Applying the Geoscience Education Research Strength of Evidence Pyramid: Developing a Rubric to Characterize Existing Geoscience Teaching Assistant Training Studies

Kelsey S. Bitting,[1,a] Rachel Teasdale,[2] and Katherine Ryker[3]

ABSTRACT

Graduate teaching assistants (GTAs) are responsible for direct instruction of geoscience undergraduate students at an array of universities and have a major effect on the knowledge, beliefs, and practices of their students. GTAs benefit from in-department training in both beliefs and practices that align with the existing literature on teaching and learning in the discipline, and such training can have long-standing effects when GTAs transition into faculty roles. However, the most recent review, in 2003, revealed little literature examining outcomes of geoscience GTA training programs. Using the framework of the GER Strength of Evidence Pyramid, this article outlines the development and application of a rubric to allow the user to analyze the existing geoscience GTA training literature and provide example study designs at each level of strength. Extending back to 1980, we discovered a total of three peer-reviewed articles describing and empirically evaluating the effect of GTA training programs in the geosciences. Thus, this article also draws from other science disciplines to provide examples for the levels of the rubric not currently represented in the geoscience literature, providing a set of contextually similar models that future designers of geoscience GTA training might draw on to maximize their strength of evidence, given specific institutional and programmatic constraints. Furthermore, we describe ways in which the use of the rubric provides a framework for characterizing the GTA training literature, which revealed areas of research and characteristics of rigor needed for future work. © *2017 National Association of Geoscience Teachers*. [DOI: 10.5408/16-228.1]

Key words: GTA training, teaching assistants, professional development, strength of evidence

LITERARY CONTEXT AND INTRODUCTION

Within this special issue on "Synthesizing Results and Defining Future Directions of Geoscience Education Research," graduate teaching assistant (GTA) training carries special importance for several reasons. First, effective undergraduate instruction at many institutions is directly dependent on the effectiveness of GTAs: at research-intensive universities, GTAs instruct most laboratory classes (Travers, 1989; Luft et al., 2004; Sundber et al., 2005), and in some science disciplines, up to 91% of undergraduate students study in laboratories or courses taught primarily by GTAs (Sundber et al., 2005). As the instructors of record for their laboratory sections, GTAs may make decisions about what should be taught and how and how to assess student performance, often without input or guidance from faculty (Kurdziel and Libarkin, 2003).

Many authors have asserted that departmental GTA training is necessary to contextualize teaching approaches within the discipline-based education research (DBER) specific to that field, as well as to signal that teaching is valued within the departmental culture (Black and Bonwell, 1991; Hammrich, 1996; Hardre, 2003; Buskist, 2013). Without that training, GTAs are likely to teach using the practices they themselves experienced as undergraduates, i.e., "teaching as they were taught" (Halpern and Hakel, 2002; Oleson and Hora, 2013). Most published geology laboratory manuals are not inquiry-oriented (Ryker and McConnell, 2017), further reinforcing the likelihood that GTAs will teach using outdated and ineffective methodologies (Ryker and McConnell, 2014) that do not align with research on effective instruction and calls for reforms in teaching undergraduate courses (AAAS, 1990; NRC, 1996, 2000, 2012).

GTAs often have smaller class sizes (especially in laboratories), may be perceived as more approachable than instructors with doctoral degrees, and are potential near-peer role models for undergraduates aspiring to take the next step in disciplinary study (Rushin et al., 1997; O'Neal et al., 2007). Therefore, GTAs have more-frequent personal interactions with students, which can influence attitudes toward, and beliefs about, the nature of science and the process of learning. In this way, GTAs have a widespread effect on undergraduate students, many of whom are not science majors, including future K–12 teachers. For future teachers, those attitudes affect the practices they rely on for teaching and learning in science contexts (Hardre and Chen, 2005).

Because of their closer relationships with undergraduates and their role in instruction, GTAs hold significant sway over determining student reactions to new instructional methods. Thus, in a department seeking to respond to the broader national movement encouraging active, evidence-based instruction in science, technology, engineering, and mathematics (STEM), GTAs can reinforce or subvert

Received: 7 November 2016, Revised: 29 June 2017, Accepted: 3 July 2017; published online 16 November 2017.

[1] Center for Advancing Teaching and Learning Through Research, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, USA

[2] Geological and Environmental Sciences Department, California State University, Chico, Physical Science, Room 217, 400 West First Street, Chico, California 95929, USA

[3] Department of Earth Science and Earth Science Education, Eastern Michigan University, 900 Oakwood Street, Ypsilanti, Michigan 48197, USA

[a] Author to whom correspondence should be addressed. Electronic mail: kelsey.bitting@gmail.com. Tel.: 732-770-8291.

attempts to shift instructional norms, depending on their understanding of, and investment in, those shifts (Wood, 2009; Bautista et al., 2014; Linenberger et al., 2014; Ryker and McConnell, 2014). For example, GTAs being asked to implement relatively unfamiliar models, such as inquiry-based laboratory classes, benefit from specific training to help them adapt to a nontraditional role in the classroom (Krystyniak and Heikkinen, 2007; Gormally et al., 2009; Sandi-Urena and Gaitlin, 2013).

Ultimately, changes in practice without supporting changes in beliefs are often short lived, inconsistent, or ineffective (e.g., Yerrick et al., 1997; Turpen and Finkelstein, 2009; Andrews et al., 2011; Henderson et al., 2011). Although beliefs about teaching and learning are remarkably resilient to change (Yerrick et al., 1997), effective professional development can affect both teaching beliefs and practices (e.g., Kane et al., 2002; Ebert-May et al., 2015). Therefore, GTA training programs must target both beliefs and practices to maximize their effect.

Finally, many GTAs are also future faculty, and their first time teaching can be a valuable opportunity to prepare them for the teaching component of their professional roles (Brownell and Tanner, 2012; Kendall et al., 2013; Schussler et al., 2015). For example, fewer inexperienced instructors use evidence-based techniques, such as peer instruction and collaborative learning (Dancy and Henderson, 2010; Budd et al., 2013; Lund et al., 2015), perhaps because of the fear of a drop in student evaluations of teaching performance, which could adversely affect tenure and promotion decisions (e.g., Bass, 1999). Experience in implementing evidence-based teaching practices as a GTA could lead to greater self-efficacy around teaching (Bandura, 1997; Hardre, 2003), which could, in turn, encourage greater comfort with trying out different practices to meet contextual student needs in the future.

Additional barriers to instructional change by STEM faculty cited in the literature have included lack of training and incentive structures that encourage faculty to spend limited available time on research, rather than on teaching (Henderson and Darcy, 2007; Brownell and Tanner, 2012; Manduca et al., 2017). Although GTAs experience similar competing demands (Luft et al., 2004), professional development in teaching and learning during the early stages of academic preparation establishes that teaching, like research, is a learned practice that benefits from attention, discussion, analysis, and continuous refinement. In fact, GTAs themselves sometimes desire greater guidance for their teaching roles (Dotger, 2010). Effective professional development for GTA teaching could provide effective (and efficient) approaches to instruction and assessment at the beginning of careers, eliminating the time a faculty member without that training might later spend on relearning or vastly revising ineffective but ingrained approaches. This, in turn, can launch GTAs into academic careers that incorporate development of their teaching practice along with development of their scientific expertise.

In light of these important reasons to implement and evaluate GTA training programs within geoscience departments and programs, Kurdziel and Libarkin (2003) evaluated the state of the existing research base on this topic. Disappointingly, they noted that "although many science departments have published articles describing their approach to developing graduate students' teaching skills, very few of these programs has even been evaluated" (Kurdziel and Libarkin, 2003, 347). Those authors called for additional research in geoscience nearly 15 y ago.

Movements within the geoscience education research (GER) community, described in this issue and explored at recent workshops on the future of GER, have also highlighted the need for the community to begin moving to greater standards of rigor in study design and analysis and to calibrate the valuing of studies for informing educational practice according to the strength of the evidence they present. This idea is neatly summarized by the GER Strength of Evidence Pyramid framework put forth by St. John and McNeal (2017; this issue). Therefore, future reviews of topics within the GER literature should also describe the strength of evidence of the studies on that topic, both to clarify for practitioners whether the evidence merits revision of practice according to the results and to identify next steps for researchers interested in exploring that topic further.

This study reevaluates the literature on geoscience GTA training using the newly-available GER Strength of Evidence Pyramid framework (St. John and McNeal, 2017; this issue, hereafter referred to as *the pyramid*), identifies and synthesizes contributions for practice, and highlights future directions for research studies in geoscience GTA training.

RESEARCH QUESTIONS

Initially, this study sought to investigate the following research questions:

(1) What is the distribution and quantity of empirical studies on GTA training programs in the geosciences at each level of strength of the pyramid framework?
(2) What learning objectives are commonly sought by the programs described in that literature?
(3) What methodologies and methods have been used to evaluate the extent to which the desired objectives are achieved by participants?
(4) What specific directions and implications does the existing empirical literature on geoscience GTA training suggest for the development of training programs and future research related to the topic?

This article also describes our methods as a process model for future systematic literature reviews based on the pyramid. Applying the pyramid in this context involved initial coding, development of a more-nuanced rubric grounded in our examination of those articles, and final calibration of two reviewers to achieve high levels of interrater reliability. To clarify the meaning of the rubric and provide close-context models for a wide range of possible future studies, we also provide examples of GTA training studies for each level of the rubric (from geoscience, where possible, and from other science disciplines when geoscience examples do not exist). Finally, we describe ways in which the systematic process of using the pyramid framework to design a rubric allows project teams to rigorously describe, compare, and analyze the state of DBER literature within and across topics.

LITERATURE SEARCH

We used a twofold approach to gather relevant studies. We began by including articles each of the researchers had

TABLE I: Search terms used to identify potential articles for inclusion in the literature review, date of search, and results of each term.

Google Scholar Search Terms	Date Searched	Number of Results[1]	No. of Results Reviewed[2]	No. of Articles Downloaded[3]
"teaching assistant" + training + "Earth science"	5 July 2016	0		
"teaching assistant" + training in "Earth science"	5 July 2016	824	200	8
"teaching assistant" + training + geoscience	5 July 2016	3,140	300	15
"teaching assistant" + training + geology (1990–2016)	5 July 2016	3,340	80	6
Research + "TA training"	5 July 2016	3,060	300	31
"teaching assistant" + training + physics	5 July 2016	12,000	300	26
"teaching assistant" + training + astronomy	5 July 2016	3,180	200	9
"teaching assistant" + training + biology	6 July 2016	15,500	100	6
"teaching assistant" + training + chemistry	13 July 2016	15,300	200	5

[1] Number of results represents the total hits provided by Google Scholar.

[2] Number of results reviewed indicates numbers of articles that were considered (80–300), depending on how far into the results relevant articles ceased to be identified.

[3] Number of articles downloaded ($n = 106$) were added to the preexisting body of 133 articles the authors began the project with, adding to the full 239 articles examined in the project. Of those, 48 met the inclusion criteria for the study and were retained for further analysis.

collected previously and used the references cited within those articles to gather additional articles ($N = 133$ publications). To ensure a complete review of the field, we also conducted a series of Google Scholar (Google, Mountain View, CA) searches in July 2016. Our search terms (Table I) prioritized science disciplines, but we downloaded any articles conducted within STEM contexts since 1990, resulting in an additional 106 studies. These two steps brought our total preliminary collection to 239 articles.

To select the subset of these 239 articles that would be of greatest value in addressing our research questions, two of us (K.S.B. and K.R.) carefully screened each article using the following inclusion criteria, resulting in the retention of 48 articles (with complete agreement between the two coders):

- Published in a peer-reviewed journal
- Published between 1990 and 2016
- Described and/or evaluated a GTA professional-development program to prepare the GTAs for their role as GTAs (i.e., excluding future faculty programs)
- Contextualized within a STEM discipline (astronomy, atmospheric science, biology, chemistry, computer science, engineering, geoscience, mathematics, psychology, and physics).

National and institutional study contexts were not always reported in the text of an article, but national context (where not reported) could be inferred from the range of locations at which study authors were employed. All retained studies took place in the U.S. ($n = 43$) or Canada ($n = 6$).

Studies conducted within science fields predominated ($n = 44$), as expected, because of our prioritization in the initial search within that context. Of those 44 articles, two also included participants from other STEM disciplines (computer science and mathematics). Two studies from engineering, one from mathematics, and one from all STEM fields were also retained.

Of the articles that met the inclusionary criteria, two were specific to GTA training in the geosciences (McManus, 2002; Dotger, 2011); one additional geoscience-related article was identified from beyond the identified time range (Schade and Bartholomew, 1980). We decided that study should be included because of the exceptionally few geoscience articles. One geoscience article (Dotger, 2010) was rejected from the review because it neither described nor evaluated the GTA training course in which the project was conducted.

METHODS

Using the original pyramid, distributed during the "Synthesizing Geoscience Education Research" session at the Earth Educators' Rendezvous 2015 (Macdonald et al., 2015), we developed a preliminary rubric representing the five original levels of evidence strength: (1) practitioner wisdom/expert opinion, (2) qualitative and quantitative case studies, (3) qualitative and quantitative cohort studies, (4) topical review articles and meta-analyses, and (5) systematic reviews).

Early in our examination of the literature, we noticed an abundance of peer-reviewed articles that fit within the original Level 1 (practitioner wisdom/expert opinion). Although many of those articles were faculty narratives describing GTA training programs without supporting data (such as some articles on other teaching topics published in *In the Trenches*), others presented some form of anecdotal evidence of the program's success. We agreed that this distinction should be reflected in the rubric, splitting Level 1 into Categories 1A (no data) and 1B (anecdotal or informally analyzed data).

Through close reading of the remaining case (Level 2) and cohort (Level 3) studies for elements of study design, instrumentation, and data analysis, one of us (K.S.B.) developed subcategories for the rubric to reflect the wide variety of strength of evidence created through those characteristics. In particular, studies using validated self-report instruments, objective measures of effect (e.g., classroom observations, analysis of instructional artifacts, student evaluations/ratings of instruction), or established theoretical frameworks were differentiated as Subcategory B articles, and studies adding comparison groups to evaluate the effect of the intervention were differentiated as Subcategory C articles. Within Level 4 (topical review articles and meta-analyses), studies analyzing the effects of a group of training programs

TABLE II: Rubric developed and used to characterize GTA articles.

General Group[1]	Category	Description	Example Articles
Practitioner wisdom/ expert opinion (program description, with or without assessment)	1A	A description of a program, programs, or multiple iterations of a program to describe how their GTA training works (but no data on effectiveness or impact)	• Druger, 1997 (biology) • Thornburg et al., 2000 (chemistry) • Holmes et al., 2013 (physics and astronomy)
	1B	• A description of a program, programs, or multiple iterations of a program to describe how their GTA training works • Some form of data or comments are collected and described, but not collected or reported or analyzed systematically; trends of data are not substantiated, or data are satisfaction data without coding for additional analysis	McManus, 2002 (geoscience)
Qualitative and quantitative case studies (single iteration of a research study)	2A	• Analysis of a single GTA training program with data that allow interpretation of change related to the intervention • Data are all student satisfaction or self-report using nonvalidated instruments	Linenberger et al., 2014 (biology, chemistry, geosciences, and psychology)
	2B	• Analysis of a single GTA training program with data that allow interpretation of change related to the intervention • Data include validated self-report instruments; codes based on established theoretical frameworks or objective measures of effect	Dotger, 2011 (geoscience)
	2C	• Analysis of a single GTA training program with data that allow interpretation of change related to the intervention • Data include validated self-report instruments and/or objective measures of effect • Study includes comparison group of some type	Hughes and Ellefson, 2013 (biology)
Qualitative and quantitative cohort studies (repeated research study)	3A	• Synthesis of multiple iterations of a single training program, either presented separately or as an aggregate data set, with data that allow interpretation of change related to the intervention • Data are all student satisfaction or self-report using nonvalidated instruments	Schade and Bartholomew, 1980 (geoscience)
	3B	• Synthesis of multiple iterations of a single training program, either presented separately or as an aggregate data set according to accepted research design conventions and with a systematic approach to data analysis • Data include validated self-report instruments; codes based on established theoretical frameworks or objective measures of effect	Komarraju, 2008 (psychology)
	3C	• Synthesis of multiple iterations of a single training program, either presented separately or as an aggregate data set, with data that allow interpretation of change related to the intervention • Data include validated self-report instruments and/or objective measures of effect • Study includes comparison group of some type	Bond-Robinson and Rodriques, 2006 (chemistry)
Syntheses and meta-analyses	4A	Synthesis of a group of training programs either presented separately or as an aggregate data set, with data that allow interpretation of change related to the interventions	DeChenne et al., 2012 (all STEM)
	4B	Meta-analysis of the data from multiple studies/publications to combine smaller data sets into a larger body of data that is synthesized in aggregate	N/A
Systematic reviews	5	Synthesis of the results of multiple studies/publications to draw broad conclusions of the group of studies	N/A

[1] General group headings are based on the GER Strength of Evidence categories (St. John and McNeal, 2017; this issue) from which categories were determined as articles were reviewed. Example articles are discussed in the text.

and related interventions were formally distinguished from meta-analyses in Subcategories A and B, respectively. This rubric was clarified and improved through discussions among all three of us based on its application to a subset of articles (10 articles randomly selected from the full sample of 48 relevant publications).

Our final rubric is presented in Table II. Note that Categories 4B and 5 were not identified in the current suite

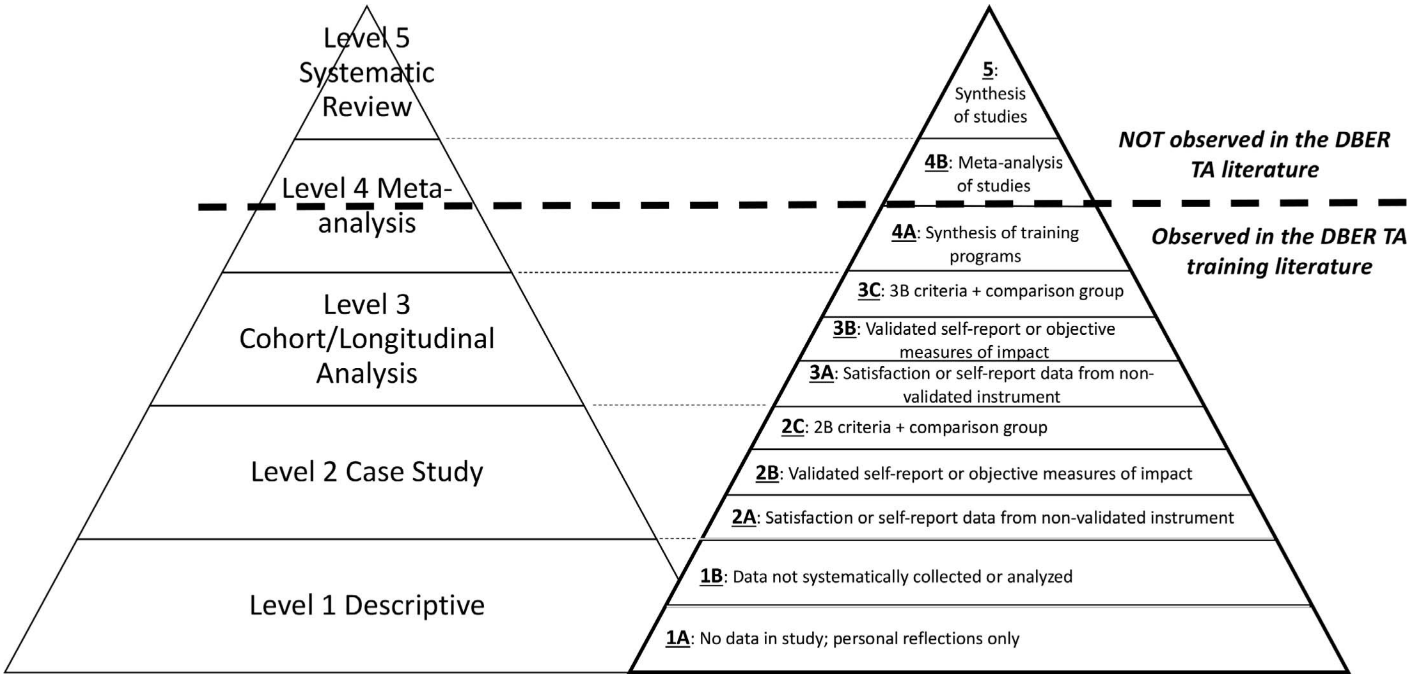
Level 5 Systematic Review
Level 4 Meta-analysis
Level 3 Cohort/Longitudinal Analysis
Level 2 Case Study
Level 1 Descriptive
5: Synthesis of studies
4B: Meta-analysis of studies
4A: Synthesis of training programs
3C: 3B criteria + comparison group
3B: Validated self-report or objective measures of impact
3A: Satisfaction or self-report data from non-validated instrument
2C: 2B criteria + comparison group
2B: Validated self-report or objective measures of impact
2A: Satisfaction or self-report data from non-validated instrument
1B: Data not systematically collected or analyzed
1A: No data in study; personal reflections only
NOT observed in the DBER TA literature
Observed in the DBER TA training literature

FIGURE 1: Comparison of GER Strength of Evidence categories at left (St. John and McNeal, 2017; this issue) with rubric categories (described in Table II) used in this work. Note the horizontal break separates categories that are not represented by the GTA literature (Category 4B and above) in the geosciences or other science disciplines. Within each level, lighter color shading indicates greater strength of evidence (addition of evidence, validated instruments, comparison groups, etc.). TA = teaching assistant.

of articles. However, because studies in those categories would represent strong contributions to this body of literature, we intentionally chose to retain the levels in the rubric as aspirational categories for future researchers.

This rubric joins the basic structure of the Strength of Evidence Pyramid (St. John and McNeal, 2017, this issue; represented by the General Group title in the left-hand column on the rubric in Table II) with the characteristics we observed within the current STEM GTA training literature (represented by the parenthetical descriptions within the General Group column and the Description column on the rubric above). The relationship between our rubric and the Strength of Evidence is presented in Fig. 1.

Interrater and Intrarater Reliability

Once the final rubric was developed, two of us (K.S.B. and K.R.) cocoded approximately 15 articles selected to represent the full range of the rubric and convened to discuss our categorizations and resolve any questions or discrepancy. We repeated this process twice more using sets of 10 articles, until no discussion was necessary for agreement. Next, we divided the articles, each coding five to eight articles individually before meeting and coding two articles together to ensure we remained calibrated. Both reviewers agreed on all 25 cocoded articles (Cohen's $\kappa = 1.0$; Cohen, 1968). During the coding period, we both changed our rating for one of the initial 15 articles, giving each of the two coders an intrarater reliability (Cohen's κ) of 0.96.

These results indicate that the rubric was applied in a consistent way to the articles described here. Supplemental File 1 (available in the online journal and at http://dx.doi.org/10.5408/16-228s1) provides a table of all retained articles, including the final rubric levels and subcategories for each.

Category Examples

To answer the research questions stated above, we initially aimed to draw specifically on articles contextualized in the geosciences that met the inclusion criteria. However, only three such articles were discovered: McManus (2002; rubric Category 1B), Dotger (2011; Category 2B), and Schade and Bartholomew (1980; Category 3A). Because these three articles alone cannot represent all categories included in the rubric, we selected an additional 12 example articles from our broader set to represent the remaining categories in contexts as similar as possible to the geosciences (Table I; Druger, 1997; Thornburg et al., 2000; Bond-Robinson and Rodriques, 2006; Komarraju, 2008; DeChenne et al., 2012; Holmes et al., 2013; Hughes and Ellefson, 2013). Each example article is described below to demonstrate the differential characteristics of articles that fall into each of the rubric categories, and also to provide examples at each level, so that future geoscience education researchers might draw upon to design GTA training studies that maximize the strength of evidence possible given their specific constraints.

Level 1: Practitioner Wisdom/Expert Opinion

Level 1 articles represent practitioner wisdom and expert opinions, which appear in the literature as peer-reviewed, published manuscripts describing a GTA training program in general terms. These articles may describe the demographics of GTAs who participated, the process of developing and gathering preliminary input on the GTA training program, general topics covered, the amount of time GTAs spent in the training, and the timing of the program relative to the start of a GTA's teaching assignment (among other details). Category 1A articles ($n = 9$) are easily identified because they include no further data than the authors' reflections on the effectiveness or value of the training program. Examples of 1A articles from across the various science disciplines include Druger (1997; biology), Thornburg et al. (2000; chemistry), and Holmes et al. (2013; physics and astronomy).

Category 1B articles ($n = 4$) generally fit the overall model of program development and implementation descriptions but also included data beyond anecdotal reflections. However, those data were not collected, reported, or analyzed in a sufficiently systematic way to allow the reader to determine the effect of the program on its participants. "Developing a Teaching Assistant Preparation Program in the School of Oceanography, University of Washington" by McManus (2002) is the selected example for this category and is one of three included manuscripts from a geoscience context. This article first describes the process of gathering input from current and former graduate students and current undergraduate majors (via questionnaires and focus groups) to inform the development of a series of college-specific interventions for GTA teaching preparation. Next, the article provides an overview of a 2-d orientation program, covering topics of immediate need, with workshops delivered by other graduate students and postdoctoral researchers. Small-group conversations facilitated by a consultant at the end of the program were used to identify the primary strengths of, and recommended changes to, the orientation. Those uncoded and unsystematically reported qualitative data were the distinguishing factor that placed the manuscript in Category 1B on the rubric. Finally, the article describes the initiation of a mentoring program, a graduate-level pedagogy course taught by the author, and a graduate-level pedagogy course taught by a lead GTA. No data on the effectiveness of any of these interventions were reported, but the author infers final implications for faculty.

Level 2: Qualitative and Quantitative Case Studies

Level 2 articles are case studies representing a single iteration of a research study, with either qualitative or quantitative data that allow determination of the effect of the training in some way. This level includes three subcategories depending on study design: Category 2A articles ($n = 3$) rely on self-report instruments that have not been validated and pre–post research designs. Category 2B articles ($n = 10$) used validated self-report instruments (e.g., self-efficacy measures) or objective measures of effect (e.g., improved alignment between objectives and assessments in instructional materials developed by GTAs) to determine the effectiveness of the program. Category 2C articles ($n = 4$) used validated instruments or objective measures of effect *and* included a comparison group of some type in the study design.

Category 2A does not currently include any geoscience-specific examples, so "Training the Foot Soldiers of Inquiry: Development and Evaluation of a Graduate Teaching Assistant Learning Community" by Linenberger et al. (2014) was selected as the example article for this category. This article describes a year-long learning community that supported 14–16 GTAs from biology, chemistry, geosciences, and psychology as they taught laboratory sections in the context of a university initiative to increase the level of inquiry in those courses. Topics for discussion, activities, and reflection prompts were selected by postdoctoral facilitators based on participants' self-reported needs and included inquiry-based instruction, student motivation, assessment, and working with faculty. Assessment of the learning community was accomplished in several ways: First, a survey asked students to provide Likert-scale assessments of 31 instructional methods according to (1) how often that method was used during their own undergraduate education, (2) how often they currently used that method, and (3) how important they believed that method to be for inquiry. That self-report revealed that students' undergraduate experience prioritized methods the GTAs did not associate with inquiry, that their teaching methods did not align with what they thought was important for inquiry (and did correlate with their prior experience of methods as undergraduates), and that even after completing the program, GTAs' assessment of methods important for inquiry differed from the authors' own valuations. Next, the end-of-semester evaluation asked GTAs to rank their level of understanding on a Likert scale, showing significant increases over time for five of eight topics. Finally, students' short-answer responses to a question asking how the learning community had influenced them as an educator revealed that four participants each semester reported a shift toward becoming a more reflective teacher. Because all quantitative data were collected using self-report instruments that were not validated in any way, the qualitative coding process was not described in any depth, and the resulting codes and themes were not fully corroborated, which classifies this article as being Category 2A research.

"Exploring and Developing Graduate Teaching Assistants' Pedagogies via Lesson Study" by Dotger (2011) represents Category 2B. This qualitative study of Earth Science GTAs collaborating during six 3-h sessions is the second of the three geoscience studies included in the review. GTAs worked together to identify their teaching goals, selected an existing lesson to improve, and reflected upon and iteratively redesigned that lesson for alignment to their identified goals. The researcher (also the instructor) avowed to have provided "minimal guidance for aligning the lesson to standards of reformed science teaching" (p. 160). That lack of instruction reflected the primary goal of the study, which was "documenting the GTAs' understandings of teaching and learning" (p. 160), not teaching development. However, the author asserted that an exchange of knowledge and perspectives among the GTA participants may have developed their thinking around teaching. Data used to evaluate the effect of participation in the program included notes and memos from the seminar meetings, student work samples, and interview transcripts. Those artifacts were inductively coded using a constant comparative methodology, which resulted in the identification of four themes. Use of that established approach for qualitative methodologies resulted in categorization of the article as 2B.

Category 2C does not currently include any geoscience examples, so "Inquiry-Based Training Improves Teaching Effectiveness of Biology Teaching Assistants" by Hughes and Ellefson (2013) was selected as an example article from the biological sciences. This study used a quantitative methodology to compare the results of one iteration each of both a traditional "best practices" GTA training and an inquiry-based training, each consisting of two 2.5-h sessions for a total of 5 h. Fifty-two participants were placed in the control or treatment groups using blind, randomized assignment. The study relied on multiple measures to compare the outcomes of the two groups, including two modified versions of the Student Evaluation of Educational Quality (Marsh, 1982), one completed by the GTAs and one completed by their students; a study-developed cognitive learning evaluation based on Bloom's Taxonomy; and student grades standardized to account for differences in grading structures (classified as an objective measure of effect within the rubric). Before the primary statistical analyses, the researchers used a logistic regression and an ANOVA to determine the equivalency of the two groups in terms of degree program (MA/MS versus PhD) and years of teaching experience, respectively. Because of both the use of objective and validated measures and the use of a control group, this article was identified as a model for Category 2C.

Level 3: Qualitative and Quantitative Cohort Studies

Level 3 articles on GTA training are cohort studies representing multiple iterations (in multiple time periods or at multiple institutions) of a research study, with either qualitative or quantitative data that allows determination of the effect of the training on participants. The three subcategories for this level mirror those of Level 2: Subcategory A ($n = 7$) lacks validated instruments or objective measures of effect, Subcategory B articles ($n = 2$) includes those more-rigorous pieces of evidence, and Subcategory C ($n = 5$) builds upon the criteria of Subcategory B articles by including a comparison group.

"Analysis of Geology Teaching Assistant Reaction to a Training Program Utilizing Video-Taped Teaching Episodes" by Schade and Bartholomew (1980) is the Category 3A example and the third identified geoscience-based study (included for its disciplinary relevance, despite falling well outside the intended time range for the overall literature review). The article describes the recording and editing of a series of brief videos of teaching behaviors performed by GTAs when teaching their laboratory sections. Next, the videos were used as centerpieces in a training intervention at three universities, each consisting of two 1-h meetings, followed by a third meeting for participants to complete the evaluation questionnaire. That nonvalidated instrument consisted of questions written by the authors to assess the participants' self-reported awareness of the highlighted techniques and reactions to participating in the program, enjoyment of the program, and views on the utility of the videos.

"A Social–Cognitive Approach to Training Teaching Assistants" by Komarraju (2008) comes from the disciplinary context of psychology, and represents Category 3B. The study includes data from four iterations of a 1-wk training program designed to cultivate GTAs' self-efficacy around teaching (based on mastery experiences, vicarious experiences, social persuasion, and managing physiological arousal, based on Bandura [1997]). Eighty-seven GTAs participated in the intervention, and the effect of the process was determined using a quantitative methodology consisting of prescores and postscores on three scales: two subscales of the Personal and Teacher Efficacy Measure (originally validated by Gibson and Dembo [1984] and Dembo and Gibson [1985]), with reported Cronbach's α values (Cronbach, 1951) for this specific study population to confirm the internal consistency of all items; and a Liking for Teaching measure, developed for the study but also with high Chronbach's α values reported in the manuscript.

"Catalyzing Graduate Teaching Assistants' Laboratory Teaching Through Design Research" by Bond-Robinson and Rodriques (2006), a mixed-methods study conducted in a chemistry context, represents the rubric Category 3C. The article integrates data from 83 GTA participants from five iterations of the semester-long "Laboratory Teaching Apprenticeship" program. Measures included an instructor-use instrument for recording 12 types of desired interactions in teaching observations and a student-evaluation measure with a high study-specific Chronbach's α (constituting some level of instrument validation). Data from the first four iterations were compared with a fifth iteration, which included a deeper emphasis on two types of knowledge shown by previous versions to be the most difficult for GTAs to master. We interpreted that comparison of data from an earlier version of the training with data from a later version as a comparison group of sorts.

Level 4: Syntheses and Meta-Analyses

Level 4 articles represent the next level of strength of evidence, including quantitative, comparative syntheses of multiple training programs (Category 4A; $n = 4$) or meta-analyses drawing from data sets produced by multiple case or cohort studies (Category 4B; $n = 0$). Although Category 4A was not described or alluded to in the original pyramid, we believe it represents a stronger data set for broad comparison across types of interventions than cohort studies do and has more in common with a meta-analytic methodology.

Category 4A is exemplified by "Science, Technology, Engineering, and Mathematics Graduate Teaching Assistants Teaching Self-Efficacy" by DeChenne et al. (2012). The study describes the development and validation of an instrument for measuring GTA teaching self-efficacy based on a preexisting instrument for use with a faculty population. Using the validation data set of 253 STEM GTAs from six universities, the authors also examined the correlations between self-efficacy and participation in professional development interventions, as well as amount of teaching experience.

No examples of Category 4B studies were identified, but we included that category within the rubric with the expectation and hope that such meta-analyses may soon be performed.

Level 5: Systematic Reviews

Level 5 ($n = 0$) of the rubric would include articles aggregating the results of multiple prior case, cohort, and synthesis or meta-analysis studies and using them to draw conclusions about a particular type of GTA training or outcome of interest. Similar to Category 4B, no existing articles met this threshold for strength of evidence, but we retained the level within the rubric for future inductive refinement as such articles are developed. We anticipate that subcategories designating different levels of rigor (A, B, etc.) might also arise inductively from an examination of future Level 5 articles. However, we cannot speculate on what those distinctions might be at this time, so the rubric remains undifferentiated at this level.

RESULTS

Our research question 1 asked "What is the distribution and quantity of empirical studies on GTA training programs in the geosciences at each level of strength of the pyramid framework?" The example articles described above include all three available geoscience GTA training studies that met the inclusionary criteria. Only one article was identified at each of the three lower categories of the rubric (one 1B, one 2B, and one 3A). One of those articles (Schade and Bartholomew, 1980) was a 37-y-old program that discussed videotapes of teaching as a novel intervention approach. In another (Dotger, 2011), the intervention itself was not designed to maximize (or even necessarily to result in) GTA learning about teaching—thus, it is not a training program, and an argument could be made that it should not be included in our study (although we chose to include it because of the few articles available).

Regarding research question 2, "What learning objectives are commonly sought by the programs described in that literature?" no data were available: none of the three geoscience GTA training articles stated any explicit learning objectives. Based on the assessments used, we can infer that GTA satisfaction or enjoyment (Schade and Bartholomew, 1980; McManus, 2002), GTA awareness and discussion of the learner (Dotger, 2011) and GTA awareness of specific teaching techniques (Schade and Bartholomew, 1980) were constructs of interest. The example articles from other science disciplines described above also suggested that perceived teaching effectiveness (Hughes and Ellefson, 2013; Bond-Robinson and Rodriques, 2006), GTA teaching self-efficacy (Komarraju, 2008; DeChenne et al., 2012), frequency of specific types of observable GTA teaching behaviors (Bond-Robinson and Rodriques, 2006), and even student grades or performance (Hughes and Ellefson, 2013) might be constructs of interest for future geoscience studies (though not all learning objectives per se).

Our research question 3 asked "What methodologies and methods have been used to evaluate the extent to which the desired objectives are achieved by participants?" The three geoscience studies examined used both qualitative and quantitative measures, as did the examples drawn from other science disciplines. Neither of the two quantitative studies (Schade and Bartholomew, 1980 [Category 3A]; McManus, 2002 [Category 1B]) made use of validated self-report instruments or objective measures of the effect of the intervention (Category 1B simply differentiates a program description that includes some anecdotal data, unlike Categories 2B and 3B, which include validated instruments). Validated instruments, such as used by the DeChenne et al. (2012) STEM GTA teaching self-efficacy instrument or the Student Evaluation of Educational Quality (Marsh, 1982; used in Hughes and Ellefson, 2013) might be drawn from the other science examples described above to lend greater rigor to future geoscience studies. Qualitative studies that use a recognized and structured, analytic process for coding, such as used by Dotger (2011) provide valuable models for evaluating GTA training in a more-holistic way than quantitative processes allow. However, Dotger (2011) used only one coder to evaluate the data; greater trustworthiness (although not distinguished explicitly in the rubric) might be achieved through the use of multiple coders working toward consensus and checking one another's findings. Finally, none of the three geoscience-specific studies included a comparison group of any kind in the research design. Comparison to a prior iteration of a training program (as in Bond-Robinson and Rodriques [2006]) would be one model for future geoscience studies, whereas concurrent control/treatment groups created through blind, random assignment (used by Hughes and Ellefson [2013]) might be considered the ideal approach where circumstances allow.

Finally, research question 4 asked "What specific directions and implications does the existing empirical literature on geoscience GTA training suggest for the development of training programs and future research related to the topic?" Given the paucity of geoscience GTA training research and the diversity of program descriptions in the other science examples, we cannot identify any specific recommendations for the development of GTA training programs at this time. However, areas for future research on geoscience GTA training abound; so little empirical literature exists on this topic and in this context that nearly any study fitting into rubric Levels 2 or 3 and building upon relevant prior theoretical or empirical work would be of value to the community. Specific directions for future research suggested by the example studies above include GTA training for inquiry-based instruction (Hughes and Ellefson, 2013; Linenberger et al., 2014), developmental progressions of GTAs' conceptions of teaching and learning (Dotger, 2011) or teaching practices (Bond-Robinson and Rodriques, 2006) with training, and approaches to guiding GTAs in developing their teaching self-efficacy (Komarraju, 2008; DeChenne et al., 2012).

DISCUSSION

Our original research questions were aimed at finding best practices in GTA training, but with the very few articles available, our expanded search included an analysis of GTA training across STEM disciplines and evolved into the development of a rubric for assessing the strength of evidence of those articles.

The process of developing a rubric grounded inductively in this subset of the DBER literature revealed a series of meaningful distinctions within, as well as differentiations from, the original pyramid framework. For example, although St. John and McNeal (2017; this issue) conceived of the bottom level of the pyramid (practitioner wisdom/expert opinion) to consist of nonpeer-reviewed materials, such as Web sites and commentary pieces, we identified many category 1 articles on this topic that had been peer reviewed. In addition, all of the Level 2 (case study) or Level 3 (cohort study) examples above were conducted by the developers and implementers of the training program being evaluated, which is representative of the broader subset of the GTA training literature across science disciplines we have examined thus far. This is contrary to the suggestion on the pyramid that stronger intervention studies would be taught and evaluated by separate parties. Taken together, we hypothesize that the science education community may have held GTA training studies to a lower standard than other research topics, although practical constraints and lack of familiarity with instruments developed outside the geoscience context likely have a role in these limitations as well. Before the recent emphasis on evidence-based approaches to instruction, communicating the organization and general design of GTA programs through more practitioner-friendly

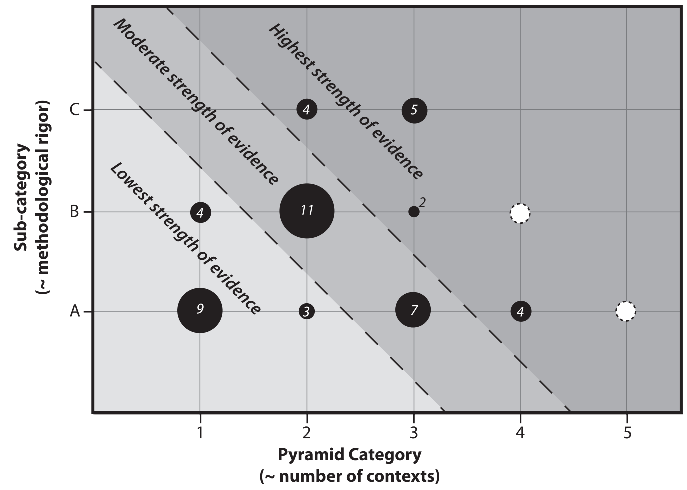

FIGURE 2: Multidimensional characterization of strength of evidence, defined by the pyramid levels (roughly equivalent to number of contexts from which data was derived) on the x-axis and rubric subcategories (roughly equivalent to methodological rigor) on the y-axis. Based on this plot, we propose that strength of evidence might be conceived of as "lowest," "moderate," and "highest" strength "zones." Size of dots plotted on the graph and associated numbers represent the number of studies identified for that category and subcategory in this study. Categories 4B and 5 have not yet been identified.

mechanisms may have initially been more significant and valuable to the community.

The initial pyramid framework measured strength of evidence along only one dimension, the study type. Above Level 1, study type in the pyramid is primarily the number of iterations or contexts over which questions were investigated, which, by definition, increases the generalizability of the findings. However, in evaluating the literature on GTA training, we observed that this unidimensional ranking of strength meant that a Level 2 case study (single iteration of a program) that compared randomly assigned control and treatment groups using multiple validated instruments and objective measures of effect would count as being "less strong" than a Level 3 examination of multiple iterations of a single training program with no comparison group, evaluated using nonvalidated, self-report measures. Elements of methodological rigor in study design and analysis may not necessarily increase the generalizability of the findings but lend greater strength of a different sort (trustworthiness, validity, reliability, etc.). The process of inductively developing subcategories to describe this second dimension of strength suggests to us that "strength of evidence" might sometimes be usefully portrayed as a two-dimensional space, with the pyramid level (roughly equivalent to number of iterations or contexts) on one axis and the subcategory (roughly equivalent to degree of methodological rigor) on the other, such as in Fig. 2. Somewhat equivalent to a metamorphic phase diagram in which a range of pressure/temperature conditions can produce the same rock type, we map onto Fig. 2 a generalized set of strength of evidence "phases" that take both axes into consideration.

Although we believe that this multidimensional characterization of strength of evidence sheds additional light on the value of individual studies for informing practice, level and category alone are not sufficient to do so. For example, a study performing a statistical pre–post comparison would be categorized the same way if it used 30 participants or 5 (which we observed in at least one case in the larger set of 48 studies; Heppner, 1994). Similarly, most of the Level 3 articles in the larger data set had data for only two semesters, a few had three to five semesters of data, and none were longitudinal data sets—although all of those would be classified together.

Subcategories may also obscure relevant detail in two ways. First, for a study to be classified as C, it needed to also meet the minimum conditions for A and B subcategories as well. This approach meant that one study in the larger data set was classified as a Category 3A, even though it used a comparison group, because it failed to use a validated self-report measure or systematic coding approach or to examine any objective measures of effect (White et al., 2012). Second, although any form of comparison group was considered sufficient for subcategory C, not all comparison groups are of equal value; some studies in the larger data set compared two iterations of a training program that happened in two separate years (e.g., Bond-Robinson and Rodriques, 2006), a design which fails to account for the effect of other differences, such as evolving departmental culture. Many studies compared the student evaluation ratings of faculty or instructors to GTAs, although the literature has shown that student expectations (and therefore ratings) of the two populations tends to differ (Kendall and Schussler, 2012). Only a few studies in the larger set (e.g., Nurrenbern et al., 1999) compared simultaneous offerings of two trainings or participating versus nonparticipating GTAs, and fewer still using random assignment procedures (e.g., Hughes and Ellefson, 2013). Nonetheless, all are classified as Subcategory C.

Thus, although categorizations according to the rubric provide general guidance, practitioners or researchers considering patterning future work on an individual study should take caution to carefully evaluate (and potentially improve upon) the specific characteristics of that study.

LIMITATIONS

In this study, we decided to exclude conference manuscripts and focus exclusively on articles published in peer-reviewed journals. We made that decision because although many conference review processes do include peer-review processes, those processes vary widely, and thus, we were unable to determine the type or rigor of review a conference article had undergone before acceptance. However, many conference articles may provide valuable insights into this topic, and certain disciplines (such as engineering) seem to share GTA training program results in that format more frequently than they do in journals. Hence, future reviews of STEM GTA training research might evaluate and include some or all of those references.

This review was limited to graduate student populations receiving training that was specifically tailored toward preparing them to teach in laboratories, recitation sections, or other instructional duties. However, the recent practice of incorporating undergraduates as near-peer teaching assistants to meet instructional demand may, in fact, place different types of pressures on existing training programs. Similarly, the dramatic rise in the prevalence of extradepartmental future-faculty programs (available to graduate students regardless of teaching duties) has resulted in additional related literature. Studies in those two areas related to GTA training may provide additional avenues for future research models, instruments, and outcomes of relevance to the future geoscience GTA training literature.

Finally, the two of us who coded the example articles described in this review achieved perfect agreement. However, agreement using the rubric herein is likely to be more challenging when applied to a larger array of articles and with additional coders added to the group.

FUTURE DIRECTIONS AND CONCLUSIONS

Although the need for high-quality professional development for geoscience GTAs is clear, the research literature on training graduate GTAs in the geosciences has not grown significantly since Kurdziel and Libarkin (2003) broached the topic in the *Journal of Geoscience Education*. A complete review of the full body of the 48 GTA training publications from across the science and other STEM disciplines is currently underway, which may provide additional clarity regarding specific training characteristics associated with specific desirable outcomes. GTA training research from those disciplinary contexts also offers validated instruments and study designs that are likely to be transferrable in some ways to geoscience studies. Nonetheless, well-designed, rigorous, empirical research on GTA training in the geoscience context will be necessary to establish where transfer of instruments, context, and other characteristics are and are not appropriate, to determine the alterations to existing interventions that allow them to have the greatest effect within our disciplinary culture and context, and to validate training approaches for training GTAs in ways of thinking, teaching, and learning that are unique to the geosciences.

Many GERs transition to the field from a geoscience background (Singer et al., 2012), may not have participated in formal courses on social science methodologies, and therefore, may be less familiar with the characteristics that lend differential strength to various study design and analysis approaches. Furthermore, the constraints of a particular institution (such as the number of available GTAs, the number of laboratory sections, the amenability of other faculty in the department to GTA teaching professional development, etc.) may limit study design in a variety of ways, e.g., comparison groups, although ideal, are challenging to implement in most normal educational settings. Thus, the specific rubric descriptions and varied example articles described here provide concrete models for a wide variety of options in study design and analysis and also categorizes those options according to the strength of evidence those study characteristics would produce.

Use of the rubric based on the pyramid benefitted our review of the GTA training literature in a number of concrete ways. First, by giving a general framework for differentiating the literature, we were able to quickly begin sorting the existing literature into those larger levels. That initial sorting allowed us to focus in on each level and, through close reading of the articles themselves, inductively develop meaningful subcategories within the overarching groups. Finally, the rubric allows us to calculate interrater reliability, adding an additional level of rigor to the systematic review process.

These outcomes reinforce the argument for the value of the pyramid's application more widely and provide a process model for systematic literature reviews based on the pyramid. Future rubrics developed for other subsets of the DBER literature may themselves be meaningful for comparing the state of the field across topics and for highlighting gaps in existing studies, depending on which elements of study design are incorporated into the rubric for each topic. Finally, use of the larger levels of the pyramid as a foundation for future rubrics allows easy comparison of the state of the literature based on the distribution of studies across those levels.

REFERENCES

American Association for the Advancement of Science (AAAS). 1990. Science for all Americans. New York: Oxford University Press.

Andrews, T.M., Leonard, M.J., Colgrove, C.A., and Kalinowski, S.T. 2011. Active learning not associated with student learning in a random sample of college biology courses. *CBE-Life Sciences Education*, 10(4):394–405.

Bass, R. 1999. The scholarship of teaching: What's the problem? *Inventio: Creative Thinking About Learning and Teaching*, 1(1):1–10.

Bautista, N.U., Schussler, E.E., and Rybczynski, S.M. 2014. Instructional experiences of graduate assistants implementing explicit and reflective introductory biology laboratories. *International Journal of Science Education*, 36(7):1184–1209.

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

Black, B., and Bonwell, C. 1991. The training of teaching assistants in departments of history. *History Teacher*, 24(4):435–444.

Bond-Robinson, J., and Rodriques, R.A.B. 2006. Catalyzing graduate teaching assistants' laboratory teaching through design research. *Journal of Chemical Education*, 83(2):313.

Brownell, S.E., and Tanner, K.D. 2012. Barriers to faculty pedagogical change: lack of training, time, incentives, and... tensions with professional identity? *CBE-Life Sciences Education*, 11(4):339–346.

Budd, D.A., van der Hoeven Kraft, K.J., McConnell, D.A., and Vislova, T. 2013. Characterizing teaching in introductory geology courses: Measuring classroom practices. *Journal of Geoscience Education*, 61(4):461–475.

Buskist, W. 2013. Preparing the new psychology professoriate to teach past, present, and future. *Teaching of Psychology*, 40(4):333–339.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provisions for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Cronbach, L.J. 1951 Coefficient alpha and the internal structure of tests, *Psychometrika*, 16(3):297–333.

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

DeChenne, S.E., Enochs, L.G., and Needham, M. 2012. Science, technology, engineering, and mathematics graduate teaching assistants teaching self-efficacy. *Journal of the Scholarship of Teaching and Learning*, 12(4):102–123.

Dembo, M.H., and Gibson, S. 1985. Teachers' sense of efficacy: An important factor in school improvement. *The Elementary School Journal*, 86(2):173–184.

Dotger, S. 2010. Offering more than "Here is the textbook": Teaching assistants' perspectives on introductory science courses. *Journal of College Science Teaching*, 39(3):71–76.

Dotger, S. 2011. Exploring and developing graduate teaching assistants' pedagogies via lesson study. *Teaching in Higher Education*, 16(2):157–169.

Druger, M. 1997. Preparing the next generation of college science teachers: Offering pedagogical training to graduate teaching assistants as part of the college reform agenda. *Journal of College Science Teaching*, 26(6):424–427.

Ebert-May, D., Derting, T.L., Henkel, T.P., Middlemis Maher, J., Momsen, J.L., Arnold, B., and Passmore, H.A. 2015. Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE–Life Sciences Education*, 14(2):1–12.

Gibson, S., and Dembo, M.H. 1984. Teacher efficacy: A construct validation. *Journal of Educational Psychology*, 76(4):569.

Gormally, C., Brickman, P., Hallar, B., and Armstrong, N. 2009. Effects of inquiry-based learning on students' science literacy skills and confidence. *International Journal for the Scholarship of Teaching and Learning*, 3(2):article 16, p. 24.

Halpern, D.F., and Hakel, M.D. 2002. Learning that lasts a lifetime: Teaching for long term retention and transfer. *New Directions for Teaching and Learning*, 2002(89):3–7.

Hammrich, P.L. 1996. An example of a discipline-specific instructional program for graduate teaching assistants. *Journal Graduate Teaching Assistant Development*, 3(2):53–57.

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

Hardré, P.L., and Chen, C.H. 2005. A case study analysis of the role of instructional design in the development of teaching expertise. *Performance Improvement Quarterly*, 18(1):34–58.

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

Henderson, C., and Dancy, M.H. 2007. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics–Physics Education Research*, 3(2):020102.

Heppner, M.J. 1994. An empirical investigation of the effects of a teaching practicum on prospective faculty. *Journal of Counseling and Development*, 72(5):500–507.

Holmes, N.G., Ives, J., and Warren, M. 2013. Teaching assistant professional development by and for TAs. *The Physics Teacher*, 51:218–219.

Hughes, P.W., and Ellefson, M.R. 2013. Inquiry-based training improves teaching effectiveness of biology teaching assistants. *PloS One*, 8(10):e78540.

Kane, R., Sandretto, S., and Heath, C. 2002. Telling half the story: A critical review of research on the teaching beliefs and practices of university academics. *Review of Educational Research*, 72(2):177–228.

Kendall, K.D., Niemiller, M.L., Dittrich-Reed, D., Chick, L.D., Wilmoth, L., Milt, A., Burt, M., Lopes, N., Cantwell, L., Rubio, L., Allison, A, and Schussler, E.E.. 2013. Departments can develop teaching identities of graduate students. *CBE–Life Sciences Education*, 12(3):316–317.

Kendall, K.D., and Schussler, E.E. 2012. Does instructor type matter? Undergraduate student perception of graduate teaching assistants and professors. *CBE–Life Sciences Education*, 11(2):187–199.

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

Krathwohl, D.R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4):212–218.

Krystyniak, R.A., and Heikkinen, H.W. 2007. Analysis of verbal interactions during an extended, open inquiry general chemistry laboratory investigation. *Journal of Research in Science Teaching*, 44(8):1160–1186.

Kurdziel, J.P., and Libarkin, J.C. 2003. Research methodologies in science education: Training graduate teaching assistants to teach. *Journal of Geoscience Education*, 51(3):347.

Linenberger, K., Slade, M.C., Addis, E.A., Elliott, E.R., Mynhardt, G., and Raker, J.R. 2014. Training the foot soldiers of inquiry: Development and evaluation of a graduate teaching assistant learning community. *Journal of College Science Teaching*, 44(1):97–107.

Luft, J.A., Kurdziel, J.P., Roehrig, G.H., and 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(3):211–233.

Lund, T.J., Pilarz, M., Velasco, J.B., Chakraverty, D., Rosploch, K., Undersander, M., and 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):article 18, p. 12.

Macdonald H., Feig, A., Lukes, L., McNeal, K., Riggs, E., St. John, K. 2015. Synthesizing geoscience education research: Where are we? What is the path forward? Available at: http://serc.carleton.edu/earth_rendezvous/2015/morning_workshops/w3/index.html (accessed 10 October 2016).

Manduca, C.A., Iverson, E.R., Luxenberg, M., Macdonald, R.H., McConnell, D.A., Mogk, D.W., and Tewksbury, B.J. 2017. Improving undergraduate STEM education: The efficacy of discipline-based professional development. *Science Advances*, 3(2):e1600193.

Marsh, H.W. 1982. SEEQ: A reliable, valid, and useful instrument for collecting students; evaluations of university teaching. *British Journal of Educational Psychology*, 52(1):77–95.

McManus, D.A. 2002. Developing a teaching assistant preparation program in the School of Oceanography, University of Washington. *Journal of Geoscience Education*, 50(2):158–168.

National Research Council (NRC). 1996. National Science Education Standards. Washington, D.C.: National Academy Press.

National Research Council (NRC). 2000. Inquiry and the national science education standards. Washington DC: National Academies Press.

National Research Council (NRC). 2012. Discipline based education research: Understanding and improving learning in undergraduate science. Washington DC: National Academies Press.

Oleson, A., and Hora, M.T. 2014. Teaching the way they were taught? Revisiting the sources of teaching knowledge and the role of prior experience in shaping faculty teaching practices. *Higher Education*, 68(1):29–45.

Nurrenbern, S.C., Mickiewicz, J.A., and Francisco, J.S. 1999. The impact of continuous instructional development on graduate and undergraduate students. *Journal of Chemical Education*, 76(1):114–119.

O'Neal, C., Wright, M., Cook, C., Perorazio, T., and Purkiss, J. 2007. The impact of teaching assistants on student retention in the sciences: Lessons for TA training. *Journal of College Science Teaching*, 36(5):24–29.

Rushin, J.W., De Saix, J., Lumsden, A., Streubel, D.P., Summers, G., and Bernson, C. 1997. Graduate teaching assistant training: A basis for improvement of college biology teaching and faculty development? *The American Biology Teacher*, 59(2):86–90.

Ryker, K., and McConnell, D. 2014. Can graduate teaching assistants teach inquiry-based geology labs effectively? *Journal of College Science Teaching*, 44(1):56–63.

Ryker, K.D., and McConnell, D.A. 2017. Assessing inquiry in physical geology laboratory manuals. *Journal of Geoscience Education*, 65(1):35–47.

Sandi-Urena, S., and Gatlin, T. 2013. Factors contributing to the development of graduate teaching assistant self-image. *Journal of Chemical Education*, 90(10):1303–1309.

Schade, W. R., and Bartholomew, R. B. 1980. Analysis of geology teaching assistant reaction to a training program utilizing video-taped teaching episodes. *Journal of Geological Education*, 28(2):96–102.

Schussler, E.E., Read, Q., Marbach-Ad, G., Miller, K., and Ferzli, M. 2015. Preparing biology graduate teaching assistants for their roles as instructors: An assessment of institutional approaches. *CBE–Life Sciences Education*, 14(3):article 31, p. 11.

Singer, S.R., Nielsen, N.R., and Schweingruber, H.A. 2012. Discipline based education research. Washington, DC: The National Academies.

St. John, K., and McNeal, K. 2017. The strength of evidence pyramid: One approach for characterizing the strength of evidence of Geoscience Education Research (GER) community claims. *Journal of Geoscience Education*, 65(4):363–372.

Sundber, M.D., Armstrong, J.E., and Wischusen, E.W. 2005. A reappraisal of the status of introductory biology laboratory education in U.S. colleges and universities. *The American Biology Teacher*, 67(9):526–529.

Thornburg, N.A., Wood, F.E., and Davis, W.E. 2000. Keeping established teaching assistant training programs vital: What does it take? *The Journal of Graduate Teaching Assistant Development*, 7(2):77–83.

Travers, P.L., 1989. Better training for teaching assistants. *College Teaching*, 37:147–149.

Turpen, C., and Finkelstein, N. D. 2009. Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction. *Physical Review Special Topics–Physics Education Research*, 5(2):020101.

White, P.J., Syncox, D., Heppleston, A., Isaac, S., and Alters, B. 2012. Putting research into practice: Pedagogy development workshops change the teaching philosophy of graduate students. *The Canadian Journal of Higher Education*, 42(1):98–111.

Wood, W.B. 2009. Innovations in teaching undergraduate biology and why we need them. *Annual Review of Cell and Developmental Biology*, 25:93–112

Yerrick, R., Parke, H., and Nugent, J. 1997. Struggling to promote deeply rooted change: The "filtering effect" of teachers' beliefs on understanding transformational views of teaching science. *Science Education*, 81(2):137–159.