

# Quotation Accuracy Matters: An Examination of How an Influential Meta-Analysis on Active Learning Has Been Cited

Amedee Marchand Martella ,  
Jane Kinkus Yacilla , Ronald C. Martella,  
Nancy E. Marchand-Martella, Zafer Ozen,  
Tugce Karatas, Helen H. Park,  
Alexandra Simpson, and Jeffrey D. Karpicke  
*Purdue University*

*When previous research is cited incorrectly, misinformation can infiltrate scientific discourse and undermine scholarly knowledge. One of the more damaging citation issues involves incorrectly citing article content (called quotation errors); therefore, investigating quotation accuracy is an important research endeavor. One field where quotation accuracy is needed is in the learning sciences given its impact on pedagogy. An integral article in pedagogical discussions surrounding how to teach at the college level is the meta-analysis on active learning by Freeman et al. The Freeman et al. meta-analysis compared active learning to traditional lecture in terms of its effects on student learning and has been important in national initiatives on STEM (science, technology, engineering, and mathematics) reform. Given its influence coupled with the impact quotation errors could have in scientific discourse, we used citation context analysis to analyze whether assertions in the citing text that related to the efficacy of lecture and active learning were supported by what was explicitly stated in the cited meta-analysis. Assertions were analyzed under supported, unsupported, or irrelevant for purposes of study categories. The most prevalent supported category related to active learning being more effective than lecture; the most prevalent unsupported category related to the effectiveness of specific activities/approaches other than the general approach of active learning. Overall, the percentage of supported assertions was 47.67%, and the percentage of unsupported assertions was 26.01%. Furthermore, the percentage of articles containing at least one unsupported assertion was 34.77%. Proactive measures are needed to reduce the incidence of quotation errors to ensure robust scientific integrity.*

**KEYWORDS:** citation accuracy, quotation accuracy, active learning, citation context analysis

Scholarly knowledge advances through the publication of scholarly research. Published articles are said to be derivative in nature in that they are connected to prior research investigations (Ziman, 1969). Authors support their own investigations by strategically citing those researchers who influenced their work (American Psychological Association, 2020). However, if current or prior research contains inaccuracies that are not corrected, scholarly knowledge can be undermined. For example, errors can be propagated unintentionally (Awrey et al., 2011; Parse, 1996) which may cause subsequent readers to become misinformed (Mogull, 2017; Schulmeister, 1998).

### *Citation Accuracy*

One way to support scientific communication and integrity is through *accurately* citing the scholarly work of others (Gasparyan et al., 2015). If, however, cited articles are the subject of errors made through the citation process, this misinformation can infiltrate scientific discourse and the dissemination of incorrect or incomplete information can occur. Errors can spread quickly in the digital era and those that persist can become *true* and have serious ramifications that could affect the integrity of scholarship and investigation (Wilson, 2015). One could also imagine how errors that lead to misinformation could find their way into policy decisions, financial investments, funding opportunities, and other large-scale initiatives. Unfortunately, citation accuracy has been and continues to be an issue in journal publishing (e.g., Mogull, 2017; Onwuegbuzie et al., 2011; Santini, 2018).

Several types of errors can affect citation accuracy.<sup>1</sup> These errors may include issues with (a) bibliographic information (e.g., misspelling author names or incorrectly listing publication year), (b) reported data (e.g., misquoting effect size or sample size), (c) secondary versus primary sources (e.g., citing a review paper rather than the primary source), and (d) claims, paraphrasing, interpretations, or directly quoted statements (e.g., making unsubstantiated claims, changing important information through paraphrasing, misinterpreting the original results, or omitting/adding/changing words in a direct quote). Citation errors that relate to claims, paraphrasing, interpretations, and directly quoted statements are termed *quotation errors* and have a negative impact on *quotation accuracy*.

### *Quotation Accuracy*

Quotation errors are serious because they can disseminate misinformation (Davids et al., 2010). The rate of these troubling errors tends to vary in the research literature.<sup>2</sup> For example, the quotation error rate was 6.7% in nursing (Schulmeister, 1998); 11.3% (Drake et al., 2013), 15.0% (Teixeira et al., 2013), and 18.3% (Todd et al., 2007) in ecology; 12.4% in manual therapy (Gosling et al., 2004); 14.5% (Mogull, 2017) to 15.0% in general medicine (de Lacey et al., 1985); 15.6% in veterinary medicine (Hinchcliff et al., 1993); 17% in otolaryngology/head and neck surgery (Fenton et al., 2000); 19.0% in anatomy (Lukić et al., 2004); 19.2%

in physical geography (Haussmann et al., 2013); 20.0% in foot and ankle surgery (Luo et al., 2013); 24.0% in drug therapy (Neihouse & Priske, 1989); 24.2% in marine biology (Todd et al., 2010); and 38.0% in pediatric orthopedics (Davids et al., 2010). Additionally, in a systematic literature review and meta-analysis of the quotation accuracy in medical journal articles, Jergas and Baethge (2015) found the error rate to be approximately 25.0%. The variation in error rates across these disciplines could be due to the types of quotation errors examined (e.g., issues with claims, paraphrasing, interpretations, or directly quoted statements), the discipline of the articles being analyzed, or other methodological considerations (e.g., random sample vs. specific articles/journals targeted). Despite these varying error rates, the common thread is that quotation errors can and do occur when scholarly works are cited. Even in the scientific enterprise, error is a human phenomenon; therefore, pinpointing what may contribute to errors is key to improving the transmission of sound scientific knowledge (Brown et al., 2018). We should not take quotation accuracy for granted, as quotation errors can lead readers to have doubt about what they are reading (Jergas & Baethge, 2015), in addition to propagating misinformation that may be cited in future research endeavors (Bareket et al., 2020). Misrepresenting the content of a cited article has been said to be “one of the potentially most damaging violations of good academic referencing” (Harzing, 2002, p. 133).

An often-cited article in the literature on quotation accuracy is by Harzing (1995). Harzing argued that a commonly held belief among the management community of high failure rates of employees living and working outside their native countries (i.e., expatriates) was actually a *myth* that stemmed from three largely misquoted articles. Her deep reading of the management literature during her doctoral studies led her to question pervasive statements about expatriate failure rates being very high as she did not observe corresponding empirical data in the literature. She traced the myth of expatriate failure rates back to the main source articles by examining references in recent articles on the topic and working her way back through the literature (i.e., examining *citation chains*). In a subsequent article, Harzing (2002) lamented that her 1995 article had done little to correct the way that expatriate failure rates continued to be discussed, indicating that it may be difficult to change inaccurate disciplinary beliefs once they have been established. In the spirit of Harzing (1995), Sanz-Martín et al. (2016) explored the chain of citations in a body of research literature to question the “fact” that jellyfish blooms were increasing globally. When citing and cited papers were analyzed, the data did not support globally increasing blooms but did support some regional trends. Sanz-Martín et al. found that almost half the assertions made were not supported by the cited articles, and they identified one review paper that had an obvious influence over subsequent papers. Relatedly, Korpela (2010) found that even when primary articles have been *publicly retracted*, affirmative citations of these works could continue for years.

### *Quotation Accuracy in the Learning Sciences: An Examination of Active Learning*

Given the seriousness of misrepresenting the content of a cited article, one important field in need of an examination of quotation accuracy is the learning

sciences due to its impact on pedagogy and student academic success. *How to teach* (i.e., instructional approaches) is an important question of study and discussion in this interdisciplinary field. In particular, an instructional approach that has gained traction in the learning sciences is *active learning*. Active learning has been defined as “instructional activities involving students in doing things and thinking about what they are doing” (Bonwell & Eison, 1991, p. iii). This approach is often considered alternative to the traditional lecture method, which may be a factor in why active learning initially gained popularity. Traditional lecture is a prominent method used in college courses (Stains et al., 2018) but has been criticized as being passive in nature with less student learning than more actively engaging methods (Deslauriers et al., 2019). Some opponents of lecture go so far as to claim that the lecture method is an unethical practice (Wieman, 2014) and to ask why lecture has refused to go away (Pickles, 2016).

When discussing active learning and/or active learning versus traditional lecture in higher education, one highly influential review article often cited is the Freeman et al. (2014) meta-analysis. The number of citations an article has is often used as a measure of how large a role it plays in scientific discourse (Lopresti, 2010). Based on citation data drawn from *Web of Science: Core Collection (WOSCC)* and *Scopus*, the Freeman et al. meta-analysis has been cited at least 2,500 times (as of August 8, 2020). In addition to being highly cited, Freeman et al. noted their paper was “the largest and most comprehensive meta-analysis of the undergraduate STEM education literature to date” (p. 8412). Their meta-analysis included 225 studies comparing student performance under active learning versus traditional lecture conditions in science, technology, engineering, and mathematics (STEM) undergraduate courses. Freeman et al. defined active learning as engaging “students in the process of learning through activities and/or discussion in class, as opposed to passively listening to an expert. It emphasizes higher order thinking and often involves group work” (pp. 8413–8414). Their results showed that student performance was, on average, almost half a standard deviation higher with active learning compared with lecture and that failure rates were about 1.5 times higher in traditional lecture courses compared with those with active learning.

When Freeman et al. (2014) is cited, it is often used as support for the adoption of active learning and the reduction or elimination of lecture in STEM courses. The meta-analysis has been cited in national and university-level initiatives and grants aimed to transform college courses to include active learning (e.g., Association of American Universities, 2017; Chasteen, 2016). The citing of Freeman et al. in these initiatives and grants is not surprising, as systematic reviews are considered a useful tool to inform policymakers and practitioners about what does and does not work especially when it comes to making decisions for public policy and practice (Gough et al., 2013). Beyond these initiatives, those involved in the institutional transformations taking place at colleges and universities across the country often cite Freeman et al. as support for their adoption of active learning.

Professional organizations and media outlets have also promoted the Freeman et al. (2014) meta-analysis and have questioned the value of lecture. For example, Bajak’s (2014) article in *Science* was titled, “Lectures Aren’t Just Boring, They’re Ineffective, Too, Study Finds.” In this article, physicist Eric Mazur of Harvard

University discussed the importance of the Freeman et al. findings stating that “the impression I get is that it’s almost unethical to be lecturing if you have this data” and there is “an abundance of proof that lecturing is outmoded, outdated, and inefficient” (para. 4). A news release from the National Science Foundation (2014) was titled, “Enough With the Lecturing.” One comment made by Scott Freeman, lead author of the meta-analysis on active learning, in this news release was, “We’ve got to stop killing student performance and interest in science by lecturing” (para. 12). Lederman (2014) noted in “A Boost for Active Learning” in *Inside Higher Education* that students are more successful in their courses when instructors use methods other than lecturing. In each of the aforementioned articles, the Freeman et al. meta-analysis was cited as evidence as to why changes were needed in higher education STEM courses. Prominent researchers such as Carl Wieman (2014; Nobel Laureate in Physics) have also made hard-hitting claims against lecture, saying that lecture is “the pedagogical equivalent of blood-letting” (p. 8320). He also said that the Freeman et al. meta-analysis makes a powerful case for adopting active learning in our teaching practices as well as for redirecting our approach to further research on active learning (e.g., dropping traditional lecture as the standard for comparison).

Despite the positive press surrounding the Freeman et al. (2014) meta-analysis, several researchers caution the conclusions that have been drawn from this study. For example, Cooper and Stowe (2018) reported that because the data were not disaggregated in the meta-analysis, it is impossible to draw any conclusions about the effectiveness of specific approaches that are categorized as active learning; one can only say the umbrella category of active learning is more effective than the umbrella category of traditional lecture. These more general categories of active learning and traditional lecture often conceal which variables actually affect student learning such as the pedagogical methods used, the amount of time devoted to these methods, and the instructor and student characteristics in these courses (Bernstein, 2018). Thus, it is unwarranted to eliminate lecture and adopt active learning when it is unclear as to which specific aspects of both general approaches are effective or ineffective for student learning (Zakrajsek, 2018).

### *Purpose of the Present Study*

Despite the fact that issues with the Freeman et al. (2014) meta-analysis have been brought to light, the prominence of the meta-analysis as support for changing courses from being lecture-based to active learning-based is quite evident throughout the literature, across institutions of higher education, and in large-scale initiatives. Further evidence of how instrumental the meta-analysis is relates to its very nature (i.e., a meta-analytic review paper)—that is, review papers have the potential to shape the field (Murphy et al., 2017) in a positive or negative direction depending on the accuracy of how they are quoted. In light of the (a) influence of the Freeman et al. meta-analysis, (b) concerns that have been raised about the meta-analysis, and (c) issues that can jeopardize quotation accuracy due to incorrect interpretations, we made the Freeman et al. meta-analysis the focus of our quotation accuracy study. The purpose of this study was to examine the quotation accuracy of those articles that cited Freeman et al. by analyzing whether the claims, paraphrasing, interpretations, and directly quoted statements (herein,

these terms are collectively called *assertions*) made in these citing articles were supported by what was explicitly stated in the meta-analysis. With regard to direct quotes, we examined these statements based on their content rather than examining their word-to-word matching with the primary text. The specific analysis we used to analyze assertions is called a *citation context analysis*; this methodology allows for the examination of how citing authors have referred to the work of others by analyzing the text surrounding a citation (Small, 1982).

We focused our efforts on examining those assertions made about active learning or lecture's efficacy (i.e., assertions related to the impact of active learning or lecture on student outcomes) given the meta-analysis' influence on moving instruction away from lecture and toward a model of active learning. These assertions could have serious ramifications with regard to course transformations, policy decisions, initiatives, and future research endeavors and thus needed to be assessed on whether or not they could be explicitly traced back to text in Freeman et al. (2014). We did not categorize and analyze assertions in our citation context analysis that were outside of our efficacy focus beyond labeling them *irrelevant for purposes of this study*. *Irrelevant* assertions included, for example, the definition of active learning, the statistical procedures used by Freeman et al., and generic overview statements about meta-analyses in education.

We had three main research questions in this citation context analysis. First, what was the percentage of assertions made by citing articles that could or could not be directly traced back to the Freeman et al. (2014) text (i.e., assertions that were or were not supported by Freeman et al.)? Second, given that Freeman et al. is cited across various disciplines, what was the percentage of supported and unsupported assertions across these different disciplines? Third, given that Freeman et al. was published in 2014, what was the percentage of supported and unsupported assertions across time?

## Method

### *Search Procedure*

To examine quotation accuracy, we conducted a citation context analysis of the Freeman et al. (2014) meta-analysis. The second author searched the two main bibliographic databases that provide citation information, *WOSCC* and *Scopus*, on June 13, 2019, to obtain articles that cited Freeman et al. These citing documents were limited to primary articles or reviews; thus, all other document types such as editorials, book chapters, letters, meeting abstracts, news stories, and proceedings papers were excluded. Results were saved in an EndNote library, and duplicates were removed using a validated method developed for systematic reviews (Bramer et al., 2016). The resulting data set included 1,124 articles. Due to 29 articles being written in a language other than English and five articles citing the meta-analysis in the reference section and not in the main text, the final data set included 1,090 articles (thus, a total of 34 articles was excluded).

### *Coding Information*

The first, fifth, sixth, and seventh authors divided the 1,090 articles into four sets, with two sets containing 273 articles and other the two sets containing 272 articles. Each of these authors took one of these sets and searched each article



PDF for the *citing text*—that is, text containing a Freeman et al. (2014) citation or text following the citation that continued the discussion of the meta-analysis. This citing text was copied and pasted into a shared Word document with columns for article ID, author(s), year published, journal title, journal discipline, and citing text. Text before and after the citing text was also copied and pasted to provide contextual information. Once all of the articles were searched and the citing text was found, the seventh and eighth authors each went through half of the articles (545 articles each) and ensured no citing text was overlooked in each article and that the copied and pasted text included the correct meta-analysis citation.

To determine quotation accuracy and whether or not the citing text included assertions related to the efficacy of lecture and active learning, overarching assertion categories were developed that covered possible *supported* and *unsupported* assertions (see Table 1 for these assertion categories). Determining these categories occurred in four steps. In Step 1, the first, fifth, sixth, and seventh authors read the Freeman et al. (2014) meta-analysis and identified all text related to the efficacy of active learning and/or lecture presented in the Introduction, Method, Results, and Discussion sections that, if cited, would be deemed *supported* (i.e., the assertion made in the citing text could be directly traced back to the information contained in the meta-analysis). Many of the statements about active learning and lecture throughout the sections of the meta-analysis were classified under a common category. For example, the meta-analysis included many statements about how active learning affected student performance compared with lecture, so we had an early category labeled *active learning's effectiveness*. If a citing article stated that active learning was more effective than lecture and cited the meta-analysis, we could identify the accompanying text from the meta-analysis to which we could refer based on the assertion category *active learning's effectiveness*. If any of the assertion categories overlapped and could be integrated, we combined them into one category.

In Step 2, we developed assertion categories related to our efficacy focus that were *unsupported* (i.e., the assertion made in the citing text could not be substantiated by the information presented in the meta-analysis). To develop these *unsupported* assertion categories, the first, fifth, sixth, and seventh authors examined the assertions that fell under the *supported* assertion categories and listed the ways in which these assertions could be misconstrued. To illustrate this, Freeman et al. (2014) found that active learning was more effective than lecture, but they did not find nor did they state that lecture was ineffective. Rather, they found and stated that overall active learning was more effective than lecture. It could be the case that lecture is effective when compared with other instructional approaches, for example. If a citing article stated that lecture is an ineffective approach and cited the meta-analysis, we created a category labeled *lecture is ineffective* and deemed it *unsupported*.

In Step 3, the first, fifth, sixth, and seventh authors read a randomly chosen set of 40 articles from our data set and coded the citing text using our previously identified *supported* and *unsupported* assertion categories. If any of the citing text did not fall under one of our assertion categories but was related to instructional efficacy, additional categories were developed.

**TABLE 1**

*Illustrative examples of assertion categories*

Category	Example assertions referring to Freeman et al. (2014) meta-analysis
Supported	<p>S1: Active learning is more effective than lecture/increases student learning</p> <p>S2: Lecture is not the best approach to teaching/needs to be abolished or left behind in favor of active learning</p> <p>S3: Second generation research should be conducted (move beyond using lecture as control condition)</p>
Unsupported	<p>US1: Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective</p>
	<p>Example assertions referring to Freeman et al. (2014) meta-analysis</p> <p>“Active learning has been shown to increase student performance in science, engineering, and mathematics [9].” (Schönwälder et al., 2017, p. 2)</p> <p>“A recent meta-analysis of 225 research studies found that test scores were higher and the number of failures was lower for active learning versus lecture classes [27].” (Wenzel, 2014, p. 5245)</p> <p>“Traditional lecture has a firm grasp in both countries [India and U.S.], something that needs to yield to active learning alternatives based on a recent <i>Proceedings of the National Academy of Sciences</i> metastudy<sup>9</sup>.” (Pienta, 2015, p. 399)</p> <p>“Based on their results [Freeman et al., 2014], the authors called for the abandonment of traditional lectures in favor of active learning in STEM courses stating that ‘if the experiments analyzed here had been conducted as randomized controlled trials of medical interventions, they may have been stopped for benefit . . . because the treatment being tested was clearly more beneficial.’” (Wilcox et al., 2018, p. 202)</p> <p>“The evidence is so clear that Freeman and colleagues [13] question ‘the continued use of traditional lecturing as a control in research studies.’” (Grissom et al., 2017, p. 1)</p> <p>“Freeman and co-workers, therefore, argue ‘it was no longer appropriate to use lecture teaching as the comparison standard.’” (Bubou et al., 2017, p. 329)</p> <p>“Despite a meta-analysis of 225 flipped STEM courses (Freeman et al., 2014) demonstrating that flipped learning can improve student exam grades by 6% and decrease failure rates by 1.5 times.” (McCollum et al., 2017, p. 1)</p> <p>“Researchers have evaluated Twitter’s use in several university programs including health care (Smith &amp; Lambert, 2014), business, management, computer systems (Evans, 2014; Menkhoff, Chay, Bengtsson, Woodward, &amp; Gan, 2015), human kinetics and recreation (Rohr &amp; Costello, 2015; Rohr et al., 2015), the hard sciences (Freeman et al., 2014), and social work (Hitchcock &amp; Young, 2016). These studies have illustrated the positive impact this social media service can have on student-student and student-instructor relationships, as well as student engagement and learning.” (Peters et al., 2018, p. 5)</p> <p>Also labeled under US4</p>

(continued)



TABLE 1 (continued)

Category	Example assertions referring to Freeman et al. (2014) meta-analysis
<i>US2</i> : Lecture is ineffective	<p>“There is overwhelming evidence that lecture-based instruction is ineffective (Bradforth et al., 2015; Freeman et al., 2014).” (Watson et al., 2018, p. 1)</p> <p>“Many health systems continue to offer one-and-done lecture formats to share information, an approach that is ineffective and increasingly redundant in an information age (Freeman et al., 2014).” (Davis &amp; McMahon, 2018, p. 892)</p> <p>“Freeman et al. (2014) carries out a meta-analysis of 225 studies comparing the effects of active and passive learning techniques in science, technology, engineering, and mathematics. They find that active learning increases exam scores and decreases failure rates relative to traditional lecturing, with particular benefit for students from disadvantaged backgrounds and for women in male-dominated fields.” (Bayer &amp; Rouse, 2016, p. 234)</p> <p>“Although nontraditional and active learning strategies have been demonstrated effective in secondary and medical education (12, 18, 19), there have been few published evaluations of the effectiveness of hands-on physiological activities in middle and high school student populations.” (Becker et al., 2017, p. 137)</p>
<i>US3</i> : Active learning is beneficial for specific populations/course topics (e.g., minorities, women, genetics material, calculus)	
<i>US4</i> : Active learning improves measures above and beyond learning/retention (e.g., motivation, engagement, attitude, interpersonal skills)	<p>“Use of active learning strategies improves student focus, interactions between peers, discovery of misconceptions and retention of the learned concepts (Freeman et al., 2014).” (Ylostalo, 2019, p. 518).</p> <p>“PBL [problem-based learning] also facilitates incorporation of active-learning techniques into the classroom, which have been shown to increase student performance, and helps students bring theory into practice and learn how to engage in life-long learning [13].” (Clyne &amp; Billiar, 2016, p. 7)</p>
Irrelevant	
Assertion does not relate to the efficacy of active learning and/or lecture	<p>“This course was designed based on a student-centered teaching approach, which emphasizes shifting the focus of activity from the teacher to the learner (e.g., Freeman et al. 2014; Weiner 2002).” (Schmitt-Harsh &amp; Harsh, 2017, p. 552)</p> <p>“In cases where articles reported multiple assessments of a single course subject, we selected the assessment that was most summative, as suggested by Freeman et al. [27].” (Hew &amp; Lo, 2018, p. 3)</p>

Note. STEM = science, technology, engineering, and mathematics.

In Step 4, the third and fourth authors examined the developed assertion categories and compared them with the Freeman et al. (2014) text to ensure what was deemed *supported* or *unsupported* was in fact correct (no issues were identified; i.e., quotations were accurate). The final assertion categories for *supported* included (a) active learning is more effective than lecture/increases student learning (labeled with an *S1* category code); (b) lecture is not the best approach to teaching/needs to be abolished or left behind in favor of active learning (labeled with an *S2* category code); and (c) second-generation research should be conducted (move beyond using lecture as control condition) (labeled with an *S3* category code; see Table 1 for illustrative examples of the three *supported* categories). The final assertion categories for *unsupported* included (a) specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective (labeled with a *US1* category code); (b) lecture is ineffective (labeled with a *US2* category code); (c) active learning is beneficial for specific populations/course topics (e.g., minorities, women, genetics material, calculus; labeled with a *US3* category code); and (d) active learning improves measures above and beyond learning/retention (e.g., motivation, engagement, attitude, interpersonal skills; labeled with a *US4* category code; see Table 1 for illustrative examples of the four *unsupported* categories). Assertions made about the meta-analysis that did not fall under any of these categories were labeled *irrelevant for purposes of this study*.

It should be noted that unless the citing authors specified what they meant by *learning* and *retention* and their specifications were outside of the student outcomes measured in the meta-analysis, we deemed the general use of these terms as *supported*. The coding of assertions as *unsupported* was conservative; that is, we attempted to keep subjective interpretation of the citing text to a minimum if the text was too general or nonspecific for us to say that the information it contained was not found in the meta-analysis. As previously stated, our focus was on quotation accuracy—what was explicitly stated and measured in the meta-analysis and whether the citing text could be directly traced back to this information.

#### *Interrater Agreement*

With the three *supported* assertion categories and four *unsupported* assertion categories, the first and third author coded all citing text. The coding process entailed tagging separate assertions as falling under one or more of the seven categories or falling under the *irrelevant* category using the category codes *S1*, *S2*, *S3*, *US1*, *US2*, *US3*, and *US4*, or *irrelevant* that were mentioned above. The first author served as the primary coder and the third author served as the secondary coder. To ensure accurate coding of assertions in the citing text, several steps were taken before the categorization of the assertions was compared with the secondary coder's categorization of the assertions. These steps included the following.

First, the fifth (Rater 1), sixth (Rater 2), seventh (Rater 3), and eighth (Rater 4) authors were initially trained on the coding system (led by the primary coder) by reviewing citing text within 25 articles and comparing their identification and subsequent categorization of the separate assertions made in the text they reviewed. Second, disagreements were discussed, and another set of 25 articles were reviewed and discussed. Third, after this training period, the raters were

assigned their official set of articles and subsequent citing text to code. Their assigned set was determined at random and counterbalanced to ensure two raters coded the assertions in the citing text. Through the process of categorizing the separate assertions made in the citing text in each article, the primary coder and raters determined there to be 1,630 separate assertions across the 1,090 articles. Raters 1 through 4 coded 48.98%, 62.36%, 39.15%, and 49.87% of these 1,630 assertions, respectively; the primary coder coded 100% of these 1,630 assertions. Thus, the assertions in the citing text had been categorized three separate times (by two raters plus the primary coder).

Fourth, the three category codes (one from the primary coder and one from each of two raters) for each assertion were then compared with each other. Interrater agreement was calculated as a percentage of agreement and as a Cohen's kappa. Finally, after the primary coder's and raters' categorizations of the assertions were compared, any coding disagreements between the primary coder and at least one of the raters were resolved by the primary coder. These final decisions on the categorization for each assertion would then be compared with the secondary coder's independent categorization decision for each assertion.

As shown in Table 2, agreement percentages between raters ranged from 77.82% (Raters 2 and 4) to 91.40% (Raters 1 and 3). Agreement percentages between the primary coder and the raters ranged from 69.00% (with Rater 3) to 82.76% (with Rater 2). Cohen's kappa levels between raters ranged from 0.61 (Raters 2 and 4) to 0.87 (Raters 1 and 3). Cohen's kappa levels between the primary coder and the raters ranged from 0.55 (with Rater 3) to 0.59 (with Rater 2). These kappa values fell into the Landis–Koch *moderate agreement*, *substantial agreement*, and *almost perfect agreement* levels (Landis & Koch, 1977). Additionally, Krippendorff's alpha was calculated for level of agreement across the primary coder and two of the raters. The range of agreement was from 0.59 (with Raters 2 and 4) to 0.67 (with Raters 1 and 4).

As previously stated, once the primary coder and two raters had agreed on the categorization of the assertions in the citing text, the primary coder compared them with the secondary coder's independently determined categorizations (see Table 2). Percentage of agreement was 95.50% with an obtained Cohen's kappa of 0.92 (*almost perfect agreement* level). Any disagreements between the primary and secondary coders were discussed. Agreements were reached on 86.00% of these disagreements; the unresolved disagreements were related to what the citing authors meant by *retention*. After discussion, a decision was reached to use the conservative method (as mentioned previously) of not counting the use of *retention* as *unsupported* unless it was defined in such a way that could not be traced back to Freeman et al. (2014). After this discussion, there was a final agreement on 99.38% of the categorizations. The primary coder resolved any remaining disagreements.

### Data Analysis

Data were organized and analyzed at two quotation accuracy levels. We conducted analyses at the (a) assertion level ( $n = 1,630$ ) and (b) article level ( $n = 1,090$ ). At the assertion level, the assertions under each assertion category were added together for an overall category total. At the article level, any article that

**TABLE 2**  
*Interrater agreement percentage (%), Cohen's kappa (k), Krippendorff's alpha ( $\alpha$ ), and percentage of assertions mutually coded (MC)*

Coder/Rater	With Rater 1	With Rater 2	With Rater 3	With Rater 4	With Raters 1 and 2	With Raters 1 and 3	With Raters 1 and 4	With Raters 2 and 3	With Raters 2 and 4	With Raters 3 and 4	With secondary coder
Primary coder	% = 69.45 $\kappa$ = 0.58 MC = 48.98%	% = 82.76 $\kappa$ = 0.59 MC = 62.36%	% = 69.00 $\kappa$ = 0.55 MC = 39.15%	% = 70.80 $\kappa$ = 0.58 MC = 49.87%	— $\alpha$ = 0.66 MC = 31.27%	— $\alpha$ = 0.65 MC = 5.86%	— $\alpha$ = 0.67 MC = 11.85%	— $\alpha$ = 0.65 MC = 3.18%	— $\alpha$ = 0.59 MC = 17.91%	— $\alpha$ = 0.61 MC = 20.11%	% = 95.50 $\kappa$ = 0.92 MC = 100%
Rater 1	— —	% = 86.90 $\kappa$ = 0.77 MC = 31.27%	% = 91.40 $\kappa$ = 0.87 MC = 5.86%	% = 89.36 $\kappa$ = 0.83 MC = 11.85%	— —	— —	— —	— —	— —	— —	— —
Rater 2	— —	— —	% = 84.95 $\kappa$ = 0.76 MC = 13.18%	% = 77.82 $\kappa$ = 0.61 MC = 17.91%	— —	— —	— —	— —	— —	— —	— —
Rater 3	— —	— —	— —	% = 82.13 $\kappa$ = 0.72 MC = 20.11%	— —	— —	— —	— —	— —	— —	— —

had multiple assertions made about the meta-analysis was assessed for whether the assertions fell under the same category. If this was the case, the assertion category was only counted once for this analysis. For example, if an article had three assertions made about the meta-analysis and the first was coded as *S3*, the second was coded as *US1*, and the third was coded as *US1*, this article would be counted once under *S3* and once under *US1*. Thus, articles could fall under more than one assertion category (under *S3* and *US1* in the above example). At both of these levels, we examined the frequency of the three *supported* assertion categories and the frequency of the four *unsupported* assertion categories in general and across journal disciplines and publication years.

We also conducted an additional analysis at the article level: Each of the 1,090 articles could have assertions made about the meta-analysis where (a) all assertions in the article fell under a *supported* assertion category, (b) all assertions in the article fell under an *unsupported* assertion category, (c) all assertions in the article fell under the *irrelevant* category, or (d) the assertions in the article fell under multiple assertion categories. For the purpose of this analysis, we focused on articles where one or more assertions fell under only the *supported* assertion categories,<sup>3</sup> only the *unsupported* assertion categories,<sup>3</sup> or under *both supported* and *unsupported* assertion categories.<sup>3</sup>

## Results

### *Supported and Unsupported Assertions: Assertion Level*

The frequency of *supported* and *unsupported* assertion categories across the 1,630 assertions is presented in Table 3. The most prevalent *supported* assertion category was *Active learning is more effective than lecture/increases student learning* (assertions with an *S1* code = 43.50%). The most prevalent *unsupported* assertion category was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (assertions with a *US1* code = 16.93%). The total percentage of assertions that were supported by the meta-analysis text was 47.67% ( $n = 777$ ). The total percentage of assertions that were not supported by the meta-analysis text was 26.01% ( $n = 424$ ).

### *Supported and Unsupported Assertions: Article Level*

The frequency of *supported* and *unsupported* assertion categories across the 1,090 articles is presented in Table 3. The most prevalent *supported* assertion category was *Active learning is more effective than lecture/increases student learning* (articles with one or more *S1* code(s) = 54.22%). The most prevalent *unsupported* assertion category was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (articles with one or more *US1* code(s) = 23.21%).

Out of the 1,090 articles, 531 (48.72%) had assertions that were categorized under only one or more of the *supported* assertion categories<sup>3</sup>; thus, these articles did not contain any *unsupported* assertions. The number of articles that had assertions categorized under only one or more of the *unsupported* assertion categories<sup>3</sup> was 289 (26.51%); thus, these articles did not contain any *supported* assertions.

**TABLE 3**  
*Frequency of assertion category codes across all assertions and articles*

Category	Analysis	
	Assertion level (out of 1,630 assertions)	Article level (out of 1,090 articles)
Supported		
<i>S1</i>	698 (42.82%)	591 (54.22%)
<i>S2</i>	40 (2.45%)	50 (4.59%)
<i>S3</i>	28 (1.72%)	24 (2.20%)
<i>S1, S2<sup>a</sup></i>	11 (0.67%)	—
Total supported	777 (47.67%)	—
Unsupported		
<i>US1</i>	237 (14.54%)	253 (23.21%)
<i>US2</i>	32 (1.96%)	33 (3.03%)
<i>US3</i>	35 (2.15%)	52 (4.77%)
<i>US4</i>	72 (4.42%)	105 (9.63%)
<i>US1, US2<sup>a</sup></i>	2 (0.12%)	—
<i>US1, US3<sup>a</sup></i>	7 (0.43%)	—
<i>US1, US4<sup>a</sup></i>	27 (1.66%)	—
<i>US2, US3<sup>a</sup></i>	2 (0.12%)	—
<i>US3, US4<sup>a</sup></i>	7 (0.43%)	—
<i>US1, US3, US4<sup>a</sup></i>	3 (0.18%)	—
Total unsupported	424 (26.01%)	—

*Note.* Irrelevant code = 429 assertions (26.32%).  
<sup>a</sup>An assertion could fall under more than one assertion category.

The number of articles that had assertions categorized under only the *irrelevant* assertion category was 180 (16.51%). Finally, the number of articles with at least one *supported* assertion and at least one *unsupported* assertion<sup>3</sup> was 90 (8.26%). Thus, 379 articles (34.77%) contained one or more assertions related to our efficacy focus that were not supported by the meta-analysis text.

*Supported and Unsupported Assertions: Journal Disciplines*

Using the same two levels of analysis discussed previously (i.e., assertion and article levels), the frequency of the seven assertion categories was compared based on the discipline of the journal in which each article was published. The disciplines included *Science, Technology, Engineering, Mathematics, STEM, non-STEM, General Education, Transdisciplinary, and Difficult to Determine* (i.e., journal title in language other than English).

*Assertion Level*

The frequency of *supported* and *unsupported* assertion categories across journal disciplines for the 1,630 assertions is presented in Table 4. The most prevalent



**TABLE 4**  
*Frequency of assertion category codes across all assertions and articles based on journal discipline*

Category	Assertion level							
	Science, <i>n</i> = 899	Technology, <i>n</i> = 134	Engineering, <i>n</i> = 131	Mathematics, <i>n</i> = 77	STEM, <i>n</i> = 74	Non-STEM, <i>n</i> = 123	General educa- tion, <i>n</i> = 185	Transdiscipli- nary, <i>n</i> = 3
Supported								
<i>S1</i>	382 (42.49%)	41 (30.60%)	66 (50.38%)	39 (50.65%)	32 (43.24%)	54 (43.90%)	81 (43.78%)	0 (0.00%)
<i>S2</i>	24 (2.67%)	1 (0.75%)	1 (0.76%)	3 (3.90%)	1 (1.35%)	4 (3.25%)	6 (3.24%)	0 (0.00%)
<i>S3</i>	18 (2.00%)	6 (4.48%)	1 (0.76%)	2 (2.60%)	0 (0.00%)	1 (0.81%)	0 (0.00%)	0 (0.00%)
<i>S1, S2<sup>a</sup></i>	5 (0.56%)	0 (0.00%)	1 (0.76%)	0 (0.00%)	2 (2.70%)	2 (1.63%)	1 (0.54%)	0 (0.00%)
Total supported	429 (47.72%)	48 (35.82%)	69 (52.67%)	44 (57.14%)	35 (47.30%)	61 (49.59%)	88 (47.57%)	0 (0.00%)
Unsupported								
<i>US1</i>	133 (14.79%)	21 (15.67%)	17 (12.98%)	11 (14.29%)	10 (13.51%)	17 (13.82%)	26 (14.05%)	1 (33.33%)
<i>US2</i>	15 (1.67%)	3 (2.24%)	7 (5.34%)	0 (0.00%)	1 (1.35%)	1 (0.81%)	4 (2.16%)	1 (33.33%)
<i>US3</i>	24 (2.67%)	1 (0.75%)	3 (2.29%)	2 (2.60%)	1 (1.35%)	1 (0.81%)	3 (1.62%)	0 (0.00%)
<i>US4</i>	33 (3.67%)	10 (7.46%)	10 (7.63%)	1 (1.30%)	4 (5.41%)	6 (4.88%)	8 (4.32%)	0 (0.00%)
<i>US1, US2<sup>a</sup></i>	2 (0.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
<i>US1, US3<sup>a</sup></i>	4 (0.44%)	0 (0.00%)	0 (0.00%)	1 (1.30%)	0 (0.00%)	1 (0.81%)	1 (0.54%)	0 (0.00%)
<i>US1, US4<sup>a</sup></i>	15 (1.67%)	4 (2.99%)	1 (0.76%)	1 (1.30%)	1 (1.35%)	3 (2.44%)	2 (1.08%)	0 (0.00%)
<i>US2, US3<sup>a</sup></i>	2 (0.22%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
<i>US3, US4<sup>a</sup></i>	4 (0.44%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (1.63%)	1 (0.54%)	0 (0.00%)
<i>US1, US3, US4<sup>a</sup></i>	2 (0.22%)	0 (0.00%)	0 (0.00%)	1 (1.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Total unsupported	234 (26.03%)	39 (29.10%)	38 (29.01%)	17 (22.08%)	17 (22.97%)	31 (25.20%)	45 (24.32%)	2 (66.67%)

(continued)

**TABLE 4 (continued)**

Category	Article level								
	Science, <i>n</i> = 600	Technology, <i>n</i> = 86	Engineering, <i>n</i> = 97	Mathematics, <i>n</i> = 52	STEM, <i>n</i> = 47	Non-STEM, <i>n</i> = 85	General education, <i>n</i> = 117	Transdisciplinary, <i>n</i> = 2	Difficult to determine, <i>n</i> = 4
Supported									
<i>SI</i>	325 (54.17%)	35 (40.70%)	56 (57.73%)	34 (65.38%)	28 (59.57%)	48 (56.47%)	62 (52.99%)	0 (0.00%)	3 (75.00%)
<i>S2</i>	28 (4.67%)	1 (1.16%)	2 (2.06%)	3 (5.77%)	3 (6.38%)	6 (7.06%)	7 (5.98%)	0 (0.00%)	0 (0.00%)
<i>S3</i>	16 (2.67%)	4 (4.65%)	1 (1.03%)	2 (3.85%)	0 (0.00%)	1 (1.18%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Unsupported									
<i>US1</i>	144 (24.0%)	22 (25.58%)	17 (17.53%)	12 (23.08%)	11 (23.40%)	19 (22.35%)	26 (22.22%)	1 (50.00%)	1 (25.00%)
<i>US2</i>	17 (2.83%)	2 (2.33%)	7 (7.22%)	0 (0.00%)	1 (2.13%)	1 (1.18%)	4 (3.42%)	1 (50.00%)	0 (0.00%)
<i>US3</i>	35 (5.83%)	1 (1.16%)	3 (3.09%)	4 (7.69%)	1 (2.13%)	3 (3.53%)	5 (4.27%)	0 (0.00%)	0 (0.00%)
<i>US4</i>	51 (8.50%)	14 (16.28%)	11 (11.34%)	3 (5.77%)	5 (10.64%)	10 (11.76%)	11 (9.40%)	0 (0.00%)	0 (0.00%)

*Note.* STEM = science, technology, engineering, and mathematics.

<sup>a</sup>An assertion could fall under more than one assertion category.

discipline from which assertions made about Freeman et al. (2014) came was Science (899 assertions) and the least prevalent was Transdisciplinary (three assertions). Transdisciplinary and Difficult to Determine had only three and four assertions, respectively; therefore, these two discipline categories were not discussed in the following results. The most prevalent *supported* assertion category across journal disciplines was *Active learning is more effective than lecture/increases student learning* (assertions with an *S1* code ranged from 30.60% [Technology] to 51.15% [Engineering]). The most prevalent *unsupported* assertion category across journal disciplines was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (assertions with a *US1* code ranged from 13.74% [Engineering] to 18.66% [Technology]). Mathematics had the highest percentage of *supported* assertions (57.14%), and Technology had the lowest percentage of *supported* assertions (35.82%); Technology had the highest percentage of *unsupported* assertions (29.10%), and Mathematics had the lowest percentage of *unsupported* assertions (22.08%).

#### Article Level

The frequency of *supported* and *unsupported* assertion categories across journal disciplines for the 1,090 articles is presented in Table 4. The most prevalent discipline from which articles came was Science (600 articles), and the least prevalent discipline was Transdisciplinary (two articles). Transdisciplinary and Difficult to Determine had only two and four articles, respectively; therefore, these two discipline categories were not discussed in the following results. The most prevalent *supported* assertion category across journal disciplines was *Active learning is more effective than lecture/increases student learning* (articles with at least one *S1* code ranged from 40.70% [Technology] to 65.38% [Mathematics]). The most prevalent *unsupported* assertion category was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (articles with at least one *US1* code ranged from 17.53% [Engineering] to 25.58% [Technology]). Similar to the assertion-level analysis, Mathematics had the highest percentage of articles with *supported* assertions, and Technology had the lowest percentage of articles with *supported* assertions; Technology had the highest percentage of articles with *unsupported* assertions, and Mathematics had the lowest percentage of articles with *unsupported* assertions.

#### *Supported and Unsupported Assertions: Publication Years*

Using the same two levels of analysis discussed previously, the frequency of the seven assertion categories was compared based on the year in which each article was published. The publication years included 2014, 2015, 2016, 2017, 2018, and half of 2019 (January 1 to June 13).

#### Assertion Level

The total number of assertions and the percentage of total *supported* and *unsupported* assertions are presented in Figure 1. Considering the data from the 4 complete years (i.e., 2015–2018) of analysis, the following can be concluded.

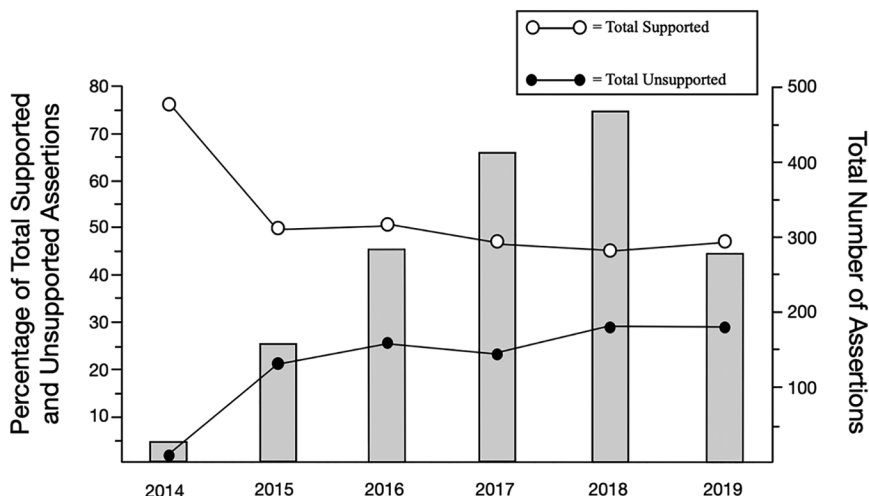


FIGURE 1. *Percentage of total supported and unsupported assertions (line graph, left scale) and total number of assertions (bar graph, right scale) across years.*

There was a general increasing trend for the number of total assertions made across years (note that 2019 was only examined from January 1 to June 13). There was no consistent, clear trend for either percentage of total *supported* and *unsupported* assertions made across the years.

The frequency of *supported* and *unsupported* assertion categories across publication years for the 1,630 assertions is presented in Table 5. The most prevalent year for assertions to be made about the meta-analysis was 2018 (468 assertions) although 2019 was only examined for part of the year and already had 276 assertions. The least prevalent year was 2014 (26 assertions); however, it is important to note that the meta-analysis was first published May 12, 2014. Although the number of assertions differed across publication years, the most prevalent *supported* assertion category across publication years was *Active learning is more effective than lecture/increases student learning* (assertions with an *S1* code ranged from 40.81% [2018] to 73.08% [2014]). Overall, the highest percentage of *supported* assertions was in 2014 (76.92%), while the lowest percentage of *supported* assertions was in 2018 (44.87%). If 2014 were excluded due to only having a total of 26 assertions, the highest number of *supported* assertions would be in 2016 (50.53%). The most prevalent *unsupported* assertion category across publication years was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (assertions with a *US1* code ranged from 0.00% [2014] to 20.51% [2018]). Overall, the highest percentage of *unsupported* assertions was in 2018 (29.27%) and the lowest percentage of *unsupported* assertions was in 2014 (3.85%). If 2014 were excluded due to only having a total of 26 assertions, the lowest percentage of *unsupported* assertions was in 2015 (22.64%).

**TABLE 5**  
*Frequency of assertion category codes across all assertions and articles based on year published*

Category	Assertion level					
	2014, <i>n</i> = 26	2015, <i>n</i> = 159	2016, <i>n</i> = 283	2017, <i>n</i> = 418	2018, <i>n</i> = 468	2019, <i>n</i> = 276
Supported						
<i>S1</i>	18 (69.23%)	69 (43.40%)	124 (43.82%)	178 (42.58%)	188 (40.17%)	121 (43.84%)
<i>S2</i>	0 (0.00%)	3 (1.89%)	9 (3.18%)	11 (2.63%)	14 (2.99%)	3 (1.09%)
<i>S3</i>	1 (3.85%)	6 (3.77%)	8 (2.83%)	5 (1.20%)	5 (1.07%)	3 (1.09%)
<i>S1, S2<sup>a</sup></i>	1 (3.85%)	1 (0.63%)	2 (0.71%)	2 (0.48%)	3 (0.64%)	2 (0.72%)
Total supported	20 (76.92%)	79 (49.69%)	143 (50.53%)	196 (46.89%)	210 (44.87%)	129 (46.74%)
Unsupported						
<i>US1</i>	0 (0.00%)	16 (10.06%)	43 (15.19%)	54 (12.92%)	81 (17.31%)	43 (15.58%)
<i>US2</i>	0 (0.00%)	3 (1.89%)	7 (2.47%)	4 (0.96%)	13 (2.78%)	5 (1.81%)
<i>US3</i>	0 (0.00%)	4 (2.52%)	7 (2.47%)	12 (2.87%)	7 (1.50%)	5 (1.81%)
<i>US4</i>	1 (3.85%)	7 (4.40%)	7 (2.47%)	19 (4.55%)	20 (4.27%)	18 (6.52%)
<i>US1, US2<sup>a</sup></i>	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.21%)	1 (0.36%)
<i>US1, US3<sup>a</sup></i>	0 (0.00%)	1 (0.63%)	0 (0.00%)	2 (0.48%)	4 (0.85%)	0 (0.00%)
<i>US1, US4<sup>a</sup></i>	0 (0.00%)	4 (2.52%)	3 (1.06%)	7 (1.67%)	9 (1.92%)	4 (1.45%)
<i>US2, US3<sup>a</sup></i>	0 (0.00%)	0 (0.00%)	1 (0.35%)	0 (0.00%)	1 (0.21%)	0 (0.00%)
<i>US3, US4<sup>a</sup></i>	0 (0.00%)	1 (0.63%)	3 (1.06%)	2 (0.48%)	0 (0.00%)	1 (0.36%)
<i>US1, US3, US4<sup>a</sup></i>	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.24%)	1 (0.21%)	1 (0.36%)
Total unsupported	1 (3.85%)	36 (22.64%)	71 (25.09%)	101 (24.16%)	137 (29.27%)	78 (28.26%)

(continued)

**TABLE 5 (continued)**

Category	Article level					
	2014, <i>n</i> = 21	2015, <i>n</i> = 102	2016, <i>n</i> = 199	2017, <i>n</i> = 259	2018, <i>n</i> = 327	2019, <i>n</i> = 182
Supported						
<i>S1</i>	16 (76.19%)	57 (55.88%)	113 (56.78%)	144 (55.60%)	169 (51.68%)	92 (50.55%)
<i>S2</i>	1 (4.76%)	4 (3.92%)	10 (5.03%)	13 (5.02%)	17 (5.20%)	5 (2.75%)
<i>S3</i>	1 (4.76%)	4 (3.92%)	5 (2.51%)	6 (2.32%)	5 (1.53%)	3 (1.65%)
Unsupported						
<i>US1</i>	0 (0.00%)	19 (18.63%)	41 (20.60%)	56 (21.62%)	92 (28.13%)	45 (24.73%)
<i>US2</i>	0 (0.00%)	3 (2.94%)	8 (4.02%)	4 (1.54%)	12 (3.67%)	6 (3.30%)
<i>US3</i>	0 (0.00%)	5 (4.90%)	10 (5.03%)	17 (6.56%)	13 (3.98%)	7 (3.85%)
<i>US4</i>	1 (4.76%)	11 (10.78%)	12 (6.03%)	29 (11.20%)	29 (8.87%)	23 (12.64%)

<sup>a</sup>An assertion could fall under more than one assertion category.



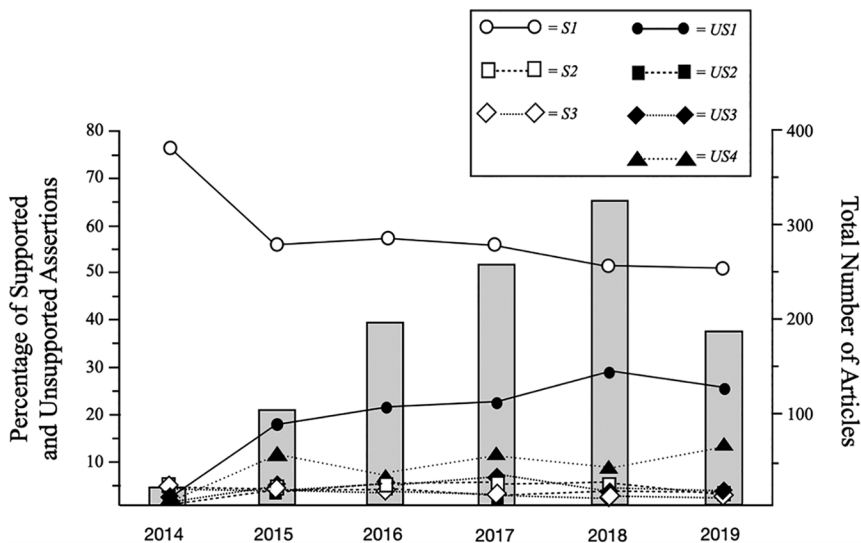


FIGURE 2. Percentage of supported (*S1*, *S2*, and *S3*) and unsupported (*US1*, *US2*, *US3*, and *US4*) assertions (line graph, left scale) and total number of articles (bar graph, right scale) with assertions across years.

### Article Level

As shown in Figure 2, the total number of articles containing assertions had an increasing trend with the exception of 2019 (for the reason explained previously). Considering the data from the 4 complete years (i.e., 2015–2018) of analysis, the following can be concluded. There was an increasing trend for *US1* (*Specific activities/approaches other than the general approach of active learning are effective*) and a slight decreasing trend for *S3* (*Second generation research should be conducted [move beyond using lecture as a control condition]*). However, there was no clear trend for the percentage of other *supported* and *unsupported* categories across the years.

The frequency of *supported* and *unsupported* assertion categories across publication years for the 1,090 articles is presented in Table 5. The most prevalent year for articles with assertions made about the meta-analysis was 2018 (327 articles) although 2019 was only examined from January 1 to June 13 and had 182 articles. The least prevalent year was 2014 (21 articles). Although the number of assertions differed across publication years, the most prevalent *supported* assertion category across publication years was *Active learning is more effective than lecture/increases student learning* (articles with at least one *S1* code ranged from 50.55% [2019] to 76.19% [2014]). The most prevalent *unsupported* assertion category across publication years was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (articles with at least one *US1* code ranged from 0.00% [2014] to 28.13% [2018]).

## Discussion

Citation accuracy has been and continues to be an issue in journal publishing (e.g., Mogull, 2017; Onwuegbuzie et al., 2011; Santini, 2018). A danger in inaccurately quoting scholarly works is the potential for misinformation to infiltrate scientific discourse. Errors can affect the integrity of scholarship and investigation (Wilson, 2015) and could potentially influence policy and practice. Thus, given how highly cited the Freeman et al. (2014) meta-analysis is in the learning sciences and how it has been used to support the adoption of active learning and the reduction or elimination of lecture in STEM courses, we investigated whether the assertions made in the citing text were or were not supported by the meta-analysis. Using citation context analysis, we examined citing text that related to the efficacy of active learning and/or lecture given the strong advocacy for the adoption and implementation of active learning in college courses. The findings from this analysis have provided insight into quotation accuracy when citing Freeman et al. as well as a more comprehensive picture of the scientific discourse surrounding active learning and instructional efficacy.

### *Supported Assertions Across All Assertions and Articles*

The most prevalent *supported* assertion category was *Active learning is more effective than lecture/increases student learning* (assertions with an *S1* category code = 43.50% [assertion level], articles with at least one *S1* category code = 54.22% [article level]). Assertions in the citing text that fell under this category generally contained information from Freeman et al. (2014) regarding how much better students did under active learning than under lecture (or subsequently, how much higher failure rates were under lecture vs. active learning). Given that the primary results from the meta-analysis were about the differential effects of active learning and lecture on student learning, it is understandable that the most common *supported* assertions in the citing text would be related to the main findings of the meta-analysis. Furthermore, the meta-analysis was the largest and most comprehensive meta-analysis on active learning in STEM education and thus the main findings were that much more important and interesting for those in the learning sciences and other education-related disciplines.

Though not as prevalent, articles did include assertions that fell under the *supported* assertion category of *Lecture is not the best approach to teaching/needs to be abolished or left behind in favor of active learning* (assertions with an *S2* category code = 3.13% [assertion level], articles with at least one *S2* category code = 4.59% [article level]). Given that lecture has been the subject of much criticism and there has been advocacy to adopt alternative approaches, the points made in Freeman et al. about questioning the continued use of lecture in light of the effects of active learning were used by a number of articles in the assertions they included about the meta-analysis. One possible reason this *supported* assertion category (*S2*) was less frequent than the *supported* assertion category that related to the effectiveness of active learning versus lecture (*S1*) could be that moving away from lecture and toward active learning is inherent in the main findings of the meta-analysis. One need not cite the specific points made in Freeman et al. about favoring active learning over lecture if the main findings can be cited that provide

evidence of active learning's impact on student learning as compared with the traditional lecture method.

Articles also included assertions that fell under the *supported* assertion category of *Second generation research should be conducted (move beyond using lecture as control condition)* (assertions with an *S3* category code = 1.72% [assertion level], articles with at least one *S3* category code = 2.20% [article level]). This category was not as common as the other two *supported* assertion categories as it has less to do with the main findings of the meta-analysis and more to do with future research directions. Thus, the recommendation in Freeman et al. about conducting second generation research may have been more relevant to researchers currently conducting or interested in conducting research in the active learning realm than to those looking to the meta-analysis for evidence of which instructional approach is more effective than the other.

### *Unsupported Assertions Across All Assertions and Articles*

Shifting the focus to *unsupported* assertion categories, the overall percentage of assertions that were not supported by the Freeman et al. (2014) meta-analysis was 26.01%. This percentage is consistent with the systematic literature review and meta-analysis of the quotation accuracy in medical journal articles (Jergas & Baethge, 2015). In our additional analysis at the article level, the overall percentage of articles that included at least one *unsupported* assertion was 34.77%. Thus, approximately one out of every three articles contained assertions that were not supported by Freeman et al. This article-level analysis is not conducted as often as the assertion-level analysis, and as such, it is more difficult to determine if this percentage is consistent with much of the quotation accuracy literature. However, based on the article-level error rate in Todd et al. (2010) and Teixeira et al. (2013)—26% and 41%, respectively—the rate of *unsupported* assertions we identified at the article level appears to be within the range of these other studies.

The most common *unsupported* assertion category was *Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective* (assertions with a *US1* code = 16.93% [assertion level], articles with at least one *US1* code = 23.21% [article level]). In these assertions, activities as diverse as Twitter, group discussions, clicker questions, modules, and inquiry exercises were deemed effective by citing Freeman et al. (2014) as support. Furthermore, approaches such as problem-based learning, flipped learning, inquiry-based learning, and cooperative learning were identified as approaches that improve student learning by citing Freeman et al. as evidence. However, Freeman et al. contained two groupings: traditional lecture and active learning. Pulling out a specific approach or activity that could fall under active learning and deem it effective due to the results found in Freeman et al. would not be supported by those results. Different forms of active learning were not compared with one another nor were specific forms of active learning compared with lecture. Thus, the aggregation of effect sizes across a variety of active learning studies creates a generalization but not a specification. In other words, we can generally say that active learning (in general) is more effective than lecture (in general), but we cannot specifically say that cooperative learning, for example, is more effective than lecture. Misconceptions can be

started if researchers are publishing papers that say, for example, “Freeman et al. (2014) showed that flipped courses led to improved student learning” or that “Twitter use in the hard sciences can impact learning, engagement, and relationships” while citing the Freeman et al. meta-analysis. Illustrative examples from articles included in our analyses are shown under *US1* in Table 1.

The second most common *unsupported* assertion category was *Active learning improves measures above and beyond learning/retention* (e.g., *motivation, engagement, attitude, interpersonal skills*) (assertions with a *US4* code = 6.69% [assertion level], articles with at least one *US4* code = 9.63% [article level]). For example, Freeman et al. (2014) was cited as support for assertions made about active learning improving motivation and attitudes to skills related to life-long learning to thinking like a scientist. In Freeman et al., the only measures examined related to student learning (i.e., assessments and failure rates). There was no investigation into or reporting of other measures. When assertions are being made about the types of student behaviors that can be improved through active learning and a meta-analysis is cited as evidence, people may adopt active learning with the intention of improving those named student behaviors. However, using the meta-analysis as evidence in support of those assertions is not accurate and can lead to an incorrect understanding of the meta-analysis and of what active learning may or may not be capable of doing for students.

Though not as prevalent as the other two *unsupported* assertion categories, articles did include assertions that fell under the *unsupported* assertion category of *Lecture is ineffective* (assertions with a *US2* code = 2.21% [assertion level], articles with at least one *US2* code = 3.03% [article level]). Although Freeman et al. (2014) provided evidence that active learning led to larger learning gains and reduced failure rates for students as compared with traditional lecture, these authors did not explicitly state that lecture is ineffective. Rather, lecture (in general) was shown to be *less* effective than active learning (in general).

Nevertheless, lecture is not necessarily ineffective in that it can be an efficient way to disseminate information (see research on explicit/direct instruction, for example (e.g., Klahr & Nigam, 2004; Martella et al., 2020). Lectures have the potential to provide context and structure for a subject, facilitate the development of an idea, promote listening and note-taking skills, give students access to new information in the field, and allow students to learn from content experts in the field, among other possible benefits (French & Kennedy, 2017). Lecture methods are multifaceted and varied (French & Kennedy, 2017) and, thus, may not all look the same in a lecture-based course or provide the same benefits to students. For example, lecture could be used for the duration of a class period or could be used periodically during a class period. Lectures could be conducted through PowerPoint presentations, whiteboards, or instructor notes. Lecturers could have received teaching awards or could have received low student evaluations. By saying that lecture is ineffective, all forms of lecture are being lumped together and criticized as one when certain forms of lecture could be more effective than certain forms of active learning. Future research on the effects of varying forms of lecture are needed to determine if and how they can be implemented effectively, especially in relation to active learning approaches.

The least common *unsupported* assertion category was *Active learning is beneficial for specific populations/course topics* (e.g., minorities, women, genetics material, calculus) (assertions with a US3 code = 3.31% [assertion level], articles with at least one US3 code = 4.77% [article level]). These assertions, similar to assertions about active learning improving more than just student learning, could resonate with those who wish to adopt strategies that could improve learning in certain populations of students or topic areas. The Freeman et al. (2014) meta-analysis did not contain specific data on any improvements for certain populations of students nor beyond STEM disciplines in general.

### *Explanations for Unsupported Assertions*

There are several possible explanations for making *unsupported* assertions. The first two explanations relate to all four *unsupported* assertion categories. First, given the communities of teaching-oriented faculty and staff in higher education or the groups of learning sciences professionals working on STEM education initiatives, for example, incorrect readings of Freeman et al. (2014) or biases influencing how information is perceived could become ubiquitous in these groups. With this misinformation in mind, authors may make assertions and cite Freeman et al. without reading or double checking the paper they are citing. Second and relatedly, authors may read others who have cited Freeman et al. (2014) and transfer that information into their own papers without being aware that those who had cited Freeman et al. had cited them incorrectly. Accepting and copying another author's interpretation of a cited article, rather than going back to the original source material, can cause the spread of incorrect information. This issue can be labeled as *citing secondary versus primary sources*.

The third explanation is specifically related to *unsupported* assertions about specific approaches being effective. There is a lack of consistent and clear operational definitions for active learning and lecture (see Martella & Demmig-Adams, 2018). Freeman et al. (2014) noted that

the active learning interventions varied widely in intensity and implementation, and included approaches as diverse as occasional group problem-solving, worksheets or tutorials completed during class, use of personal response systems with or without peer instruction, and studio or workshop course designs. (p. 8410)

However, deeper insight into which forms of active learning (or active learning features) are most effective was not provided in Freeman et al.—the different active learning interventions were grouped under the broad category of active learning and the effect sizes were statistically combined. Those reading Freeman et al. may have read the statement about how the active learning interventions varied (with various example features listed) and, coupled with the finding that active learning was more effective than lecture, made the assumption that all forms of active learning were more effective than lecture and thus specific active learning features could be deemed more effective than lecture.

The fourth explanation is specifically related to *unsupported* assertions about active learning improving measures above and beyond learning/retention. Active learning is generally thought of as a more engaging approach than the traditional

lecture method. In fact, in Freeman et al. (2014), the authors discuss analyses that indicated better student performance (passing the class, getting higher grades) and increased engagement help students persist in STEM and cite Goodman et al. (2002), Seymour and Hewitt (1997), and Watkins and Mazur (2013). Those who cited Freeman et al. (2014) may not have closely read that part of the meta-analysis and subsequently made the assumption that (a) Freeman et al. (2014) found active learning to increase student engagement or (b) because students had higher performance in active learning courses, they must have been more engaged and motivated.

The fifth and sixth explanations are specifically related to *unsupported* assertions about lecture being ineffective. The methodological decisions made in Freeman et al. (2014) created a false dichotomy between active learning and traditional lecture (it should be noted that when Freeman et al. tried to categorize and analyze different types of active learning, the sample sizes were highly unbalanced). There is danger in creating false dichotomies in that they appear as though they are mutually exclusive or that you must choose between them (see LaBoskey, 1998). Due to the false dichotomy of active learning versus traditional lecture created in the meta-analysis, it appears that active learning is devoid of any lecture. However, active learning can and often does contain lecture components (Martella et al., in press; Zakrajsek, 2018). Freeman et al. (2014) found active learning courses could contain lecture, stating that active learning activities could range from taking 10% to 100% of the class time—this percentage range led us to assume that lecture could range from 0% to 90% of class time in active learning courses. Therefore, based on this range, it is possible that lecture courses could contain active learning components for up to 9% of the class time. However, by creating these two instructional bins (i.e., active learning bin and traditional lecture bin), readers may have seen these two approaches as mutually exclusive and made the assumption that lecture was ineffective.

Additionally, the popular press may influence the takeaways from Freeman et al. The interpretations written in these articles such as “It Puts Kids to Sleep—but Teachers Keep Lecturing. Here’s What to Do About It,” from *The Washington Post* (Strauss, 2017) may influence others’ reading of the meta-analysis or may lead them to cite Freeman et al. as saying something that was an exaggeration found in a popular press article. In yet another example, an article in *Science* was titled, “Lectures Aren’t Just Boring, They’re Ineffective, Too, Study Finds” (Bajak, 2014). These popular press and other media-coverage articles frequently mentioned lecture as an ineffective or boring approach.

The seventh and last explanation is specifically related to *unsupported* assertions about active learning being beneficial for specific populations or course topics. In Freeman et al. (2014), the authors mention that other research indicates active learning can have disproportionate benefits for STEM students from disadvantaged backgrounds and female students; they cite Haak et al. (2011) and Lorenzo et al. (2006). Perhaps it was the case that authors who cited Freeman et al. as saying that active learning confers greater advantages for students from diverse backgrounds pulled this information from the discussion section in the meta-analysis without directly going back to the Haak et al. or Lorenzo et al. sources to ensure the information was correct and to cite them directly. Readers



could have also assumed Freeman et al. had made those findings if a careful read was not conducted.

### *Supported and Unsupported Assertions Across All Assertions and Articles Based on Journal Discipline*

In addition to examining the prevalence of the seven assertion categories across assertions and articles, we examined the general frequency of the seven categories across journal disciplines. The most common journal discipline citing Freeman et al. (2014) was *Science* followed by *General Education*, and the least common journal discipline was *STEM* (excluding *Transdisciplinary* and *Difficult to Determine* discipline categories). Given that active learning is generally discussed in discipline-based education research and the learning sciences, it is not surprising that discipline-specific journals and general education–related journals cite this meta-analysis more than journals that encompass an array of disciplines. It is also not surprising that *Science* was the most common discipline publishing studies that cited Freeman et al. as biology, chemistry, and physics have the status of “parent disciplines” for discipline-based education research and have a long history in education research (Singer et al., 2012).

Furthermore, the typical most common supported category across journals and publication years was “Active learning is more effective than lecture/increases student learning (labeled *SI*),” and the typical most common *unsupported* category across years and journals was “Specific activities/approaches (e.g., group discussions, flipped classes, inquiry-based learning) other than the general approach of active learning are effective (labeled *USI*).” The most common publication year for citing Freeman et al. (2014) in our analysis was 2018. It should be noted based on frequency counts in *WOSCC* and *Scopus*, each year since the meta-analysis was published has experienced an increase in the number of citations with 2019 (the full year) surpassing 2018 by almost 40%. This trend appears to be similar for 2020. It appears that discussions surrounding active learning continues to garner interest, which is not unexpected given current national conversations and initiatives surrounding student achievement and retention in STEM disciplines.

### *Future Directions and Implications*

The Freeman et al. (2014) meta-analysis is a prominent paper that has helped fuel the discussion of STEM educational improvements and course transformations at the college level. As Willingham (2014) stated, “Perhaps the best news is that the effectiveness of college instruction is on people’s minds” (para. 9). However, ensuring the dissemination of correct information is of the utmost importance given the policy decisions, funding opportunities, institutional changes, and large-scale initiatives that involve active learning. Citation practices should not be taken for granted (Leatham, 2015), and there are many steps that can be taken to promote quotation accuracy (see recommendations by Bareket et al., 2020; Jergas & Baethge, 2015, for details). Five overarching suggestions can be made to ensure quotation accuracy.

First, mistakes are often made in citing the works of others due to a failure to closely read articles and/or reading and citing secondary sources. Authors should

always go back to primary sources to further reduce the spread of overgeneralizations and inaccurate information rather than trusting secondary sources or popular press articles. They should carefully read these primary articles to make sure all pertinent information is obtained.

Second, because there are a lack of consistent and clear operational definitions for active learning and lecture, it is not possible to state with any certainty what specific components constitute active learning. As is, these are general categories rather than specific procedures. Future research should focus on clear and precise operational definitions of instructional variables. Components of active learning including the inclusion of various forms and durations of lectures and the order they are delivered should be specified. In other words, more than two bins should be developed for categorizing variables. Those citing the works of others should also consider the possible range of components that may be present in a general category such as *active learning*. When citing the works of others, a consideration of these other components should be made and indicated.

Third, it is critical to restrict one's own conclusions or those noted by others to the dependent variables measured and to eliminate any inferences related to these conclusions. For example, Freeman et al. (2014) found that active learning improved learning and retention of students. However, they did not measure an increase in student motivation, for example. Conclusions of some who cited the meta-analysis inferred there were improvements in this area when there were no improvement.

Fourth, researchers should eliminate false dichotomies when investigating complex areas of study. For example, due to the overlap of active learning and lecture (i.e., lecture can comprise 0% to 90% of class time in active learning classrooms), a false and overly simplistic dichotomy was created where active learning was interpreted as inconsistent with lecture and was considered mutually exclusive. Also, such a false dichotomy permits the interpretation that lecture is ineffective or damaging. Researchers should clearly specify the range of possible components in an instructional approach such as active learning in a manner that makes it more salient (e.g., table or other visual display). Also, authors citing the works of others should pay particular attention to descriptions of instructional variables as cited in the literature.

Finally, authors who cite the works of others should consider the specific populations and settings where the studies took place compared with those statements made by the researchers. For example, Freeman et al. (2014) did not analyze the effects of active learning on students from disadvantaged backgrounds and females. However, because Freeman et al. claimed, with noted study citations, that active learning affects these students, others indicated this was a finding from Freeman et al. when it was not.

Overall, a crucial step in disseminating accurate information is prioritizing quotation accuracy when writing and publishing papers. Double (and triple) checking each assertion and subsequent research paper being used as evidence can help reduce the creation of incorrect assertions; journals might even require a signed statement from authors that quotation accuracy was ensured on manuscript submission. For researchers, it is important to use

operational definitions, present important information in a salient manner, and refrain from making claims that go beyond the purpose of the study and the data obtained.


### Conclusion

Given the propensity for humans to make errors, even those who work in the scientific arena, it is our hope that the results of this study illuminate those errors that can easily be made in the scientific communication process so that producers and consumers of scientific writing can avoid or notice (and fix) these errors in the future. Perhaps quotation accuracy has been more frequently studied in medicine due to the serious consequences that errors can have such as affecting a person's health and well-being. However, errors in the learning sciences literature can also have serious consequences in that they can negatively affect how students are educated (relating to, e.g., pedagogy and educational resources or opportunities) or can hinder their success in present or future environments (relating to, e.g., STEM fatigue and attrition). Unfortunately, studies that focus on citation errors are not common in the learning sciences. Our citation context analysis served to fill this gap in the literature by providing insight into how accurately a highly influential article relating to the learning sciences has been cited. As evidenced by our citation context analysis, quotation errors were common in studies citing Freeman et al. (2014), and these errors could affect the scholarly discourse surrounding active learning and lecture. It is important to note that we did not examine the accuracy of the *irrelevant* assertions; and thus, there could be additional errors in the citing text that could also affect the scholarly discourse surrounding Freeman et al.

Future research examining quotation accuracy could be conducted across learning sciences, education-related journals, or on other highly cited articles. Furthermore, researchers could examine if and to what extent quotation errors have been propagated. For example, for those unsupported assertions we identified, did other authors read these assertions and use them in their own papers? Researchers could also investigate the frequency of the lack of citation where one is actually needed (Jergas & Baethge, 2015). Through these research investigations, issues in the literature can be revealed and myths may not continue to be promulgated. "Robust science needs robust corrections" (Allison et al., 2016, p. 29).

### ORCID iDs

Amedee Marchand Martella  <https://orcid.org/0000-0002-9207-4300>

Jane Kinkus Yacilla  <https://orcid.org/0000-0003-0660-4983>

### Notes

The first author acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1842166. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

<sup>1</sup> Although *citations* and *references* are sometimes used interchangeably (e.g., Masic, 2013), whereas others make a clear distinction (e.g., Bryson, 2012), we use *citations* to refer to scholarly works mentioned in the main text and *references* to refer to scholarly works in a reference list at the end of the paper (this distinction is consistent with the American Psychological Association, 2020).

<sup>2</sup> The research literature on quotation accuracy tends to be in the health sciences. We were not able to find a similar research base in psychology or education.

<sup>3</sup> If an article had assertions that fell under a *supported* assertion category and the *irrelevant* category, an *unsupported* assertion category and the *irrelevant* category, or a *supported*, *unsupported*, and *irrelevant* category, the article was still categorized as *supported only*, *unsupported only*, or *both supported and unsupported*, respectively (in other words, the *irrelevant* category did not affect categorization).

## References

*References marked with an asterisk indicate studies included in the analysis.*

- Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. *Nature*, 530(7588), 27–29. <https://doi.org/10.1038/530027a>
- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). <https://doi.org/10.1037/0000165-000>
- Association of American Universities. (2017). *Progress toward achieving systemic change: A five-year status report on the AAU undergraduate STEM education initiative*. <https://www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/STEM-Status-Report.pdf>
- Awrey, J., Inaba, K., Barmparas, G., Recinos, G., Teixeira, P. G. R., Chan, L. S., Talving, P., & Demetriades, D. (2011). Reference accuracy in the general surgery literature. *World Journal of Surgery*, 35(3), 475–479. <https://doi.org/10.1007/s00268-010-0912-7>
- Bajak, A. (2014, May 12). Lectures aren't just boring, they're ineffective, too, study finds. *Science*. <https://www.sciencemag.org/news/2014/05/lectures-arent-just-boring-theyre-ineffective-too-study-finds>
- Bareket, R., Schonberg, M. A., & Schonmann, Y. (2020). Make quotations great again: A proposal to reduce false knowledge. *BMJ Evidence-Based Medicine*, 25(1), 12–14. <https://doi.org/10.1136/bmjebm-2018-111109>
- \*Bayer, A., & Rouse, C. E. (2016). Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives*, 30(4), 221–242. <https://doi.org/10.1257/jep.30.4.221>
- \*Becker, B. K., Schiller, A. M., Zucker, I. H., Eager, E. A., Bronner, L. P., & Godfrey, M. (2017). A day of immersive physiology experiments increases knowledge and excitement towards physiology and scientific careers in Native American students. *Advances in Physiology Education*, 41(1), 137–144. <https://doi.org/10.1152/advan.00165.2016>
- Bernstein, D. A. (2018). Does active learning work? A good question, but not the right one. *Scholarship of Teaching and Learning in Psychology*, 4(4), 290–307. <https://doi.org/10.1037/stl0000124>
- Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom* [Monograph]. <http://files.eric.ed.gov/fulltext/ED336049.pdf>
- Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote.

- Journal of the Medical Library Association*, 104(3), 240–243. <https://doi.org/10.3163/1536-5050.104.3.014>
- Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, 115(11), 2563–2570. [www.pnas.org/cgi/doi/10.1073/pnas.1708279115](http://www.pnas.org/cgi/doi/10.1073/pnas.1708279115)
- Bryson, D. (2012). Using research papers: Citations, referencing and plagiarism. *Journal of Visual Communication in Medicine*, 35(2), 82–84. <https://doi.org/10.3109/17453054.2012.690128>
- \*Bubou, G. M., Offor, I. T., & Bappa, A. S. (2017). Why research-informed teaching in engineering education? A review of the evidence. *European Journal of Engineering Education*, 42(3), 323–335. <https://doi.org/10.1080/03043797.2016.1158793>
- Chasteen, S. (2016). *TRESTLE mini seed grant proposals: Transforming education, supporting teaching and learning excellence*. Center for STEM Learning, University of Colorado Boulder. [https://www.colorado.edu/csl/sites/default/files/attached-files/trestle\\_rfp\\_minigrant\\_0.pdf](https://www.colorado.edu/csl/sites/default/files/attached-files/trestle_rfp_minigrant_0.pdf)
- \*Clyne, A. M., & Billiar, K. L. (2016). Problem-based learning in biomechanics: Advantages, challenges, and implementation strategies. *Journal of Biomechanical Engineering*, 138(7), Article 070804. <https://doi.org/10.1115/1.4033671>
- Cooper, M. M., & Stowe, R. L. (2018). Chemistry education research: From personal empiricism to evidence, theory, and informed practice. *Chemical Reviews*, 118(12), 6053–6087. <https://doi.org/10.1021/acs.chemrev.8b00020>
- Dauids, J. R., Weigl, D. M., Edmonds, J. P., & Blackhurst, D. W. (2010). Reference accuracy in peer-reviewed pediatric orthopaedic literature. *Journal of Bone & Joint Surgery*, 92(5), 1155–1161. <https://doi.org/10.2106/JBJS.I.00063>
- \*Davis, D. A., & McMahon, G. T. (2018). Translating evidence into practice: Lessons for CPD. *Medical Teacher*, 40(9), 892–895. <https://doi.org/10.1080/0142159X.2018.1481285>
- de Lacey, G., Record, C., & Wade, J. (1985). How accurate are quotations and references in medical journals? *British Medical Journal (Clinical Research Ed.)*, 291(6499), 884–886. <https://doi.org/10.1136/bmj.291.6499.884>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences of the United States of America*, 116(39), 19251–19257. [www.pnas.org/cgi/doi/10.1073/pnas.1821936116](http://www.pnas.org/cgi/doi/10.1073/pnas.1821936116)
- Drake, D. C., Maritz, B., Jacobs, S. M., Crous, C. J., Engelbrecht, A., Etale, A., Fourie, M. J., Furniss, D. G., Scott, S. L., Parusnath, S., & Tye, D. R. (2013). The propagation and dispersal of misinformation in ecology: Is there a relationship between citation accuracy and journal impact factor? *Hydrobiologia*, 702(1), 1–4. <https://doi.org/10.1177/0309133313482135>
- Fenton, J. E., Brazier, H., de Souza, A., Hughes, J. P., & McShane, D. P. (2000). The accuracy of citation and quotation in otolaryngology/head and neck surgery journals. *Clinical Otolaryngology*, 25(1), 40–44. <https://doi.org/10.1046/j.1365-2273.2000.00322.x>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of*

- the United States of America*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>
- French, S., & Kennedy, G. (2017). Reassessing the value of university lectures. *Teaching in Higher Education*, 22(6), 639–654. <https://doi.org/10.1080/13562517.2016.1273213>
- Gasparyan, A. Y., Yessirkepov, M., Voronov, A. A., Gerasimov, A. N., Kostyukova, E. I., & Kitas, G. D. (2015). Preserving the integrity of citations and references by all stakeholders of science communication. *Journal of Korean Medical Science*, 30(11), 1545–1552. <https://doi.org/10.3346/jkms.2015.30.11.1545>
- Goodman, I. F., Cunningham, C. M., Lachapelle, C., Thompson, M., Bittinger, K., Brennan, R. T., & Delci, M. (2002). *Final report of the Women's Experiences in College Engineering (WECE) project*. National Science Foundation and Alfred P. Sloan Foundation. [http://www.grginc.com/documents/WECE\\_FINAL\\_REPORT.pdf](http://www.grginc.com/documents/WECE_FINAL_REPORT.pdf)
- Gosling, C. M., Cameron, M., & Gibbons, P. F. (2004). Referencing and quotation accuracy in four manual therapy journals. *Manual Therapy*, 9(1), 36–40. [https://doi.org/10.1016/S1356-689X\(03\)00056-0](https://doi.org/10.1016/S1356-689X(03)00056-0)
- Gough, D., Oliver, S., & Thomas, J. (2013). *Learning from research: Systematic reviews for informing policy decisions: A quick guide*. Alliance for Useful Evidence. <https://www.alliance4usefulevidence.org/assets/Alliance-FUE-reviews-booklet-3.pdf>
- \*Grissom, S., McCauley, R., & Murphy, L. (2017). How student centered is the computer science classroom? A survey of college faculty. *ACM Transactions on Computing Education*, 18(1), 1–27. <https://doi.org/10.1145/3143200>
- Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, 332(6034), 1213–1216. <https://doi.org/10.1126/science.1204820>
- Harzing, A. W. (1995). The persistent myth of high expatriate failure rates. *International Journal of Human Resource Management*, 6(2), 457–474. <https://doi.org/10.1080/09585199500000028>
- Harzing, A. W. (2002). Are our referencing errors undermining our scholarship and credibility? The case of expatriate failure rates. *Journal of Organizational Behavior*, 23(1), 127–148. <https://doi.org/10.1002/job.125>
- Hausmann, N. S., McIntyre, T., Bumby, A. J., & Loubser, M. J. (2013). Referencing practices in physical geography: How well do we cite what we write? *Progress in Physical Geography*, 37(4), 543–549. <https://doi.org/10.1177/0309133313482135>
- Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A meta-analysis. *BMC Medical Education*, 18(1), 1–12. <https://doi.org/10.1186/s12909-018-1144-z>
- Hinchcliff, K. W., Bruce, N. J., Powers, J. D., & Kipp, M. L. (1993). Accuracy of references and quotations in veterinary journals. *Journal of the American Veterinary Medical Association*, 202(3), 397–400.
- Jergas, H., & Baethge, C. (2015). Quotation accuracy in medical journal articles: A systematic review and meta-analysis. *PeerJ: Life & Environment*, 3, 1–20. <https://doi.org/10.7717/peerj.1364>
- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661–667. <https://doi.org/10.1111/j.0956-7976.2004.00737.x>



- Korpela, K. M. (2010). How long does it take for the scientific literature to purge itself of fraudulent material? The Breuning case revisited. *Current Medical Research and Opinion*, 26(4), 843–847. <https://doi.org/10.1185/030077991003603804>
- LaBoskey, V. K. (1998). The danger of false dichotomies. *Teacher Education Quarterly*, 25(4), 39–43. <https://www.jstor.org/stable/23478155>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Leatham, K. R. (2015). Observations on citation practices in mathematics education research. *Journal for Research in Mathematics Education*, 46(3), 253–269. <https://doi.org/10.5951/jresmetheduc.46.3.0253>
- Lederman, D. (2014, May 13). A boost for active learning. *Inside Higher Education*. <https://www.insidehighered.com/news/2014/05/13/stem-students-fare-better-when-professors-dont-just-lecture-study-finds>
- Lopresti, R. (2010). Citation accuracy in environmental science journals. *Scientometrics*, 85(3), 647–655. <https://doi.org/10.1007/s11192-010-0293-6>
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118–122. <https://doi.org/10.1119/1.2162549>
- Lukić, I. K., Lukić, A., Glunčić, V., Katavić, V., Vučenik, V., & Marušić, A. (2004). Citation and quotation accuracy in three anatomy journals. *Clinical Anatomy*, 17(7), 534–539. <https://doi.org/10.1002/ca.10255>
- Luo, M., Li, C. L., Molina, D., Andersen, C. R., & Panchbhavi, V. K. (2013). Accuracy of citation and quotation in foot and ankle surgery journals. *Foot & Ankle International*, 34(7), 949–955. <https://doi.org/10.1177/1071100713475354>
- Martella, A. M., & Demmig-Adams, B. (2018). Combining effective instructional approaches in a large introductory biology classroom: A research review and illustrative case study. *Journal on Excellence in College Teaching*, 29(2), 121–146.
- Martella, A. M., Klahr, D., & Li, W. (2020). The relative effectiveness of different active learning implementations in teaching elementary school students how to design simple experiments. *Journal of Educational Psychology*, 112(8), 1582–1596. <https://doi.org/10.1037/edu0000449>
- Martella, A. M., Lovett, M., & Ramsay, L. (in press). Implementing active learning: A critical examination of sources of variation in active learning science courses. *Journal on Excellence in College Teaching*.
- Masic, I. (2013). The importance of proper citation of references in biomedical articles. *ACTA Informatica Medica*, 21(3), 148–155. <https://doi.org/10.5455/aim.2013.21.148-155>
- \*McCullum, B. M., Fleming, C. L., Plotnikoff, K. M., & Skagen, D. N. (2017). Relationships in the flipped classroom. *Canadian Journal for the Scholarship of Teaching and Learning*, 8(3), 1–19. <https://doi.org/10.5206/cjsotl-rcacea.2017.3.8>
- Mogull, S. A. (2017). Accuracy of cited “facts” in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLoS One*, 12(9), Article e0184727. <https://doi.org/10.1371/journal.pone.0184727>
- Murphy, P. K., Knight, S. L., & Dowd, A. C. (2017). Familiar paths and new directions: Inaugural call for manuscripts. *Review of Educational Research*, 87(1), 3–6. <https://doi.org/10.3102/0034654317691764>
- National Science Foundation. (2014, May 12). *Enough with the lecturing*. [https://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=131403](https://www.nsf.gov/news/news_summ.jsp?cntn_id=131403)

- Neihouse, P. F., & Priske, S. C. (1989). Quotation accuracy in review articles. *DICP: The Annals of Pharmacotherapy*, 23(7–8), 594–596. <https://doi.org/10.1177/1060028089023007-813>
- Onwuegbuzie, A. J., Combs, J. P., Frels, R. K., & Slate, J. R. (2011). Editorial: Citation errors revisited: The case for *Educational Researcher: Research in the Schools*, 18(1), i–xxxv.
- Parse, R. R. (1996). Hear ye, hear ye, novice and seasoned authors! *Nursing Science Quarterly*, 9(1), 1. <https://doi.org/10.1177/089431849600900101>
- \*Peters, A. M. F., Costello, J., & Crane, D. (2018). Deviating from the traditional instructional tools: Integrating Twitter in a sociology of deviance course. *Canadian Journal of Learning and Technology*, 44(3), 1–19. <https://doi.org/10.21432/cjlt27792>
- Pickles, M. (2016, November 23). Shouldn't lectures be obsolete by now? *BBC News*. <http://www.bbc.com/news/business-38058477>
- \*Pienta, N. J. (2015). Innocents abroad: Part II. A glimpse at chemical education in India. *Journal of Chemical Education*, 92(3), 399–400. <https://doi.org/10.1021/acs.jchemed.5b00101>
- Santini, A. (2018). The importance of referencing. *Journal of Critical Care Medicine*, 4(1), 3–4. <https://doi.org/10.2478/jccm-2018-0002>
- Sanz-Martín, M., Pitt, K. A., Condon, R. H., Lucas, C. H., Novaes de Santana, C., & Duarte, C. M. (2016). Flawed citation practices facilitate the unsubstantiated perception of a global trend toward increased jellyfish blooms. *Global Ecology and Biogeography*, 25(9), 1039–1049. <https://doi.org/10.1111/geb.12474>
- Schmitt-Harsh, M., & Harsh, J. A. (2017). Engaging nonscience majors in urban ecology: Recommendations for course design. *Journal of Environmental Studies and Sciences*, 7(4), 550–561. <https://doi.org/10.1007/s13412-016-0410-9>
- \*Schönwälder, J., Friedman, T., & Pras, A. (2017). Using networks to teach about networks. *ACM SIGCOMM Computer Communication Review*, 47(3), 1–5. <https://ccronline.sigcomm.org/wp-content/uploads/2017/08/sigcomm-ccr-final119.pdf>
- Schulmeister, L. (1998). Quotation and reference accuracy of three nursing journals. *Journal of Nursing Scholarship*, 30(2), 143–146. <https://doi.org/10.1111/j.1547-5069.1998.tb01269.x>
- Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Westview Press. <http://www-personal.umich.edu/~glarose/dept/teaching/resources/talkingAboutLeavingSeymourHewittOverview.pdf>
- Singer, S. R., Nielsen, N. R., & Schweingruber, H. A. (Eds.). (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press. <http://eduinfo.cchem.berkeley.edu/pdf/NAPunderstandingandimprovinglearning.pdf>
- Small, H. (1982). Citation context analysis. In B. Dervin & M. J. Voigt (Eds.), *Progress in communication sciences* (Vol. 3, pp. 287–310). Ablex.
- Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., . . . Young, A. M. (2018). Anatomy of STEM teaching in North American universities: Lecture is prominent, but practices vary. *Science*, 359(6383), 1468–1470. <http://chemistry.as.virginia.edu/sites/chemistry.as.virginia.edu/files/2018-Stains%20et%20al-Science-%20COPUS%20profiles.pdf>

- Strauss, V. (2017, July 12). It puts kids to sleep—but teachers keep lecturing anyway: Here's what to do about it. *The Washington Post*. <https://www.washingtonpost.com/news/answer-sheet/wp/2017/07/11/it-puts-kids-to-sleep-but-teachers-keep-lecturing-anyway-heres-what-to-do-about-it/>
- Teixeira, M. C., Thomaz, S. M., Michelan, T. S., Mormul, R. P., Meurer, T., Fasolli, J. V. B., & Silveira, M. J. (2013). Incorrect citations give unfair credit to review authors in ecology journals. *PLoS One*, 8(12), Article e81871. <https://doi.org/10.1371/journal.pone.0081871>
- Todd, P. A., Guest, J. R., Lu, J., & Chou, L. M. (2010). One in four citations in marine biology papers is inappropriate. *Marine Ecology Progress Series*, 408(June), 299–303. <https://www.jstor.org/stable/10.2307/24873970>
- Todd, P. A., Yeo, D. C. J., Li, D., & Ladle, R. J. (2007). Citing practices in ecology: Can we believe our own words? *Oikos*, 116(9), 1599–1601. <https://doi.org/10.1111/j.2007.0030-1299.15992.x>
- Watkins, J., & Mazur, E. (2013). Retaining students in science, technology, engineering, and mathematics (STEM) majors. *Journal of College Science Teaching*, 42(5), 36–41. <https://www.jstor.org/stable/43631580>
- \*Watson, R. M., Willford, J. D., & Pfeifer, M. A. (2018). A cultured learning environment: Implementing a problem- and service-based microbiology capstone course to assess process- and skill-based learning objectives. *Interdisciplinary Journal of Problem-Based Learning*, 12(1), 1–22. <https://doi.org/10.7771/1541-5015.1694>
- \*Wenzel, T. J. (2014). Active learning materials for molecular and atomic spectroscopy. *Analytical and Bioanalytical Chemistry*, 406(22), 5245–5248. <https://doi.org/10.1007/s00216-014-7991-5>
- Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23), 8319–8320. <https://doi.org/10.1073/pnas.1407304111>
- \*Wilcox, R. L., Adem, P. V., Afshinnekoo, E., Atkinson, J. B., Burke, L. W., Cheung, H., Dasgupta, S., DeLaGarza, J., Joseph, L., LeGallo, R., Lew, M., Lockwood, C. M., Meiss, A., Norman, J., Markwood, P., Rizvi, H., Shane-Carson, K. P., Sobel, M. E., Suarez, E., . . . Haspel, R. L. (2018). The undergraduate training in genomics (UTRIG) initiative: Early and active training for physicians in the genomic medicine era. *Personalized Medicine*, 15(3), 199–208. <https://doi.org/10.2217/pme-2017-0077>
- Willingham, D. (2014, May 20). “Active learning” in college STEM courses: Meta-analysis. *RealClearEducation.com*. <http://www.danielwillingham.com/daniel-willingham-science-and-education-blog/active-learning-in-college-stem-courses-meta-analysis>
- Wilson, M. L. (2015). The persistence of misinformation. *American Journal of Clinical Pathology*, 144(3), 359–360. <https://doi.org/10.1309/AJCPMBV10SW4QHGU>
- \*Ylostalo, J. H. (2019). Engaging students into their own learning of foundational genetics concepts through the 5E learning cycle and interleaving teaching techniques. *Journal of Biological Education*, 54(5), 514–520. <https://doi.org/10.1080/00219266.2019.1620311>
- Zakrajsek, T. (2018). Reframing the lecture versus active learning debate: Suggestions for a new way forward. *Education in the Health Professions*, 1(1), 1–3. [https://doi.org/10.4103/EHP.EHP\\_14\\_18](https://doi.org/10.4103/EHP.EHP_14_18)
- Ziman, J. M. (1969). Information, communication, knowledge. *Nature*, 224, 318–324. <https://doi.org/10.1038/224318a0>

## Authors

AMEDEE MARCHAND MARTELLA is a National Science Foundation graduate research fellow and doctoral student in cognitive psychology at Purdue University, West Lafayette, IN 47907-2098, USA; email: *martella@purdue.edu*. She received her bachelor's degree from the University of Colorado Boulder with a double major in ecology and evolutionary biology and psychology and received her master's degree in psychology from Carnegie Mellon University. Her research interests center on active learning and improving the teaching of STEM at the college level.

JANE KINKUS YATCILLA is an associate professor in the Purdue Libraries & School of Information Studies, West Lafayette, IN 47907-2098, USA; email: *janeyat@purdue.edu* and a health and life sciences information specialist. Her research focuses on the analysis of research publication metadata (bibliometrics) and content analyses to discover characteristics of research fields.

RONALD C. MARTELLA is a professor of educational studies at Purdue University, West Lafayette, IN 47907-2098, USA; email: *rcmartella@purdue.edu*. He has more than 30 years of research experience in the areas of effective instructional and behavior management procedures for struggling students in K–12 settings. He is author of several peer-reviewed research articles and curricular reading programs. His more current area of interest is in multitiered system of supports.

NANCY E. MARCHAND-MARTELLA is the Suzi and Dale Gallagher Dean of Education and a professor in the Department of Educational Studies at Purdue University, West Lafayette, IN 47907-2098, USA; email: *nmarchand-martella@purdue.edu*. She has more than 30 years of research experience in the areas of effective instructional and behavior management procedures for struggling students in K–12 settings. Her primary area of interest is in reading remediation and effective instructional procedures within a multitiered system of supports model.

ZAFER OZEN is a doctoral student in the Gifted Education Research and Resource Institute at Purdue University, West Lafayette, IN 47907-2098, USA; email: *zozen@purdue.edu*. He received his bachelor's degree in physics education at Dokuz Eylul University in Turkey and his master's degree in gifted education at Purdue University in the USA. His research interests are STEM integration and gifted education.

TUGCE KARATAS is a doctoral student in Gifted, Creative, and Talented Studies at Purdue University, West Lafayette, IN 47907-2098, USA; email: *tkaratas@purdue.edu*. She received her Bachelor's degree in physics education from Dokuz Eylul University in Turkey, her first Master's degree in curriculum and instruction from Cukurova University in Turkey, and second Master's degree from Gifted, Creative, and Talented Studies at Purdue University. Her research interests are STEM education and curriculum development in gifted education.

HELEN H. PARK is a 2018–2019 Undergraduate Research Training (URT) scholar, 2019–2020 Office of Undergraduate Research (OUR) scholar, and an undergraduate student majoring in both elementary and special education (mild P-12) with a concentration in reading as well as a minor in global studies and learning science at Purdue University, West Lafayette, IN 47907-2098, USA; email: *park907@purdue.edu*. She is

*Martella et al.*

also a 2017–2021 recipient of the Next Generation Hoosier Educators Scholarship from the State of Indiana Commission for Higher Education as well as a recipient of the Indiana Retired Teacher Education Scholarship for the 2020-2021 school year. She will receive her bachelor's degree from Purdue University in May 2021.

ALEXANDRA SIMPSON graduated as a psychological sciences major at Purdue University, West Lafayette, IN 47907-2098, USA; email: *simpso86@purdue.edu*, and served as a research assistant in the Cognition and Learning Lab run by Dr. Jeffrey Karpicke at Purdue. She plans to pursue a PhD Fall 2021.

JEFFREY D. KARPICKE is the James V. Bradley Professor and head of the Department of Psychological Sciences at Purdue University, West Lafayette, IN 47907-2098, USA; email: *karpicke@purdue.edu*. He is interested in the science of learning, education, and technology with particular interests in retrieval processes and retrieval-based learning; complex learning, comprehension, and knowledge application; and metacognition and self-regulated learning.