a question	Apprehension reduction, practice, accountability	Dallimore *et al.*, 2004, 2006, 2012
Debrief	Analyzing reasons correct answer was correct and incorrect answers were wrong	Logic development	deWinstanley and Bjork, 2002; Turpen and Finkelstein, 2010; Smith *et al.*, 2011; Nielsen *et al.*, 2012
No apology	Demonstrating belief in importance of the instructional methods and curriculum	N/A	Roney *et al.*, 1995
Normalize error	Framing errors as natural and beneficial to learning	Apprehension reduction	Keith and Frese, 2005, 2008; Bell and Kozlowski, 2008
Praise effort	Explicitly recognizing and praising student effort	Apprehension reduction	Dweck and Leggett, 1988; Dweck, 2007; Bell and Kozlowski, 2008
Praise improvement	Explicitly recognizing and praising student growth	Apprehension reduction	
Right is right*	Setting high standards for accuracy in student responses	Practice, accountability	Epstein *et al.*, 2002
Stretch it: explain logic*	Asking students to explain the reasoning behind an answer	Practice, accountability, logic development	Willoughby *et al.*, 2000; Dunlosky *et al.*, 2013
Stretch it: follow-up	Asking related follow-up questions to stretch boundaries of knowledge and check for integration	Practice, accountability	Lemov, 2010

[a]Technique names and descriptions are derived from (Lemov, 2010). Selected references are given; however, the same or very similar teaching practices may be referred to by various names in the literature. Techniques marked with an asterisk were drilled during training sessions.

Drills were designed to allow GTAs to practice techniques in a highly simplified mock classroom environment in order to make the techniques routine and habitual. GTAs were informed throughout training that drills were not meant to be accurate representations of authentic classroom experiences. Complications arising in the GTAs' classrooms and advice for applying the techniques to specific classroom situations were discussed throughout the quarter. After techniques had been explained and modeled by the training leaders, drills were conducted as follows:

Cold call—GTAs were provided with a topic and given 3 minutes to develop questions on that topic. Each GTA then had 3 minutes to cold-call the other members of the group, with the goal of calling on as many people as possible. GTAs playing the role of students were instructed to always provide correct and complete answers.

Stretch it: Explain logic—Same as the cold-call drill, with the addition that, for each answer provided, the GTA asked the "student" to explain his or her reasoning before moving to a new question.

Right is right—Same as the cold-call drill, with the exception that specific "students" were selected to provide incorrect or incomplete answers. The GTA doing the drill was unaware of which "students" would give faulty answers. Upon detecting an incorrect or incomplete answer, the GTA was instructed to restate the question in a way that helped the "student" identify his or her error. "Students" were instructed to give the correct answer after redirection.

Stretch it: Follow-up—Same as the cold-call drill, with the addition that each "student" called on was asked two to three questions of increasing difficulty on the same topic to check for depth of understanding.

Circulate/check for understanding—Before the drill, training leaders collaboratively identified GTAs needing additional skill development ("rookies") and GTAs with strongly developed skill in this technique ("peer modelers"), based on classroom observations. Each training leader formed a drill team composed of two rookies and one peer modeler. GTAs were not informed whether they were rookies or peer modelers. While the remaining GTAs worked in small groups to solve mock warm-up problems, the drill teams circulated. First, the training leader stopped at a group and demonstrated proper circulation technique: entry into the group's conversation, asking for an explanation of the group's solution or thought process, and ensuring verbal participation from all group members. Each GTA completed this process with different drill groups.

During drill sessions, the training leader and other GTAs provided short (10- to 15-second) positively framed feedback for each GTA. Public feedback focused on positive aspects of individual performance, while targeted corrective feedback was reserved for regularly scheduled private feedback meetings (see *Observations and Feedback*).

Observations and Feedback

Owing to the reported importance of timely, goal-directed feedback in influencing teaching practices (Rezler and Anderson, 1971; O'Reilly and Renzaglia, 1994; Fedor and Buckley, 1987; Gormally *et al.*, 2014), we incorporated three cycles of feedback into our training program. In-person classroom observations were conducted for each GTA in the second, fourth, and seventh weeks of the course. The goal of these observations was to monitor adoption of target techniques and provide frequent, near-term, and goal-oriented feedback to the GTAs (Rezler and

Anderson, 1971; O'Reilly and Renzaglia, 1994; Fedor and Buckley, 1987). To ensure that performance goals were clear, feedback focused on a recognized task standard (Hattie and Timperley, 2007), using an in-house classroom observation protocol (Supplemental Document 2) that was provided to GTAs at the beginning of the term. This protocol also facilitated the development of quantitative personalized improvement goals for each GTA. Upcoming observations were announced in the weekly training meeting, at which point GTAs completed written self-evaluations (Supplemental Document 3) that asked them to assess their own strengths and weaknesses in implementing the target techniques. Weaknesses were phrased as "instructional techniques that you would like to improve," and it was explained that personalized goals would be set during follow-up meetings. Individualized goal setting is thought to improve motivation, leading to increased effort toward the desired task (Gormally *et al.*, 2014).

Each GTA was responsible for three discussion sections. In-person observations were conducted during the second of each GTA's discussion sections, allowing the GTAs to practice the techniques unobserved in their first discussion section. The observation period was shortened as the course progressed, with median observation length of 40 minutes in week 2, 30 minutes in week 4, and 15 minutes in week 7. This decrease in the length of observation was intentional, as observers needed less time to assess classroom practice after becoming familiar with each GTA's basic teaching habits. Observations were conducted such that each GTA was observed at least once by each of the two training leaders (E.A.B. and E.J.E.). As necessary, observations were also completed by a third observer (C.P.), who was part of the study design and had been trained on the observation protocol.

At the end of each observation week, the observers met to discuss GTA progress and come to consensus on what personal feedback should be provided to each GTA. This feedback included a written evaluation followed by a face-to-face meeting. The written evaluation included one to four observed strengths and one to three areas to improve, paired with specific, quantitative goals, and was sent to each GTA by email. Each GTA then attended a 15-minute individual face-to-face meeting with one of the observers to discuss their feedback and any issues they were facing in their classrooms. Written summaries for the second and third iterations of the feedback process included whether the previous goal(s) had been met. For the third iteration, written summaries were more informal, representing a summary of the GTAs' progress over the quarter, and in-person feedback meetings were not held. Instead, open office hours for discussing feedback were offered; however, none of the GTAs attended these nonmandatory meetings.

Video Coding

For assessment of the frequency of GTA implementation of target techniques and how GTAs elicited student participation across classrooms throughout the quarter, video recordings were taken of all 45 classrooms in the second, fourth, seventh, and 10th weeks of the course. A total of 160 (89%) of the targeted classroom sessions were successfully recorded. Eighty-six percent of missing data (19 videos) were from week 7. All GTAs were recorded between eight and 12 times, with 80% of the GTAs being recorded 10 or more times. Data from 159 videos were included in analysis of student learning outcomes; one video was excluded where one GTA substituted for another. Two of the recorded classroom sessions were led by ULAs and were excluded from analysis of GTA instructional practices but included in analysis of student learning outcomes. Only the first hour of class time was coded, roughly corresponding to the GTAs' individually designed interactive questions and answer ("warm-up") sessions. In five instances, observations began after the start of class, with a median start time of 53 minutes into class. In three instances, less than 1 hour of video was recorded, with a mean recorded time for incomplete videos of 52 minutes.

Videos were coded using a modified version of the in-house classroom observation protocol, along with a coding manual (Supplemental Documents 4 and 5). Video coding was done by three trained undergraduate observers. Initial training consisted of two 1-hour-long sessions in which codes were defined, potential scenarios discussed, and short clips of observation videos coded to consensus. All videos were coded using one of four progressive strategies. Initially, observers coded videos in tandem with a partner to provide opportunities to discuss any discrepancies in their understanding of the codes ("paired-tandem"). After solidifying understanding of the codes, observer pairs coded independently and compared results before entering the data in order to resolve major discrepancies ("paired-checked"). After developing expertise in coding, observer pairs coded independently of one another and did not compare results ("paired-independent"). Finally, after confirming high levels of interrater reliability (IRR; see next paragraph), individual observers coded the remaining videos independently ("single"). Videos were randomly assigned to the undergraduate observers, and observer pairs were rotated throughout the first three stages.

IRR was assessed separately for paired-checked (23 videos) and paired-independent (19 videos) video sets. For both sets, IRR was assessed independently for each code using a one-way mixed model, absolute agreement, single-measures, interclass correlation (Shrout and Fleiss, 1979; McGraw and Wong, 1996) to assess the degree of consistency in observer ratings of classroom practices. Interclass correlation coefficient (ICC) calculations were carried out using the irr R package (Gamer *et al.*, 2015). The resulting ICC values are provided in Supplemental Table 2.

Only codes with an ICC value >0.75 ("excellent") in the paired-independent video set were included in further analyses. High ICC indicates that these classroom practices and participation types were rated with a high degree of fidelity across observers, suggesting that a low degree of measurement error was introduced. Consequently, statistical power for further analysis was not significantly impacted by independent coding. These 42 codes were therefore deemed to be suitable to use in further hypothesis testing. Owing to this high degree of consistency, the remaining videos were coded by a single observer. Note that ULA classroom practices are included only in analysis of student learning outcomes, and not analysis of GTA practice. Code frequencies were averaged between observers. Codes for class sessions with less than 1 hour of recorded video were time normalized to 60 minutes.

Student participation events were coded as one of five types: 1) student question (student asks a question of the GTA or ULA), 2) cold call–individual (GTA asks an individual student a question without the opportunity to discuss with group members), 3) cold call–group (GTA asks an individual student a question

after that student has had the opportunity to discuss the question with group members), 4) volunteer–individual (student answers a question without prompting by GTA or ULA), and 5) volunteer–group (GTA calls on a group of students to answer a question, of whom one of the students volunteers to provide the answer). In investigating student participation levels, all categories were analyzed independently and in appropriate combinations.

For GTA interactions with small groups (circulate), a distinction was made between interactions initiated by the GTA ("active" circulate) and by the students ("moderate" circulate). Instances in which the GTA was moving through the classroom but not interacting with students were captured by the "passive" circulate code. For more details on code definitions, see Supplemental Document 5.

Longitudinal Analyses of Classroom Practice

To understand the dynamics of GTA instructional practices, we analyzed changes in frequency of technique use throughout the quarter and changes in student participation in class discussions. For each week, the frequency of each coded technique or participation type was averaged for each GTA across his or her recorded sessions (usually three). For each technique, we investigated whether there were differences in GTA practice between the beginning (week 2) and end (week 10) of the course. Because the frequency for most techniques did not meet assumption of normality, statistical significance of difference in means was tested using the nonparametric Wilcoxon rank-sum test in R (R Core Team, 2014). Effect sizes were calculated using Cliff's *d*, which is an ordinal statistic describing the frequency with which an observation from one group is higher than an observation from another group compared with the reverse situation (Cliff, 1993). Cliff's *d* can be interpreted as the degree to which two distributions (*x* and *y*) overlap, with values ranging from –1 to 1. A Cliff's *d* value of 0 represents no difference in the sample distributions, a Cliff's *d* value of –1 indicates that all samples in distribution *x* are lower than all samples in distribution *y*, and a Cliff's *d* value of 1 indicates the opposite. Threshold values for Cliff's *d* used throughout are defined in Romano *et al.* (2006) as implemented in Torchiano (2015). This method has been shown to be quite robust to violations of normality and heterogeneity of variance (Cliff, 1993). Cliff's *d* calculations were done using the effsize R package (Torchiano, 2015).

Analysis of Feedback

Trainers met with individual GTAs the week after their classrooms were observed to discuss strong points and potential areas of improvement in classroom management, technique implementation, and content knowledge. Each GTA was also given a written summary of feedback before the in-person meeting. The feedback given to the GTAs was characterized independently by the training leaders (E.A.B. and E.J.E.), who analyzed the written feedback summaries for each GTA each week and coded them for presence or absence of appreciation feedback (in which the GTA was praised or acknowledged for proper technique use or high technique frequency) or coaching feedback (in which GTA was prompted to improve fidelity or frequency of technique implementation; Stone and Heen, 2014) for each of the target techniques. Discrepancies in independently derived codes were discussed by the training leaders until consensus was reached. For each technique, each GTA was coded as having received neither type of feedback, appreciation only, coaching only, or (in rare cases) both types of feedback, for each week in which feedback was given.

For each technique, the change in technique frequency between the observation immediately preceding feedback and immediately following feedback was compared between 1) GTAs receiving appreciation feedback, 2) GTAs receiving coaching feedback, and 3) GTAs receiving no feedback for that technique. Change in technique frequency, rather than absolute frequency, was used as an outcome metric due to longitudinal changes in technique implementation. Statistical significance of difference in means was tested using the nonparametric Wilcoxon rank-sum test, and effect sizes were calculated as Cliff's *d* (Cliff, 1993), using the effsize R package (Torchiano, 2015).

Relationship between Classroom Practice and Student Learning Outcomes

We investigated the relationship between GTA/ULA classroom practices and student learning outcomes using multiple linear regression with overall course exam points as the response variable. We used exam points, rather than GTA-awarded points, to avoid confounding GTA teaching effectiveness with GTA grading leniency. As discussion enrollment was not randomized, student demographics were incorporated into statistical modeling as potential confounding variables for their success in the course. The importance of correcting for differential student demographics in nonrandomized educational studies has been demonstrated by Theobald and Freeman (2014).

To determine appropriate student demographic variables for inclusion in the model, we used the previous and subsequent Fall terms (Fall 2013 and Fall 2015) as training data sets. These terms were taught by the same lecture instructor as the study term (Fall 2014). First, we established that prior academic achievement (cumulative institutional GPA) and demographics (gender, transfer status, first-generation student status, underrepresented minority (URM) status, and course-repeater status) were similar across all three terms (Supplemental Table 3). To enable inclusion of new transfer students and freshmen, we used the cumulative institutional GPA from the end of the term rather than from the beginning. Grade earned for the study course was excluded from the calculated GPA. First-generation students were defined as students whose parent(s) or legal guardian(s) had not completed a bachelor's degree. URM students were defined as anyone who self-identified ethnically as African American/Black, Puerto Rican, American Indian/Alaskan Native, Mexican-American/Mexican/Chicano, Latino/Other Spanish, or Hispanic-Other. Transfer students were defined as anyone who had transferred to the university from an accredited community college. Demographic data were compiled from centralized university databases.

Model selection for the two training data sets was automated using the cross-validate model selection function in the glmnet R package (Hastie and Qian, 2014). For each training data set, the selected model was the one that had the lowest mean-squared error within 1 SD of the minimum value for the regularization parameter lambda. Both models included GPA, gender, and first-generation status but did not include students' URM or repeater status. The model for Fall 2015 also included students' transfer status. We used these models as a starting point for developing an appropriate model for the study term.

To model student learning outcomes as a function of GTA behavior, we employed a data reduction technique, reducing all coded behaviors to a series of aggregate variables. The technique used was principal component analysis (PCA), a data reduction technique for summarizing the information contained in several variables (in our case, coded behaviors), via a smaller number of aggregate variables (Cudeck, 2000). A total of nine coded behaviors (cold-call rate [%], volunteer rate [%], student question rate [%], circulate [passive], circulate [moderate] circulate [active], right is right, stretch-it: explain logic, and stretch-it: follow-up) were used in the final PCA. Statistically significant outliers (assessed using Mahalanobis distance [Mahalanobis, 1930], cutoff chi-square >20) were removed from the data set to yield a final analytical sample of nine coded GTA/ULA behaviors and 132 total classroom observations. Note that, although a PCA analysis was conducted, due to the relatively small sample size of our observations ($n = 132$), a statistical reduction technique such as PCA may not be appropriate and may lead to inaccurate component estimations. PCA was conducted only as an exploratory analysis to see which potential components could be extracted.

PCA analysis was run using the psych R package (Revelle, 2015). An oblimin rotation was used to extract a total of two components. An oblimin rotation was selected to allow the extracted components to have some potential correlation between them (Kim and Mueller, 1978). The two components were named "accountability" and "volunteer rate" and together accounted for 48% of the observed variation in classroom practice. Accountability included cold-call rate (%), right is right, stretch-it: explain logic, and stretch-it: follow-up. Volunteer rate was the percent of unique students who volunteered to answer a question. For creating an aggregate variable for accountability, the individual variables for that component were normalized (mean = 0, SD = 1) and then averaged, creating a z-score for use in the regression model.

Based on models recovered from the training data sets and components recovered from PCA analysis, the starting model for the study term was as follows:

Total_exam_points ~ cumulative_GPA + gender + first_generation_status + accountability + percent_volunteer

Accountability and percent volunteer metrics were constant for all students in a particular discussion section, while other variables represent values for individual students. Thus, we first tested whether a multilevel model with random effects for GTAs' experience level (number of times during study term a GTA had taught that discussion material before the discussion in question) and/or GTA identity (i.e., which GTA taught the section) or a single-level model was more appropriate. GTA experience and identity were individually added as random effects to an unconditional model using the nlme R package (Pinheiro *et al.*, 2014), and improvement to the model was tested by analysis of variance (ANOVA). Neither significantly improved the model (cutoff ANOVA $p < 0.05$). Thus, the single-level model was used.

To this base model, we tested addition of students' transfer status. Student transfer status was added as a fixed effect and tested by ANOVA. Although addition of transfer status significantly improved the model ($p = 0.030$), adjusted R^2 increased only marginally ($\Delta = 0.0026$), and therefore this term was not incorporated into the model. The final model was thus identical to the starting model.

Students who were missing demographic information for variables included in final model were excluded from the analysis ($n = 31$ students, 3.1%). Students who did not take all required examinations were also excluded ($n = 13$ students, 1.3%). A total of 946 students (95.5%) were included in the final analysis.

The final model was tested for conformity to assumptions of linear modeling: normality of residuals (by Shapiro-Wilk test [R Core Team, 2014], cutoff value $W > 0.95$ or $p > 0.05$), constant variance (by Breusch-Pagan test [bptest in R package lmtest; Hothorn *et al.*, 2011], cutoff value $p > 0.05$), and lack of overly influential data points (by Cook's distance [cooks.distance; R Core Team, 2014], cutoff value 1). As the model displayed heteroskedasticity, a heteroskedasticity-corrected covariance matrix was calculated using hccm in the R car package (Fox and Weisberg, 2011), with the classical White correction. This corrects SE estimates to account for heteroskedasticity but does not affect model coefficients.

IRB

This study was deemed exempt from full IRB review and determined to not be research involving human subjects as defined by the Department of Health and Human Services (IRB ID #513796-1).

RESULTS

GTA Classroom Practices and Student Participation Levels

We first sought to describe overall GTA classroom practice in terms of frequency of use of the target techniques and levels of student participation. Although each of the target techniques listed in Table 3 was included in the GTA information packet at the start of the term and discussed in training sessions, only five techniques were drilled (circulate, cold call, right is right, stretch it: explain logic and stretch it: follow-up). Significantly, techniques that were not drilled occurred rarely, at frequencies too low to pass our filter for IRR. Overall frequency (across all 158 classroom observations) for the 10 target techniques is given in Table 4. Five number summaries for each of the drilled techniques are given in Supplemental Table 4.

TABLE 4. Target technique total observed frequency[a]

Technique	Total observed
Circulate/ check for understanding*	2142
Cold call*	2907
Debrief	Not coded
No apology	7
Normalize error	4
Praise effort	17
Praise improvement	5
Right is right*	710
Stretch it: explain logic*	812
Stretch it: follow-up*	1677

[a]Only GTA (not ULA) activities are included in counts. Passive circulation (in which GTA moved through the room but did not interact with students) is excluded. The debrief technique was not coded due to inability to reach consensus coding criteria. Counts represent raw sums from all 158 classroom sessions in which a GTA acted as the primary instructor (excluding two observations in which a ULA acted as instructor). Techniques marked with an asterisk were drilled during training sessions.

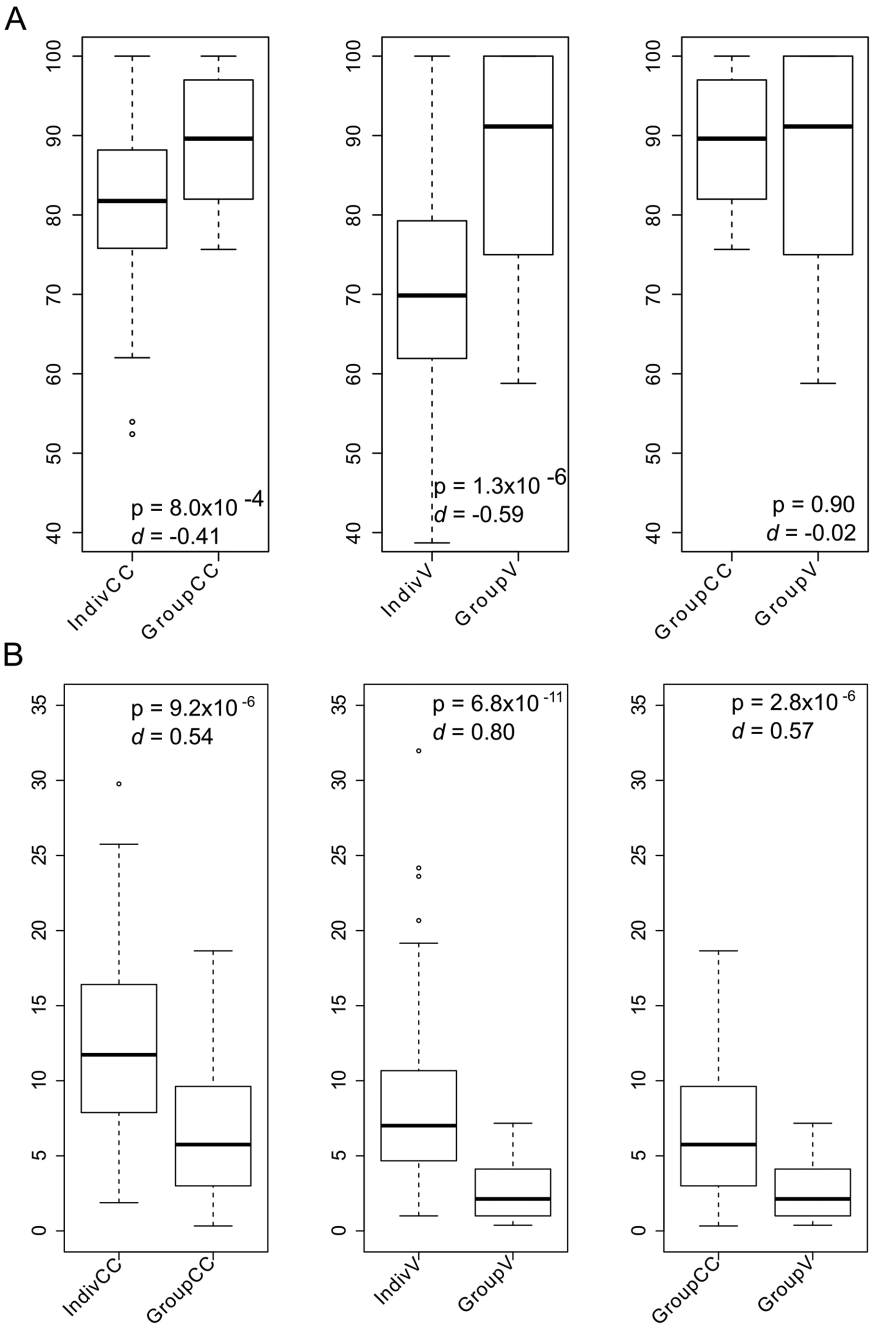

FIGURE 1. Student participation types and unique responder rates. (A) Percent of responders who were unique for each type of participation. (B) Frequency of different participation types. All values are averaged across observed classroom sessions for each individual section ($n = 45$ sections). IndivCC, individual cold call; GroupCC, group cold call; IndivV, individual volunteer; GroupV, group volunteer. Negative values of Cliff's *d* indicate technique on the right of each pair occurs at higher frequency (B) or has higher proportions of unique responders (A). Positive values indicate the opposite. For more information about interpreting Cliff's *d* values, see *Methods*.

Cold call was the most frequently used technique (Supplemental Table 4), with a median frequency of 18 events per hour observed. Circulate was the next most common, at 13 events per hour when both moderate (student-initiated) and active (GTA-initiated contact) were considered. The other techniques used often enough to quantify accurately (right is right, stretch it: follow-up, and stretch it: explain logic) were implemented with lower frequencies. This was not surprising, given that these techniques, by definition (see Supplemental Document 5), could only follow a cold-call event and could not be initiated independently. When occurrence of these three techniques are summed, their frequency is similar to cold call (median of 19.73 compared with 18.17, p value of difference between means = 0.81), with strong positive correlations between the number of right is right, stretch it: follow-up and stretch it: explain logic events and the percent of students in a classroom who were cold called (Supplemental Figure 1). This relationship indicates that GTAs using cold call were doing so in conjunction with additional techniques designed to improve students' accountability for their knowledge and provide opportunities for practice and logic development.

In investigating how participation was manifested in our active-learning classrooms (i.e., which types of participation were present and what proportion of students participated), we sought to answer the intertwined questions of 1) which of the four participation mechanisms we measured (individual volunteer, group volunteer, individual cold call, and group cold call) were most successful in prompting high proportions of unique students to respond, and 2) whether the more successful mechanisms were preferentially used by GTAs. These questions were addressed to determine whether GTAs were making optimal use of these techniques for eliciting student participation.

First, for the two response mechanisms for which GTAs called on individual students, we investigated whether GTAs were more likely to call on unique students without providing time for students to discuss a problem with their groups (individual cold call) or after such a discussion (group cold call). Although the percentage of unique responders was high for each response type (mean of 81% and 89%, respectively), GTAs were significantly more likely to call on new students after group work than when no group work time was provided (Cliff's $d = -0.41$, 95% CI [–0.6, –0.18], $p = 8 \times 10^{-4}$; Figure 1A). Despite being more effective in producing unique participants, the group cold-call participation mechanism was used less frequently by GTAs than individual cold call (Cliff's $d = 0.54$, 95% CI [0.33, 0.71], $p = 9.2 \times 10^{-6}$; Figure 1B).

Next, we asked whether GTA targeting of specific groups to answer a question (group volunteer) led to higher levels of unique participants than relying on individual students to volunteer. Individual volunteers showed a strong bias toward repeated response from the same students (mean unique responder rate of 70%) compared with GTA-targeted procurement of volunteers from a specific group (mean of 87% unique; Cliff's $d = -0.59$, 95% CI [−0.74, −0.38], $p = 1.3 \times 10^{-6}$; Figure 1A). Although calling on groups led to a more diverse set of participants, GTAs were significantly more likely to take volunteer responses from the class as a whole than to request an answer from a particular group (Cliff's $d = 0.80$, 95% CI [0.64, 0.89], $p = 6.8 \times 10^{-11}$; Figure 1B).

Finally, for instances in which students were asked to respond to questions following group work, we asked which of two strategies—cold-calling an individual student or asking for volunteers from the group—was likely to lead to a greater proportion of unique responders. We found no significant difference in unique responder rate between the two strategies (Cliff's $d = -0.02$, 95% CI [−0.26, 0.23], $p = 0.89$), although GTAs were more likely to ask for volunteers than to cold-call students after group work (Cliff's $d = 0.57$, 95% CI [0.35, 0.74], $p = 2.6 \times 10^{-6}$).

In addition to counting frequency of implementation for each of the 10 target techniques, we were also interested in measuring the fraction of students in each classroom who were given the opportunity to participate in the whole-class discussion. Three-quarters of classrooms had an average participation level above 70% (range: 29–100%) and a cold-call participation level above 46% (range: 14–100%; Table 5). When GTAs are considered as the unit of interest, three-quarters averaged overall student participation levels above 64% and a cold-call participation level above 50%. In the median classroom, only 15% of students asked a question during whole-class discussion (range: 3–37%), highlighting the inadequacy of relying on spontaneous student questions for gauging understanding. See Supplemental Table 5 for a breakdown of participation types. For this analysis, we were interested in GTA attempts to implement the target techniques; therefore, we included here students who were called on to answer a question but did not respond. Overall nonresponse to cold call was 4.4%. Note that this measure does not capture student–student interactions in small groups or interactions between GTAs and students during small-group work; the latter factor is captured in the circulate (active) and circulate (moderate) codes.

Changes in Classroom Practice over the Course of the Training Program

To characterize possible impacts of the training program, we investigated whether GTA classroom practices changed over the course of the term. Of the 16 techniques and participation types measured, six showed significant changes between the first and last observation periods (Figure 2 and Supplemental Table 6). Frequency of cold call, and stretch it: explain logic significantly decreased over the course of the term, as did overall participation rate and cold-call levels. In contrast, the frequency of student-initiated contact with the GTA during small-group work (circulate [moderate]) increased, as did the frequency of group volunteer events (in which GTA called on a specific group but allowed students within that group to decide who would answer the question).

Changes in Classroom Practice Following Feedback

To explore the potential impact of technique-specific feedback on GTA practice, we looked at changes in technique use between the observation immediately before and immediately after each feedback session. We compared the change in use of specific techniques for GTAs who received either coaching feedback or appreciation feedback against those who received no feedback for that technique. The stretch it: explain logic technique was not included in this analysis, due to a difference in how the two variants of this technique were coded, preventing the pooling of observations for instances in which the same or a different student was asked to provide an explanation (see Supplemental Document 5). Circulation (moderate; i.e., student-initiated GTA–small group contact) was also excluded, because GTAs had no direct control over how frequently this type of interaction occurred. As discussed previously, the no apology, normalize error, praise effort, and praise improvement techniques were not used often enough to accurately quantify. Thus, only four of the 10 target techniques were included in analysis of feedback impact.

Of these four, a significant positive relationship between coaching feedback and technique frequency was found for right is right ($p = 0.04$) and circulate (active; i.e., GTA-initiated small-group contact; $p = 0.03$; Figure 3), indicating that GTAs receiving coaching feedback for these techniques subsequently increased their implementation frequency, at least transiently, compared with GTAs who received no feedback on these techniques. Coaching feedback was only provided once for stretch it: explain logic; we are therefore unable to make any claims about the effectiveness of this type of feedback for

TABLE 5. Summary statistics for student participation levels across all observed classroom sessions for each classroom (left) or each GTA (right)[a]

	By classroom		By GTA	
	% participation (SQ + V + CC)	% participation (CC only)	% participation (SQ + V + CC)	% participation (CC only)
Min	28.59	13.56	46.29	29.00
Q1	70.02	45.72	64.23	49.92
Median	79.47	62.50	79.49	58.59
Q3	85.12	74.91	84.98	74.67
Max	100.00	99.64	94.17	89.32

[a]Numbers represent percent of students in class on the day observed who participated in whole-class discussion. Overall levels of participation (including student questions [SQ], volunteer responses [V], and Cold Call [CC]) are shown, with Cold-Call levels also shown separately. For more detailed breakdown of participation types, see Supplemental Table 5. $n = 45$ classrooms, 15 GTAs.

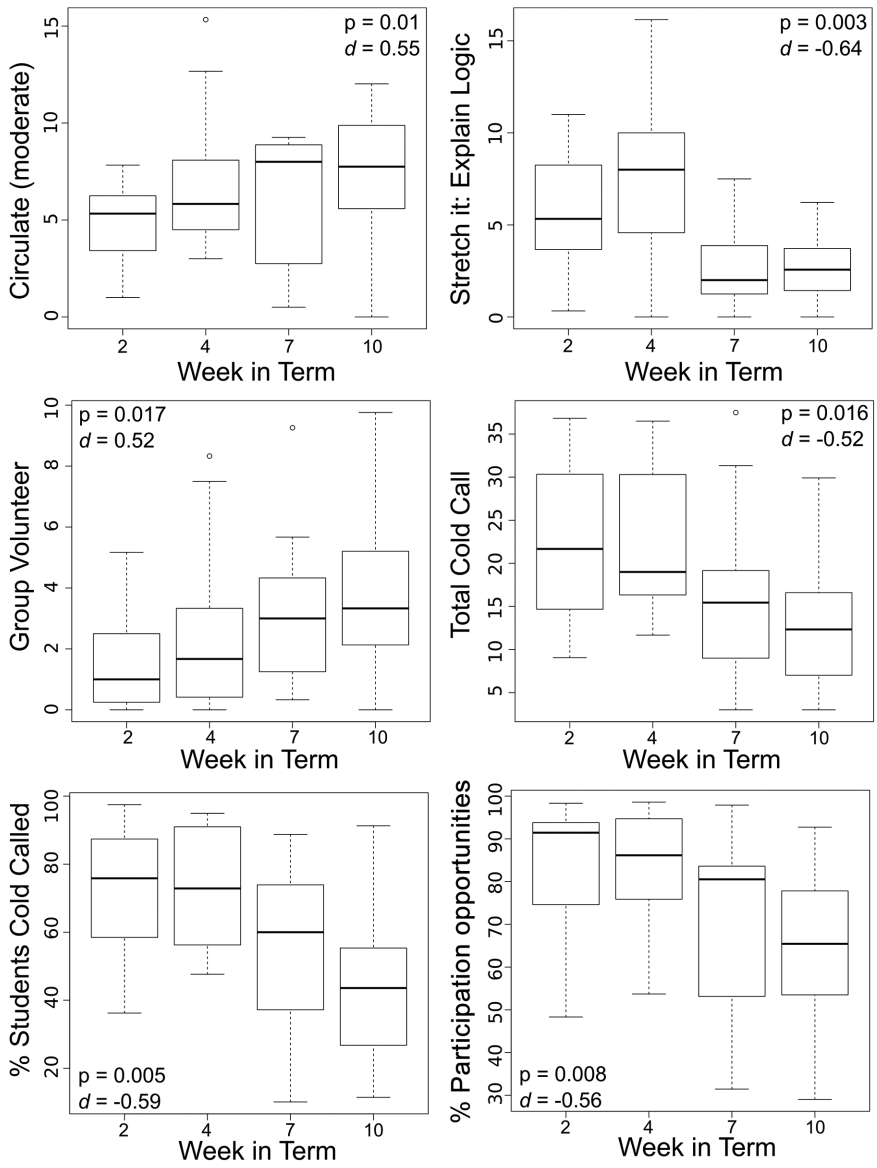

FIGURE 2. Longitudinal changes in GTA classroom practices. Box plots showing distribution of technique and participation type frequency for each week in which classroom observations were done, averaged across all observed sections for each GTA (usually three). Box sections represent second and third quartiles. Whiskers represent first and fourth quartiles. Thick line represents medians. Outliers are shown with open circles. *p* values and Cliff's *d* values are for difference between weeks 2 and 10. Negative values of Cliff's *d* indicate decrease in frequency of practice; positive values indicate increase in frequency. For more information about interpreting Cliff's *d* values, see *Methods*. *n* = 15 each for weeks 2, 4, and 10 and 11 for week 7. Box plots for techniques and participation types not shown here are available in Supplemental Document 6. See Supplemental Table 6 for all longitudinal comparisons.

this technique. We also observed a negative relationship between appreciation feedback and technique frequency for stretch it: explain logic ($p = 0.009$). However, no significant relationship was found between appreciation feedback and implementation frequency for other techniques. In contrast to the other three techniques investigated, no significant change in use of cold call was observed regardless of the type of feedback provided.

Student Learning Outcomes

In an attempt to determine whether any of the teaching practices we measured were associated with student learning, we modeled the relationship between these practices and student exam scores. Classroom behaviors were condensed into components via PCA. The two components that most fully explained variability in classroom practices were 1) accountability (which included % cold call, right is right, stretch it: explain logic, and stretch it: follow-up; see Table 3) and 2) % volunteer rate. Although neither of these two components was significantly correlated with student learning outcomes at the $p < 0.05$ confidence level (Table 6), the indication of a possible negative trend between student exam scores and volunteer rates warrants further research.

DISCUSSION

Evidence has shown improved student learning outcomes in classroom environments where students take an active role in interacting with the material (Freeman *et al.*, 2014). Effective implementation of evidence-based teaching practices, including active learning, is a highly complex skill requiring deliberate practice to master. Here, we sought to transfer a practice-based training framework from its original context in K–12 teacher training to a higher education context and to demonstrate its ability to help novice higher education instructors (here GTAs) learn and implement evidence-based practices.

Although there are many different aspects of active learning, important components with demonstrated positive effects on student learning have recently been reviewed (Eddy *et al.*, 2015). In this study, we focused on two major hurdles for the implementation of these identified components—expanding the penetrance of active learning–based teaching methods through GTA training and determining how feedback and structured practice influenced the adoption of these techniques in the classroom. To begin to address these broad goals, we assessed whether a practice- and feedback-based training program could be used to train GTAs to implement specific active-learning practices with high fidelity. We also examined whether student achievement tracked with any of these specific teaching practices as implemented by our GTAs.

We found that, given practice-based training, GTAs were capable of implementing evidence-based teaching practices. However, although initial use of drilled practices was high,

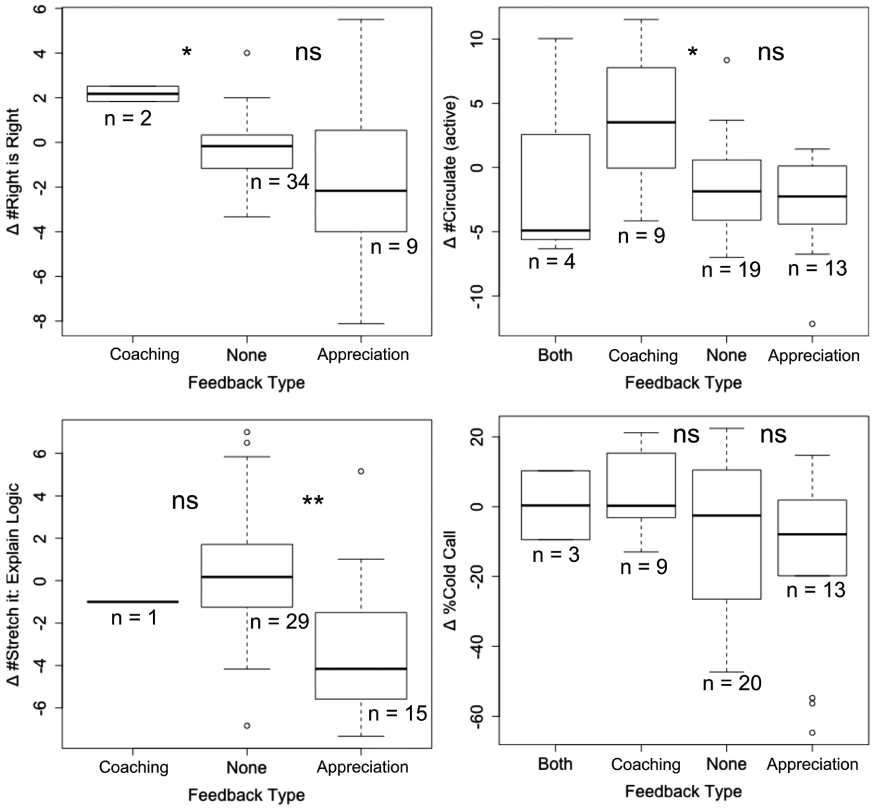

FIGURE 3. Changes in GTA classroom practices following feedback. Changes in technique frequency between prefeedback and postfeedback observation sessions. Box sections represent second and third quartiles. Whiskers represent first and fourth quartiles. Thick line represents median. Outliers are shown with open circles. Coaching, feedback intended to bring practice closer to defined task standard; appreciation, feedback that recognizes performance in line with task standard. *n* = number of instances where particular type of feedback was given. *, *p* value <0.05; **, *p* value <0.01; ns, *p* value >0.05. See the text for *p* values.

adoption was not stable, particularly for participation enforcement techniques (e.g., cold call), which decreased over the course of the term. These same techniques were not amenable to formal feedback, as GTAs' use of these techniques declined despite coaching. Analysis of student participation suggests specific changes in our training methods may ameliorate this effect and increase student engagement. For example, given documented instructor resistance to cold calling due to a fear of embarrassing students (Dallimore *et al.*, 2004, 2006, 2012), encouraging GTAs to cold-call groups rather than individuals may have led to more persistent implementation of this technique. Additionally, although drilled practices were initially adopted with some declining over the course of the term, nondrilled practices were not used with any appreciable frequency. In light of this result, we recommend incorporating drills for all targeted teaching practices into the training program.

Utility of the feedback program was supported by direct evidence of the effectiveness of feedback for at least two of the four assessed techniques (right is right and circulate [active]; Figure 3). These results indicate that at least some of the targeted instructional behaviors were responsive to feedback. We did not detect any significant change in GTA implementation of cold call after feedback. This may indicate that this instructional practice is more resistant to external influence. Instructor resistance to cold call has been previously documented (Dallimore *et al.*, 2006) and matches the authors' experiences in a variety of instructional coaching settings. However, due to wide variance in GTA practice and our small sample size (only nine instances of coaching feedback for cold call), this result is inconclusive. A priori estimation of required sample size suggests that more than 200 independent feedback events would be required to detect a true effect, assuming a moderate Cohen's *d* of 0.5.

GTAs did not maintain consistently high levels of some of the targeted classroom practices, specifically cold call and the closely-linked stretch it: explain logic technique, which each decreased significantly over the course of the term (Figure 2). This result is particularly intriguing, given that cold call was the only technique for which an explicit performance target (100%) was set. In addition to providing more evidence for the difficulty of influencing instructor use of cold call and related accountability techniques, these results highlight the importance of directly assessing classroom practice. However, asking GTAs about their perceptions of the utility of each of the individual target techniques may have enabled us to more directly assess both buy-in for individual techniques and the extent to which this buy-in translated into adoption of these practices.

The four metrics that decreased over the course of the term were all directly or indirectly tied to cold-call use (Supplemental Table 6 and Figure 2). Both the total number of times that cold call was implemented and the proportion of students in a classroom who were asked to respond via cold call decreased. One technique that could only be implemented as a follow-up to cold call (stretch it: explain logic) also decreased. The decrease in cold call and stretch it: explain logic correlated with

TABLE 6. Student learning outcomes[a]

Parameter	Regression coefficient ± SE	*p* value
Intercept	48.40 ± 1.96	$<2 \times 10^{-16}$
GPA	9.90 ± 0.56	$<2 \times 10^{-16}$
Female	−2.26 ± 0.65	5.24×10^{-4}
First generation	−2.98 ± 0.66	7.59×10^{-6}
Accountability	0.26 ± 0.38	0.49
% Volunteer	−0.048 ± 0.026	0.061

[a]Student exam performance was not significantly associated with either classroom volunteer response levels or accountability (combination of % cold call, right is right, stretch it: explain logic, and stretch it: follow-up). Bolded *p* values are significant at the $p < 0.05$ level. Coefficients are in terms of percentage of exam points.

the cessation of drilling after week 5. Although we cannot causally link these events, this fact, in combination with the lack of GTA use of nondrilled techniques, supports the importance of deliberate practice for transforming instructional practices.

This decrease in cold call was linked to an overall decrease in student participation levels, as, despite the increase in one type of volunteer response (group volunteer—GTAs asking for a volunteer respondent from a particular group), overall shifts in volunteer rates were not sufficient to compensate for decreased levels of enforced participation. We did not observe an increase in total volunteer rates after exposure to cold call, as had been previously reported (Dallimore *et al.*, 2012). However, we did find that students became more likely to initiate contact with their GTAs during group-work time as the term progressed. This result provides some evidence for the idea that students in classrooms with high levels of enforced participation are more likely to take an active role in their education, although the mode of action of this general principle may differ across instructional environments (i.e., increased volunteer rates in Dallimore *et al.* [2012] vs. initiating interaction with instructor in this study).

One of the reasons for increasing classroom participation rates was to increase the likelihood for participation from the maximum possible number of students. Therefore, we assessed whether GTAs preferentially used participation mechanisms that were more effective at eliciting unique respondents. We found that GTAs' preferred implementations of the cold-call and volunteer participation types favored individual rather than group responses, despite the fact that the group response mechanism in both cases turned out to be more successful at eliciting unique respondents (Figure 1). In terms of concrete classroom practices, this meant that GTAs were more likely to ask questions of students without providing group work time, and that in cases where group-work time was provided, GTAs were more likely to ask for volunteers from the room at large rather than from a particular group. We speculate that this first preference may be an artifact of the way drills were conducted, as the drills for Cold Call and related techniques did not involve group work.

We found no indication that calling on individual students following group work (group cold call) was more successful at eliciting unique respondents than calling on groups (group volunteer). As the group volunteer participation technique was one of only two practices to increase significantly over the course of the term, this result opens up the potential for shifting toward this perhaps less intimidating form of participation enforcement. By emphasizing use of group volunteer rather than cold call in future iterations of the training program, we may be able to improve classroom participation rates, as this engagement mechanism was organically favored by GTAs.

This analysis of GTA use of participation mechanisms suggests two concrete changes to our training program to increase student participation. First, cold-call drills should provide time for mock students to discuss questions as a group before being called on. Second, a variation of the cold-call drill should be added that prompts instructors to call on groups of students rather than individuals. The data suggest that these two modifications of the cold-call technique would help overcome instructor resistance to cold call, while maximizing unique student participation.

Along with investigating longitudinal changes in GTA instructional practices, we also questioned whether these changes were likely to be attributable to the training program or to organic changes in teaching practices as GTAs became more experienced. The results of our feedback analysis provide evidence that at least a subset of the changes we observed were responses to the feedback component of the training program (Figure 3). We therefore propose that longitudinal changes observed in GTA instructional practices are not the result of GTAs simply gaining more teaching experience, but are related to their experiences in the training program. These experiences include both the intentional aspects of the program (e.g., drills and feedback) and unintended exposure to alternate conceptions of appropriate teaching behaviors (e.g., via conversations with other GTAs and interactions with students).

Most GTAs adopted, at least temporarily, the five drilled target techniques. However, the techniques that were discussed and modeled in training sessions but not drilled were nearly completely absent from the observed classrooms (Table 4). These techniques (no apology, normalize error, praise effort, and praise improvement) are all components of the apprehension reduction dimension of active learning (Eddy *et al.*, 2015), which focuses on reducing students' fear of participation, thereby lowering the barrier to a highly participatory classroom. We speculate that our failure to emphasize these techniques in training may have hampered GTA adoption of enforced participation techniques, due to the perception (whether by GTAs or students) that enforced participation was threatening. In addition, the lack of adoption by our GTAs of techniques that were discussed and modeled but never practiced suggests that current methods of GTA training focusing on either literature-based or modeling-based exposure to instructional practices are not sufficient to transform GTA-led instruction.

In analyzing the relationship between GTA use of the targeted teaching behaviors and student exam performance, we found neither of the two components investigated were significantly correlated with student learning outcomes at the $p < 0.05$ confidence level. However, we speculate that a negative relationship between the rate of volunteer responses and student performance may exist, due to reduced class-wide attentiveness when volunteers are used as the primary mechanism for eliciting responses. Although our data offer only weak support for this claim, they suggest that further investigation of this research question may be illuminating.

CONCLUSIONS

Our results suggest that, although GTAs are capable of using evidence-based instructional practices given substantive practice-based training and regular guidance, adoption of these practices can be unstable and dependent on factors outside the training program. This work can provide a starting model for practice-based training of GTAs, but further work is required to understand how existing GTA attitudes toward teaching and interactions between GTAs and students influence adoption of evidence-based teaching behaviors.

Our observation that GTAs did not use nondrilled techniques suggests that discussing and modeling instructional behaviors is insufficient for modifying GTA classroom practices. Just as student mastery of academic content is aided by practice (e.g., Freeman *et al.*, 2014), so too, targeted practice appears to be a prerequisite for mastery of an instructional skill set. This result has implications for revising current GTA-training practices, which often do not provide opportunities for practice.

For techniques that were practiced, we found that adoption varied depending on the specific instructional practice being targeted. Some techniques were adopted readily and consistently and were easily influenced by specific, goal-oriented, and timely feedback. Other practices (primarily those involved in participation enforcement) were not stably adopted. Thus, the effectiveness of formal feedback programs for instruction may be dependent on the particular instructional practices being targeted. We suggest future work focus on understanding the complex relationship between attitudes (both of students and of instructors) toward evidence-based teaching practices, particularly enforced participation, and instructor readiness to adopt such techniques.

LIMITATIONS AND FUTURE DIRECTIONS

This work was carried out in a specific instructional setting and was influenced by the institutional culture present in this setting. The pre-existing structure of this course greatly facilitated our study, as GTAs for this course were already expected to attend weekly training sessions, and our training program did not increase overall GTA time commitment. In the absence of established training requirements, introduction of a training program may cause issues with GTA buy-in. Our course also has a dedicated full-time staff member who is responsible for training and overseeing GTAs (E.J.E.). The existence of this resource enabled the time-intensive, repeated, in-person classroom observations and one-on-one meetings called for in our training program. Courses lacking this resource may have difficulty in implementing a similar training program.

Our ability to analyze changes to GTA practices in our study is limited by the lack of a formal control group or measurement of GTAs' instructional practices before the beginning of the training program. It is, however, the experience of our GTA coordinator (E.J.E.) from years of classroom observations, that GTAs for our course do not spontaneously practice these behaviors.

We also note that, although some GTA classroom practices did appear to be responsive to feedback, this response may have been transient, with GTAs eventually returning to their previous teaching practices. Future studies investigating the persistence of feedback-responsive change in teaching methods would help to understand the optimal frequency with which to deliver such feedback.

We were unable to uncover any relationship between GTA instructional practices and student exam performance. Although the sample size for our study was quite large in the context of classroom observational studies, due to high amounts of natural variation in instructional practices across GTAs, it may be necessary to collect data on an even larger number of class sessions to resolve these relationships. In particular, the possibility of a negative relationship between student volunteer rate and exam performance warrants further investigation.

ACKNOWLEDGMENTS

We thank the GTAs and ULAs for their patience with us in conducting this research, the undergraduate research assistants who helped transform many hours of raw video footage into data for analysis, and the many Bis2A instructors who have been instrumental in continued improvements of curricular materials for this course. We also thank Dr. Marco Molinaro for his involvement in securing funding for this project. This work was funded by the Bill & Melinda Gates Foundation Adaptive Learning Market Acceleration Program Grant. Funders had no role in the design and conduct of the study; collection, analysis, or interpretation of data; or preparation, review or approval of the article.

REFERENCES

Bell, B. S., & Kozlowski, S. W. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. Retrieved June 9, 2016, from Cornell University, ILR School site http://digitalcommons.ilr.cornell.edu/articles/410

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. doi: 10.1037/0033-2909.114.3.494

Cudeck, R. (2000). In Tinsley, H. E. A. & Brown, S D. (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp.265–296). San Diego, CA: Academic. doi: https://doi.org/10.1016/B978-012691360-6/50011-2

Dallimore, E. J., Hertenstein, J. H., & Platt, M. B. (2004). Classroom participation and discussion effectiveness: Student-generated strategies. *Communication Education*, *53*(1), 37–41. doi: 10.1080/0363452032000135805

Dallimore, E. J., Hertenstein, J. H., & Platt, M. B. (2006). Nonvoluntary class participation in graduate discussion courses: Effects of grading and cold calling. *Journal of Management Education*, *30*(2), 354–377. doi: 10.1177/1052562905277031

Dallimore, E. J., Hertenstein, J. H., & Platt, M. B. (2012). Impact of cold-calling on student voluntary participation. *Journal of Management Education*, *37*(3), 305–341. doi: 10.1177/1052562912446067

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Wiley.

Davis, S. F., & Kring, J. P. (2001). A model for training and evaluating graduate teaching assistants. *College Student Journal*, *35*, 45–51.

deWinstanley, P. A., & Bjork, R. A. (2002). Successful lecturing: Presenting information in ways that engage effective processing. *New Directions for Teaching and Learning*, *89,* 19–32.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

Dweck, C. S. (2007). Boosting achievement with messages that motivate. *Education Canada*, *47*(2), 6–10.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273.

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. doi: 10.1187/cbe-14-06-0095

Englert, C. S., & Sugai, G. (1983). Teacher training: Improving trainee performance through peer observation and observation system technology. *Teacher Education and Special Education*, *6*(1), 7–17.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *Psychological Record*, *52,* 187–201.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. doi: 10.1037/0033-295X.100.3.363

Farrell, J. J., Moog, R. S., & Spencer, J. N. (1999). A guided inquiry general chemistry course. *Journal of Chemical Education*, *76*(4), 570.

Fedor, D. B., & Buckley, M. R. (1987). Providing feedback to organizational members: A reconsideration. *Journal of Business and Psychology*, *2*(2), 171–182.

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the*

National Academy of Sciences USA, *111*(23), 8410–8415. doi: 10.1073/pnas.1319030111

Gamer, M., Lemon, J., & Singh, I. F. P. (2015). Irr: Various coefficients of interrater reliability and agreement (R package, Version 0.84).

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Gormally, C., Evans, M., & Brickman, P. (2014). Feedback about teaching in higher ed: Neglected opportunities to promote change. *CBE—Life Sciences Education*, *13*(2), 187–199. doi: 10.1187/cbe.13-12-0235

Grossman, P., Hammerness, K., & Mcdonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching: Theory and Practice*, *15*(2), 273–289. doi: 10.1080/13540600902875340

Hardré, P. L. (2003). The effects of instructional training on university teaching assistants. *Performance Improvement Quarterly*, *16*(4), 23–39.

Hastie, T., & Qian, J. (2014). Glmnet vignette. 1–30. Retrieved June 9, 2016, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.455.6350&rep=rep1&type=pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi: 10.3102/003465430298487

Hothorn, A. T., Zeileis, A., Millo, G., & Mitchell, D. (2011). Package "Lmtest."

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology*, *90*(4), 677–691.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology*, *93*(1), 59–69.

Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Sage: Thousand Oaks, CA.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*, Berrett-Koehler: Oakland, CA.

Komarraju, M. (2008). A social-cognitive approach to training teaching assistants. *Teaching of Psychology*, *35*(4), 327–334. doi: 10.1080/00986280802374344

Lampert, M., Beasley, H., Ghousseini, H., Kazemi, E., & Franke, M. (2010). Using designed instructional activities to enable notices to manage ambitious mathematics teaching. In Stein, M. K. & Kucan, L. (Eds.), *Instructional explanations in the disciplines* (pp. 129–141). New York: Springer. doi: 10.1007/978-1-4419-0594-9

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., … Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, *64*(3), 226–243. doi: 10.1177/0022487112473837

Lemov, D. (2010). *Teach like a champion: 49 techniques that put students on the path to college* (1st ed.). San Francisco, CA: Jossey-Bass.

Liden, R. C., & Mitchell, T. R. (1985). Reactions to feedback: The role of attributions. *Academy of Management Journal*, *28,* 291–308. doi: 10.2307/256202

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

Mahalanobis, P. C. (1930). On the generalized distance in statistics. *Journal of the Asiatic Society of Bengal*, *26,* 541–588.

Marbach-Ad, G., Schaefer, K. L., Kumi, B. C., Friedman, L. A., Thompson, K. V., & Doyle, M. P. (2012). Development and evaluation of a prep course for chemistry graduate teaching assistants at a research university. *Journal of Chemical Education*, *89,* 865–872. doi: 10.1021/ed200563b

McGraw, K. O., & Wong, S. P. (1996). "Forming inferences about some intraclass correlations coefficients": Correction. *Psychological Methods*, *1*(4), 390–390. doi: 10.1037/1082-989X.1.4.390

Nielsen, K. L., Hansen-Nygård, G., & Stav, J. B. (2012). Investigating peer instruction: How the initial voting session affects students' experiences of group discussion. *ISRN Education*, *2012,* 1–8.

O'Reilly, M., & Renzaglia, A. (1994). A systematic approach to curriculum selection and supervision strategies: A preservice practicum supervision model. *Teacher Education and Special Education*, *17*(3), 170–180. doi: 10.1177/088840649401700304

Pentecost, T. C., Langdon, L. S., Asirvatham, M., Robus, H., & Parson, R. (2012). Graduate teaching assistant training that fosters student-centered instruction and professional development. *Journal of College Science Teaching*, *41,* 68–75.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Team, R. C. (2014). Nlme: Linear and nonlinear mixed effects models.

Prieto, L. R., & Scheel, K. R. (2008). Teaching assistant training in counseling psychology. *Counselling Psychology Quarterly*, *21*(1), 49–59. doi: 10.1080/09515070801900780

R Core Team (2014). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Reeves, T. D., Marbach-Ad, G., Miller, K. R., Ridgway, J., Gardner, G. E., Schussler, E. E., & Wischusen, E. W. (2016). A conceptual framework for graduate teaching assistant professional development evaluation and research. *CBE—Life Sciences Education*, *15*(2), es2. doi: 10.1187/cbe.15-10-0225

Revelle, W. (2015). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University.

Rezler, A. G., & Anderson, A. S. (1971). Focused and unfocused feedback and self-perception. *Journal of Educational Research*, *65*(2), 61–64. doi: 10.1080/00220671.1971.10884253

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In Florida Association of Institutional Research, held February 1–3, 2006, Cocoa Beach, FL (pp. 1–33).

Roney, C. J. R., Higgins, E. T., & Shah, J. (1995). Goals and framing: How outcome focus influences motivation and emotion. *Personality and Social Psychology Bulletin*, *21*(11), 1151–1160.

Schussler, E. E., Read, Q., Marbach-Ad, G., Miller, K., & Ferzli, M. (2015). Preparing biology graduate teaching assistants for their roles as instructors: An assessment of institutional approaches. *CBE—Life Sciences Education*, *14,* ar31. doi: 10.1187/cbe-14-11-0196

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi: 10.1037/0033-2909.86.2.420

Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE Life Sciences Education*, *10,* 55–63.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., & Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide No. 8. *Medical Teacher*, *28*(6), 497–526. doi: 10.1080/01421590600902976

Stone, D., & Heen, S. (2014). *Thanks for the feedback: The science and art of receiving feedback well*, New York: Viking.

Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, *13*(1), 41–48. doi: 10.1187/cbe-13-07-0136

Torchiano, M. (2015). Effsize: Efficient effect size computation (R package, Version 0.5.4).

Turpen, C., & Finkelstein, N. D. (2010). The construction of different classroom norms during peer instruction: Students perceive differences. *Physical Review Special Topics–Physics Education Research*, *6*(2), 20123.

Willoughby, T., Wood, E., McDermott, C., & McLaren, J. (2000). Enhancing learning through strategy instruction and group interaction: Is active generation of elaborations critical? *Applied Cognitive Psychology*, *14*(1), 19–30.

Wyse, S. A., Long, T. M., & Ebert-May, D. (2014). Teaching assistant professional development in biology: Designed for and driven by multidimensional data. *CBE—Life Sciences Education*, *13,* 212–223. doi: 10.1187/cbe.13-06-0106

Young, S. L., & Bippus, A. M. (2008). Assessment of graduate teaching assistant (GTA) training: A case study of a training program and its impact on GTAs. *Communication Teacher*, *22*(4), 116–129. doi: 10.1080/17404620802382680

Zeichner, K. (2012). The turn once again toward practice-based teacher education. *Journal of Teacher Education*, *63*(5), 376–382. doi: 10.1177/0022487112445789