Educational Psychology Review (2023) 35:107
https://doi.org/10.1007/s10648-023-09826-1

REVIEW ARTICLE

How Rigorous is Active Learning Research in STEM Education? An Examination of Key Internal Validity Controls in Intervention Studies

Amedee Marchand Martella[1] · Ronald C. Martella[2] · Jane K. Yatcilla[3] · Alexandra Newson[4] · Eric N. Shannon[5] · Charissa Voorhis[5,6]

Accepted: 14 October 2023 / Published online: 4 November 2023

Abstract

Active learning is a popular approach to teaching and learning that has gained traction through research on STEM educational improvement. There have been numerous university- and national/international-level efforts focused on transitioning courses from the lecture method to active learning. However, despite these large-scale changes, the active learning literature has not been assessed on its methodological rigor to ensure instructional recommendations are rooted in rigorous research studies. The purpose of the present review was to determine areas of strengths and areas in need of improvement and to provide specific recommendations on how to continue or improve active learning research to strengthen the respective literature base and increase confidence in results. We assessed the articles included in the Freeman et al. (PNAS, 111:8410–8415, 2014) meta-analysis as well as a random sample of more recent active learning articles (2015–2022) on 12 internal validity controls (i.e., control procedure used to prevent a threat to the internal validity of a study). Results indicated that there were high percentages of articles that did not meet each internal validity control. In fact, no articles from the Freeman et al. meta-analysis and no sampled 2015–2022 articles met each of the 12 internal validity controls. Therefore, the active learning literature contains numerous internal validity control issues that need to be addressed if we are to determine the extent to which active learning interventions are effective and if there are any boundary conditions for when particular active learning interventions are or are not effective.

Keywords Methodological rigor · Internal validity · Active learning · STEM education · Intervention studies

✉ Amedee Marchand Martella
amedeemartella@ucsb.edu

Extended author information available on the last page of the article

With STEM pipeline issues and STEM inequity, there have been numerous university- and national-level efforts (e.g., Association of American Universities [AAU], 2017; Carl Wieman Science Education Initiative, n.d.; Center for STEM Learning, 2016) focused on transitioning courses from the lecture method to active learning. As such, active learning, as an umbrella term, continues to gain political and instructional interest (Hartikainen et al., 2019). Active learning is a popular approach to teaching and learning that focuses on involving students in the learning process and affording them agency for their learning (Lombardi et al., 2021). Rather than transmit information to students for them to absorb during a class lecture, for example, instructors implement more student-centered learning where students can participate in class activities such as responding to clicker questions, working on a problem-solving worksheet with their peers, or digging into a case study (Martella, Lovett, & Ramsay, 2021).

The focus on participatory activities/discussion is represented in many definitions of active learning such as those presented by Freeman et al. (2014; see p. 8413–8414) and Lombardi et al. (2021; see p. 16) as well as reflected in the Interactive Constructive Active Passive (ICAP) framework (Chi & Wylie, 2014). One of the most seminal definitions of active learning presented by Bonwell and Eison (1991) is "instructional activities involving students in doing things and thinking about what they are doing" (p. iii). Unsurprising given the open-endedness of active learning definitions throughout the educational literature is the variation in active learning implementations (see discussions by Martella, Lovett, & Ramsay, 2021 and Martella & Schneider, in press). For example, in the Freeman et al. (2014) meta-analysis on active learning, active learning courses "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). This variation has resulted in active learning being called a "curious construct" (Lombardi et al., 2021, p. 8) and being described as "an easy thing to prescribe as a cure but difficult to put into practice" (Eyler, 2018, para. 5).

Despite this variation, the common thread in these active learning implementations is the reduction or elimination of lecture in favor of more participatory activities assigned during class given that lecture is often considered a less effective mode of teaching (e.g., Deslauriers et al., 2019; Stains et al., 2018; Wieman, 2014). Active learning, often touted as a constructivist approach to teaching and learning (e.g., Freeman et al., 2014), focuses on students constructing/building their own knowledge. To do this, active learning instruction often moves students away from passive behaviors and toward more active, constructive, and interactive behaviors (see ICAP Framework—Chi & Wylie, 2014). Researchers who are invested in studying the "active versus passive" contrast may assume that being more behaviorally active leads students to be more cognitively active—an assumption that has been referred to as the *constructivist teaching fallacy* because it equates active learning (cognitively active) with active teaching (behaviorally active; Mayer, 2004, p. 15). According to the science of learning, active learning involves appropriate *cognitive* processing during learning, where during a lesson, information is attended to, organized into a coherent structure, and integrated with relevant prior knowledge (Mayer, 2011, 2022). When active teaching leads to active learning, more meaningful learning can occur—a likely goal of active learning researchers.

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 that included 225 studies comparing student performance under active learning versus traditional lecture in STEM undergraduate courses. One way to determine the impact an article has on scientific discourse is to examine its number of citations (Lopresti, 2010). Based on citation data drawn from Web of Science: Core Collection (WOSCC), the Freeman et al. (2014) meta-analysis has been cited 3182 (WOSCC) times (as of February 2, 2023). When authors cite Freeman et al., they typically use the article as support for reducing or eliminating lecture and adopting active learning (Martella, Yatcilla, et al., 2021). In fact, since the release of Freeman et al. (2014), many calls have been made to switch to active learning instruction with such strong statements as "in undergraduate STEM education, we have the curious situation that, although more effective teaching methods have been overwhelmingly demonstrated, most STEM courses are still taught by lectures—the pedagogical equivalent of bloodletting" (Wieman, 2014, p. 8320); "the impression I get is that it's almost unethical to be lecturing if you have this data…[there is] an abundance of proof that lecturing is outmoded, outdated, and inefficient" (Eric Mazur as quoted in Bajak, 2014, para. 4); and "to put it bluntly, everyone should be taken off the control (i.e., traditional lecture) and switched to the treatment (i.e., carefully considered active learning methodologies)" (Pienta, 2015, p. 963).

Furthermore, when examining national, international, and university-level initiatives and grants that are focused on active learning adoption in college classrooms (e.g., AAU, 2017; Carl Wieman Science Education Initiative, n.d.; Chasteen, 2023; Massachusetts Institute of Technology, 2021; University of Georgia, 2022), the Freeman et al. (2014) meta-analysis is often cited as support for this department or campus-wide adoption. As justification, faculty or administrators often make statements such as "active learning is empirically demonstrated to improve student retention of content" (University of Georgia, 2022, p. 6). For policymakers and practitioners, systematic reviews (like the Freeman et al. meta-analysis) can prove useful when making decisions for public policy and practice by providing efficacy information (Gough et al., 2013).

Given that the fervent efficacy claims, pedagogical decisions, and national and international university initiatives surrounding active learning adoption have been rooted in active learning research, particularly Freeman et al. (2014), it is important to examine the methodological strengths and weaknesses of the literature base to ensure these research-based recommendations are rooted in rigorous research. There have been many approaches taken to assess the methodological rigor in a number of disciplines (e.g., Bratt & Moons, 2015; Garavan et al., 2019; Pennington et al., 2021; Ramirez et al., 2017; Yang et al., 2012). Methodological rigor refers to the "thoroughness and accuracy with which research is conducted and it therefore involves elements such as empirical validity, technical quality, statistical significance, and the generalizability of results" (Flickinger et al., 2014, p. 105). Researchers typically choose methodological controls within categories of validity from which to assess the rigor of their respective literature. For example, Garavan et al. (2019) investigated four categories of validity—each with several dimensions—to determine the methodological rigor of empirical studies on the training and organizational

performance relationship. These categories included internal validity, external validity, construct validity, and statistical conclusion validity. Yang et al. (2012) developed six criteria for their "rigor checklist" to determine the methodological rigor of clinical studies. These criteria included population, design, data and sampling, measure instruments, analysis, and interpretation.

Much of the research conducted on methodological rigor has been conducted in medicine/health sciences which may be unsurprising given the direct impact medical interventions can have on the well-being of human subjects (e.g., Bratt & Moons, 2015; Ramirez et al., 2017). However, another research area in which interventions can have a direct impact on the well-being of human subjects is the learning sciences which includes interdisciplinary research focused on "instruction in the social, organizational, and cultural dynamics of learning; learning and cognition; learning strategies; educational psychology; educational testing and measurement; instructional design and technology; and statistical design of educational research" (National Center for Education Statistics, 2010, p. 1). Assessing the methodological rigor of studies relating to educational topics is critical given these studies' potential to impact course instruction and student academic success.

Methodological rigor assessments in the learning sciences have spanned educational disciplines such as computer science education (Lishinski et al., 2016; Randolph et al., 2007), school psychology (Burns et al., 2012), special education, and experimental STEM education (Avcu & Avcu, 2022) and topics such as literacy (Finch, 2022), reading fluency (Naveenkumar et al., 2022), and concept mapping (Rosas & Kane, 2012). These studies used established methodological quality indicators in their respective fields (i.e., design quality/type of design, sampling procedures, instrument technical adequacy, intervention description, bias and confounding issues, outcome assessment, analytic approach, etc.) and/or have used professional quality design standards such as those from What Works Clearinghouse (WWC) and the Council for Exceptional Children (CEC) (for those studies in education and special education).

For the methodological reviews that have been conducted in the broader educational literature (see examples above), the results have been mixed. For example, Naveenkumar et al. (2022) examined the methodological rigor using the CEC quality indicators of reading fluency intervention studies and found that 22 of the 26 (85%) studies met the quality indicators for internal validity. Sulu et al. (2023) reviewed science education articles focused on school-age students with developmental disabilities and found that 18 of 27 (67%) studies met all of the CEC quality indicators. Sulu et al. (2022) reviewed self-monitoring interventions for students with disabilities and found that 18 of 24 (75%) studies met all WWC design standards fully or with reservations. Alternatively, Avcu and Avcu (2022) examined the methodological rigor of experimental STEM education articles published in Turkish journals and found that all of the selected articles suffered from some form of serious methodological flaw.

Given (a) the limited number of reviews conducted on methodological rigor in the learning, (b) the mixed results of methodological studies depending on educational discipline/topic, and (c) the strong advocacy for active learning adoption, it is important to make active learning the subject of a specific methodological review.

Although there are many categories of methodological controls, the focus of the present review was on methodological controls used to prevent threats to the internal validity of a study. To make valid inferences, researchers have to ensure something other than the independent variable(s) did not produce the outcome. One of the major goals of scientific research is to identify the cause(s) of particular phenomena. Causal explanations allow us to better understand the relationships between or among different variables. However, there are many extraneous factors or confounding variables that can cast doubt over the causality identified in experimental research and weaken the overall internal validity of a study (Marquart, 2017; Martella et al., 2013). These factors, considered to be threats to the internal validity of a study, may include maturation, selection, selection by maturation interaction, statistical regression, mortality, instrumentation, testing, history, and resentful demoralization of the control group (Martella et al., 2013).

To assess the current methodological rigor of the active learning literature as it relates to internal validity, we took two approaches. Our first approach was to examine all of the studies included in the Freeman et al. (2014) meta-analysis to determine the quality of active learning studies on which this popular and influential meta-analysis was based. Our second approach was to examine more current active learning research on which a new meta-analysis could be based by examining a random sample of active learning studies published after the release of Freeman et al. (i.e., years 2015–2022). The year *2015* was selected as a starting point for more recent research to provide an overview of the internal validity of articles that were published after the publication of Freeman et al. in 2014, particularly given that Freeman et al. called for future research to be conducted on active learning (see p. 8413). For each of these approaches, we coded articles according to several controls we deemed critical in establishing methodological rigor as they relate to internal validity (i.e., where cause-and-effect relationships could be demonstrated).

Method

Article Obtainment Procedure

Figure 1 summarizes the search and screening processes for the Freeman et al. (2014) articles and the 2015–2022 sampled articles.

Freeman et al. (2014) Articles

Search Strategy. The second author reviewed Supplemental Table S4 of Freeman et al. (2014) to identify the 225 studies included in the meta-analysis. The second author of the present review identified 187 distinct articles in this table; he contacted the lead author of the Freeman et al. meta-analysis to gain access to one article (representing two studies) that was referred to as "pers comm" in the original reference list (i.e., Mays, n.d.). The second author was unsuccessful in receiving the Mays (n.d.) article; therefore, the present review moved forward with 186 articles (which

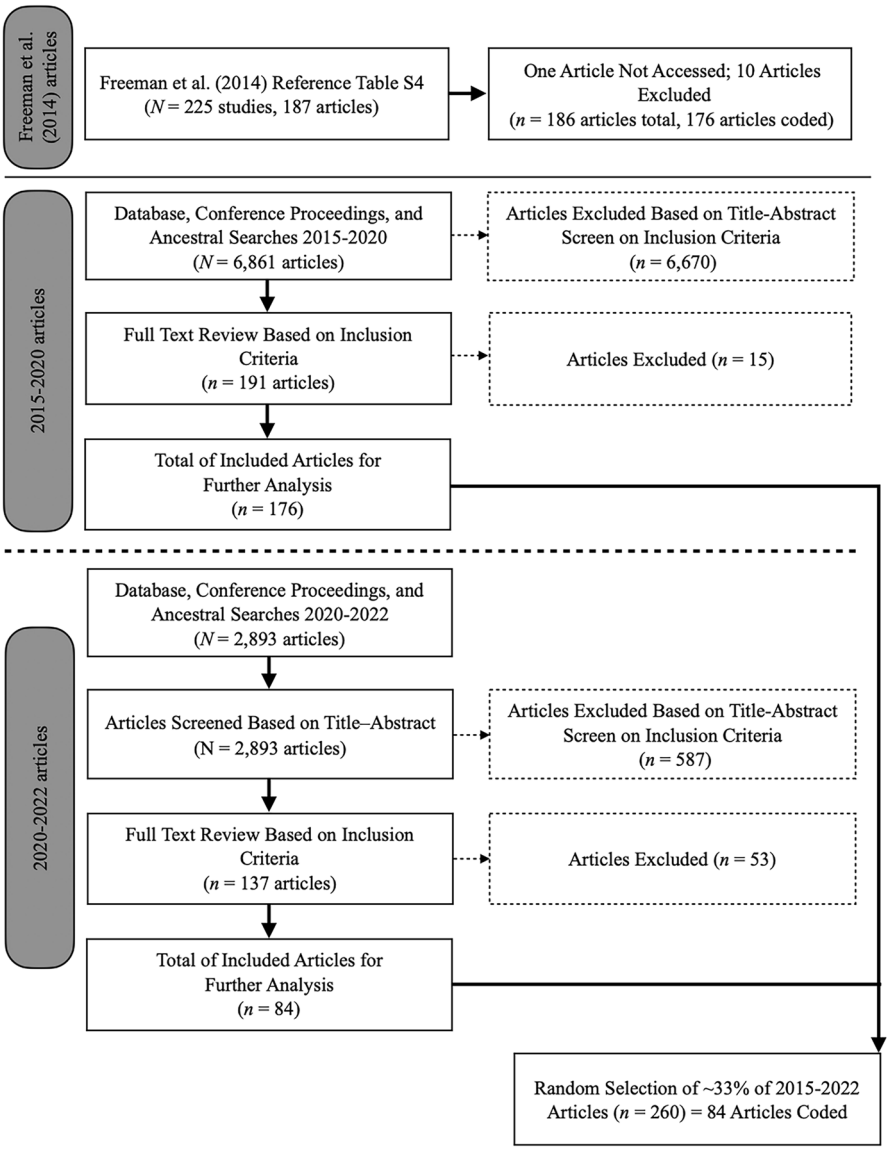

Fig. 1 Search strategies for articles reviewed and coded

represented 223 studies) that were located using the third author's university library subscriptions and interlibrary loan service.

2015–2022 Sampled Articles

Search Strategy. The complete pre-registered search strategy is available on the Open Science Framework at (see supplemental material for blinded review purposes). It is

important to note that our preregistered search strategy was limited to articles with 2010+ publication dates, and we decided after this search to manually limit our search to 2015+ to focus on articles published after the publication of Freeman et al. in 2014. Searches were executed (a) on February 19, 2020 and results were limited to publication years January 1, 2015 to February 19, 2020 and (b) on September 9, 2022 and results were limited to publication years February 20, 2020 to September 9, 2022. The following procedure was identical for our 2015–2020 and 2020–2022 searches.

To gather articles that could be included in a more recent meta-analysis on active learning in STEM, we started with the search terms used by Freeman et al. (2014) and then made a few changes to their original search strategy given changes that have occurred since the release of Freeman et al. (2014). Two of these changes included excluding studies conducted in psychology courses and expanding our search to include studies conducted in more health discipline courses (i.e., STEMM). We made the decision to be more stringent in categorizing STEM disciplines through our exclusion of psychology by aligning with the STEM categories (see Appendix C for the STEM disciplines included in this report—*psychology* categorized as social/behavioral sciences) used in the large report on STEM attrition conducted by the National Center for Education Statistics (Chen, 2013; note: this report was published after the Freeman et al., 2014 articles were sampled). This decision was made given the critical focus on reducing STEM fatigue and attrition by moving toward more active learning-based college interventions. We also made the decision to include more health disciplines (e.g., nursing, biomedical sciences, anatomy and human biology) in our categorization of STEM disciplines for three primary reasons. First, traditional STEM disciplines form the foundation for public health (CDC, 2021). Second, medical science has been the subject of a large number of studies on active learning (see McCoy et al., 2018). Third, medical science has now been added to the STEM acronym (i.e., STEMM; note: this addition was largely discussed after the Freeman et al., 2014 meta-analysis had been published) to reflect the importance of removing barriers in science, technology, engineering, mathematics, and medicine to encourage equitable participation (see National Academies of Sciences, Engineering, and Medicine, 2020) and to improve the innovation ecosystem in the USA (see The White House, 2022).

An additional change we made to the Freeman et al. (2014) search strategy was adding several new search terms that have more recently been used to refer to active learning or lecture interventions. We identified these new search terms by reviewing a sample of active learning articles identified in an initial review of the literature. These search terms included personal response system, team-based learning, inquiry based learning, discovery learning, discovery based learning, enquiry based learning, peer tutor*, interactive lectur*, flipped, passive learning, group activit*, and student led. We also made more use of "*" with search terms from Freeman et al. in case a paper included traditional lecture rather than traditional lecturing, for example (i.e., we included "traditional lectur*"). Search strategies were run in the databases *Web of Science Core Collection*, *PubMed*, *Education Research Information Center* (EBSCOhost), *Engineering Village* (Scopus), and *Dissertations and Theses Global* (ProQuest) and were limited to undergraduate or baccalaureate education. Results

were limited to publications starting in year 2015 (the year after Freeman et al. was published).

After duplicate records were removed using an iterative procedure outlined in Bramer et al. (2016), article citations and abstracts were uploaded into a project on the systematic review screening platform Rayyan (rayyan.ai). Furthermore, an undergraduate research assistant and the first author searched the *American Educational Research Association* for conference proceedings that may have been missed in our search of the literature. No relevant papers were found beyond those already captured in our search strategy. During full-text review (as described below), five research assistants reviewed the reference list to identify papers that were relevant to our review but missed in our search strategy. One article was captured from this hand-searching process and added to the 2020–2022 set. PDFs for included articles were obtained from the university's library holdings or through Interlibrary Loan. The total number of articles available for screening from 2015 to 2020 was 6861. The total number of articles available for screening from 2020 to 2022 was 2894.

Criteria for Inclusion. To be included in our set of 2015–2022 sampled articles, the following criteria had to be met: The articles needed to include (a) a comparison of at least one active learning intervention versus a lecture intervention or versus another active learning intervention; (b) a course(s) conducted at the undergraduate college level; (c) a general face-to-face college course(s) (not honors or remedial courses); (d) a course(s) conducted during the main academic year (i.e., fall/spring semester or fall/winter/spring quarter); (e) a course(s) in a STEM discipline; (f) the intervention(s) scheduled during the main class session (i.e., not in a laboratory or recitation); (g) data on student academic performance included; and (h) a control group (if included) that was specifically labeled to ensure it reflected either a lecture intervention or an active learning intervention.

Screening. For articles in the years 2015–2020, the first author screened 100% of the 6861 articles based on title–abstract. For purposes of interrater agreement, five research assistants received an approximately equal number of the 6861 articles (~1372) to screen. The interrater agreement level was 93.67%. The first author discussed all conflicts with the research assistants and reached resolutions for each. After the screening process, 191 articles moved on to the next phase of screening: full-text review. The same five research assistants each received an approximately equal subset (~38 articles) of the 191 articles to review and sent any they believed should be excluded to the first author, who made the final decision based on the inclusion criteria outlined above. After full-text review, the number of articles that were included in the final set was 176 articles.

For the years 2020–2022, the first author screened 100% of the 2893 articles based on title–abstract. For purposes of interrater agreement, the second and third authors randomly screened 724 of the 2893 articles (~25%). The interrater agreement level was 94.6%. The first author discussed all conflicts with the co-authors and reached a resolution for each. After the screening process, 137 articles moved on to the next phase of screening: full-text review. The first author reviewed the 137 articles; after full-text review, the number of articles that were included in the final set was 84 articles. Thus, a total of 260 articles (176 + 84) were available for review from the years 2015 to 2022.

Coding Procedure

Articles were coded across six basic features related to the method and results section (e.g., course discipline, outcome of article; see Table 1) as well as across 12 internal validity controls (see Table 2). These internal validity controls will be described in further detail below.

Internal Validity Controls

Several internal validity controls were selected based on a review of the research literature on accepted critical areas for article quality. Sources for internal validity controls included APA Publications and Communications Board (2008), Cook et al. (2015), Cook and Campbell (1979), Martella et al. (2013), Theobald et al. (2020), and What Works Clearinghouse Standards Handbook Version 4.1 (2020). The goal was to determine critical variables in an article that demonstrated the level of control of the article. The selected 12 internal validity controls are shown in Table 2; each internal validity control will be briefly described below.

Research Design. Gold standard of research designs is the randomized control group design (Martella et al., 2013). However, a quasi-experimental design may provide experimental control if certain conditions are met such as quality of groups.

Group Equivalence for Both/All Groups—Directly Related to Dependent Variable. If the groups are not equal on measures directly related to the dependent variable (i.e., pretests, examination scores on the same or similar skills, professional content-specific assessments), valid comparisons cannot be made due to a selection confound.

Group Equivalence for Both/All Groups—Indirectly Related to Dependent Variable. Preferably, there are measures used that are directly related to the dependent variable; however, without such measures, measures indirectly related to the dependent variable should be used at a minimum. These measures may include, for example, high school/college GPAs, prerequisite classes taken (possibly with grades), and SAT/ACT scores. If the groups are not equal on certain measures, valid comparisons cannot be made due to a selection confound.

Attrition Equivalency. If there is a differential loss of participants (both number and characteristics of participants), a confound of mortality is introduced (Cook & Campbell, 1979) which makes valid comparisons between/among groups difficult, if not impossible.

Matched Sample Size. If groups are different in size by more than 25%, comparisons become more difficult because class size may affect the effectiveness of group discussions, lecture engagement, responding to instructor questions, etc. (Theobald et al., 2020), leading to the introduction of a potential confound.

Matched Instructor. If the groups had different instructors, the comparison is not only on the type of instruction provided but who provided it (Theobald et al., 2020), leading to a potential confound.

Operational Definition of Variables Provided. Without an operational definition of the critical variable, it is unclear how the groups differed from one another

Table 1 Results of article coding for basic article features

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
Course discipline			
1—Engineering	19 (10.8%)	28 (33.3%)	47 (18.1%)
2—Mathematics	32 (18.2%)	9 (10.7%)	41 (15.8%)
3—Chemistry	25 (14.2%)	12 (14.3%)	37 (14.2%)
4—Biology	31 (17.6%)	9 (10.7%)	40 (15.4%)
5—Physics	30 (17.0%)	8 (9.5%)	38 (14.6%)
6—Computer Science	14 (8.0%)	10 (11.9%)	24 (9.2%)
7—Health Sciences	2 (1.1%)	3 (3.6%)	5 (1.9%)
8—Geology	3 (1.7%)	1 (1.2%)	4 (1.5%)
9—Other (e.g., psychology, astronomy)	17 (9.7%)	2 (2.4%)	19 (7.3%)
Combinations:			
1 and 2	—	2 (2.4%)	2 (0.8%)
1, 2, and 5	1 (0.6%)	—	1 (0.4%)
2 and 3	1 (0.6%)	—	1 (0.4%)
3 and 4	1 (0.6%)	—	1 (0.4%)
Comparison groups			
1—AL vs. L	157 (89.2%)	72 (85.7%)	229 (88.1%)
2—AL vs. AL	2 (1.1%)	1 (1.2%)	3 (1.2%)
3—Other (e.g., AL vs. AL. vs. L; AL vs. L vs. L; L vs. AL [20% flipped] vs. AL [50% flipped] vs. AL [100% flipped])	16 (9.1%)	11 (13.1%)	27 (10.4%)
Combinations:			
1 and 2 (control group had access to AL at times)	1 (0.6%)	—	1 (0.4%)
Intervention time frame?			

Table 1 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
1—Full semester/quarter	147 (83.5%)	67 (79.8%)	214 (82.3%)
2—Multiple weeks	21 (11.9%)	9 (10.7%)	30 (11.5%)
3—1 Week	3 (1.7%)	2 (2.4%)	5 (1.9%)
4—1 Class	5 (2.8%)	5 (6.0%)	10 (3.8%)
Combination:			
1 (AL) and 2 (L)	—	1 (1.2%)	1 (0.4%)
Assessment type			
1—Researcher/teacher developed (or book test bank)	76 (43.2%)	44 (52.4%)	120 (46.2%)
2—Grades	22 (12.5%)	3 (3.6%)	25 (9.6%)
3—Standardized/normed	12 (6.8%)	7 (8.3%)	19 (7.3%)
4—Failure rates	7 (4.0%)	1 (1.2%)	8 (3.1%)
Combinations:			
1 and 2	20 (11.4%)	9 (10.7%)	29 (11.2%)
1, 2, and 3	4 (2.3%)	3 (3.6%)	7 (2.7%)
1, 2, 3, and 4	—	1 (1.2%)	1 (0.4%)
1, 2, and 4	2 (1.1%)	2 (2.4%)	4 (1.5%)
1 and 3	15 (8.5%)	3 (3.6%)	18 (6.9%)
1, 3, and 4	1 (0.6%)	2 (2.4%)	3 (1.2%)
1 and 4	5 (2.8%)	3 (3.6%)	8 (3.1%)
2 and 3	4 (2.3%)	—	4 (1.5%)
2, 3, and 4	1 (0.6%)	—	1 (0.4%)
2 and 4	4 (2.3%)	4 (4.8%)	8 (3.1%)
3 and 4	2 (1.1%)	2 (2.4%)	4 (1.5%)

Table 1 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
Other:			
Accuracy of students taking health-related measurements on live subjects	1 (0.6%)	—	1 (0.4%)
Results: group that resulted in the highest performance			
1—Statistical significance experimental	52 (29.5%)	33 (39.3%)	85 (32.7%)
2—Statistical significance control	1 (0.6%)	0 (0.0%)	1 (0.4%)
3—Non-significance experimental	26 (14.8%)	4 (4.8%)	30 (11.5%)
4—Non-significance control	5 (2.8%)	2 (2.4%)	7 (2.7%)
5—Equal	2 (1.1%)	0 (0.0%)	2 (0.8%)
6—Depends on DM examined (mixed)	35 (19.9%)	18 (21.4%)	53 (20.4%)
7—No stats analysis (favoring experimental group)	28 (15.9%)	13 (15.5%)	41 (15.8%)
7—No stats analysis (favoring control group)	1 (0.6%)	1 (1.2%)	2 (0.8%)
7—No stats analysis (no differences)	1 (0.6%)	0 (0.0%)	1 (0.4%)
Combinations:			
1 and 2	1 (0.6%)	1 (1.2%)	2 (0.8%)
1 and 3	8 (4.5%)	4 (4.8%)	12 (4.6%)
1 and 6	12 (6.8%)	4 (4.8%)	16 (6.2%)
1, 3, and 6	1 (0.6%)	—	1 (0.4%)
1, 5, and 6	1 (0.6%)	1 (1.2%)	2 (0.8%)
2 and 6	1 (0.6%)	—	1 (0.4%)
3 and 4	—	3 (3.6%)	3 (1.2%)
3 and 6	1 (0.6%)	—	1 (0.4%)
Majority student preference			
0—N/A	111 (63.1%)	57 (67.9%)	168 (64.6%)

Table 1 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
1—Experimental	48 (27.3%)	16 (19.0%)	64 (24.6%)
2—Control	4 (2.3%)	2 (2.4%)	6 (2.3%)
3—Equal	12 (6.8%)	5 (6.0%)	17 (6.5%)
Combinations:			
1 and 2	—	1 (1.2%)	1 (0.4%)
1, 2, and 3	—	1 (1.2%)	1 (0.4%)
2 and 3	1 (0.6%)	1 (1.2%)	2 (0.8%)
Mixed based on question	—	1 (1.2%)	1 (0.4%)

Table 2 Results of article coding for the key internal validity controls

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
Research design			
1—Randomized	18 (10.2%)	7 (8.3%)	25 (9.6%)
2—Quasi-Experimental	155 (88.1%)	74 (88.1%)	229 (88.1%)
3—Other (within)	2 (1.1%)	3 (3.6%)	5 (1.9%)
Combination:			
1 and 2 (two different universities)	1 (0.6%)	—	1 (0.4%)
Group equivalence for both/all groups: directly related to dependent variable			
1—Yes, equivalent	41 (23.3%)	26 (31.0%)	67 (25.8%)
2—Yes, nonequivalent	4 (2.3%)	10 (11.9%)	14 (5.4%)
3—No	128 (72.7%)	46 (54.8%)	174 (66.9%)
4—N/A	2 (1.1%)	2 (2.4%)	4 (1.5%)
Combination:			
1 and 2 (equivalent for three experimental groups but comparison group different)	1 (0.6%)	—	1 (0.4%)
Group equivalence for both/all groups: indirectly related to dependent variable			
1—Yes, equivalent	21 (11.9%)	9 (10.7%)	30 (11.5%)
2—Yes, nonequivalent	4 (2.3%)	2 (2.4%)	6 (2.3%)
3—No	84 (47.7%)	55 (65.5%)	139 (53.5%)
4—N/A	2 (1.1%)	2 (2.4%)	4 (1.5%)
5—Demographic	6 (3.4%)	7 (8.3%)	13 (5.0%)
Combinations:			
1 and 2	1 (0.6%)	—	1 (0.4%)
1, 2, and 5	4 (2.3%)	2 (2.4%)	6 (2.3%)
1 and 3	1 (0.6%)	—	1 (0.4%)

Table 2 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
1 and 5	33 (18.8%)	6 (7.1%)	39 (15.0%)
2 and 5	19 (10.8%)	1 (1.2%)	20 (7.7%)
2, 4, and 5	1 (0.6%)	—	1 (0.4%)
Attrition equivalency			
Attrition occurred from pre- to posttest in any group (greater than 30% loss)			
Yes	27 (15.3%)	9 (10.7%)	36 (13.8%)
No	46 (26.1%)	28 (33.3%)	74 (28.5%)
Not specified	103 (58.5%)	47 (56.0%)	150 (57.7%)
Differential attrition between/among groups			
Yes	36 (20.5%)	5 (6.0%)	41 (15.8%)
No	35 (19.9%)	14 (16.7%)	49 (18.8%)
Not specified	105 (59.7%)	65 (77.4%)	170 (65.4%)
Matched sample size (i.e., not different by more than 25%)			
1—Yes	72 (40.9%)	42 (50.0%)	114 (43.8%)
2—No	81 (46.0%)	29 (34.5%)	110 (42.3%)
3—Not specified	20 (11.4%)	11 (13.1%)	31 (11.9%)
Combination:			
1 and 2 (depends on year/phase, course pairs/ comparisons made)	3 (1.7%)	2 (2.4%)	5 (1.9%)
Matched instructor			
1—Yes	58 (33.0%)	38 (45.2%)	96 (36.9%)
2—No	81 (46.0%)	22 (26.2%)	103 (39.6%)
3—Not specified	34 (19.3%)	23 (27.4%)	57 (21.9%)
Combination:			

Table 2 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
1 and 2 (two AL taught by different instructors and two L taught by the same instructors; depends on years/experiment; same instructor but classes led by different TAs)	3 (1.7%)	1 (1.2%)	4 (1.5%)
Operational definition of variables provided			
Operational definition of experimental variable(s)			
1—Yes	88 (50.0%)	43 (51.2%)	131 (50.4%)
2—Partial	70 (39.8%)	38 (45.2%)	108 (41.5%)
3—No	18 (10.2%)	3 (3.6%)	21 (8.1%)
Operational definition of control variable(s)			
1—Yes	35 (19.9%)	21 (25.0%)	56 (21.5%)
2—Partial	57 (32.4%)	29 (34.5%)	86 (33.1%)
3—No	84 (47.7%)	34 (40.5%)	118 (45.4%)
Dosage of variables described and matched			
Dosage of the experimental variable(s) specified			
1—Yes	71 (40.3%)	21 (25.0%)	92 (35.4%)
2—Partial	49 (27.8%)	34 (40.5%)	83 (31.9%)
3—No	56 (31.8%)	29 (34.5%)	85 (32.7%)
Dosage of the control variable(s) specified			
1—Yes	56 (31.8%)	18 (21.4%)	74 (28.5%)
2—Partial	40 (22.7%)	31 (36.9%)	71 (27.3%)
3—No	80 (45.5%)	35 (41.7%)	115 (44.2%)
Dosage equal between (among) groups			
1—Yes	52 (29.5%)	16 (19.0%)	68 (26.2%)
2—No	23 (13.1%)	6 (7.1%)	29 (11.2%)

Table 2 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
3—Not specified	101 (57.4%)	62 (73.8%)	163 (62.7%)
Matched implementation time frame			
1—Yes	83 (47.2%)	36 (42.9%)	119 (45.8%)
2—No	77 (43.8%)	43 (51.2%)	120 (46.2%)
3—Same semester different weeks	1 (0.6%)	3 (3.6%)	4 (1.5%)
4—Not specified	10 (5.7%)	0 (0.0%)	10 (3.8%)
Combination:			
1 and 2 (depends on university involved, compare within and across years)	5 (2.8%)	2 (2.4%)	7 (2.7%)
Matched content			
1—Yes	112 (63.6%)	41 (48.8%)	153 (58.8%)
2—No	24 (13.6%)	4 (4.8%)	28 (10.8%)
3—Not specified	39 (22.2%)	39 (46.4%)	78 (30.0%)
Combination:			
1 and 2 (depends on semester compared)	1 (0.6%)	—	1 (0.4%)
Dependent measure(s): equal for both/all groups			
1—Yes	138 (78.4%)	55 (65.5%)	193 (74.2%)
2—No	15 (8.5%)	8 (9.5%)	23 (8.8%)
3—Partial	10 (5.7%)	6 (7.1%)	16 (6.2%)
4—Unsure	13 (7.4%)	12 (14.3%)	25 (9.6%)
Combination:			
1 and 4 (depends on measure)	—	3 (3.6%)	3 (1.2%)
Implementation fidelity			

Table 2 (continued)

Coding description	Freeman et al. (2014) ($N = 176$) $n = (\%)$	2015–2022 Sample ($N = 84$) $n = (\%)$	Combined ($N = 260$) $n = (\%)$
Experimental group fidelity			
1—Yes	13 (7.4%)	5 (6.0%)	18 (6.9%)
2—No	163 (92.6%)	79 (94.0%)	242 (93.1%)
Control group fidelity			
1—Yes	11 (6.3%)	5 (6.0%)	16 (6.2%)
2—No	165 (93.8%)	79 (94.0%)	244 (93.8%)

and what specifically led one group to outperform (or not outperform) the other (Martella et al., 2020); it is also difficult to replicate studies without this information.

Dosage of Variables Described and Matched. Dosage refers to how much time students spent with class content (both outside of class and inside of class), and controlling the dosage between/among conditions is important to ensure the independent variable caused the results and not a difference in time-on-task/exposure to class content (Mason & Smith, 2020).

Matched Implementation Time Frame. Implementing the experimental and control groups during the same time period is a control procedure to protect against history effects (events that occur during a study that can affect the outcome) (Campbell & Stanley, 1963).

Matched Content. Valid comparisons cannot be made between groups if the same content was not taught to the different groups compared (Deslauriers et al., 2019).

Dependent Measure(s)—Equal for All Groups. To examine differences in learning gains between/among conditions, students need to be assessed on measures that are deemed equivalent (Martella et al., 2013); otherwise, it is unclear if differences in student learning are due to the independent variable or are due to assessment differences.

Implementation Fidelity. Implementation fidelity allows one to assess if the interventions were implemented as intended and/or described (Sanetti et al., 2021)—without it, claims as to the effectiveness of one group over another are difficult to validate.

Coding of Freeman et al. (2014) Articles

A total of 176 of the 187 articles were coded (see Fig. 1) for the following reasons. For the present review, if an article had two or more experiments/studies, the article was counted as one overall article rather than as several separate studies. Multiple (combination) codes may have been provided if two or more studies within an article used different designs, measures, procedures, etc. Thus, we use the terms "article" or "articles" versus "study" or "studies" throughout. Furthermore, the personal communication (i.e., Mays, n.d.) article could not be located and 10 articles (representing 15 studies) were removed from the present analysis because they were (a) implemented in an introduction to exceptional children course (i.e., Kellum et al., 2001); (b) a review of different university curriculum approaches or a review of published studies (i.e., Al-Holou et al., 1999; Gosser, 2011); (c) a description of different in-class learning experiences (some students on "hot seat") or methods of taking notes (i.e., Crider, 2004; Davis & Hult, 1997); (d) a description of anecdotal observations of differences in including web-based activities (i.e., Marrs & Novak, 2004); or (e) a comparison of frequent and/or unannounced quizzes versus no quizzes before examinations/tests (i.e., Barbarick, 1998; Graham, 1999; Haberyan, 2003; Steele, 2003). Therefore, the number of articles in our coding set was 176.

The fourth, fifth, and sixth authors each received the 176 Freeman et al. (2014) articles to code according to the 12 internal validity controls. The fourth author

coded 61 (34.7%) of the articles, the fifth author coded 58 (33.0%) of the articles, and the sixth author assistant coded 57 (32.4%) of the articles. The second author independently coded 100% of these articles and was the primary rater.

Interrater agreement was calculated by dividing the number of agreements by disagreements and multiplying by 100. A strict level of agreement was required for interrater agreement analysis. For example, if the sixth author coded results as "7" (no statistical analysis provided) but the primary rater coded the same article as "7" and "3" (no statistical analysis provided but the results favored the experimental group), it was scored as a complete disagreement. Likewise, if a code was left blank by the fifth author and the primary rater provided a code, it was scored as a disagreement. In other words, the codes had to be identical to be counted as an agreement. The interrater agreement was as follows: the primary rater and the fourth author had an agreement of 96.3% (range 82.6 to 100%); the primary rater and the fifth author had an agreement of 92.9% (range 73.9 to 100%); the primary rater and the sixth author had an agreement of 88.1% (range 65.2 to 100%). The overall interrater agreement was 92.5% (range 65.2 to 100%). When there were disagreements in scoring, the coding of the primary rater (who coded all articles) was used.

Coding of 2015–2022 Sampled Articles

The first author used a simple random sampling process to obtain a sample of ~33% (84) of the 260 articles from 2015 to 2022 to be coded according to the 12 internal validity controls (see Fig. 1). The purpose of conducting a random sample was twofold. First, the primary focus of the review was on the articles included in the Freeman et al. (2014) meta-analysis due to its far-reaching influence. However, given that the meta-analysis was published in 2014, we deemed it important to gather a snapshot of the quality of the more recent active learning literature that could be included in a new meta-analysis on active learning post-Freeman et al. Second, due to the time-intensive coding procedure (including double coding all articles) and the large, combined number of active learning studies from Freeman et al. (2014) and post-Freeman et al. (436 articles in total), a one-third sample of articles was chosen for this snapshot to ensure the recency of the review. Random sampling methods have been used previously in internal validity control studies (e.g., Han et al., 2022; Randolph et al., 2007) as well as in other types of systematic reviews (e.g., Lazonder & Janssen, 2022; Luo et al., 2013) due to the time-intensive coding procedures and/or number of articles identified through the screening process.

The first author coded 69 (82.1%) of these 84 articles and the fourth author coded 15 (17.9%) of the 84 articles. The second author independently coded 100% of the 84 articles and was the primary rater. Interrater agreement was calculated as described above. The same strict level of agreement described above was required for interrater agreement analysis. The interrater agreement between the primary rater and the first author was 93.8% (range 87.0 to 100%), and the interrater agreement between the primary rater and the fourth author was 89.3% (range 78.2 to 100%). The overall interrater agreement was 93.0% (range 78.2 to 100%). When there were disagreements in scoring, the coding of the primary rater (who coded all articles) was used.

Data Analysis

Frequency counts were determined for each code that fell under each internal validity control. These frequencies were then turned into a percentage of articles that received a particular code. If an article received a combined code—e.g., 3 and 4—it did not get double counted in the "3" frequency count and the "4" frequency count; rather, a combined code category was created.

Results

The data for the Freeman et al. (2014) articles and the 2015–2022 sampled articles are presented separately.

Freeman et al. (2014) Articles

Tables 1 and 2 show the results from each of the coded "basic article features" and "internal validity controls" for the articles reviewed in the Freeman et al. (2014) meta-analysis.

Basic Article Features

The codings for the basic article features are shown in Table 1.

Course Discipline. As shown in the table, the greatest number of articles (18.2%) were conducted in mathematics classes. Biology and physics were next with 17.6% and 17.0%, respectively. Articles in chemistry classes were the fourth most frequent at 14.2%. Finally, engineering was the fifth most frequent at 10.8%. Therefore, 77.8% of all articles were conducted in mathematics, biology, physics, chemistry classes, and/or engineering. The least frequent course discipline was health sciences (1.1%).

Comparison Groups. The most frequent comparison (89.2%) made across articles was active learning versus lecture. Comparisons between two active learning approaches occurred in just 1.1% of the articles. Other comparisons were made in 9.1% of the articles and involved >2 course/intervention comparisons (e.g., active learning 1 vs. active learning 2 vs. lecture). One article (0.6%) was coded alone since it compared an active learning group to a lecture group that had access to active learning components at various times.

Intervention Time Frame. The vast majority of articles (83.5%) implemented comparison groups over a full semester/quarter. A smaller number of articles implemented the groups over multiple weeks (11.9%), one week (1.7%), or during a single class session (2.8%).

Assessment Type. Most of the assessments used in the articles (43.2%) were those developed by the researcher/teacher or from a book test bank. These assessments were used alone or in combination with other assessments in 69.9% of the

articles. Grades were solely used in 12.5% of the articles or solely used and used in combination with other assessments in 32.4% of the articles. Standardized/normed assessments were solely used in 6.8% of the articles or solely used and used in combination with other assessments in 22.2% of the articles. Failure rates were solely used in 4.0% of the articles or solely used and used in combination with other assessments in 12.5% of the articles. A combination of assessments occurred in 33.0% of the articles. Other assessments reported were health related measures on live subjects (0.6%).

Results: Group That Resulted in the Highest Performance. The experimental group (i.e., active learning) was found to produce statistically significantly greater improvements in academic performance in 29.5% of the articles and greater performance in 14.8% of the articles but without statistically significant differences. The control group (i.e., lecture) was found to produce statistically significantly greater improvements in academic performance in 0.6% of the articles and greater performance in 2.8% of articles but without statistically significant differences. The experimental and control groups were equal in 1.1% of articles, and mixed results were found in 19.9% of the articles depending on the dependent measure reported. Inferential statistics were not used in 17.0% of the articles; of these articles, 93.3% showed results favoring the experimental group, 3.3% showed results favoring the control group, and 3.3% showed results favoring neither the experimental or control group (i.e., no differences between them).

Given additional nuances of articles, we used multiple codes to indicate combinations of results due to multiple dependent variables, multiple studies in an article, and/or multiple groups that were compared. Of the 14.2% of articles that had a combination of results, 32.0% had statistically significant or nonsignificant results favoring just the experimental group(s). Therefore, overall, 64.8% of the articles showed only the experimental group to be more effective than the control group, and 4.0% of the articles showed only the control group to be more effective than the experimental condition(s).

Majority of Student Preference. The majority of articles (63.1%) did not report a student preference for either group. In the articles that did present this information, there was a student preference for the experimental condition in 27.3% of these articles and a student preference for the control condition in 2.3% of these articles. Equal preference was reported in 6.8% of these articles, and one article (0.6%) reported that there was a preference for both depending on the question asked.

Internal Validity Controls

The codings for the internal validity controls are shown in Table 2.

Research Design. The vast majority of articles (88.1%) used a quasi-experimental design (i.e., when different years were compared or when existing classrooms were used). Only 10.2% of the articles used randomized control designs. Few articles used other designs such as a within-subjects designs (1.1%) or a combination of designs when there were multiple experiments and/or multiple group comparisons (0.6%).

Group Equivalence for Both/All Groups: Directly Related to Dependent Variable. The following results were obtained despite the fact that Freeman et al. (2014) reported they excluded articles that did not have student equivalence. A group equivalence measure directly related to the dependent variable was not reported in 128 (72.7%) of the articles. Of the articles that did include/report this information, only 23.3% of these articles had at least one group equivalence measure directly related to the dependent variable that showed the groups to be equivalent, and only 2.3% had at least one group equivalence measure directly related to the dependent variable that showed the groups to be nonequivalent. A small percentage (0.6%) of articles reported that at least one group equivalence measure directly related to the dependent measure showed some of the groups to be equivalent and others to be nonequivalent (i.e., three experimental groups were equivalent but the comparison group was nonequivalent). Finally, including a group equivalence measure directly related to the dependent variable was considered "not applicable" in 1.1% of the articles (e.g., within-subjects design).

Group Equivalence for Both/All Groups: Indirectly Related to Dependent Variable. The following results were obtained despite the fact that Freeman et al. (2014) reported they excluded articles that did not have student equivalence. A group equivalence measure indirectly related to the dependent variable (e.g., ACT scores, grade point average before the study) was not included in 47.7% of the articles. At least one group equivalence measure indirectly related to the dependent variable that showed the groups to be equivalent was reported in 11.9% of the articles, and 2.3% of the articles had at least one group equivalence measure indirectly related to the dependent variable that showed the groups to be nonequivalent. Including a group equivalence measure indirectly related to the dependent variable was considered "not applicable" in 1.1% of the articles (e.g., within-subjects design). Finally, several articles reported demographic information only (3.4%) or some combination of measures (33.5%).

Attrition Equivalency. There was attrition that occurred between the pretest and posttest of greater than 30.0% in 15.3% of the articles. There was no attrition or attrition was less than 30.0% in 26.1% of the articles. However, information on the level of attrition, if any, was absent in 58.5% of the articles.

Of the articles that did have attrition greater than 30.0% (15.3%), the attrition was differential in 92.6% of these articles and not specified in 7.4% of these articles. Overall (i.e., across all articles), differential attrition between/among groups occurred in 20.5% of the articles, did not occur 19.9% of the articles, and was not reported in 59.7% of the articles.

Matched Sample Size. The groups were comparable in size (i.e., not different by more than 25%) in 40.9% of the articles; however, they were not comparable in size in 46.0% of the articles. In addition, 1.7% of the articles reported that groups were both comparable and not comparable in size depending on which comparisons were made between/among the multiple groups. Group sizes were not reported in 11.4% of the articles.

Matched Instructor. The following results were obtained despite the fact that Freeman et al. (2014) reported they excluded articles that did not have instructor equivalence. The same instructor was used for the experimental and control

groups in 33.0% of the articles. However, 46.0% of the articles did not use the same instructor for the comparison groups. In addition, there was a combination of instructors in 1.7% of articles where two active learning groups were taught by different instructors and two lecture groups were taught by the same two instructors or the same instructors were used in one year and different instructors were used in another year. Information on whether the instructor was the same between/among groups was not reported in 19.3% of the articles.

Operational Definition of Variables Provided. All features of the experimental variable(s) were operationally defined in 50.0% of the articles. These features were partially operationally defined in 39.8% of the articles. Experimental features were not operationally defined in 10.2% of the articles. For the control condition, all features of the control variable(s) were operationally defined in 19.9% of the articles, and these features were partially operationally defined in 32.4% of the articles. Control features were not operationally defined in 47.7% of the articles.

Dosage of Variables Described and Matched. The dosage of the experimental variable(s) was specified in 40.3% of the articles and was partially specified in 27.8% of the articles. The dosage of the experimental variable(s) was not specified in 31.8% of the articles. The dosage of the control variable(s) was specified in 31.8% of the articles and was partially specified in 22.7% of the articles. The dosage of the control variable(s) was not specified in 45.5% of the articles.

When comparing the dosage between/among groups, the dosage was equal in 29.5% of the articles, was not equal in 13.1% of the articles, and was not specified in 57.4% of the articles.

Matched Implementation Time Frame. Groups were not implemented at the same time in 43.8% of the articles while 47.2% of the articles did so. In addition, 0.6% of the articles implemented the groups during the same semester but during different weeks, and 2.8% of the articles had a combination of implementing the groups at the same time and at different times when different groups were compared during the same year and/or across different years.

Finally, it was not specified if the groups were implemented at the same time in 5.7% of the articles.

Matched Content. The same content was taught across groups in 63.6% of the articles. However, the same content was not taught across groups in 13.6% of the articles. In addition, some content was the same and some was different depending on which semesters were compared in 0.6% of the articles. Whether the content was the same was not specified in 22.2% of the articles.

Dependent Measure(s): Equal for All Groups. The following results were obtained despite the fact that Freeman et al. (2014) reported they excluded articles that did not have examination equivalence. The dependent measure(s) was equal for both/all groups in 78.4% of the articles but was not equal in 8.5% of the articles. Dependent measures were partially equal (i.e., some but not all of the measures were equal) in 5.7% of the articles. It was not possible to determine if the dependent measure(s) was equal in 7.4% of the articles.

Implementation Fidelity. Implementation fidelity for the experimental group occurred in 7.4% of the articles but did not occur in 92.6% of the articles.

Implementation fidelity for the control condition occurred in 6.3% of the articles but did not occur in 93.8% of the articles.

Rigor Across All Internal Validity Controls. Only 2.3% articles (i.e., Basili & Sanford, 1991; Bilgin, 2006; Carmichael, 2009; Randolph, 1992) successfully met each of the 11 internal validity controls; this analysis (a) excluded both implementation fidelity and measures indirectly related to the dependent variable (if they already had measures directly related to the dependent variable that showed equivalence); (b) allowed for partial ratings on operationally defined variables and dosage of variables; and (c) gave articles the benefit of the doubt when not enough information or no information was provided for an internal validity control (i.e., was coded as "not specified"). If articles that were coded as "not specified" were counted as not meeting the internal validity control, only 0.6% (i.e., Randolph, 1992) would have met all controls, excluding implementation fidelity. If implementation fidelity were included in the analysis, no articles would have met all of the controls. Therefore, 100% of the articles had at least one internal validity control issue that could interfere with the conclusion of the effects of the independent variable.

Figure 2 shows the percentage of articles that fully met one or more of the internal validity controls. Note: credit was provided for having a measure that was *either* directly or indirectly related to the dependent variable that showed the groups were equivalent; therefore, there were 11 as opposed to 12 controls shown. The majority (61.9%) of the articles met half or fewer of the controls, only 24.4% of the articles met 7 to 10 of the controls, and no articles met all 11 of the controls.

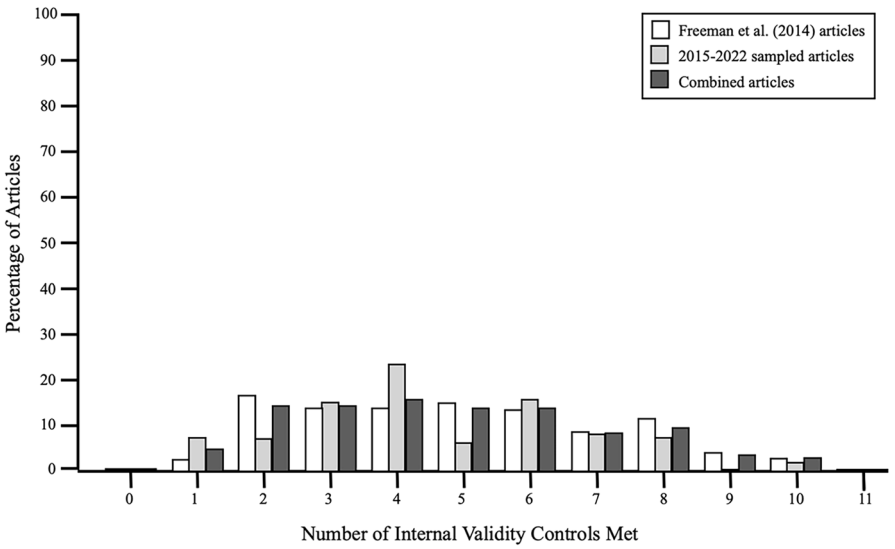

Fig. 2 Percentage of articles meeting 0 to 11 internal validity controls

2015–2022 Sampled Articles

Tables 1 and 2 show the results from each of the coded "basic article features" and "internal validity controls" for the sample of articles (and their subsequent studies) reviewed since the Freeman et al. (2014) meta-analysis was published (years 2015 to 2022).

Basic Article Features

The codings for the basic article features are shown in Table 1.

Course Discipline. As shown in the table, the greatest number of articles (33.3%) were conducted in engineering classes. Chemistry and computer science were next with 14.3% and 11.9%, respectively. Articles in mathematics and biology classes were the fourth most frequent at 10.7% each. Therefore, 81.0% of all articles were conducted in engineering, chemistry, computer science, mathematics, and biology classes. The least frequent course discipline was geology (1.2%).

Comparison Groups. The most frequent comparison (85.7%) made across articles was active learning versus lecture. Comparisons between two active learning approaches occurred in just 1.2% of the articles. Other comparisons were made in 13.1% of the articles and involved >2 course/intervention comparisons (e.g., active learning 1 vs. active learning 2 vs. lecture).

Intervention Time Frame. The vast majority of articles (79.8%) implemented comparison groups over a full semester/quarter. A smaller number of articles implemented the groups over multiple weeks (10.7%), one week (2.4%), or during a single class session (6.0%). Only 1.2% of the articles implemented multiple time frames by providing an active learning group over a full semester but the control group over multiple weeks.

Assessment Type. Most of the assessments used in the articles were those developed by the researcher/teacher or from a test bank (52.4%). These assessments were used alone or in combination with other assessments in 79.8% of articles. Grades were solely used in 3.6% of the articles or solely used and used in combination with other assessments in 26.2% of the articles. Standardized/normed assessments were solely used in 8.3% of the articles or solely used and used in combination with other assessments in 21.4% of the articles. Failure rates were solely used in 1.2% of the articles or solely used and used in combination with other assessments in 17.9% of articles. A combination of assessments occurred in 34.5% of the articles.

Results: Group That Resulted in the Highest Performance. The experimental group was found to produce statistically significantly greater improvements in academic performance in 39.3% of the articles and greater performance in 4.8% of the articles but without statistically significant differences. The control group was found to produce statistically significantly greater improvements in academic performance in 0.0% of the articles and greater performance in 2.4% of the articles but without statistically significant differences. There were no articles where the groups were equal, but mixed results were found in 21.4% of the articles depending on the dependent measure reported. Inferential statistics were not used in 16.7% of articles;

of these articles, 92.9% showed results favoring the experimental group and 7.1% showed results favoring the control group.

Given additional nuances of articles, we used multiple codes to indicate combinations of results due to multiple dependent variables, multiple studies in an article, and/or multiple groups that were compared. Of the 15.5% of articles that had a combination of results, 30.8% had statistically significant or nonsignificant results favoring just the experimental group(s). Thus, 64.3% of the articles showed the experimental group (i.e., active learning) to be more effective than the control group (i.e., lecture) in some way, and 3.6% of the articles showed only the control group to be more effective than the experimental condition(s).

Majority of Student Preference. The majority of articles (67.9%) did not report a student preference for either group. In the articles that did present this information, there was a student preference for the experimental condition in 19.0% of these articles and a student preference for the control condition in 2.4% of these articles. Equal preference was reported in 6.0% of these articles. There were combinations of student preference in four (4.8%) of the articles.

Internal Validity Controls

The codings for the internal validity controls are shown in Table 2.

Research Design. The vast majority of articles (88.1%) used a quasi-experimental design (i.e., when different years were compared or when existing classrooms were used). Only 8.3% of the articles used randomized control designs. Within-subjects designs were used in 3.6% of the articles.

Group Equivalence for Both/All Groups: Directly Related to Dependent Variable. A group equivalence measure directly related to the dependent variable was not included in 54.8% of the articles. Of the articles that did include/report this information, only 31.0% of these articles had at least one group equivalence measure directly related to the dependent variable that showed the groups to be equivalent, and only 11.9% of these articles had at least one group equivalence measure directly related to the dependent variable that showed the groups to be nonequivalent. Finally, including a group equivalence measure directly related to the dependent variable was considered "not applicable" in 2.4% of the articles (e.g., within-subjects design).

Group Equivalence for Both/All Groups: Indirectly Related to Dependent Variable. A group equivalence measure indirectly related to the dependent variable (e.g., ACT scores, grade point average before the study) was not included in 65.5% of the articles. At least one group equivalence measure indirectly related to the dependent variable that showed the groups to be equivalent was reported in 10.7% of the articles, while 2.4% of the articles had at least one group equivalence measure indirectly related to the dependent variable that showed the groups to be nonequivalent. Including a group equivalence measure indirectly related to the dependent variable was considered "not applicable" in 2.4% of the articles (e.g., within-subjects design). Finally, several articles reported demographic information only (8.3%) or in some combination with other measures (10.7%).

Attrition Equivalency. There was attrition that occurred between the pretest and posttest of greater than 30.0% in 10.7% of the articles. There was no attrition or attrition was less than 30% in 33.3% of the articles. However, information on the level of attrition, if any, was absent in 56.0% of the articles.

Of the articles that did have attrition greater than 30.0% (10.7%), the attrition was differential in 22.2% of these articles and not specified in 77.8% of these articles. Overall (i.e., across all articles), differential attrition between/among groups occurred in 6.0% of the articles, did not occur in 16.7% of the articles, and was not reported in 77.4% of the articles.

Matched Sample Size. The groups were comparable in size (i.e., not different by more than 25%) in 50.0% of the articles; however, they were not comparable in size in 34.5% of the articles. In addition, 2.4% of the articles reported that groups were both comparable and not comparable in size depending on which comparisons were made between/among the multiple groups. Group sizes were not reported in 13.1% of the articles.

Matched Instructor. The same instructor was used for the experimental and control groups in 45.2% of the articles. However, 26.2% of the articles did not use the same instructor for the comparison groups. In addition, 1.2% of the articles had the same instructor but the classes were led by different teaching assistants. Information on whether the instructor was the same between/among groups was not reported in 27.4% of the articles.

Operational Definition of Variables Provided. All features of the experimental variable(s) were operationally defined in 51.2% of the articles. These features were partially operationally defined in 45.2% of the articles. Experimental features were not operationally defined in 3.6% of the articles. For the control condition, all features of the control variable(s) were operationally defined in 25.0% of the articles, and these features were partially operationally defined in 34.5% of the articles. Control features were not operationally defined in 40.5% of the articles.

Dosage of Variables Described and Matched. The dosage of the experimental variable(s) was specified in 25.0% of the articles and was partially specified in 40.5% of the articles. The dosage of the experimental variable(s) was not specified in 34.5% of the articles. The dosage of the control variable(s) was specified in 21.4% of the articles and was partially specified in 36.9% of the articles. The dosage of the control variable(s) was not specified in 41.7% of the articles.

When comparing the dosage between/among groups, the dosage was equal in 19.0% of the articles, was not equal in 7.1% of the articles, and was not specified in 73.8% of the articles.

Matched Implementation Time Frame. Groups were not implemented at the same time in 51.2% of the articles while 42.9% of the articles did so. In addition, 3.6% of the articles implemented the groups during the same semester but during different weeks, and 2.4% of the articles had a combination of implementing the groups at the same time and at different times when different groups were compared during the same year and across different years.

Matched Content. The same content was taught across groups in 48.8% of the articles. However, the same content was not taught across groups in 4.8% of the articles. Whether the content was the same was not specified in 46.4% of the articles.

Dependent Measure(s): Equal for All Groups. The dependent measure(s) was equal for both/all groups in 65.5% of the articles but was not equal in 9.5% of the articles. Dependent measures were partially equal (i.e., some but not all of the measures were equal) in 7.1% of the articles. In addition, only 3.6% of the articles had results that depended on the measure examined. It was not possible to determine if the dependent measure(s) was equal in 14.3% of the articles.

Implementation Fidelity. Implementation fidelity for the experimental group occurred in 6.0% of the articles but did not occur in 94.0% of the articles. Implementation fidelity for the control condition occurred in 6.0% of the articles but did not occur in 94.0% of the articles.

Rigor Across All Internal Validity Controls. Only 1.2% of the articles (i.e., Lape et al., 2016) successfully met each of the 11 internal validity controls; this analysis (a) excluded both implementation fidelity and measures indirectly related to the dependent variable if they already had measures directly related to the dependent variable that showed equivalence; (b) allowed for partial ratings on operationally defined variables and dosage of variables; and (c) gave articles the benefit of the doubt when not enough information or no information was provided for an internal validity control (i.e., was coded as "not specified"). If articles that were coded as "not specified" were counted as not meeting the internal validity control, no articles would have met all of the controls. Therefore, 100% of the articles had at least one internal validity control issue that could interfere with the effects of the independent variable.

Figure 2 shows the percentage of the articles that fully met one or more of the internal validity controls. Note: credit was provided for having a measure that was *either* directly or indirectly related to the dependent variable that showed the groups were equivalent; therefore, there were 11 as opposed to 12 controls shown. The majority (64.3%) of the articles met half or fewer of the controls, only 19.0% of the articles met 7 to 10 of the controls, and no articles met all 11 of the controls.

Freeman et al. (2014) and 2015–2022 Sampled Articles Combined

Tables 1 and 2 shows the results from each of the coded "basic article features" and "internal validity controls" for the combination of Freeman et al. (2014) and 2015–2022 sampled articles.

Basic Article Features

The codings for the basic article features are shown in Table 1.

Course Discipline. As shown in the table, the top five disciplines in order from most to least were engineering (18.1%), mathematics (15.8%), biology (15.4%), physics (14.6%), and chemistry (14.2%). Overall, 78.1% of all articles were conducted in one of these five disciplines. The least frequent course discipline was geology (1.5%).

Comparison Groups. The most frequent comparison was active learning and lecture (88.1%). Comparisons between two active learning approaches occurred

in just 1.2% of the articles. Other comparisons were made in 10.4% of the articles involving at least three comparison conditions.

Intervention Time Frame. Over 82% of the articles were conducted over a full semester/quarter. Only 17.3% of the articles were conducted over multiple weeks, one week, or one class.

Assessment Type. Researcher/teacher developed or test bank assessments were used in 46.2% of the articles. These assessments were used alone or in combination with other assessments in 73.1% of the articles. Grades and standardized/normed assessments were solely used in 9.6% and 7.3% of the articles, respectively. Grades were used alone or in combination with other assessments in 30.4% of the articles. Standardized/normed assessments were used alone or in combination with other assessments in 21.9% of the articles. Failure rates were solely used in 3.1% of the articles and used in combination with other assessments in 14.2% of the articles. A combination of assessments occurred in 33.5% of the articles.

Results: Group That Resulted in the Highest Performance. The experimental group was found to produce statistically significantly greater improvements in academic performance in 32.7% of the articles and greater performance in 11.5% of the articles but without statistically significant differences. The control group was found to produce statistically significantly greater improvements in academic performance in 0.4% of the articles and greater performance in 2.7% of the articles but without statistically significant differences. Mixed results were found in 20.4% of the articles depending on the dependent measure reported. Inferential statistics were not used in 16.9% of the articles; of these articles, 93.2% showed results favoring the experimental group, 4.5% showed results favoring the control group, and 2.3% showed results favoring neither the experimental nor control group (i.e., no differences between them).

We used multiple codes to indicate combinations of results; most favored the experimental group. Therefore, overall, 64.6% of the articles showed the experimental group (i.e., active learning) outperform the control group (i.e., lecture), and 3.8% of the articles showed only the control group to be more effective than the experimental condition(s).

Majority of Student Preference. Student preference was reported favoring the experimental condition in 24.6% of the articles and the control condition in 2.3% of articles. Student preference was equal for both conditions in 6.5% of the articles. Finally, student preference was not reported in 64.6% of the articles.

Internal Validity Controls

The codings for the internal validity controls are shown in Table 2.

Research Design. A quasi-experimental design (i.e., when different years were compared or when existing classrooms were used) was used in 88.1% of the articles. Randomized control designs were used in only 9.6% of the articles. Within-subjects designs were used in 1.9% of the articles.

Group Equivalence for Both/All Groups: Directly Related to Dependent Variable. A group equivalence measure directly related to the dependent variable was not included in the majority (i.e., 66.9%) of the articles. Almost 26% of the

articles provided such a measure that showed the groups to be equivalent, and 5.4% of the articles showed the groups to be nonequivalent on such a measure.

Group Equivalence for Both/All Groups: Indirectly Related to Dependent Variable. Over 11% of the articles reported at least one group equivalence measure indirectly related to the dependent variable that showed the groups to be equivalent, and only 2.3% of the articles that reported a group equivalence measure indirectly related to the dependent variable showed the groups to be nonequivalent. However, a group equivalence measure indirectly related to the dependent variable was not included in 53.5% of the articles.

For the Freeman et al. (2014) and 2015–2022 sampled articles, several different directly and indirectly related assessments were used to determine equivalency. For example, articles in both sets reported assessments that demonstrated the groups were equivalent on skills common to what was being taught such as pretests that were the same or similar to posttests, subject matter or professional skills tests, and initial examination scores (before the active learning approach was provided) in a series of examinations. In addition, articles in both sets reported assessments that demonstrated the groups were equivalent on measures that were not directly related to what was being taught such as SAT or ACT scores, GPAs, and prerequisite courses taken (possibly with grades). Demographic information presented in both sets included courses taken in high school or college, race/ethnicity, gender, and year in college/age.

Attrition Equivalency. Information on the level of attrition, if any, was absent in 57.7% of the articles. There was attrition that occurred between the pretest and posttest of greater than 30.0% in 13.8% of the articles. There was no attrition or attrition was less than 30% in 28.5% of the articles.

Of the articles that did have attrition greater than 30.0% (13.8%), the attrition was differential in 75.0% of these articles and not specified in 25.0% of these articles. Overall (i.e., across all articles), differential attrition between/among groups occurred in almost 16% of the articles, did not occur 18.8% of the articles, and was not reported in 65.4% of the articles.

Matched Sample Size. The groups were comparable in size (i.e., not different by more than 25%) in 43.8% of the articles; however, they were not comparable in size in 42.3% of the articles. Group sizes were not reported in 11.9% of the articles.

Matched Instructor. It was not possible to determine if the instructor was the same between/among groups in 21.9% of the articles. However, the same instructor was used for the experimental and control groups in 36.9% of the articles but not in 39.6% of the articles.

Operational Definition of Variables Provided. All features of the experimental variable(s) were operationally defined in 50.4% of the articles and were partially operationally defined in 41.5% of the articles. The experimental features were not operationally defined in 8.1% of the articles. For the control condition, all features of the control variable(s) were operationally defined in 21.5% of the articles and were partially operationally defined in 33.1% of the articles. Control features were not operationally defined in 45.4% of the articles.

Dosage of Variables Described and Matched. The dosage of the experimental variable(s) was specified in 35.4% of the articles and was partially specified in

31.9% of the articles. The dosage of the experimental variable(s) was not specified in 32.7% of the articles. The dosage of the control variable(s) weas specified in 28.5% of the articles and was partially specified in 27.3% of the articles. The dosage of the control variable(s) was not specified in 44.2% of the articles.

When comparing the dosage between/among groups, the dosage was equal in 26.2% of the articles and was not equal in 11.2% of the articles. A comparison between/among groups could not be made in 62.7% of the articles.

Matched Implementation Time Frame. Approximately the same percentage of articles had the same implementation time frame (45.8%) as with a different implementation time frame (46.2%). Groups were implemented during the same semester but during different weeks in 1.5% of the articles, and 2.7% of the articles had a combination of implementation times. Finally, the implementation time frame was not specified in 3.8% of the articles.

Matched Content. The same content was taught across groups in 58.8% of the articles. The same content was not taught across groups in 10.8% of the articles. It was not possible to determine if the content matched across groups in 30.0% of the articles.

Dependent Measure(s): Equal for All Groups. The vast majority of articles reported that the dependent measure(s) was equal for both/all groups (74.2%) or was partially equal (i.e., some but not all of the measures were equal) (6.2%). However, 8.8% of the articles reported that the dependent measure(s) was not equal. It was not possible to determine if the dependent measure(s) was equal in 9.6% of the articles.

Implementation Fidelity. Implementation fidelity for the experimental group occurred in 6.9% of the articles but did not occur in 93.1% of the articles. Implementation fidelity for the control condition occurred in 6.2% of the articles but did not occur in 93.8% of the articles.

Rigor Across All Internal Validity Controls. Only 1.9% of the articles (i.e., Basili & Sanford, 1991; Bilgin, 2006; Carmichael, 2009; Lape et al., 2016; Randolph, 1992) met each of the 11 internal validity controls; this analysis (a) excluded both implementation fidelity and measures indirectly related to the dependent variable (if they already had measures directly related to the dependent variable that showed equivalence); (b) allowed for partial ratings on operationally defined variables and dosage of variables; and (c) gave articles the benefit of the doubt when not enough information or no information was provided for an internal validity control (i.e., was coded as "not specified"). If articles that were coded as "not specified" were counted as not meeting the internal validity control, only 0.4% (i.e., Randolph, 1992) would have met all controls (excluding implementation fidelity). If implementation fidelity were included in the analysis, no articles would have met all of the controls. Therefore, 100% of the articles had at least one internal validity control issue that could interfere with the effects of the independent variable.

Figure 2 shows the percentage of articles that fully met one or more of the internal validity controls. Note: credit was provided for having a measure that was either directly or indirectly related to the dependent variable that showed the groups were equivalent; therefore, there were 11 as opposed to 12 controls. The majority (62.7%) of the articles met half or fewer of the controls, only 22.7% of the articles met 7 to 10 of the controls, and no articles met all 11 of the controls.

Discussion

A concern in the USA and in other countries is STEM fatigue (or STEM attrition) wherein students are either dropping out of STEM fields or are not entering them during college (Martella & Demmig-Adams, 2018). To a large extent, the continued success of a nation is dependent on highly educated citizens, especially those in STEM fields (National Science Board, 2010). As a result, there have been calls for changing the manner in which we educate individuals who enter these professions (Martella & Demmig-Adams, 2018; Wieman, 2012). This instructional change involves moving from the "traditional" method of lecture to an active learning-based approach.

An issue with the term "active learning" is that there is not a unified definition on what it involves (Lombardi et al., 2021), other than the reduction or elimination of lecture. A problem ensues when an approach is defined by what it is not versus what it is. In addition, the overwhelming majority of active learning conditions in the science education literature have been found to devote at least 20% of class time or 30 min per week to lecture (Martella et al., 2021a). In addition to the issue of a missing operational definition for active learning, there is not a unified definition of what is meant by the "traditional lecture method" (Zakrajsek, 2018). Despite these issues, active learning continues to be viewed as not only a more effective approach than lecture but an even more ethical choice (Wieman, 2014).

One reason for the emphasis on adopting active learning in STEM courses is the findings in the research literature (e.g., Freeman et al., 2014) that show active learning (however it is defined) to outperform more "traditional" methods (whatever those might be). In our analysis of the articles reviewed by Freeman et al. (2014), 64.8% of the articles showed that active learning was more effective than lecture while only 3.4% of the articles showed lecture to be more effective than active learning. Similarly, in our updated sample of 2015–2022 articles, 64.3% of the sampled articles showed that the active learning groups outperformed the lecture groups while only 3.6% of the sampled articles showed the opposite effect.

In addition to performance outcomes, some researchers have started to examine student preferences for active learning or lecture, although these studies are not as numerous to date. These outcomes relate to a second reason for the emphasis on active learning, which seems to be that active learning is preferred by students over more "traditional" methods, although this conclusion seems to be in doubt by active learning advocates such as Deslauriers et al. (2019). Of the articles in Freeman et al. (2014) that had a direct comparison of preference, 73.8% of the articles showed student preference for the active learning condition(s). However, only 27.3% of *all* Freeman et al. articles showed this preference so this outcome remains less studied than learning outcomes. For the 2015–2022 sampled articles that did directly assess the comparison of preference, 59.3% showed student preference for the active learning conditions(s). However, only 19.0% of *all* sampled 2015-2022 articles showed this preference.

It is important to note that these types of performance and preference comparisons reflect a false dichotomy given that active learning and lecture are not

necessarily devoid of each other. In fact, most active learning interventions contain some aspect of lecture (Martella et al., 2021a, b, Zakrajsek, 2018). In Freeman et al. (2014), active learning courses could involve up to 90% of the class period spent on lecture and still be deemed an active learning course and lecture courses could involve just under 10% of the class period spent on active learning and still be deemed a lecture course.

Although the evidence seems clear based on the overwhelming number of articles showing the effectiveness of active learning in all variations and dosages, we must determine the quality of the literature base to determine the degree of confidence we can have that the independent variable was responsible for the results or that the conclusions made in these articles were just. This type of analysis is not uncommon in other fields where there is a range of articles that meet professional quality control standards such as those set by WWC. However, we know of no assessment of the methodological rigor of articles in the active learning STEM literature. It seems that such an analysis is critical given the focus on and push for active learning in STEM courses across university campuses. In the present analysis, there were 12 internal validity controls by which we assessed the articles included in the Freeman et al. (2014) meta-analysis as well as in a sample of the 2015–2022 articles that could be included in an updated meta-analysis.

Although we assessed each article against the 12 internal validity controls, one must not assume that just because one article has fewer controls present than another that it is overall weaker. We view all 12 controls as important considerations in an article; however, it is the severity of one or more specific controls that may threaten cause-and-effect claims. For example, an article may have all internal validity controls in place with the exception of matched groups. This one internal validity flaw could be severe enough to threaten internal validity claims. On the other hand, if an article has all controls in place with the exception of a matched instructor, internal validity may not be threatened if it was shown that the different instructors did not introduce bias into their instruction (this may be required even if the instructor was the same for each group). Thus, the presented data should be viewed as an indication of where there are internal validity flaws in the literature versus the severity of each flaw which can only be determined on a study-by-study basis. The following presents a discussion based on the findings for each internal validity control.

Research Designs and Group Equivalence

The gold standard of research designs is the randomized control group design (Martella et al., 2013). These are considered gold standard because they allow for the control of many potential confounding variables and allow for a high degree of internal validity or experimental control. The vast majority of articles did not use a randomized group design but used a quasi-experimental design. Of particular importance is ensuring group equivalence when a quasi-experimental design is used, preferably on a measure directly related to the dependent variable. If the groups are not equal on critical variables, valid comparisons cannot be made due to a selection confound (Martella et al., 2013) given that the groups are made up of different

types of participants (Cook & Campbell, 1979). Unfortunately, most articles that used a quasi-experimental design did not establish that the groups were equivalent on a measure directly related to the dependent variable. This was also the case even for measures indirectly related to the dependent variable.

It is not surprising that the majority of articles were conducted using quasi-experimental designs given the applied nature of the instructional implementations. However, it is critical for researchers to demonstrate that the groups are comparable on key variables. At a minimum, groups should be given an assessment that is related to the dependent variable such as a pretest or a parallel measure to show the groups are at similar skill levels before instruction is provided. In addition, indirect measures should also be provided to show equality on variables of interest based on research questions such as age, gender, and/or race/ethnicity.

Attrition

The loss of participants in a study is problematic, especially if this loss is differential between/among groups (Martella et al., 2013). If there is a loss of participants, it is important to show that the amount of loss was equal between/among groups and that the participants who left the study were essentially the same for both/all groups. If these conditions are not met, a confound of mortality is introduced which makes valid comparisons between/among groups difficult if not impossible because each group may be composed of different types of participants (Cook & Campbell, 1979). Unfortunately, the majority of articles did not provide information on attrition in general and differential attrition in particular.

At a minimum, researchers should explicitly document if and how much attrition occurred. Even if attrition did not occur, this should be stated or shown in the number of participants who began the study and how may finished or the number of students whose data were used compared to the number of students who were initially in the study. In addition, if attrition occurred, researchers should determine if this was differential between groups. In other words, even if there was a loss of participants in one or more groups, we need to determine whether the participants who did not finish the study were different from those who did. If there were no or minimal differences, there likely would not be a concern. However, if the attrition was differential, an internal validity concern would be present.

Sample Size Equivalence

If groups are not equal in size (i.e., by more than 25%), comparisons become more difficult because class size may affect the effectiveness of groups discussions, lecture engagement, responding to instructor questions, etc., leading to the introduction of a potential confound based on the impact of the size of classes (Theobald et al., 2020). For all articles combined, over half of all articles reported group size differences or did not report group size information to be able to determine if the groups were equal in size.

Researchers should make every attempt to compare groups that are equivalent in size while also reporting the size of each group. If the groups are not approximately equal in size, an explanation should be provided on if and how differences in group sizes may have affected the effects of the instruction while also taking these differences into account in the interpretation and discussion of the results.

Instructor of Groups

Another potential confound is if there were differences in who taught the classes compared in the article. One instructor may be more or less effective than another instructor for a variety of reasons such as excitement about the subject matter, the manner in which students are interacted with and provided with feedback, organization, pacing, etc. If the groups had different instructors, the comparison is not only on the type of instruction provided but who provided it, including instructor ability or experience (Theobald et al., 2020). Preferably, the instructor for both/all groups would be the same to remove the instructor as an alternative explanation for the results. Unfortunately, the majority of articles reported having different instructors or did not include enough information to determine if the same instructor taught both/all groups.

A complicating issue relating to the instructor is that having the same instructor for both or all groups may still introduce a potential confound. It is unlikely that a singular instructor would not be blind to the instructional methods used in each class and would not have a preference for one instructional approach over another. Therefore, it is important to point out that having the same instructor for both (all) groups may remove one confound (e.g., skill level of different instructors) but could add another one (e.g., bias for or against one instructional method over another one). Therefore, researchers should either use the same instructor for both/all groups who was not influenced by personal bias or expectations or provide information that the instructors were equal or similar on all critical variables related to teaching. This issue is also dependent on another control feature that will be discussed below, namely, implementation fidelity. Researchers can demonstrate that bias or expectations did not affect the manner in how one provided instruction by presenting information (i.e., data) that instruction for the different groups was provided as described.

Operational Definition of Experimental and Control Variables

A critical aspect of any study is the provision of an operational definition of the independent variable (Klahr, 2013). Without an operational definition of the critical variable, it is difficult to compare groups given that it is unclear, specifically and comprehensively, how the groups differed from one another with regard to what they received (Martella et al., 2020). This is an issue for a study because knowing why one group outperformed (or did not outperform) the other will be difficult to deduce. It is also difficult, if not impossible, to replicate articles without having the procedures of the interventions described in detail. Fortunately, all features of the

experimental variable(s) were operationally defined or partially defined in most of the articles.

Despite this positive finding for the literature base, operational definitions for control variable(s) were much less common. Nearly half of all articles did not provide an operational definition for the control variable(s). Even though most of the articles had control conditions that involved lecture, there were differences in how lectures were given in several articles. Some lectures only involved lecturing to the students with no interactions while others involved active learning components such as clicker questions, class discussions, and group work. Note that in the Freeman et al. (2014) meta-analysis, active learning could contain up to 90% lecture with 10% active learning activities. In other words, lecture classes could involve just less than 10% active learning components and still be considered lecture classes. Thus, just indicating that the comparison group used lecture or traditional methods tells us nothing about what occurred during those classes.

Researchers should provide operational definitions of what occurred in each group whether that was termed active learning or lecture/traditional methods. For example, researches could code what occurred during class (e.g., lecture, discussion, group activities) while also tracking the amount of time spent on each activity. Without such definitions, all we can conclude in a study is that what is labeled lecture/traditional instruction is less effective than what is labeled active learning without knowing what specifically occurred in each.

Dosage of Experimental and Control Variables

Similar to the issue of operational definitions and treatment fidelity (see below), it is critical to determine the dosage level of each instructional approach and ensure that the dosage of instruction students receive is similar because additional instruction time can increase the amount learned (Anderson et al., 2016). Dosage refers to how much time students spend with class content (both outside of class and inside of class) (i.e., the number of opportunities to respond to class content; Mason & Smith, 2020). This is an important issue given that if one group has a larger dosage (i.e., spends more time on class content either inside or outside of class), a head-to-head comparison cannot be made between instructional methods. It is not possible to know if differences between groups was due to the differential effects of instructional approaches or due to the time students spent engaged with class material. Perhaps most importantly, controlling the dosage between/among conditions is critical in ensuring the independent variable caused the results and not a difference in time-on-task/exposure to class content. The majority of articles either did not report the dosage of the experimental variable(s) or reported partial dosage information. More importantly, it was found that most articles did not have the same dosage between/among groups.

Researchers must ensure that the amount of time spent with class material is equal and the only difference between the groups is the type of instruction received. This can be accomplished by documenting the amount of work required outside of class, keeping the time in class equal or making up for differences of in-class time

with additional activities such as watching videos, and/or having students document the amount of time they spent on class assignments and attendance.

Time of Implementation

Implementing the experimental and control groups during the same time period is a control procedure, and it relates to history effects (Campbell & Stanley, 1963). History effects are events that occur during a study that can affect the outcome (Cook & Campbell, 1979). If groups receive the experimental and control variables at the same time, it is less likely that an extraneous event will differentially affect one group (Campbell & Stanley, 1963; Martella et al., 2013). The majority of articles had groups that received instruction at different times such as different years or semesters. For example, in Burnham et al. (2017), the time difference between the two groups was 5 years (2007 vs. 2012). The difficulty with this implementation time-frame discrepancy is that the students in the groups can be quite different as can the course, the technology, etc. If one were to compare the college campus before and after Covid or even compared the college campus a year ago to the current year where there is an increased awareness of diversity, equity, and inclusion, significant differences may be seen that likely affected the learning environment in one way or another. Thus, comparisons are difficult to make across time periods.

Researchers should implement groups during the same time period to allow for a valid comparison of groups based on time of implementation. If implementing the groups at the same time is not possible, there should be information on if and how there were differences in societal and educational environments that may have affected the results.

Content Taught

Valid comparisons cannot be made between groups if the same content was not taught to the different groups because the results may be based on differences in the content taught as opposed to how it was taught—this issue was discussed by Deslauriers et al. (2019). If the experimental group is taught content that differs from what the control group is taught, valid comparisons cannot be made on an assessment of their learning. Over one-third of the articles in our analysis either indicated the same content was not taught or did not specify if the same content was taught.

Researchers should demonstrate that the same content was taught in all groups to ensure a fair comparison of learning gains. One way to demonstrate this content would be to provide and/or compare course descriptions and objectives. Professional standards may also be shown that are addressed in each course. At a minimum, there should be a statement that the same content was taught in each course.

Dependent Measure(s)

Equality of measures is important since student performance must be measured in the same way and with the same assessment, otherwise a valid comparison cannot

be made (Martella et al., 2013) and would be considered an instrumentation threat (Campbell & Stanley, 1963). Nearly one-fourth of the articles had measures that were not or may not have been the same—either the measure(s) was not equal, were partially equal (i.e., components of the measure were the same and others were different), or had a lack of information on the equality of the measure(s).

Researchers should ensure that the measures used to determine student outcomes are equal. This can be achieved by providing the same measures or parallel forms of the measure. Either way, researchers should state explicitly that the same or similar measures were used to compare groups.

Implementation Fidelity

A critical aspect of any study that includes cause-and-effect claims or that includes claims of external validity is the assessment of implementation fidelity (Capin et al., 2018). Implementation fidelity allows one to assess if the interventions were implemented as intended and/or described (Sanetti et al., 2021). Unfortunately, fidelity of the experimental group and control group implementation occurred in less than 1 in 10 articles. As stated previously, implementation fidelity can aid in removing instructor effect as an alternative explanation for group differences. We believe this assessment should be a standard in all educational research that compares instructional methods. Without it, claims as to the effectiveness of one group over another are difficult to validate given that there is a lack of evidence that one or both of the groups were actually implemented as intended and described. In addition, it is difficult, if not impossible, to generalize the results of an investigation if there is a lack of information on how the instructional methods were implemented.

Researchers should take steps to ensure that fidelity of implementation occurs for all groups compared in a study. This can be done in several ways. For example, published implementation fidelity forms or checklists can be used or forms can be constructed based on idiosyncratic aspects of an instructional approach (Marchand-Martella & Lignugaris Kraft, 1997). These forms can be used to document if the various aspects of instruction were implemented. Direct observation and recording of various aspects of the different instructional approaches can be utilized in a similar manner as the direct observational recording of the dependent variable(s) (Lane et al., 2004). Finally, recordings of classes during instruction can be made and coded at a later time to document the various aspects of instruction that occurred (Gresham et al., 2000). (See Capin et al., 2018, for guidelines on enhancing treatment fidelity.)

Final Note About Internal Validity Controls

Freeman et al. (2014) included criteria to weed out articles that had certain internal validity control issues. For example, they reported that they excluded articles that did not have examination equivalence, student equivalence, or instructor equivalence. However, in our coding of these articles, we still found issues with these internal validity controls. Furthermore, we expanded our review to include additional internal validity controls that are important for rigorous educational research.

No articles in Freeman et al. met each of the 12 internal validity controls (if implementation fidelity were included). If we were to conduct a new meta-analysis using the 2015–2022 sampled articles, there would be no articles (based on our random sample) included in our meta-analysis as 0% of the articles met each internal validity control (if implementation fidelity were included).

We realize that it is unrealistic to expect articles to meet all internal validity controls. These articles are in applied settings. Decisions are made for legitimate reasons that may create a methodological flaw (Martella et al., 2013) such as comparing a current year's class to one or more classes from previous years. However, researchers should attempt to implement as many of the controls as are feasible. For those controls that are not feasible, researchers should provide an indication that these flaws were taken into consideration in the interpretation of the results and the conclusions that were reached based on the results.

Limitations

There are six limitations to mention surrounding the procedures used in this systematic review. First, one of the Freeman et al. (2014) articles representing two studies could not be located and, thus, was not included in this review. An additional 10 articles representing 15 studies were removed for various reasons (described previously). However, given that 94.1% of the articles (representing 92.4% of the studies) were included in the current review, it is unlikely the results would have changed to any meaningful degree with their inclusion.

Second, it may be seen as a limitation that articles rather than studies were coded. The decision was made to code entire articles so that each article could be weighted equally as opposed to weighting one article heavier than others if it included multiple studies. For example, it was possible for an article to include three studies; thus, rather than weighting such an article three times as much as another article, it was weighted the same but may have been subject to combination codes.

Third, we decided to include sampled articles beginning in 2015 rather than in 2010 which was the end of the Freeman et al. meta-analysis. This decision was made because one purpose of this review was to examine research published after the Freeman et al. meta-analysis had been published as a snapshot of the overall quality of the active learning research since its release.

Fourth, many changes have taken place since the initial release of the Freeman et al. (2014) meta-analysis. As such, we decided to widen our search strategy to include articles in medical fields given that active learning has become popular in medical science, STEM disciplines are foundational for public health, and medical science is now part of a new STEM acronym (i.e., STEMM) to reflect its importance, both for equity and for societal reasons. We also narrowed our search strategy by excluding psychology to align with our focus on improving college courses to prevent/reduce STEM attrition (as discussed by the National Center for Education Statistics [Chen, 2013]; see Appendix C in this report). This revised search strategy coupled with our analysis provides a more in-depth look into the quality of the current STEM active learning literature.

Fifth, all articles from 2015 to 2022 were not coded in our review. To reiterate, the purpose of conducting a random sample was to (a) gather a snapshot of the quality of the more recent active learning literature that could be included in a more recent meta-analysis on active learning and to (b) ensure the recency of the review given the time-intensive coding process (including double coding all articles) based on our specific internal validity control categories and the number of articles obtained from our search procedures. Given that the sampling was random in nature, it is likely that the sample was representative of all of the articles. However, as with random sampling, error can be introduced. In this current review, the sampling error was approximately 8.82%. However, given that the results of the 2015–2022 articles were consistent with the results of the Freeman et al. (2014) review, we have confidence in our sample of articles. Future research could include a larger sample or all articles to verify these results and build on the present review.

Sixth and finally, extensive steps were made to assess the level of agreement and, thus, the accuracy of the codings. The levels of agreement were above the acceptable levels; however, it is possible that others would code the articles differently than the research group on this project. Therefore, the results should be interpreted as based on our assessment of the articles that may or may not be representative of how other researchers would view or code the articles.

Moving Forward

Based on the present review, we outline three primary future research directions. First, we recommend active learning researchers design their studies to minimize methodological issues, specifically as they relate to internal validity controls. For any controls that are not able to be incorporated into the study design, we recommend explicitly pointing out these potential areas of weakness as study limitations such that results can be interpreted with this information in mind.

Second, we recommend future reviews be conducted on other areas of methodological controls such as statistical validity. The present review focused strictly on internal validity controls but statistical validity, for example, is also an important category for methodological rigor as it relates to the extent to which a certain level of confidence can be reached that the results of a study were due to a systematic variable (e.g., independent variable) as opposed to unsystematic variables (e.g., measurement error, sampling bias). In addition, other methodological factors may be considered in the future such as p-hacking (i.e., continuously analyzing data until a significant result emerges), including underpowered studies (i.e., not including a sufficiently large sample size to detect the effects of a systematic variable), and publication bias (i.e., only publishing studies that support a certain perspective).

Third, we recommend active learning researchers move beyond the active learning versus traditional lecture contrast and focus more on researching strategies we can implement in college classrooms to promote cognitive engagement such as self-explanations and retrieval practice (see Dunlosky et al., 2013 for a review of learning strategies). Researchers should also focus on investigating how lecture can be effectively integrated with participatory class activities (see discussion by Martella

& Schneider, in press). "Lecture or active learning" reflects a false dichotomy and an unneeded choice. As some researchers have noted, "there are still times when lectures will be needed" (Noah Finkelstein quoted in Bajak, 2014, para. 7) and it is "a matter of *both*, not one or the other" (Opdal, 2021, p. 16). Therefore, moving away from the constructivist teaching fallacy (Mayer, 2004) where being behaviorally active is equated with being cognitively active, we need to refocus our efforts on studying how to design learning environments that cultivate cognitive engagement.

Conclusion

The active learning literature contains a number of internal validity control issues that need to be addressed if we are to determine the extent to which active learning interventions are effective and if there are any boundary conditions for when particular active learning interventions are or are not effective. Consider that half of the articles conducted in both databases (i.e., Freeman et al., 2014 and 2015–2022 sampled articles) compared groups from different years, with the control group almost always reflecting the prior year(s). It is possible that biases can be introduced in articles when instructors have the goal of improving their future course by changing multiple variables at once. Furthermore, the interventions are generally not observed to ensure fidelity of implementation, and groups often differ on multiple factors such as the instructor and the class size. Without controlling for confounding variables, including this bias, it is possible that the results obtained in these articles are not reflecting the true effect of active learning interventions (nor lecture interventions).

Unfortunately, we are still left with several questions that were not answered by the current literature base. First, we do not know what type of active learning methods work best or what comprises active learning (i.e., observable motoric behaviors or cognitive [thinking] behaviors). Second, we do not know if the quality of lectures (however that is defined) would change the outcomes of these articles. Third, we do not know when active learning approaches should be introduced in the instructional sequence. Finally, we do not know what the appropriate dosage levels are for both active learning and lecture components. In our review of 260 articles, we cannot answer any of these questions with the current literature base. It is our hope that future research will begin to address methodological control issues to gain a better understanding of the effects of different active learning interventions and move toward answering questions that will develop the construct of active learning and provide more practical implementation recommendations.

Funding

The first author has received support from the National Science Foundation Graduate Research Fellowship Program under grant number DGE-1842166 and the National Science Foundation STEM Education Postdoctoral Research Fellowship Program under grant number 2222208. 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-023-09826-1.

Data Availability Data are available upon request and the complete search strategy is available on the Open Science Framework at 10.17605/OSF.IO/GFT4B.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Al-Holou, N., Bilgutay, N. M., Corleto, C., Demel, J. T., Felder, R., Frair, K., Froyd, J., Hoit, M., Morgan, J., & Wells, D. L. (1999). First-year integrated curricula: Design alternatives and examples. *Journal of Engineering Education, 88*(4), 435–448. https://doi.org/10.1002/j.2168-9830.1999.tb00471.x

Anderson, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences, 113*, 7481–7484. https://doi.org/10.1073/pnas.1516686113

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Association of American Universities (AAU). (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

Avcu, R., & Avcu, S. (2022). The methodological quality of experimental STEM education articles published in scholarly journals from 2014 to 2020. *International Journal of Assessment Tools in Education, 9*(2), 290–318. https://doi.org/10.21449/ijate.946743

Bajak, A. (2014). Lectures aren't just boring, they're ineffective, too, study finds. *Science Insider.* https://www.science.org/content/article/lectures-arent-just-boring-theyre-ineffective-too-study-finds

Barbarick, K. A. (1998). Exam frequency in introductory soil science. *Journal of Natural Resources and Life Sciences Education, 27*(1), 55–58. https://doi.org/10.2134/jnrlse.1998.0055

Basili, P. A., & Sanford, J. P. (1991). Conceptual change strategies and cooperative group work in chemistry. *Journal of Research in Science Teaching, 28*(4), 293–304. https://doi.org/10.1002/tea.3660280403

Bilgin, I. (2006). Promoting pre-service elementary students' understanding of chemical equilibrium through discussions in small groups. *International Journal of Science and Mathematics Education, 4*, 467–484. https://doi.org/10.1007/s10763-005-9015-6

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom.* ASHE-ERIC Higher Education Report No. 1. 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

Bratt, E. L., & Moons, P. (2015). Forty years of quality-of-life research in congenital heart disease: Temporal trends in conceptual and methodological rigor. *International Journal of Cardiology, 15*(195), 1–6. https://doi.org/10.1016/j.ijcard.2015.05.070

Burnham, N. A., Kadam, S. V., & DeSilva, E. (2017). In-class use of clickers and clicker tests improve learning and enable instant feedback and retest via automated grading. *Physics Education, 52*(6), 1–7. https://doi.org/10.1088/1361-6552/aa8833

Burns, M. K., Klingbeil, D. A., Ysseldyke, J. E., & Petersen-Brown, S. (2012). Trends in methodological rigor in intervention research published in school psychology journals. *Psychology in Schools, 49*(9), 843–851. https://doi.org/10.1002/pits.21637

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Capin, P., Walker, M. A., Vaughn, S., & Wanzek, J. (2018). Examining how treatment fidelity is supported, measured, and reported in K-3 reading intervention research. *Educational Psychology Review, 30*, 885–919. https://doi.org/10.1007/s10648-017-9429-z

Carl Wieman Science Education Initiative. (n.d.). *Carl Wieman Science Education Initiative*. The University of British Columbia. https://cwsei.ubc.ca

Carmichael, J. (2009). Team-based learning enhances performance in introductory biology. *Journal of College Teaching, 38*(4), 54–61 https://eric.ed.gov/?id=EJ838347.

Center for STEM Learning. (2016). *TRESTLE mini seed grant proposals: Transforming education, supporting teaching and learning excellence*. University of Colorado Boulder https://www.colorado.edu/csl/sites/default/files/attached-files/trestle_rfp-_minigrant_0.pdf.

Centers for Disease Control and Prevention (CDC). (2021). *Public health in STEM education.* https://www.cdc.gov/stem/education/stem_in_public_health.html

Chasteen, S. (2023). *How can I set clear expectations, and motivate students, so that they engage in active learning?* PhysPort https://www.physport.org/recommendations/Entry.cfm?ID=101200

Chen, X. (2013). *STEM attrition: College students' paths into and out of STEM fields (NCES 2014-001)*. Department of Education. Washington, DC https://nces.ed.gov/pubs2014/2014001rev.pdf.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*(4), 220–234. https://doi.org/10.1177/0741932514557271

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand McNally.

Crider, A. (2004). "Hot Seat" questioning: A technique to promote and evaluate student dialogue. *Astronomy Education Review, 3*(2), 137–147. https://doi.org/10.3847/AER2004020

Davis, M., & Hult, R. E. (1997). Effects of writing summaries as a generative learning activity during note taking. *Teaching of Psychology, 24*(1), 47–49. https://doi.org/10.1177/009862839702400112

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, 116*(39), 19251–19257 www.pnas.org/cgi/doi/10.1073/pnas.1821936116.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Eyler, J. (2018). *"Active Learning" has become a buzzword (and why that matters).* Rice University Center for Teaching Excellence. https://cte.rice.edu/blog/2018/active-learning

Finch, M. (2022). Complexities of practitioner research: Seeking hallmarks of quality. *Impacting Education: Journal on Transforming Professional Practice, 7*(3), 1–10. https://doi.org/10.5195/ie.2022.256

Flickinger, M., Tuschke, A., Gruber-Muecke, T., & Fiedler, M. (2014). In search of rigor, relevance, and legitimacy: What drives the impact of publications? *Journal of Business Economics, 84*, 99–128. https://doi.org/10.1007/s11573-013-0692-2

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, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Garavan, T., McCarthy, A., Sheehan, M., Lai, Y., Saunders, M. N. K., Clarke, N., Carbery, R., & Shanahan, V. (2019). Measuring the organizational impact of training: The need for greater methodological rigor. *Human Resource Development Quarterly, 30*(3), 291–309. https://doi.org/10.1002/hrdq.21345

Gosser, D. K. (2011). The PLTL boost: A critical review of research. *The Journal of Peer-Led Team Learning, 14*(1), 2–14.

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.betterevaluation.org/sites/default/files/2022-12/systematic%20review%20for%20informing%20policy%20decisions.pdf.

Graham, R. B. (1999). Unannounced quizzes raise test scores selectively for mid-range students. *Teaching of Psychology, 26*(4), 271–273. https://doi.org/10.1207/S15328023TOP260406

Gresham, F., MacMillan, D. L., Beebe-Frankenberger, M. B., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research and Practice, 15*, 198–205 https://www.scinapse.io/papers/2077759688.

Haberyan, K. A. (2003). Do weekly quizzes improve student performance on general biology exams? *The American Biology Teacher, 65*(2), 110–114. https://doi.org/10.2307/4451449

Han, H., Youm, J., Tucker, C., Teal, C., Rougas, S., Oark, Y. S., Mooney, C., Hanson, J., & Berry, A. (2022). Research methodologies in health professions education publications: Breadth and rigor. *Academic Medicine, 97*(11S), S54–S62. https://doi.org/10.1097/ACM.0000000000004911

Hartikainen, S., Rintala, H., Pylväs, L., & Nokelainen, P. (2019). The concept of active learning and the measurement of learning outcomes: A review of research in engineering higher education. *Education Sciences, 9*(4), 1–19. https://doi.org/10.3390/educsci9040276

Kellum, K. K., Carr, J. E., & Dozier, C. L. (2001). Response-card instruction and student learning in a college classroom. *Teaching of Psychology, 28*(2), 101–104. https://doi.org/10.1207/S15328023TOP2802_06

Klahr, D. (2013). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences, 110*(3), 14075–14080. https://doi.org/10.1073/pnas.1212738110

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential but often forgotten component of school-based interventions. *Preventing School Failure: Alternative Education for Children and Youth, 48*(3), 36–44 https://www.tandfonline.com/doi/abs/10.3200/PSFL.48.3.36-43.

Lape, N. K., Levy, R., Yong, D. H., Hankel, N., & Eddy, R. (2016). *Probing the flipped classroom: Results of a controlled study of teaching and learning outcomes in undergraduate engineering and mathematics*. In *ASEE Annual Conference & Exposition* https://peer.asee.org/probing-the-flipped-classroom-results-of-a-controlled-study-of-teaching-and-learning-outcomes-in-undergraduate-engineering-and-mathematics.

Lazonder, A. W., & Janssen, N. (2022). Quotation accuracy in educational research articles. *Educational Research Review, 35*(2), 1–10. https://doi.org/10.1016/j.edurev.2021.100430

Lishinski, A., Good, J., Sands, P., & Yadav, A. (2016). Methodological rigor and theoretical foundations of CS education research. *International Computing Education Research Conference, 161–169*. https://doi.org/10.1145/2960310.2960328

Lombardi, D., Shipley, T. F., & Astronomy Team, Biology Team, Chemistry Team, Engineering Team, Geography Team, Geoscience Team, and Physics Team. (2021). The curious construct of active learning. *Psychological Science in the Public Interest, 22*(1), 8–43. https://doi.org/10.1177/1529100620973974

Lopresti, R. (2010). Citation accuracy in environmental science journals. *Scientometrics, 85*(3), 647–655. https://doi.org/10.1007/s11192-010-0293-6

Luo, M., Li, C. C., 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

Marchand-Martella, N. E., & Lignugaris Kraft, B. (1997). Reliability of observations done by cooperating teacher supervisors in a Direct Instruction practicum. *Effective School Practices, 16*(4), 46–57 https://www.nifdi.org/research/esp-archive/volume-16/375-effective-school-practices-vol-16-no-4-fall-1997/file.html.

Marquart, F. (2017). Methodological rigor in quantitative research. In J. Matthes, C. S, Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods*. https://doi.org/10.1002/9781118901731.iecrm0221

Marrs, K. A., & Novak, G. (2004). Just-in-time teaching in biology: Creating and active learning classroom using the internet. *Cell Biology Education, 3*(1), 49–61. https://doi.org/10.1187/cbe.03-11-0022

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 https://eric.ed.gov/?q=source%3A%22Journal+on+Excellence+in+College+Teaching%22&ff1=subCase+Studies&id=EJ1185678.

Martella, A. M., Klahr, D., & Li, W. (2020). The relative effectiveness of different active learning implementations in teaching elementary students how to design simple experiments. *Journal of Educational Psychology, 112*, 1582–1596. https://doi.org/10.1037/edu0000449

Martella, A. M., Lovett, M., & Ramsay, L. (2021a). Implementing active learning: A critical examination of sources of variation in active learning science courses. *Journal on Excellence in College Teaching, 32*(1), 67–96 https://files.eric.ed.gov/fulltext/EJ1310521.pdf.

Martella, A. M., & Schneider, D. W. (in press). A reflection on the current state of active learning research. *Journal on the Scholarship of Teaching and Learning*.

Martella, A. M., Yatcilla, J., Martella, R. C., Marchand-Martella, N. E., Karatas, T., Ozen, Z., Park, H., Simpson, A., & Karpicke, J. D. (2021b). Quotation accuracy matters: An examination of how an influential meta-analysis on active learning has been cited. *Review of Educational Research, 9*(2), 272–308. https://doi.org/10.3102/0034654321991228

Martella, R. C., Nelson, J. R., Morgan, R. L., & Marchand-Martella, N. E. (2013). *Understanding and interpreting educational research*. .

Mason, E. N., & Smith, R. A. (2020). Tracking intervention dosage to inform instructional decision making. *Intervention in School and Clinic, 56*(2), 92–98 https://files.eric.ed.gov/fulltext/EJ1271201.pdf.

Massachusetts Institute of Technology. (2021). *Technology-enhanced active learning*. MIT Press. https://web.mit.edu/edtech/casestudies/teal.html

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14

Mayer, R. E. (2011). *Applying the science of learning*. Pearson.

Mayer, R. E. (2022). *Multimedia learning* (3rd ed.). Cambridge University Press.

McCoy, L., Pettit, R. K., Kellar, C., & Morgan, C. (2018). Tracking active learning in the medical school curriculum: A learning-centered approach. *Journal of Medical Education and Curricular Development, 5*(2), 1–9. https://doi.org/10.1177/2382120518765135

National Academies of Sciences, Engineering, and Medicine. (2020). *Promising practices for addressing the underrepresentation of women in science, engineering, and medicine: Opening doors*. https://nap.nationalacademies.org/catalog/25585/promising-practices-for-addressing-the-underrepresentation-of-women-in-science-engineering-and-medicine

National Center for Education Statistics. (2010). *Detail for CIP Code 13.0607*. Institute of Education Sciences https://nces.ed.gov/ipeds/cipcode/cipdetail.aspx?y=55&cipid=89199.

National Science Board. (2010). *Preparing the next generation of STEM innovators: Identifying and developing our nation's human capital*. National Science Foundation. http://www.nsf.gov/nsb/publications/2010/nsb1033.pdf

Naveenkumar, N., Georgiou, G. K., Vieira, A. P. A., Romero, S., & Parrila, R. (2022). A systematic review on quality indicators of randomized control trial reading fluency intervention studies. *Reading & Writing Quarterly, 38*(4), 359–378. https://doi.org/10.1080/10573569.2021.1961647

Opdal, P. A. (2021). To do or to listen? Student active learning vs. the lecture. *Studies in Philosophy and Education, 1196*, 1–19. https://doi.org/10.1007/s11217-021-09796-3

Pennington, C. R., Jones, A., Bartlett, J. E., Copeland, A., & Shaw, D. J. (2021). Raising the bar: Improving methodological rigour in cognitive alcohol research. *Addiction, 116*(11), 3243–3251. https://doi.org/10.1111/add.15563

Pienta, N. J. (2015). Understanding our students in general chemistry. *Journal of Chemical Education, 92*, 963–964. https://doi.org/10.1021/acs.jchemed.5b00330

Ramirez, F. D., Motazedian, P., Jung, R. G., Di Santo, P., MacDonald, Z. D., Moreland, R., Simard, T., Clancy, A. A., Russo, J. J., Welch, V. A., Wells, G. A., & Hibbert, B. (2017). Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation. *Circulation Research, 120*(12), 1916–1926. https://doi.org/10.1161/CIRCRESAHA.117.310628

Randolph, J. J., Julnes, G., Bednarik, R., & Sutinen, E. (2007). A comparison of the methodological quality of articles in computer science education journals and conference Proceedings. *Computer Science Education, 17*(4), 263–274. https://doi.org/10.1080/08993400701483517

Randolph, W. M. (1992). *The effects of cooperative learning on academic achievement in introductory college biology* [unpublished doctoral dissertation]. Washington State University.

Rosas, S., & Kane, M. (2012). Quality and rigor of the concept mapping methodology: A pooled study analysis. *Evaluation and Program Planning, 35*(2), 236–245. https://doi.org/10.1016/j.evalprogplan.2011.10.003

Sanetti, L., Cook, B. G., & Cook, L. (2021). Treatment fidelity: What it is and why it matters. *Learning Disabilities Research and Practice, 36*(1), 5–11 https://onlinelibrary.wiley.com/doi/epdf/10.1111/ldrp.12238.

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., et al. (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.

Steele, J. E. (2003). Effect of essay-style lecture quizzes on student performance on anatomy and physiology exams. *Bioscene: Journal of College Biology Teaching, 29*(4), 15–20 https://c2ip.insa-toulouse.fr/_attachment/des-pedagogies-actives-article/Steele_2003.pdf?download=true.

Sulu, M. D., Martella, R. C., Aydin, O., Bolshakova, V. L. J., & Erden, E. (2023). A meta-analysis of science education studies for students with intellectual and developmental disabilities (IDD). *Journal of Developmental and Physical Disabilities.* https://doi.org/10.1007/s10882-023-09890-z

Sulu, M. D., Martella, R. C., Grimmet, K., Austin, A., & Erden, E. (2022). Investigating the effects of self-monitoring interventions with students with disabilities on the maintenance and generalization of on-task behavior: A systematic literature review. *Review Journal of Autism and Developmental Disorders, 10*, 458–476. https://doi.org/10.1007/s40489-022-00304-y

The White House. (2022), *Equity and excellence: A vision to transform and enhance the U.S. STEMM ecosystem.* https://www.whitehouse.gov/ostp/news-updates/2022/12/12/equity-and-excellence-a-vision-to-transform-and-enhance-the-u-s-stemm-ecosystem/

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., et al. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences, 117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

University of Georgia. (2022). *Quality enhancement plan: Active learning at UGA.* https://provost.uga.edu/oaie/accreditation/reaffirmation-2022/PDFS/QEP-Active-Learning-At-UGA.pdf

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, version 4.1.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Wieman, C. (2012). Applying new research to improve science education. *Issues in Science and Technology, 29*(1), 25–32 http://www.jstor.org/stable/43315691.

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

Yang, L. J. S., Chang, K. W. C., & Chung, K. C. (2012). Methodology rigor in clinical research. *Plastic and Reconstructive Surgery, 129*(6), 979e–988e. https://doi.org/10.1097/PRS.0b013e31824eccb7

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

Authors and Affiliations

Amedee Marchand Martella[1] · Ronald C. Martella[2] · Jane K. Yatcilla[3] · Alexandra Newson[4] · Eric N. Shannon[5] · Charissa Voorhis[5,6]

1 Department of Psychological & Brain Sciences, UC Santa Barbara, Santa Barbara, CA, USA

2 Department of Teaching & Learning, University of Colorado at Colorado Springs, Colorado Springs, CO, USA

3 Libraries and School of Information Sciences, Purdue University, West Lafayette, IN, USA

4 Department of Special Education and Clinical Sciences, University of Oregon, Eugene, OR, USA

5 Department of Educational Studies, Purdue University, West Lafayette, IN, USA

6 Department of Special Education, University of Missouri, Columbia, MO, USA