Does Context Matter? Convergent and Divergent Findings in the Cross-Institutional Evaluation of Graduate Teaching Assistant Professional Development Programs

Todd D. Reeves,[†‡*] Laura E. Hake,[§] Xinnian Chen,[‖¶] Jennifer Frederick,[¶#] Kristin Rudenga,[¶#@] Larry H. Ludlow,[†] and Clare M. O'Connor[§]

[†]Department of Measurement, Evaluation, Statistics, and Assessment and [§]Biology Department, Boston College, Chestnut Hill, MA 02467; [‖]Department of Physiology and Neurobiology, University of Connecticut, Storrs, CT 06269; [#]Center for Teaching and Learning, Yale University, New Haven, CT 06520

ABSTRACT

Graduate teaching assistants (GTAs) play important instructional roles in introductory science courses, yet they often have little training in pedagogy. The most common form of teaching professional development (PD) for GTAs is a presemester workshop held at the course, department, or college level. In this study, we compare the effectiveness of presemester workshops at three northeastern research universities, each of which incorporated scientific teaching as the pedagogical content framework. The comparison of GTA PD program outcomes at three different institutions is intended to test theoretical assertions about the key role of contextual factors in GTA PD efficacy. Pretest and posttest surveys were used to assess changes in GTA teaching self-efficacy and anxiety following the workshops, and an objective test was used to assess pedagogical knowledge. Analysis of pretest/posttest data revealed statistically significant gains in GTA teaching self-efficacy and pedagogical knowledge and reductions in teaching anxiety across sites. Changes in teaching anxiety and self-efficacy, but not pedagogical knowledge, differed by training program. Student ratings of GTAs at two sites showed that students had positive perceptions of GTAs in all teaching dimensions, and relatively small differences in student ratings of GTAs were observed between institutions. Divergent findings for some outcome variables suggest that program efficacy was influenced as hypothesized by contextual factors such as GTA teaching experience.

INTRODUCTION

Professional development (PD) has been shown to positively affect the characteristics of teachers, leading to improvements in both teaching practices and student outcomes (reviewed in Reeves *et al.*, 2016). Although continuing PD is now an established feature of K–12 education, PD is a much less consistent feature in the preparation of faculty to teach undergraduate courses in science, technology, engineering, and mathematics (STEM). Faculty in STEM disciplines often report that they received little or no formal pedagogical training outside the training that prepared them for roles as graduate teaching assistants (GTAs; Tanner and Allen, 2006). The effectiveness of PD for GTAs therefore has far-reaching implications for student learning. In addition to potential future roles as faculty members, GTAs are directly responsible for much of the instruction in introductory STEM courses at many research universities (Sundberg *et al.*, 2005; Schussler *et al.*, 2015). Fortunately, the importance of PD programs for GTAs is increasingly recognized by universities. A 2013 survey of 71 biology programs (Schussler *et al.*, 2015) revealed that 96% of institutions provided some kind of mandatory GTA training.

Elizabeth Schussler, *Monitoring Editor*
Submitted March 6, 2017; Revised November 6, 2017; Accepted November 8, 2017
CBE Life Sci Educ March 1, 2018 17:ar8
DOI:10.1187/cbe.17-03-0044

[¶]These authors contributed equally to the project and are listed alphabetically.

Present addresses: [†]Department of Educational Technology, Research and Assessment, Northern Illinois University, De Kalb, IL 60115; [@]Kaneb Center for Teaching and Learning, University of Notre Dame, Notre Dame, IN 46556.
*Address correspondence to: Todd D. Reeves (treeves@niu.edu).

By their very nature, GTA course assignments are transitory, and introductory laboratory courses are often taught by GTAs with varied research and teaching experiences. GTAs must also balance their teaching responsibilities with their research activities, and they may encounter some pressure to minimize the time spent on teaching (Austin, 2002; Gardner and Jones, 2011; DeChenne *et al*., 2012; Hardré and Burris, 2012). The literature suggests that GTAs in STEM disciplines can gain pedagogical knowledge with a variety of different approaches to PD. GTAs have been reported to gain knowledge about teaching and learning by participating in presemester workshops (Pentecost *et al.*, 2012; Bauer *et al.*, 2013), course-specific weekly meetings (Wyse *et al.*, 2014), mentoring relationships with experienced teachers (Bond-Robinson and Rodriques, 2006; Page *et al.*, 2011), and seminar courses on pedagogy (Hammrich, 2001; Baumgartner, 2007; Sales *et al.*, 2007; Miller *et al.* 2008; Marbach-Ad *et al.*, 2012). Primary responsibility for PD programs may be relegated to course instructors, while departments or centers for teaching and learning may assume this responsibility at other institutions.

The existing literature on the design of GTA PD includes reports on program content, structure, and activities. With respect to PD *content*, PD programs have covered topics such as assessment, pedagogical methods, policies and procedures, and multicultural issues (e.g., Luft *et al*., 2004; Prieto *et al.*, 2007). With respect to PD *structure*, GTA PD programs have often taken the form of a onetime workshop (Gardner and Jones, 2011; Schussler *et al.*, 2015). Other designs or design elements, such as GTA mentoring or receipt of teaching feedback, are much less common (Austin, 2002; DeChenne *et al.*, 2012). With respect to PD *activities*, prior research has examined the effectiveness of activities such as microteaching (Gilreath and Slater, 1994) and teaching skits (Marbach-Ad *et al.*, 2012). The literature suggests that some PD design variables (e.g., training length) positively enhance changes in GTA cognition (e.g., Prieto and Meyers, 1999; Hardré, 2003; Young and Bippus, 2008).

Scientific Teaching

A prominent pedagogical content framework for undergraduate STEM teaching–related PD, including the PD of GTAs, is that of scientific teaching (ST; Handelsman *et al*., 2007; Couch *et al*., 2015). The ST framework distills evidence-based educational practices and principles into an abbreviated format for college STEM instructors who have little or no pedagogical training, drawing instead on their experiences as scientists. Toward this end, ST draws parallels between lab research approaches and teaching. Instructors engaged in ST draw from the available evidence base to hypothesize teaching approaches or interventions that will be effective, and subsequently collect and analyze student outcome data to test their hypotheses.

ST emphasizes three core principles: active learning, assessment, and diversity. ST has been used effectively as the curriculum at the National Academy of Sciences Summer Institutes (now Summer Institutes on Scientific Teaching) to train more than 1500 undergraduate STEM educators. Research has shown lasting changes in participants' reported instructional practices over a 2-year period (Pfund *et al*., 2009), although self-reported instructional practices may not necessarily cohere with those observed through observation (Ebert-May *et al*., 2011). ST has also been used as the framework for a seminar course for graduate students who aspired to future faculty positions (Miller *et al.*, 2008). There are only a few published examples of ST being explicitly integrated into PD for GTAs assigned to STEM courses. In a pilot study, Chen *et al.* (2013) conducted an ST-based workshop for GTAs who were teaching a laboratory course in anatomy and physiology (A&P). Students in the A&P lab class reported higher learning gains than students in a similar A&P lab class in which GTAs had not received ST training, suggesting that ST training improved instructional practices. Course observers also noted enhanced use of learner-centered approaches (e.g., group discussion) in sections led by ST-trained GTAs. The results could only be considered preliminary, however, due to differences in the course structures and student populations. Wyse *et al.* (2014) incorporated ST training for GTAs as part of a major restructuring of an introductory lab class. In that study, analysis of structured observation data indicated that GTAs in the reformed course demonstrated improved teaching skills. In the present study, we report on and compare the effectiveness of three PD programs for GTAs assigned to introductory STEM courses—all of which infused the principles of ST (Handelsman *et al*., 2007).

Conceptual Framework

The present study is grounded in the recent conceptual framework advanced by Reeves *et al.* (2016) for GTA PD research and evaluation studies (see Figure 1), which was itself informed by earlier-cited work by DeChenne *et al*. (2015) and Hardré (2003). The framework articulates three key PD program outcomes variables: GTA cognition; GTA teaching practice; and undergraduate student outcomes. The PD outcome of GTA cognition pertains to cognitive changes in GTAs' knowledge, skills, and attitudes toward or beliefs about teaching that directly result from the GTA PD. GTA teaching practice concerns GTAs' behavior related to planning, instruction, and assessment. Undergraduate student outcomes center on gains in knowledge and skills achieved by students of GTAs, as well as more distal student outcomes such as retention and graduation. The framework posits that these outcome variable categories are linearly (sequentially) related, in that PD directly impacts GTA cognition, which in turn impacts GTA teaching practice, which then impacts undergraduate student outcomes. It is also worth pointing out that the relationships posited are probabilistic rather than wholly deterministic; indeed, prior K–12 research indicates that a teacher's cognition (e.g., attitudes) is not always in alignment with his or her practice (Crawford, 1999; Luft *et al.*, 2011).

Of the three GTA program outcomes within the framework, the first (GTA cognition) has been examined most often in GTA evaluation and research. For instance, various studies have reported evidence for a relationship between participation in PD and GTA knowledge, self-efficacy, and/or anxiety (Hardré, 2003; Harris *et al*., 2009; DeChenne *et al*., 2012, 2015; Bauer *et al*., 2013; Pelton, 2014). The relationship between PD and GTA instructional planning, delivery, and assessment (GTA teaching practice) has also been examined (Baumgartner, 2007; Marbach-Ad *et al.*, 2012; Chen *et al.*, 2013; Wyse *et al.*, 2014). In several posttest-only studies, GTAs reported that their instructional practices improved with the training; or that they implemented practices taught during the training (Baumgartner, 2007; Hardré and Burris, 2012; Pentecost *et al*., 2012).

However, GTAs' actual implementation of training content is not always straightforward. Addy and Blanchard (2010)

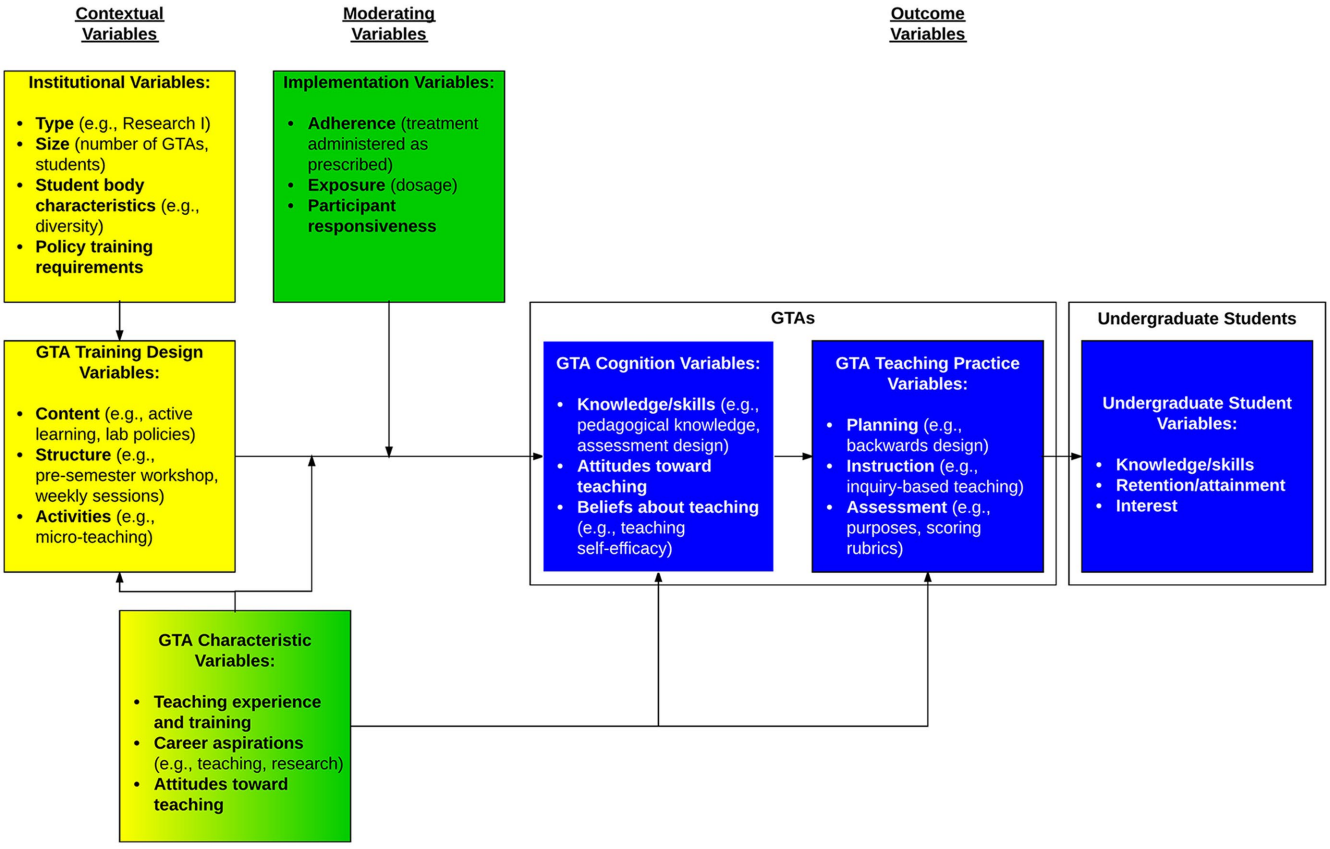

FIGURE 1. Conceptual framework for research and evaluation related to GTA PD. Framework shows relationships among graduate teaching assistant (GTA) teaching professional development (PD) outcome variables (blue), contextual variables (yellow), and moderating variables (green). The framework contains three main categories of outcomes at two levels: GTA and undergraduate student. These impacts (blue) are linearly (sequentially) related: GTA cognition; GTA teaching practice; and undergraduate student outcomes. GTA cognition pertains to GTAs' knowledge, skills, attitudes, or beliefs about teaching. GTA teaching practice concerns the GTAs' approach to planning, instruction, and assessment. Undergraduate student outcomes center on the GTAs' students' knowledge and skills, as well as more distal student outcomes such as retention and graduation. The framework supposes that GTA PD directly promotes changes in GTA cognition, which in turn impacts their instructional behavior (GTA teaching practice) and subsequent outcomes for undergraduates (undergraduate student outcomes). The framework contains three categories of contextual variables (yellow): GTA training design, institutional, and GTA characteristics. GTA training design variables pertain to the nature of the GTA training, and are hypothesized to drive the most direct outcomes of GTA PD: GTA cognition. Institutional and GTA characteristic variables are hypothesized to have effects on GTA training design. GTA characteristics are also hypothesized to directly impact GTA cognition (e.g., knowledge/skills, attitudes, and beliefs), as well as GTA teaching practice, independent of PD (Enochs and Riggs, 1990). The final category of variables in the framework are moderating variables, that is, variables that impact or modify the relationship between two other variables (in this case, the relationship between GTA training design and GTA cognition). The framework invokes Dane and Schneider's (1998) implementation concepts of program adherence, exposure, and participant responsiveness. The framework secondly includes GTA characteristics as moderators of the relationship between GTA training design and GTA cognition, given that some GTAs may change more than others during PD. GTA characteristics serve as both contextual variables and moderating variables in the model. Reproduced with permission from Reeves *et al.* (2016). We include GTA attitudes toward teaching as both an example GTA characteristic variable and an example outcome variable (specifically GTA cognition variable), because GTA attitudes toward teaching can have effects on GTA training design and the relationship between GTA training design and GTA cognition and also represent a potential cognitive change in GTAs that directly results from PD.

observed inconsistent effects of PD when they assessed the teaching practices of biology GTAs using the Reformed Teaching Observation Protocol (Sawada *et al.*, 2002). Some GTA teaching practices were in accord with the specific GTA PD program, while others were not. The authors suggested that the constraints of teaching within a laboratory context with a highly procedural structure might have explained this discrepancy. Bond-Robinson and Rodriquez (2006) demonstrated large GTA-to-GTA variation even after participation in PD.

The undergraduate student outcomes of GTA PD are less often examined, although the Chen *et al.* (2013) study cited earlier provided some evidence for higher learning gains among students taught by trained GTAs. In the present study, we examine the relationship between PD participation and changes in two outcome categories within the framework: GTA cognition (by way of self-reported self-efficacy, anxiety, and pedagogical knowledge measures) and GTA practice (by way of student ratings of instruction).

The GTA PD framework also highlights the importance of contextual variables (Tanner and Allen, 2006; Gardner and Jones, 2011; DeChenne *et al.*, 2012, 2015; Schussler *et al.*, 2015) that may impact the effectiveness of GTA PD and that are necessary to consider when comparing results across programs. These include variables associated with GTA training design, institutional setting, and GTA characteristics. In the framework, GTA training design variables (e.g., content, structure, and activities) affect the degree to which PD changes GTA cognition. Indeed, there is considerable variation in the design of GTA PD programs, which can differ significantly in their duration (e.g., number of hours) and structure (e.g., presemester workshop; Schussler *et al.*, 2015). The framework also includes institutional variables such as institutional type, culture, size, student body characteristics, and policy training requirements, which may influence the nature of the PD provided to GTAs (Park, 2004; Schussler *et al.*, 2015). GTA characteristic variables in the framework reflect the fact that the characteristics of GTAs themselves (e.g., prior teaching experience) can impact the design of GTA training.

Finally, the framework also includes two categories of moderator variables, or variables that affect or are related to relationships among other variables: implementation variables and GTA characteristic variables. Inclusion of implementation variables as moderators reflects the fact that the impact of a given GTA program hinges on its successful implementation (Stains and Vickrey, 2017). The inclusion of GTA characteristic variables as moderators reflects the fact that GTAs may respond differently to a training, based on their characteristics, such as prior teaching experience (Addy and Blanchard, 2010; Marbach-Ad *et al.*, 2012; DeChenne *et al.*, 2015). Indeed, studies by Hardré and Chen (2005) and French and Russell (2002) found that prior teaching knowledge and/or experience were related to implementation of training content (implying differential changes in their cognition). It is the latter hypothesis of the framework—that contextual variables such as GTA characteristics can affect the degree to which a given PD program promotes successive changes in GTA cognition and teaching practice—that primarily motivates the present study.

The Importance of Multidimensional Evaluation of GTA PD

In response to arguments for multidimensional evaluation of GTA PD (Wyse *et al.*, 2014) and theory on how GTA PD operates (Reeves *et al.*, 2016), several outcome variables were considered in this analysis of ST-infused GTA PD programs. In addition to GTA pedagogical knowledge and instructional practice, we particularly investigated changes in GTA teaching self-efficacy and teaching anxiety, given their importance in social cognitive theory. Social cognitive theory predicts that self-efficacy, defined as "beliefs in one's capabilities to organize and execute courses of action required to produce given attainments" (Bandura, 1997), affects behavior. A large body of K–12 research supports social cognitive theory, demonstrating a positive relationship between teacher self-efficacy and student achievement (e.g., Caprara *et al.*, 2006).[1] Other research has highlighted the importance of teacher beliefs, including self-efficacy beliefs, in science teaching specifically (Enochs and Riggs, 1990). The effect of PD on GTA teaching self-efficacy has been investigated in several studies with a variety of qualitative and quantitative methods, including interviews, surveys, pre/posttests, and/or observations (Hardré, 2003; Harris *et al.*, 2009; DeChenne *et al.*, 2012, 2015; Bauer *et al.*, 2013). Consistent with social cognitive theory, these studies generally found that self-efficacy was positively correlated with PD participation. Drawing from this literature, DeChenne *et al.* (2015) constructed a model that identified GTA PD, together with teaching experience and the departmental climate, as major sources of GTA teaching self-efficacy. While this study found that K–12 teaching experience rather than GTA teaching experience was associated with teaching self-efficacy, other studies have linked the latter with teaching self-efficacy as well (Prieto and Altmaier, 1994). Finally, based on her comparison of control and trained GTA groups, Hardré (2003) posited that GTA training affects knowledge, which in turn affects both teaching self-efficacy and instructional practice.

Anxiety about teaching has been demonstrably related to teaching behavior in K–12 settings (Coates and Thoreson, 1976), but has received less attention in research on GTA training programs. In one study of sociology GTAs, Pelton (2014) used precourse and postcourse evaluations to measure the effect of a teaching seminar on GTA anxiety. Sociology GTAs who completed the 6-week seminar reported both significant reductions in anxiety and increases in confidence. In another large quantitative survey of GTAs at a research university, Cho *et al.* (2011) identified and categorized a variety of GTA concerns that might produce anxiety. Chief GTA concerns included class control, external evaluation, teaching tasks, student impact, and time management. Several of these concerns were found to be associated with GTA confidence.

Present Study

In this study, we investigated how GTA training programs at three northeastern research universities prepared GTAs to teach sections of large introductory science classes. The PD programs at all three institutions were anchored in a common pedagogical content framework: scientific teaching (Handelsman *et al.*, 2007). All of the programs also entailed a presemester workshop, a popular model for GTA PD programs (Schussler *et al.*, 2015). Given cross-institutional differences in institutional characteristics and GTA characteristics, there were some variations in the designs of the three PD training programs (as is hypothesized by the GTA PD framework; Figure 1). In particular, we systematically compared the outcomes of these three GTA PD programs to understand the role of contextual characteristics in GTA PD impact. In doing so, this study affords testing of the GTA PD framework, specifically its assertion that the effects of participation in PD will likely vary as a function of contextual factors, notably the characteristics of GTAs. The study focused on the following research question: To what extent do changes in GTA self-efficacy, anxiety, and pedagogical knowledge, as well as GTA instructional practices, vary across ST GTA PD programs at three different institutions?

METHODS

We employed a pretest–posttest pre-experimental design involving treatment implementation at the three sites during the Fall semester of 2012 and the Spring semester of 2013. We used

[1]While Bandura's (1977) social cognitive theory also positions outcome expectancy—the belief that a behavior will promote a specific outcome—as another important determinant of behavior; that variable was not considered in the present study.

quantitative data collected from GTAs and undergraduate students to describe and compare changes in GTAs' teaching self-efficacy, teaching anxiety, and pedagogical knowledge during the PD programs, and GTAs' instructional practices related to the training. Each university's institutional review board approved the research. All data were maintained and are reported here anonymously.

Institutional Context

GTAs at all three sites were being prepared to teach sections of introductory laboratory courses. Each of the training programs began with a precourse teaching workshop of approximately 12 hours in length over 2–3 days. At least one instructor at each site had attended or facilitated a Summer Institute on Scientific Teaching, which provided a model for ST activities that could be incorporated into GTA training. Workshop directors from the three sites met twice and shared ideas in advance of the Fall training workshops, but there was no attempt to fully standardize the workshops. All workshops included training in the ST framework by emphasizing active learning, diversity, and assessment. Specific activities, however, varied among programs.

At Universities A and B, workshops were led by laboratory course instructors and the participants included GTAs with various degrees of experience. GTAs at University A were all doctoral students with previous teaching experience. More than 80% of the GTAs were in year 3 or later of their degree programs. Workshops were repeated at the beginning of the Spring semester, but returning GTAs were not required to repeat PD activities. Approximately 40% of the GTAs at University B were master's students, and the remainder were doctoral students, most of whom were in the first or second year of their degree programs. The PD program at University C was designed and administered by a centralized teaching and learning center for all incoming first-year graduate students in the chemistry department, by request from the department. GTAs at University C were subsequently assigned to different introductory courses in general and organic chemistry. Table 1 provides a comprehensive comparison of the characteristics of the GTA training programs, GTA teaching contexts, and institutions. Data on these contextual factors should help the reader interpret the programmatic comparisons reported here and may help others understand the applicability of the findings to their own GTA PD contexts.

University A. GTAs teach a single section of an introductory biology laboratory course that meets twice per week and involves a semester-long research investigation. Approximately half of the presemester workshop was devoted to pedagogy using the ST framework. The other half was devoted to the course research project. GTAs modeled student teams as they designed, conducted, and presented the results of experiments that students would perform during the semester. At weekly GTA meetings during the semester, a team of GTAs introduced the topic for week, incorporating learning objectives and student activities that they had designed for the topic.

University B. GTAs teach two laboratory courses of either the basic A&P course for pre–health majors or an enhanced A&P course designed for biology majors (Chen *et al.*, 2013). In Fall 2012 and Spring 2013, GTAs from both courses attended the same presemester workshop, which was primarily devoted to pedagogical instruction. Role playing was used to simulate lab dynamics. During the semester, pairs of GTAs developed two teachable units that were peer reviewed at weekly course meetings before their implementation in lab sections.

University C. GTAs are assigned to one of several general chemistry or organic chemistry laboratories, which are co-requisites for introductory lecture courses. In addition to leading their laboratory sessions, GTAs occasionally led discussion sections. GTAs were required to attend the presemester workshop, which included a general introduction to ST and evidence-based teaching practices, microteaching with peer feedback, and community-building activities. Each GTA also completed a written self-reflection about strengths and opportunities for improvement, which he or she revisited at the end of the term. Most of the laboratory courses also held weekly GTA meetings to discuss course-specific content and policies, student activities, and grading. The center for teaching and learning held follow-up meetings at midsemester and at the end of the term.

Participants

Eighty-one GTAs participated in the research project, 68 in Fall 2012 (84.0%) and 13 in Spring 2013 (16.0%).[2] The overall GTA sample composition was 59.7% male, 35.6% nonnative English speaker, and 26.0% international student, and the mean GTA age was 24.7 years (SD = 3.6). About 75% of the GTA sample had some form of prior teaching experience (37.0% in a lecture course, 52.1% in a laboratory course, and 15.1% in a combined lecture–laboratory course). Thirty-one percent indicated that they had received some prior formal training in how to teach. Table 2 presents characteristics of the GTA participants at each institution (all contextual factors in the GTA PD framework). As can be seen in the table, GTAs at University C were somewhat younger on average than GTAs at University A and University B; GTAs at University A had the most prior teaching experience (whereas GTAs at University C had the least, though prior teaching experience was still common), and GTAs at all three institutions had generally not received prior formal teacher training. The presence of salient differences among GTA characteristics at each site afforded testing of the role of contextual factors in GTA PD impact. Notably, 100% of GTAs trained during the study period at each institution participated in the study.

One thousand, six hundred and sixty-five undergraduate students from two of the three institutions also contributed study data via end-of-semester evaluations of their GTAs (70.7% in Fall 2012 and 29.3% in Spring 2013). In total, 19.0% of the students were from University A, and 81.0% were from University B. The 316 undergraduate student respondents at University A and 1349 undergraduate student respondents at University B constituted ~88 and 64%, respectively, of the total number of undergraduate students served by the GTAs at these institutions.

Instrumentation

Each semester, GTAs responded to a maximum of three surveys: 1) immediately before the training (henceforth the "pretest"); 2) immediately after the training (the "posttest"); and 3) at the end

[2]We excluded Spring 2013 data for three GTAs who participated in both the Fall 2012 and Spring 2013 training programs at University A. University C offered a single training program in Fall 2012.

TABLE 1. Comparison of GTA training programs, GTA teaching contexts, and institutions[a]

Characteristics	University A	University B	University C
GTA training program			
Content (e.g., active learning, lab policies)	Active learning, diversity, assessment, lab policies, course experiments	Active learning, diversity, assessment, lab policies, course experiments, classroom control, lab manual, logistics (i.e., GTA assignment, Blackboard)	Active learning, diversity, assessment, critical thinking, teaching problem solving, grading and feedback
Structure (e.g., presemester workshop, weekly sessions)	Presemester workshop, 12 hours total distributed across 2 days, followed by weekly GTA meetings	Presemester workshop, 10 hours total distributed across 2 days, followed by weekly GTA meetings and separate meetings with course instructors if presenting in a given week	Presemester workshop, 13 hours total distributed across 3 days, follow-up and end-of-semester PD meetings, and weekly course-specific GTA meetings for some GTAs
Activities (e.g., microteaching)	Lectures, design of lesson (objectives, activities), receipt of feedback on lesson	Design and demonstration of two student-centered teaching lessons (objectives, activities), receipt of feedback on lessons	Lectures, microteaching with peer feedback, community-building activities, self-reflection writing
Training provider/facilitator	Course instructors	Course instructors	Teaching center staff
Mandatory or optional	Mandatory	Mandatory	Mandatory
GTAs trained during study period	19	34	28
GTA teaching context			
Course	Investigations in Molecular Cell Biology	Human Anatomy and Physiology (A&P)/Enhanced A&P	Introductory Biology/Chemistry laboratories (several)
Credit hours	3	1	1
Enrolled undergraduate students	180 (Fall 2012); 180 (Spring 2013)	750/360 (Fall 2012); 690/302 (Spring 2013)	~500 (Fall 2012)
Number of sections	12	36/26	32
Undergraduate student level	Mainly sophomores	Sophomores to seniors	Freshmen and sophomores
Undergraduate student course enrollment during study period	360	2102	~500
GTA role(s) in course	Supervise one laboratory section of ~15 students, grade student assignments	Supervise two laboratory sections of between 14 and 21 students, grade student assignments	Supervise one or two laboratory sections of up to 16 students, or lead a discussion section of between 20 and 25 students for a lecture course
Institutional			
Type (control, Carnegie classification)	Private, doctoral university—higher research activity	Public, doctoral university—very high research activity	Private, doctoral university—highest research activity
Size (total number of students, total number of undergraduates, total number of graduate and professional students)	14,359 total; 9110 undergraduate; 4673 graduate and professional	30,256 total; 22,301 undergraduate; 7955 graduate and professional	12,385 total; 5505 undergraduate; 6880 graduate and professional
Student body characteristics (i.e., percent white, Black or African American, and Hispanic or Latino undergraduate students)	68% white; 5% Black or African American; 5% Hispanic or Latino	63% white; 6% Black or African American; 9% Hispanic or Latino	52% white; 8% Black or African American; 13% Hispanic or Latino
Policy training requirements	Yes	Yes	No

[a]Characteristics adapted from Reeves *et al.* (2016) framework. Owing to changes over time in GTA training programs, GTA teaching contexts, and institutions since data collection in 2012–2013, these data may not necessarily reflect current GTA training program, GTA teaching context, or institutional characteristics. Some data (i.e., student course-enrollment totals) are estimated from historical data.

of the semester (the "follow-up"). Of the 81 total GTAs, 65 (80.2%) completed the pretest, 70 (86.4%) completed the posttest, and 57 (70.4%) completed the follow-up.[3] Seventeen GTA participants at University A (89%), 17 GTA participants at University B (50%), and 27 GTA participants at University C (96%) completed *both* the pretest and the posttest, which constituted the primary data sources for answering our research question.

The pretest survey collected demographic, educational, and teaching-related data about the GTAs via closed-ended survey questions. As described in the following sections, the pretest and posttest surveys both included measures of GTA self-efficacy and anxiety about teaching, as well as pedagogical knowledge. Student surveys gathered data on the GTAs' instructional

[3]Insufficient institutional sample sizes at the follow-up time point precluded statistical power to incorporate end-of-semester GTA self-efficacy, anxiety, and pedagogical knowledge into the analysis. Future research should also examine GTA self-efficacy, anxiety, and pedagogical knowledge in the months and years following GTA training programs to examine the long-term effects of such programs.

TABLE 2. GTA participant characteristics by institution[a]

GTA characteristic	University A	University B	University C
Age (M, SD)	27.75, 3.02	25.88, 4.55	22.26, 1.29
Sex (percent)			
Male	56.25	47.06	59.26
Female	43.75	52.94	40.74
Prior teaching experience (percent)			
Yes	100.00	76.47	62.96
No	0.00	23.53	37.04
Prior formal teacher training (percent)			
Yes	29.41	35.71	26.92
No	70.59	64.29	73.08
Year in graduate program	First- to fifth-year biology graduate students	First- to fifth-year biology graduate students	First-year chemistry graduate students
Native English speaker (percent)			
Yes	70.59	47.06	74.07
No	29.41	52.94	25.93
International student status (percent)			
Yes	17.65	35.29	22.22
No	82.35	64.71	77.78

[a]GTA participant characteristics estimated based on the 17 GTAs at University A, 17 GTAs at University B, and 27 GTAs at University C who responded to corresponding GTA survey questions. We did not collect data on GTA career aspirations or attitudes toward teaching, per the Reeves *et al.* (2016) framework.

practices at Universities A and B only, because of feasibility challenges at University C. In general, the instrumentation used was developed by the researchers to bolster its alignment with the GTA PD program aims and teaching contexts (these data were collected during an evaluation study). One instrument, the Anxiety and Confidence in Teaching scale, was a modification of an established instrument, the Teaching Economic Literacy: Confidence and Anxiety scale. While the reliance on researcher-developed or researcher-adapted instruments may be considered an important limitation of this study, reliability and validity evidence is quite promising.

For our measures of GTA teaching self-efficacy, teaching anxiety, and pedagogical knowledge, we estimated reliability with Cronbach's alpha (α) or Kuder-Richardson formula 20. As other psychometric analyses, we also investigated test–retest reliability by examining the correlations between the measures administered at different time points; and, as validity evidence, how these measures (and gains in the measures) were related to one another within and across time (Reeves and Marbach-Ad, 2016). Given the exploratory nature of this study, we used listwise deletion to handle missing data. Copies of the assessment instruments are included in the Supplemental Material.

Self-Efficacy and Anxiety

The Anxiety and Confidence in Teaching scale, a generalized version of the Teaching Economic Literacy: Confidence and Anxiety scale (TELCA; Ludlow *et al.*, 2012), assessed self-efficacy and anxiety about teaching on the pretest and posttest.[4] Eighteen items were intended to measure self-efficacy for teaching (e.g., "When I am confronted with teaching a new concept in biology, I know I can cope with it"), and 12 items were intended to measure anxiety about teaching (e.g., "Thinking about teaching biology topics makes me anxious"). Across all items, the response format was a five-point Likert scale: 1 = strongly disagree; 2 = disagree; 3 = uncertain; 4 = agree; 5 = strongly agree. Common factor analysis previously provided evidence of a two-factor structure with one factor underlying the self-efficacy item responses and another underlying the anxiety item responses (correlation between factors was –0.62; Ludlow *et al.*, 2012). The original TELCA scale was adapted for use in this study by supplanting language related to economics with language appropriate for the respective context (e.g., biology). In this study, the internal score consistencies of the self-efficacy and anxiety scales were very high, 0.94 and 0.97 at pretest and 0.90 and 0.95 at posttest, respectively. The correlations (p values < 0.001) between the pretest and posttest measures were 0.85 (self-efficacy) and 0.75 (anxiety), respectively. For both of these measures, we took the mean of the corresponding items at a given time point to yield a score for each person.

TELCA was selected as the measure of self-efficacy for teaching on the basis of its alignment with the GTA PD programs and teaching contexts studied and the soundness of the available reliability and validity evidence concerning its scores. At the time of the evaluation, TELCA was also an active focus of one evaluation team member's (L.H.L.'s) research program. Certainly, other self-report instruments designed to measure teacher self-efficacy among GTAs could have been used, such as the Self-Efficacy Toward Teaching Inventory—Adapted (SETI-A; Prieto and Altmaier, 1994). However, select item content from the SETI-A was deemed misaligned with the GTA PD programs and teaching contexts considered, in that some items referenced facets of teaching practice such as preparing teaching materials

[4]While the TELCA instrument includes the term "confidence" in its title, its "confidence" subscale is not intended to measure confidence/self-confidence in the colloquial or general sense. The subscale is instead intended as a task-specific, self-efficacy measure (self-efficacy related to teaching). In addition, the TELCA is intended to measure only self-efficacy and not outcome expectancy, another key belief in Bandura's (1977) social cognitive theory.

and identifying course objectives. A similar argument motivated the use of the TELCA as a measure of anxiety about teaching. Other self-efficacy instruments, such as the STEM GTA-Teaching Self-Efficacy Scale (DeChenne *et al*., 2012), were unavailable at the time of this research.

Pedagogical Knowledge

A researcher-developed pedagogical knowledge objective test, embedded in both the pretest and posttest surveys, featured five dichotomously scored multiple-choice items aligned with the ST framework (Handelsman *et al.*, 2007). The objective test was "researcher developed" in the sense that it was designed by the research team specifically for use in the present study, because no existing instrument was available to measure knowledge of core facets of the ST pedagogical framework. The pedagogical knowledge test items pertained to active learning (2), behavioral objectives, assessment, and Bloom's taxonomy. For example, the item about assessment was, "Which of the following is *not* a strategy used to ensure that grading is fair? A) De-identify work products; B) Consider each student's typical performance (keyed response); C) Score papers in random order; and D) Use a rubric or scoring guide." The item about Bloom's taxonomy was: "Which of the following is a reasonable ordering in terms of cognitive complexity (ordered from *least to most* complex)? A) Factual knowledge, knowledge application, conceptual understanding, analytical thinking; B) Factual knowledge, conceptual understanding, knowledge application, analytical thinking (keyed response); C) Conceptual understanding, factual knowledge, knowledge application, analytical thinking; and D) Conceptual understanding, factual knowledge, analytical thinking, knowledge application." The systematic design of items to represent key facets of the ST framework constitutes content-related evidence for the validity of inferences drawn from the instrument (Reeves and Marbach-Ad, 2016). For the pedagogical knowledge measures, we took the sum of each item score at a given time point to yield pretest and posttest pedagogical knowledge scores for each person.[5]

Student Surveys

In two of three research sites, undergraduate students completed post-semester surveys about their GTAs. Students provided ratings of instruction in response to 19 researcher-developed items. A researcher-developed instrument was necessary to solicit evidence related to *both* ST practices specifically and the general teaching quality and pedagogy dimensions of interest to the PD programs' stakeholders. Without these constraints related to the inquiry's scope and purpose, alternative instruments such as the College and University Classroom Environment Inventory (Treagust and Fraser, 1986) or the Questionnaire of Teacher Interaction (Coll *et al.*, 2002) could have been used, as they were by Kendall and Schussler (2013). Nevertheless, the student ratings of instruction instrument was systematically designed to represent a variety of general aspects of teaching quality and pedagogy, derived from theory and research on effective teaching (e.g., Brophy, 1986), and specific aspects of instructional quality represented within the ST framework (Handelsman *et al*., 2007). In particular, the items were broadly organized according to the following categories: general teaching quality and pedagogy (e.g., "My TA articulated the goals for student learning"), assessment practices (e.g., "My TA evaluated our work fairly and impartially"), differentiation and inclusive teaching practices (e.g., "My TA knew how to teach students with different backgrounds, needs, and interests"), and active-learning practices (e.g., "My TA implemented class exercises that were interesting and stimulating"). Given the thrust of the training programs, five questions pertained to GTAs' use of discussion questions (e.g., "My TA asked discussion questions that require students to think"). Explicit mapping of items to the training content framework constitutes content-related validity evidence for the inference drawn from the student ratings of instruction (Reeves and Marbach-Ad, 2016). For all items, the response format was a Likert scale: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, and 5 = strongly agree.

As our focus was largely on documentation of differential GTA PD outcomes as a function of contextual factors, we opted to use student ratings of instruction as opposed to systematic observations of teaching. This design choice was primarily made on the basis of feasibility, in light of the fact that technically sound observations are extremely resource-intensive in terms of personnel, time, and cost (especially in multi-institutional studies). While student ratings of instruction constitute an indirect (albeit efficient) measure of GTA instructional practices, there is evidence that such ratings are moderately correlated with student learning outcomes (Cohen, 1981), which is another form of validity evidence (Reeves and Marbach-Ad, 2016). Similarly, Kendall and Schussler (2013) found that such ratings can differentiate more and less experienced instructors. The use of student ratings of instruction—replete with associated score reliability and validity threats—nevertheless constitutes a critical limitation of the present study (Kulik, 2001; Spooren *et al.*, 2013).

We expected to reduce these ratings to a smaller set of GTA instructional subdimensions (e.g., diversity, assessment), but a common factor analysis with principal axis factoring indicated only a single underlying factor. Therefore, we analyzed institutional and GTA differences in instruction overall, by taking the mean of the set of 19 student ratings of instruction $\alpha = 0.97$. As a supplemental analysis, we also examined institutional and GTA differences in each item separately to examine particular aspects of GTA instruction. The latter analysis afforded fine-grained exploration of GTA instruction differences in relation to GTA training program, as well as exploration of facets of instruction on which GTAs were most variable. The results of the common factor analysis also provide validity evidence, namely that the set of 19 student ratings of instruction tap a common instructional quality construct.

Analytic Approach

We used quantitative methods for data analysis. With our quantitative data, we began by computing descriptive statistics—either means and standard deviations or counts

[5]The internal consistency (lower bound of reliability) of the pedagogical knowledge test was very low (–0.24 at pretest and –0.07 at posttest); however, the correlation between the pretest and posttest measures was 0.41 ($p < 0.01$). Poor internal consistency may be explained on account of sample homogeneity with respect to the pedagogical knowledge construct, multidimensionality, and/or other sources of random error.

and percentages, depending on the nature of the data. These analyses allowed us to understand the distributional properties of the variables and to describe our sample(s) at fixed and/or successive points in time (e.g., pedagogical knowledge, teaching anxiety, teaching self-efficacy, and instructional practices).

We also used inferential statistical methods. First, we used mixed-model repeated-measures analyses of variance (ANOVAs) to examine pretest–posttest changes in self-efficacy, anxiety about teaching, pedagogical knowledge, and differential changes by institution. The full models were estimated to include the within-subjects time variable, the between-subjects institution variable, and their interaction. Second, using fixed-effect ANOVAs, we examined end-of-semester differences in each GTA instructional practice item by institution. Third, to contextualize the institutional fixed effects, we estimated GTA random effects in each instructional practice rating to examine their variance across all GTAs. We also similarly examined institutional and GTA differences in instruction overall. Partial eta-squared $\left(\eta_p^2\right)$ values were taken as evidence of practical significance, given the very large student sample size. Cohen's (1988) *d* was used to report mean differences (e.g., pretest–posttest self-efficacy gains) on a standard metric.

Limitations

The present study had a number of limitations that might threaten its external and internal validity. Our results should also be interpreted in light of the institutional contexts at which we collected data, all research institutions in the Northeast; the characteristics of these institutional sites (i.e., research cultures) might constrain generalization from our work. In a related vein, in general, the GTAs in our study were from traditional scientific disciplines and served as GTAs in introductory science courses. The structured, standardized, and procedural nature of some laboratory courses might to some extent limit the GTAs' opportunities to implement knowledge/skills learned during a workshop (e.g., discussion questioning). While the overall research participation and survey response rates were rather high, nonresponse (particularly at University B) too might mean that the findings are somewhat unreliable.

In terms of internal validity specifically, this study had a number of common research design and measurement limitations. In particular, possible threats to internal validity that we were unable to address through design or analysis include testing, maturation, and history. For example, examination of GTA instructional practices at the end of a semester might reflect both the influence of training *and* experience. Threats related to the instrumentation used may also be present (e.g., reliability and validity), although it bears noting the high reliabilities of the teaching self-efficacy and anxiety measures, and the objective nature of the pedagogical knowledge measure. Of course, the study's pre-experimental design necessarily precludes any causal inferences. Although the primary contribution of our manuscript centers on the role of contextual factors in GTA PD, we offer our findings with these caveats in mind.

RESULTS

Pretest and posttest self-efficacy, anxiety, and pedagogical knowledge measures for each site are summarized in Table 3.

TABLE 3. Summary of changes in self-efficacy, anxiety, and knowledge by institution

		Pretest		Posttest		
Group	N^a	M	SD	M	SD	d^b
Self-efficacy						
Overall	58	3.79	0.51	3.97	0.54	0.34
University A	17	3.68	0.55	3.75	0.65	0.11
University B	14	4.03	0.43	4.04	0.46	0.02
University C	27	3.74	0.50	4.07	0.48	0.68
Anxiety						
Overall	57	2.16	0.77	1.83	0.82	–0.42
University A	17	2.04	0.71	1.87	0.85	–0.20
University B	13	2.06	1.03	1.96	0.85	–0.10
University C	27	2.30	0.67	1.74	0.81	–0.73
Knowledge						
Overall	51	2.96	0.98	3.41	0.94	0.47
University A	10	2.90	0.74	3.00	0.94	0.12
University B	14	2.29	0.99	2.79	0.80	0.55
University C	27	3.33	0.88	3.89	0.75	0.68

[a]*N* = group size.
[b]Cohen's *d* standardized mean difference for dependent means calculated using Wilson's (2001) Excel macro.

Self-Efficacy

Repeated-measures ANOVA showed an overall difference in the self-efficacy means over time, such that the posttest mean was higher, F (1, 55) = 16.20, $p < 0.001$, $\eta_p^2 = 0.23$. There was no main effect of institution, $F(2, 55) = 1.69$, $p = 0.19$. However, the interaction between time and institution was significant, $F(2, 55) = 9.92$, $p < 0.01$, $\eta_p^2 = 0.27$, which implies differential growth by institution. The overall standardized mean difference (*d*) was 0.34, which is small to moderate in magnitude (Cohen, 1988). However, GTAs at University C reported 0.68 of a SD change in self-efficacy from immediately before to immediately after the workshop, whereas GTAs at University B reported almost no change.

Anxiety

Repeated-measures ANOVA showed an overall difference in the anxiety means over time, such that the posttest mean was lower, $F(1, 54) = 13.94$, $p < 0.001$, $\eta_p^2 = 0.21$. There was no main effect of institution, $F(2, 54) = 0.04$, $p = 0.96$. However, the interaction between time and institution was significant, $F(2, 54) = 4.54$, $p < 0.05$, $\eta_p^2 = 0.14$, which again implies differential GTA growth by institution. The overall reported change (decrease) in anxiety amounted to 0.42 of an SD, which is not trivial, especially for a relatively brief training. GTAs at University C reported the largest decrease in anxiety ($d = –0.73$), and GTAs at University B reported the least ($d = –0.10$).

Pedagogical Knowledge

Repeated-measures ANOVA showed an overall difference in the knowledge means over time, such that the posttest mean was higher, $F(1, 48) = 5.79$, $p < 0.05$, $\eta_p^2 = 0.11$. There was a main effect of institution, $F(2, 48) = 12.90$, $p < 0.001$, $\eta_p^2 = 0.35$, such that pedagogical knowledge differed across institutions regardless of time point. However, the interaction between time and institution was not significant, $F(2, 48) = 0.71$, $p = 0.50$,

meaning that the degree of GTA knowledge change across the three institutions was statistically indistinguishable. The magnitude of the overall increase in knowledge ($d = 0.47$) was somewhat larger than reported changes in self-efficacy (0.34) and anxiety (–0.42), with the largest changes at Universities B and C, and the smallest change at University A.

Instructional Practice

Table 4 summarizes the student ratings of GTAs at Universities A and B. All mean agreement ratings were between "agree" (4) and "strongly agree" (5), indicating favorable perceptions of the GTAs overall at the two universities. While the highest rating referred to a GTA's knowledge of the course content and the lowest rating related to assessment, mean ratings for all teaching dimensions were spread throughout a small range. Table 4 also includes the results of fixed-effect ANOVAs in which GTA instructional ratings were compared between the two institutions where such data were collected. In the overall measure of instructional quality, there was only a marginally significant difference between GTAs from University A and GTAs from University B ($p = 0.06$). However, there were statistically significant differences between GTAs at the two institutions for 10 of the 19 individual items. For nine of those 10 items, GTAs at University A were rated higher than GTAs at University B (note b in Table 4); for the 10th item, GTAs at University B were rated higher than GTAs at University A (note d in Table 4). In terms of the magnitudes of the differences, however, the strength-of-effect measures imply that institutional differences are quite small given that the largest η_p^2 estimate was 0.017.

Much greater variation was observed when the ANOVA treated GTA as a random effect. Random GTA effects were statistically significant for instructional practice overall and for all individual instructional practice items. The amount of variance across GTAs was moderate to large, with η_p^2 values ranging from 0.10 to 0.23. The largest GTA-to-GTA differences pertained to the GTAs' ability to provide clear explanations, clearly answer questions, and teach students with diverse backgrounds, needs, and interests. Overall, the least variable dimensions across GTAs were in asking of questions that were stimulating, and that require students to think. The magnitude of even these smallest random effects were still considerable, however, suggesting a wide range in instructional quality among the GTA population across sites.

TABLE 4. Student ratings of GTA instruction[a]

			Institution fixed effects		GTA random effects	
Item	**M**	**SD**	***F***	η_p^2	***F***	η_p^2
My TA knew the subject matter well.[b,c]	4.45	0.85	27.40***	0.017	4.74***	0.18
My TA evaluated our work fairly and impartially.[c,d]	4.42	0.89	5.85*	0.007	3.93***	0.16
My TA was enthusiastic about the subject matter.	4.37	0.83	1.03	0.001	3.22***	0.13
My TA asked discussion questions that require students to think.[b]	4.36	0.75	5.15*	0.004	2.38***	0.10
My TA asked discussion questions that reflect the goals for student learning.[c]	4.35	0.78	0.38	0.000	3.19***	0.13
My TA articulated the goals for student learning.	4.34	0.88	0.13	0.000	3.52***	0.14
My TA was well organized.	4.32	0.86	2.78	0.002	4.13***	0.16
My TA asked discussion questions that are appropriate given our knowledge of the topic.[c]	4.31	0.81	0.23	0.000	3.86***	0.16
My TA asked discussion questions that drew out the class's existing knowledge.	4.27	0.80	2.68	0.002	2.67***	0.11
My TA provided good examples of concepts.[b]	4.25	0.90	4.50*	0.004	4.30***	0.17
My TA provided valuable feedback on our work.[b]	4.22	0.97	8.20**	0.007	4.69***	0.18
My TA regularly asked stimulating discussion questions.	4.22	0.88	1.33	0.001	2.39***	0.10
My TA provided engaging activities for the class to do.[c]	4.20	0.91	0.59	0.001	2.87***	0.12
My TA answered students' questions clearly.[b]	4.17	1.07	10.22**	0.008	6.07***	0.22
MY TA provided clear explanations.[b]	4.14	1.07	5.63*	0.005	6.26***	0.23
My TA made him/herself available outside of class.[b]	4.13	0.94	9.95**	0.008	4.05***	0.16
My TA implemented class exercises that were interesting and stimulating.	4.13	0.95	0.13	0.000	3.28***	0.14
My TA knew how to teach students with different backgrounds, needs, and interests.[b]	4.11	1.03	3.96*	0.003	5.01***	0.19
My TA knew how well his/her students understood the material.[b]	4.08	1.01	15.24***	0.012	3.85***	0.16
Overall[e]	4.25	0.74	3.68	0.003	4.98***	0.19

[a]Item *N*s ranged from 1208 to 1214. Items sorted by grand mean. Full ANOVA results are available from the authors. *F*, *F*-statistic. η_p^2, partial eta squared.

[b]Ratings of GTA at University A were higher than those for GTAs at University B for this item.

[c]ANOVA test based on Welch statistic because Levene's test found that homogeneity of variance assumption violated ($p < 0.05$).

[d]GTAs at University B were rated higher than GTAs at University A for this item.

[e]Overall is mean of 19 student ratings of instruction.

*$p < 0.05$.

**$p < 0.01$.

***$p < 0.001$.

DISCUSSION

In this paper, we reported and compared changes in GTAs' self-efficacy, anxiety about teaching, and pedagogical knowledge during ST-based GTA training programs at three institutions. Additionally, we examined and compared undergraduate student perceptions of GTAs' instruction. Our study was explicitly grounded in the GTA PD framework (Figure 1), which notably posits that contextual factors such as GTA characteristics can moderate GTA PD impact. By enacting similar GTA PD programs in three contexts and comparing them along multiple outcome dimensions—GTA cognition and GTA teaching practice—the present study afforded a preliminary test of this "contextual factors" hypothesis. In this section, we summarize our findings, discuss their potential implications for the GTA PD framework and GTA PD more broadly, and identify possible directions for additional research on GTA training.

Across the three participating institutions, we found small but statistically significant increases in self-efficacy ($d = 0.34$) and pedagogical knowledge ($d = 0.47$), as well as a decrease in anxiety ($d = -0.42$). While changes in pedagogical knowledge were statistically indistinguishable across the three institutions, the magnitudes of changes in anxiety and self-efficacy differed. In both cases, GTAs at University C reported the most change and GTAs at University B reported the least. Our self-efficacy and anxiety findings comport with the results of other pretest–posttest studies of GTA training programs, which have also reported changes in GTA self-efficacy and anxiety (e.g., Young and Bippus, 2008; Page *et al*., 2011; Pelton, 2014). It is interesting that the largest changes in self-efficacy and anxiety were observed at University C, which was attended solely by incoming graduate students with no prior teaching experience as graduate students (Prieto and Altmaier, 1994). This finding is consistent with the GTA literature showing previous experience to be an important source of GTA teaching self-efficacy (Prieto and Altmaier, 1994) and arguments in the literature that much teacher learning occurs on the job (e.g., Sykes *et al.*, 2010).

Posttraining, we found that, on average, GTAs were rated very highly by students. In this respect, the findings support those of Kendall and Schussler (2013), who observed that student perceptions of GTA instructional characteristics improved over the course of a semester, such that GTA ratings at the end of the semester were similar to those that students gave to professors in the same roles. However, we must acknowledge that positive student ratings of GTA instruction at the end of the semester could be a function of the teaching experience GTAs gained during the semester, rather than the activities of the GTA training programs. At the same time, we also found some small, but statistically significant institution-to-institution differences in GTA teaching dimensions. In general, GTAs at University A were rated more highly than GTAs at University B. (Similar data were not available from University C for comparison.) Some dimensions where we observed such differences corresponded to the ST framework (e.g., diversity), and others were more general (e.g., knowledge of subject matter). It is particularly interesting that GTAs at University A generally received high ratings, because these GTAs all had significant previous GTA experience and more contact hours with their students than GTAs at University B. GTAs at University A also taught a single section, while GTAs at University B taught two sections. This is consistent with Kendall and Schussler's (2013) finding that instructors with more teaching experience tended to receive higher ratings from their students.

Of course, observed institutional differences have many possible explanations—some related to actual training quality and some related to other factors. One GTA group might simply stand to gain more or less than another based on their characteristics at training program entry, such as their prior teaching experience and training. Unfortunately, the present study's sample sizes overall and by institution did not afford reliable statistical consideration of whether GTAs with different characteristics responded differently to the PD. However, such moderating variables are an important area for future inquiry (Reeves *et al.*, 2016). In addition, differences in GTA instructional ratings between institutions might reflect institutional differences in the GTA populations themselves, or differences in the contexts in which those GTAs serve (e.g., institutional, teaching placement, or student characteristics). In their model of GTA teaching self-efficacy, DeChenne *et al*. (2015) identified the departmental teaching climate as a major contributing factor. Finally, the training programs offered at the three sites shared a common framework in ST, but instructors at the three sites designed their own programs and incorporated activities suitable for their own situations.

Despite very high student ratings of GTAs overall and small institutional differences, we found large GTA-to-GTA differences in every aspect of instructional quality assessed. There were no discernible patterns, however, in the categories of instructional quality (e.g., assessment, active learning, diversity) in which performance was most variable. Indeed, the significantly and largely variable GTA-to-GTA rating differences suggest wide variation in (even trained) GTAs. Addy and Blanchard (2010) and Bond-Robinson and Rodriquez (2006) observed similar findings after the implementation of their GTA training programs. In addition to putting institutional GTA teaching practice differences in context, these consistent findings about large variation among GTAs in their teaching practices underscore the GTA PD framework's assertion of the import of considering GTA characteristics. Our findings might also highlight for GTA training stakeholders those aspects of teaching that deserve the most attention during training programs (i.e., those GTA instructional dimensions that are most variable). Given nonrandom assignment of undergraduate students to GTAs, we also recognize that observed variation among GTAs might at least in part be explained by factors other than institutional quality (e.g., sectional student composition).

The present study offers implications for the GTA PD framework. Most importantly, our findings suggest the importance of considering contextual factors in both research on and the evaluation of GTA PD impact. The cross-institutional differences in gains observed in self-efficacy and anxiety, and differences in GTA instructional practices after PD, imply that the impact of a given PD program may hinge on relevant contextual factors such as GTA characteristics (e.g., prior teaching experience, student status [master's or doctoral level], prior teacher training, and attitudes toward teaching; Prieto and Altmaier, 1994; Serow *et al*., 2002).[6] A practical corollary

[6]We recognize that observed outcome differences among the three GTA PD programs may be at least partly explained by their differential structural and activity designs. A sounder test of the role of contextual factors would implement GTA PD programs identical in not only their content, but also their structures and activities.

of this finding is that identification of "best practices" or the like for GTA PD may prove elusive, as one-size-fits-all approaches to GTA PD may not adequately accommodate contextual differences relative to the characteristics of institutions, GTA teaching contexts, and GTAs. Indeed, the PD programs studied here were themselves somewhat differentiated in their activities, consistent with the expectations of the GTA PD framework. The convergence of our cross-institutional findings about gains in GTA pedagogical knowledge, on the other hand, might suggest that contextual factors matter or more less for different GTA PD outcomes. Our study, then, illustrates that training programs anchored in the same content framework (e.g., ST) might manifest differently in practice and produce divergent impacts because of contextual factors.

This study also offers evidence for the impact of three GTA training programs based on the ST framework. Notwithstanding the differences in program settings, with the general ST approach, all three training programs were associated with gains in content knowledge and GTA self-efficacy and decreases in teaching anxiety. Nonetheless, this study had a number of limitations that warrant additional research, and many questions remain about GTA training programs. In particular, more randomized studies are sorely needed to demonstrate unequivocally the effects of GTA training programs and which models are most effective in changing GTA characteristics, teaching practices, and student outcomes (Reeves *et al*., 2016). Future longitudinal research should also examine the long-term effects of GTA training programs. The field would also benefit from standardized instruments that could be used to assess GTA training program effects, including knowledge tests and observational measures of teaching practice. In this respect, Couch *et al*. (2015) have constructed a taxonomy of ST classroom situations. Research is also needed on particular design elements, for example, the most effective ways to facilitate GTA interaction or structure training programs. Given the importance of contextual factors, it would behoove those in the GTA PD community to consider not only what works, but for whom and under what conditions. Building on original findings presented here, which suggest that ST-focused GTA PD can indeed be effective, such future work will serve to amplify the impact of these programs.

ACKNOWLEDGMENTS

We thank all the GTAs and students who participated in this project and Jill Gomolka, Beatrice Frederique, Robert Polacheck, and Chenda Hong for research assistance. This work was supported by National Science Foundation Grant 1140428 to C.M.O. and L.E.H. and a grant from the Howard Hughes Medical Institute to Yale University.

REFERENCES

Addy, T. M., & Blanchard, M. R. (2010). The problem with reform from the bottom up: Instructional practises and teacher beliefs of graduate teaching assistants following a reform-minded university teacher certificate programme. *International Journal of Science Education*, *32,* 1045–1071.

Austin, A. E. (2002). Preparing the next generation of faculty: Graduate school as socialization to the academic career. *Journal of Higher Education*, *73,* 94–122.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84,* 191–215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bauer, C., Libby, R. D., Scharberg, M., & Reider, D. (2013). Transformative research-based pedagogy workshops for chemistry graduate students and postdocs. *Journal of College Science Teaching*, *43,* 36–43.

Baumgartner, E. A. (2007). A professional development teaching course for science graduate students. *Journal of College Science Teaching*, *36,* 16–21.

Bond-Robinson, J., & Rodriques, R. A. B. (2006). Catalyzing graduate teaching assistants' laboratory teaching through design research. *Journal of Chemical Education*, *83,* 313–323.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41,* 1069–1077.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology*, *44,* 473–490.

Chen, X., Kimball, K., Frederick, J., & Graham, M. J. (2013). Training TAs in scientific teaching for the human physiology and anatomy laboratory. *Advances in Physiology Education*, *37,* 436–439.

Cho, Y., Kim, M., Svinicki, M. D., & Decker, M. L. (2011). Exploring teaching concerns and characteristics of graduate teaching assistants. *Teaching Higher Education*, *16,* 267–279.

Coates, T. J., & Thoreson, C. E. (1976). Teacher anxiety: A review with recommendations. *Review of Educational Research*, *46,* 159.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51,* 281–309.

Coll, R. K., Taylor, N., & Fisher, D. L. (2002). An application of the Questionnaire on Teacher Interaction and college and university classroom environment inventory in a multicultural tertiary context. *Research in Science & Technological Education*, *20,* 165–183.

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

Crawford, B. A. (1999). Is it realistic to expect a preservice teacher to create an inquiry-based classroom? *Journal of Science Teacher Education*, *10,* 175–194.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18,* 23–45.

DeChenne, S. E., Enochs, L. G., & Needham, M. (2012). Science, technology, engineering, and mathematics graduate teaching assistants teaching self-efficacy. *Journal of the Scholarship of Teaching and Learning*, *12,* 102–123.

DeChenne, S. E., Koziol, N., Needham, M., & Enochs, L. (2015). Modeling sources of teaching self-efficacy for science, technology, engineering, and mathematics graduate teaching assistants. *CBE—Life Sciences Education*, *14,* ar32.

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

Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics*, *90,* 694–706.

French, D., & Russell, C. (2002). Do graduate teaching assistants benefit from teaching inquiry-based laboratories? *BioScience*, *52,* 1036–1041.

Gardner, G. E., & Jones, M. G. (2011). Pedagogical preparation of science graduate teaching assistant: Challenges and implications. *Science Educator*, *20,* 31–41.

Gilreath, J. A., & Slater, T. F. (1994). Training graduate teaching assistants to be better undergraduate physics educators. *Physical Education*, *29,* 200.

Hammrich, P. L. (2001). Preparing graduate teaching assistants to assist biology faculty. *Journal of Science Teacher Education*, *12,* 67–82.

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

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

Hardré, P. L., & Burris, A. O. (2012). What contributes to teaching assistant development: Differential responses to key design features. *Instructional Science*, *40,* 93–118.

Hardré, P. L., & Chen, C. H. (2005). A case study analysis of the role of instructional design in the development of teaching expertise. *Performance Improvement Quarterly*, *18,* 34–58.

Harris, G., Froman, J., & Surles, J. (2009). The professional development of graduate mathematics teaching assistants. *International Journal of Mathematical Education in Science and Technology*, *40,* 157–172.

Kendall, K. D., & Schussler, E. E. (2013). Evolving impressions: Undergraduate perceptions of graduate teaching assistants and faculty members over a semester. *CBE—Life Sciences Education*, *12,* 92–105.

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research*, *27,* 9–25.

Ludlow, L. H., Rollison, J. M., Cronin, J., & Wallingford, T. (2012). Development of the Teaching Economic Literacy: Confidence and Anxiety (TELCA) instrument. *International Journal of Educational and Psychological Assessment*, *9,* 82–103.

Luft, J. A., Firestone, J. B., Wong, S. S., Ortega, I., Adams, K., & Bang, E. (2011). Beginning secondary science teacher induction: A two-year mixed methods study. *Journal of Research in Science Teaching*, *48,* 1199–1224.

Luft, J. A., Kurdziel, J. P., Roehrig, G. H., & Turner, J. (2004). Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching*, *41,* 211–233.

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.

Miller, S., Pfund, C., Pribbenow, C. M., & Handelsman, J. (2008). Scientific teaching in practice. *Science*, *322,* 1329–1330.

Page, M., Wilhelm, M. S., & Regens, N. (2011). Preparing graduate students for teaching: Expected and unexpected outcomes from participation in a GK–12 classroom fellowship. *Journal of College Science Teaching*, *40,* 32–37.

Park, C. (2004). The graduate teaching assistant (GTA): Lessons from the North American experience. *Teaching and Teacher Education*, *9,* 349–361.

Pelton, J. A. (2014). Assessing graduate teacher training programs: Can a teaching seminar reduce anxiety and increase confidence? *Teaching Sociology*, *42,* 40–49.

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.

Pfund, C., Miller, S., Brenner, K., Bruns, P., Chang, A., Ebert-May, D., … Handelsman, J. (2009). Summer institute to improve university science teaching. *Science*, *324,* 470–471.

Prieto, L. R., & Altmaier, E. M. (1994). The relationship of prior training and previous teaching experience to self-efficacy among graduate teaching assistants. *Research in Higher Education*, *35,* 481–497.

Prieto, L. R., & Meyers, S. A. (1999). The effects of training and supervision on the self-efficacy of psychology graduate teaching assistants. *Teaching of Psychology*, *26,* 264–266.

Prieto, L. R., Yamokoski, C. A., & Meyers, S. A. (2007). Teaching assistant training and supervision: An examination of optimal delivery modes and skill emphases. *Journal of Faculty Development*, *21,* 33–43.

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

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

Sales, J., Comeau, D., Liddle, K., Perrone, L., Palmer, K., & Lynn, D. (2007). Preparing future faculty. *Journal of College Science Teaching*, *36,* 24–30.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, *102*(6), 245–253.

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.

Serow, R. C., Van Dyk, P. B., McComb, E. M., & Harrold, A. T. (2002). Cultures of undergraduate teaching at research universities. *Innovative Higher Education*, *27,* 25–37.

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

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16,* rm1.

Sundberg, M. D., Armstrong, J. E., & Wischusen, E. W. (2005). A reappraisal of the status of introductory biology laboratory education in U.S. colleges and universities. *American Biology Teacher*, *67,* 526–529.

Sykes, G., Bird, T., & Kennedy, M. (2010). Teacher education: Its problems and some prospects. *Journal of Teacher Education*, *61,* 464–476.

Tanner, K. D., & Allen, D. (2006). Approaches to biology teaching and learning: On integrating pedagogical training into the graduate experiences of future science faculty. *Cell Biology Education*, *5,* 1–6.

Treagust, D. F., & Fraser, B. J. (1986). Validation and application of the College and University Classroom Environment Inventory (CUCEI). *Paper presented at: Annual Meeting of the American Educational Research Association*. San Francisco, CA.

Wilson, D. B. (2001). *Effect size determination program (Version 2.0)*. College Park: University of Maryland.

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.

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,* 116–129.