Journal of Behavioral Education (2022) 31:157–185
https://doi.org/10.1007/s10864-021-09452-3

ORIGINAL PAPER

A Systematic Review and Quantitative Analysis of Interteaching

Camilo Hurtado-Parrado[1] · Nicole Pfaller-Sadovsky[2] · Lucia Medina[3] · Catherine M. Gayman[4] · Kristen A. Rost[4] · Derek Schofill[4]

Accepted: 16 August 2021 / Published online: 27 September 2021

Abstract

Interteaching is a behavioral teaching method that departs from the traditional lecture format (Boyce & Hineline in BA 25:215–226, 2002). We updated and expanded previous interteaching reviews and conducted a meta-analysis on its effectiveness. Systematic searches identified 38 relevant studies spanning the years 2005–2018. The majority of these studies were conducted in undergraduate face-to-face courses. The most common independent variables were manipulations of the configuration of interteaching or comparisons to traditional-lecture format. The most common dependent variables were quiz or examination scores. Only 24% of all studies implemented at least five of the seven components of interteaching. Prep guides, discussions, record sheets, and frequent assessments were the most commonly implemented. Meta-analyses indicated that interteaching is more effective than traditional lecture or other control conditions, with an overall large effect size. Furthermore, variations in the configuration of the interteaching components do not seem to substantially limit its effectiveness, as long as the discussion component is included. Future research informed by the present review includes: (a) investigating the efficacy of interteaching in additional academic areas, online environments, workplace training, and continuing education, (b) testing alternative outcome measures, generalization, and procedural integrity, (c) conducting systematic component analyses, and (d) measuring social validity from the instructor's perspective.

Keywords Behavior analysis · Behavioral teaching methods · College education · Higher education · Interteaching · Meta-analysis

✉ Camilo Hurtado-Parrado
camilo.parrado@siu.edu

Extended author information available on the last page of the article

Introduction

Lecture-based methods continue to be the predominant college pedagogy (Stains et al., 2018). Research has demonstrated the relative inefficacy of lecture-based teaching at improving student learning, which has resulted in the search for superior teaching methods (Saville & Zinn, 2011). Early behavior analytic efforts (late 1960s and 1970s) in this direction involved the development of different approaches based on well-established behavioral principles (e.g., reinforcement, discrimination, generalization, and shaping; Moran & Malott, 2004). Overall, these early behavioral teaching methods (a) focused on restructuring the classroom environment for enhancing student learning and enjoyment (e.g., Programmed Instruction Holland & Skinner, 1961; Personalized System of Instruction Keller, 1968) and (b) produced better student learning outcomes compared to traditional methods (Moran & Malott, 2004).

Notwithstanding these promising findings, behavioral teaching methods failed to gain widespread popularity. A strong tradition of lecture-based teaching, and perhaps more importantly, the typical structure of most educational settings, likely limited the implementation of these methods. For instance, the original version of the Personalized System of Instruction (Keller, 1968), one of the most popular approaches, incorporated absolute self-paced student progress, which does not fit well into traditional semester-based courses (however, see some alternative developments by Pear et al., 2011). Additionally, early behavioral teaching methods were often initially time-consuming to prepare (Boyce & Hineline, 2002), and academics frequently struggle to balance the many obligations contending for their time (e.g., administrative duties, service, research, and mentorship). It is common for academics to be released from some portion of their responsibilities (e.g., course releases) when undertaking large projects in service or research. However, such allowances are rarely given to instructors for making substantial changes in classroom pedagogy. These obstacles may explain why instructors would not be willing or able to transition to behavioral teaching methods.

Interteaching

Boyce and Hineline (2002) introduced interteaching as an approach that could address the limitations that previously hindered the implementation of behavioral teaching methods. The authors described seven components of interteaching: (1) preparation guides (or prep guides; sets of 10–15 questions of varying complexity), (2) in-class discussions between two or more students, (3) record sheets (students list the prep guide items they would like the instructor to expand/clarify during the next lecture), (4) brief clarifying lectures (less than half of the session focused on difficult topics), (5) reinforcement contingencies for discussion/prep guide completion, (6) frequent assessments, and (7) quality points (a cooperative contingency aimed at improving discussion quality).

The growing interest in interteaching since its inception has resulted in a continued increase in studies describing the application of a wide range of variations in its components and tests of its effectiveness. Reviews of this literature (Querol et al., 2015; Saville et al., 2011b; Sturmey et al., 2015) have generally reported that (a) interteaching contributes to student success (e.g., in terms of quizzes, homework, in-class participation, cumulative final examinations, long-term recognition memory) when compared with traditional lecture-based instruction; (b) interteaching has been successfully implemented across a wide range of academic disciplines (e.g., political science, engineering, business, nutrition, special education, psychology) and class formats (e.g., classes that differ in frequency, duration, size, and media format, such as face-to-face, online, and blended classes); (c) interteaching has been successfully implemented in a variety of higher-education settings (i.e., both inside and outside of the USA and in undergraduate and graduate courses); (d) social validity measures indicate that students often rate interteaching favorably and prefer it over traditional lecture; and (e) substantial efforts have been dedicated to testing a wide range of variations of the components of interteaching (e.g., impact of quality points, effect of discussions, scheduling of clarifying lectures, discussion-group size). Notwithstanding this supporting evidence on the versatility and efficacy of interteaching, the most recent reviews noted limitations of the identified studies and recommended several areas for future research (Querol et al., 2015; Sturmey et al., 2015). These included the systematic replication of laboratory and applied studies in work training settings, and across other modes of classroom delivery (e.g., online, hybrid courses) and populations (e.g., different age-groups and academic institutions).

Purpose and Overview of the Present Study

In addition to the specific limitations and areas for future research noted above, we have identified factors that provide impetus for a re-examination of the interteaching literature as a whole. First, a survey of the literature indicates the absence of systematic quantitative analyses (i.e., meta-analyses) of the efficacy of interteaching across all the available literature. Second, recent reviews (Querol et al., 2015; Sturmey et al., 2015) included interteaching studies published through 2014, and a cursory review of literature published after this date indicates a substantial increase in relevant research in recent years. Thus, an updated literature review and meta-analysis would serve to fill these gaps. Accordingly, the purpose of the present study was twofold: update and expand previous reviews on interteaching and conduct a meta-analytic review of its effectiveness.

To accomplish our purpose, we aimed to address the following research objectives: (1) complete systematic searches for interteaching research across a wide range of academic databases, including those holding dissertations and theses; (2) evaluate empirical studies across various methodological and outcome variables (e.g., research design, use of interteaching components, procedural integrity measures); (3) conduct a quantitative synthesis of observed effects by computing overall mean effect sizes (i.e., Hedges' g); and (4) conduct moderator analysis to identify variables that modulate the efficacy of interteaching (e.g., class size, number of examinations/quizzes, contingencies for completion of preparation guides, interteaching frequency).

Method

Literature Search and Study Selection Process

The search procedures for relevant records followed the recommendations of Petticrew and Roberts (2006) for conducting systematic literature reviews and meta-analyses and complied with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati et al., 2009; Moher et al., 2009).

Searches using the keyword *interteaching* were conducted in the following academic databases (conducted in March 2019; repeated in January 2020 and produced no new records): EBSCO, ERIC, MEDLINE, PsycARTICLES, PsycINFO, SCOPUS, and Web of Science. As recommended by Higgins et al. (2020) and Petticrew and Roberts (2006), to minimize the risk of bias, including publication bias (the issue that statistically significant studies are more likely to be published than those with nonsignificant outcomes; Higgins et al., 2020), additional searches using the same keywords were conducted in the following databases specialized in theses and dissertations (first conducted in April 2019; repeated in January 2020 and produced no new records): EBSCO Open Dissertations, ProQuest Dissertations & Theses Global, PQDT Open, and British Library EThOS, and DART-Europe E-theses Portal. Supplementary ancestry searches (i.e., an examination of reference lists) were conducted on the most recent interteaching reviews (Querol et al., 2015; Sturmey et al., 2015), and all the manuscripts that were published since 2014, which was the last year covered by those reviews. No restrictions regarding publication dates were used during these searches. A total of 159 records were identified through database searching and 24 through ancestry search (Supplementary Fig. 1 shows the flow of information through the different stages of the search, screening, and selection procedures, based on PRISMA guidelines; Liberati et al., 2009; Moher et al., 2009).

After duplicate records were removed, the last author screened the abstracts of the remaining 79 records using the following exclusion criteria: (a) language different than English, Spanish, or German; (b) theoretical/review manuscripts or book chapters with no original (previously unpublished) data; (c) non-peer-reviewed publications (except dissertations and theses); (d) conference abstracts; (e) records with incomplete information for which no full text was possible to obtain; (f) dissertations/theses that resulted in articles published elsewhere in peer-reviewed journals; and (g) studies that did not report the implementation of at least one interteaching component.

The full texts (pdfs) of the 46 records that were not excluded during the first stage of screening were obtained. A second round of screening was conducted on these full texts using the same exclusion criteria, which resulted in the exclusion of eight additional records. During a quality-control test of the screening process, two authors (first and last) independently screened the 46 records. The agreement score was computed by number of agreements divided by the number of agreements plus the number of disagreements multiplied by 100. The two authors reached 98% agreement on record exclusion. Disagreements were resolved by re-examining the records and reaching a consensus. Finally, 38 full-text records were deemed eligible for further analysis (Supplementary Fig. 2 shows the distribution of these references over time).

Coding of Studies and Data Extraction

A specially designed Microsoft® Excel™ matrix was used for the coding process (a copy is available at Open Science Framework [OSF] repository https://osf.io/ejxrq/). Based on the information reported in earlier reviews on interteaching (Querol et al., 2015; Saville et al., 2011b; Sturmey et al., 2015), 56 variables were selected and defined for coding of the 38 records that were identified. The complete list of variables, definitions, and codes are available in a supplemental Excel file available in the OSF repository described above. Variables were classified in the following categories: (a) bibliographic information (e.g., title, authors, date, publication, etc.); (b) participant information (e.g., gender, sample size, student major); (c) instructor information (e.g., teaching assistants or tutors); (d) study methodology (e.g., design, dependent variables, independent variables, class size); (e) details of the interteaching components (e.g., contents and length of the prep guides, scheduling of the lectures, duration and facilitation of the discussions, number of quizzes/examinations/probes); (f) student ratings on different aspects of interteaching (i.e., preference for interteaching or traditional lecture, perceived enjoyment of discussions, ratings of knowledge acquired); (g) study outcomes (e.g., effect sizes on quiz/examination/probe scores); (h) details of procedural integrity (present or absent and percentage reported by authors); and (i) statistical information (e.g., means, standard deviations, *p*-values). If effect sizes were not reported, they were computed, when possible, using data provided in the sources.

To ensure the reliability of data extraction, the second author independently coded eleven of the records (29%). These articles were selected via a random function in a Microsoft® Excel™ application. The inter-coder agreement (ICA) score was determined by the number of agreements divided by the number of agreements plus the number of disagreements multiplied by 100. This computation was done for each of the 56 variables coded, which yielded an ICA of 97.9%.

Assessing the methodological quality of eligible studies (critical appraisal of a study and the extent to which authors reported their research to the highest possible standard) is recommended for systematic reviews and meta-analyses (Dreier, 2013; Kratochwill & Levin, 2014; Wendt & Miller, 2012). The second and third authors conducted a quality assessment of all the selected records by using instruments developed for single-case experimental methods (SCMs; Logan et al., 2008) and group designs (Downs & Black, 1998). Quality assessment showed that the majority of SCMs and group design studies were classified as moderate to strong, as per the corresponding scales (for group designs, *moderate* = 33–66.9%, *strong* ≥ 70%, Logan et al., 2008; for SCMs, *moderate* = 50–70%, *strong* > 70%, Downs & Black, 1998). Accordingly, information on quality assessment was not further used in the present study. An ICA score for quality assessments was calculated by adding the number of agreements between the coders, divided by the number of agreements plus the number of disagreements multiplied by 100. An overall 94% ICA for quality assessments was found across all research designs (91% for group designs and 96% for SCMs). Disagreements were resolved by re-examining the records and reaching a consensus.

Meta-statistical Analyses

Out of the 38 records identified, 24 studies reported sufficient information (i.e., design, sample size, means, and standard deviations were at least required) for processing of the meta-statistical analyses. Three research designs were included (between-groups designs, within-subjects designs, and SCMs). Intervention versus control comparisons (eleven between-groups designs) examined at least two groups and compared each groups' mean change measures (Germain et al., 2018). In the intervention versus control comparisons, the control groups were either true controls without exposure to any interventions (e.g., "Participants in the control condition had no exposure to the information contained in the article. Rather, they reported only once to the laboratory and took the quiz"; Saville et al., 2005, p. 162) or were alternative interventions, such as attending traditional lectures or completing anagrams (e.g., Saville et al., 2014). Pre- versus post-comparisons (two within-subjects designs) either investigated a single group comparing a change in outcome measures (e.g., examination scores; Felderman, 2014) or compared pre- versus post-changes in two groups, including a control group (e.g., interteaching versus traditional lecture on examination scores; Slezak & Faas, 2017). The eleven studies that reported implementation of SCMs (e.g., alternating-treatment designs; Felderman, 2016; Gayman et al., 2018) were included in the between-groups meta-analyses in a similar approach to that reported by Zelinsky and Shadish (2018) because the data reported in them were already aggregated per courses, groups and/or sections (i.e., no individual/single-case data were available).

A comparison across records was feasible if there was a minimum of three records for a given type of research design. Data were then combined across records based on the type of intervention used (e.g., comparison of traditional lecture to interteaching or comparisons among variations of interteaching), and their respective outcome measures (e.g., examinations, probes, or quiz scores). If a given record reported several outcome measures (e.g., quizzes, student satisfaction, number of assignments completed), the primary outcome measure was selected based on the most complete information.

Due to across-record differences (e.g., differences in the composition of participants or variation in the interventions used; Borenstein et al., 2009), it was assumed that the true effect size varied from study to study. Therefore, a random effects model was applied (Borenstein et al., 2015), with efficacy as the primary outcome variable. Application of the random effects model also allowed for an analysis of potential moderator variables (e.g., class size or number of participants in the discussions; Higgins & Green, 2011). However, Borenstein et al. (2009) pointed out that a small number of records, as was the case in the current analysis, are a limitation for conducting moderator analyses and should thus be interpreted with caution.

Data Analysis and Overall Effect Size

The Comprehensive Meta-Analysis© software Version 3.3 (Biostat, 2019) was used to analyze extracted data and to compute a standardized mean difference (SMD), namely Hedges' g. The SMD Hedges' *g* includes a correction factor for bias, resulting in more accurate and conservative estimates when sample sizes are small (Borenstein et al., 2009; Lipsey & Wilson, 2001). The computer software used was specifically designed to calculate an effect size and other statistical information (e.g., confidence intervals [CI] and *p*-values) for each of the included records and weight them to provide an overall mean effect size (Borenstein et al., 2015). For the current analysis, Hedges' *g* estimates were computed according to the different study designs used (e.g., intervention versus control group, if any) and were weighted with respect to sample size. Hedges' *g* can be interpreted using the following guidelines proposed by Lipsey and Wilson (2001): small (<0.49), medium (0.50–0.79), and large (≥ 0.80). Weighted mean effect size computations were calculated for all outcome measures (e.g., examinations, probes and quiz scores, homework submission).

Between-Study Variation or Heterogeneity

The second step in the analysis was to assess the heterogeneity of studies (i.e., between-study variation calculating Q test, I^2, and T^2) by using the Comprehensive Meta-Analysis© software (Biostat, 2019). Heterogeneity assessments show the degree to which each study's effect size varies within the distribution of effect sizes (Borenstein et al., 2009).

The Q test provides information about whether the included studies show unaccounted variance, and if specific covariates moderate the effect, assuming a random error (Germain et al., 2018). The Q statistic was calculated as the weighted sum of squared differences between each study's effects and the pooled effect across studies (Cochran, 1954). If the Q test yields a statistically significant result ($p<.05$), the included studies do not share a common effect size (Borenstein et al., 2009) and are said to show statistical heterogeneity that cannot be accounted for by sampling error (Ahn & Kang, 2018; Littell et al., 2008). The I^2 statistic is an index to quantify the dispersion of effect sizes within a meta-analysis. It reports the proportion of the observed variance reflecting the variation in true effect sizes rather than sampling error (Borenstein et al., 2017). General benchmarks to guide the interpretation of I^2 are as follows: I^2 values of $<30\%$ are considered small, values of approximately 50% are average, and values $>75\%$ indicate high levels of heterogeneity (Ahn & Kang, 2018; Higgins, & Green, 2011). However, Borenstein et al. (2017) note that I^2 is ideally reported in combination with forest plots and the T^2 statistic to provide the reader with maximum information about heterogeneity and the true effects. Hence, T^2 was calculated, which is interpreted as an estimate of the actual variance of the true effect sizes across the population of studies (Borenstein et al., 2009, 2010).

Moderator Analyses

The next component of data analysis encompassed moderator analyses, which aim to assess the relationship between study-level covariates and effect size (Borenstein et al., 2009). Put differently, moderator analyses evaluate whether study or participant characteristics (i.e., moderator variables, such as the number of examinations, probes, or quizzes or student level) modified the effectiveness of interteaching. A meta-regression model was applied based on the moderator variables of interest. As recommended by Deeks et al. (2019), the variables tested in the model were selected prior to the analysis based on their assumed relevance for the effectiveness of the interventions. Such relevance was assumed based on the components put forward by Boyce and Hineline (2002) and the findings reported and discussed in earlier systematic reviews (Querol et al., 2015; Saville et al., 2011b; Sturmey et al., 2015). The unit of analysis was the individual study, as well as variables that had data from at least four studies (e.g., class size; Germain et al., 2018; Higgins & Green, 2011).

Publication Bias

Publication bias refers to the fact that studies with statistically significant results are more likely to be published, cited, and reprinted than those with nonsignificant outcomes (Borenstein et al., 2009; Littell et al., 2008; Petticrew & Roberts, 2006). In an attempt to address the issue of publication bias, specific databases and repositories were searched (e.g., ERIC and ProQuest Dissertations & Theses Global) and "published" (i.e., peer-reviewed) and "unpublished" records (i.e., "documents that were not independently edited or were not refereed"; White, 2019, p. 61) were included if relevant. For example, one unpublished doctoral dissertation (Gutierrez, 2017) and four unpublished masters' theses (e.g., Bethke, 2016) were eligible for inclusion in the systematic review. However, two of them were not eligible for inclusion in the meta-statistical analysis due to lack of suitable data (e.g., no detailed statistical information was available; Gutierrez, 2017; Wright & Wright, 2011).

A funnel plot was produced and visually inspected to assess the presence of publication bias. A funnel plot is a scatterplot of the estimated effect size versus standard error or sample sizes of the studies (Higgins & Green, 2011; White, 2019). If publication bias exists, the funnel plot will be skewed to one side (i.e., asymmetrical with a gap in a bottom corner of the plot), a pattern that is identifiable with visual inspection (Borenstein, 2005; Higgins & Green, 2011). The funnel plot can be a helpful initial visual diagnostic tool for assessing publication bias (White, 2019). However, funnel plots should be interpreted with caution. Meta-analyses consisting of a representative number of studies reporting statistically significant results and large sample sizes may produce a funnel plot that appears symmetrical, indicating that publication bias is not present (Keenan, 2016). Thus, to aid visual analysis, a linear regression of the effect sizes and the standard errors of individual studies was calculated (i.e., Egger's regression; Egger et al., 1997). A statistically significant regression ($p < .05$), together with an asymmetrical funnel plot, provides strong evidence of publication bias (Sterne et al., 2000). Egger's regression method has been shown to be more powerful than other methods (e.g., rank correlation) in detecting publication bias in meta-analyses, especially those comprising less than 30 studies, as was the case in the present effort (Sterne et al., 2000).

Results

Descriptive Findings

Study Characteristics

The systematic search yielded 38 eligible records with publication dates ranging between 2005 and 2018. Twenty-one (55%) records were published from 2014 to 2018 (see Supplementary Fig. 2). The majority of records were peer-reviewed articles (87%), and the remaining were theses and dissertations (13%). No other gray literature, such as book chapters, conference proceedings, or government reports, was eligible (e.g., excluded because records were not peer-reviewed, or no original data were provided). Nearly half (48%) of the identified peer-reviewed articles were published in one of the following journals: *Journal of the Scholarship of Teaching and Learning* (*n* = 6), *Teaching of Psychology* (*n* = 6), and *Journal of Applied Behavior Analysis* (*n* = 4).

Supplementary Table 1 shows the details of each record (e.g., year, country, study duration, independent and dependent variables, design, media, etc.) using a corresponding DocID number. Using the approach described by Lipsey and Wilson (2001) and Littell and Corcoran (2010), each record's information was extracted and labeled with a subID depending on the design implemented. Specifically, subIDs were generated on the basis of the comparisons implemented (e.g., interteaching versus traditional lecture, baseline versus interteaching), and independent and dependent variables analyzed (e.g., social validity, examination scores, quizzes). Accordingly, a single manuscript could have resulted in several subIDs, depending on the complexity of the design and independent and dependent variables analyzed. For instance, data extracted from DocID no. 32 (Rehfeldt et al., 2010) were further classified in four subIDs that resulted from the cross section of the two levels of the independent variable (points contingent on prep guides or no points) and the two dependent variables measured in the study (percentage of assignments submitted and quiz scores; see Supplementary Table 1). With this approach, the 38 identified studies were subcategorized into 136 subIDs (i.e., 136 separate combinations corresponding to 136 individual rows in Supplementary Table 1). Except for results related to research designs, independent and dependent variables, and components of IT, descriptive and meta-statistic results were based on the information extracted at the subID level (Lipsey, 2019; Lipsey & Wilson, 2001; Littell & Corcoran, 2010). The descriptive results focused on aspects of interteaching (e.g., characteristics of each of the components of interteaching or student perception of interteaching) excluded subIDs related to control and traditional lecture independent variables.

Participants and Settings

The majority of the studies (92%) were conducted in North America (the USA and Canada), with the remaining 8% of studies distributed across different countries, including Australia, Colombia, Norway, and Vietnam (see Supplementary Table 1). Only 57.7% of the records provided the ages of the participating students. Of these records, 51.8% reported participation of students 20–26 years old, and the remaining reported an equal distribution for younger (2.9%) or older (2.9%) samples. The majority of the studies (69%) reported that more than half of the participants were female (see Supplementary Table 1). Information provided in several records (7%) allowed the identification of mixed female and male groups, but no clear proportions were possible to establish. In the remaining 24% of the records, no information on the gender of participants was available. Regarding collegiate level, 83.9% of all the records reported undergraduate samples, and 16.1% reported graduate samples (see Supplementary Table 2).

All the studies identified for the present review were conducted in higher-education settings. Only 55.5% provided information on student majors (academic discipline to which undergraduate students formally commit). The most frequently reported majors were in the areas of social science (24.1%; e.g., psychology or social work; Arntzen & Hoium, 2010; Felderman, 2014, 2016; Rehfeldt et al., 2010), education (13.1%; e.g., Cannella-Malone et al., 2009; Mason, 2012; Rieken et al., 2018), and health (10.9%; e.g., nursing; Byrne & Guy, 2016; Goto & Schneider, 2010; Rosales et al., 2018; Soldner et al., 2015). Graduate-level studies were only conducted with participants majoring in areas of social science (5.1%) and education (10.9%; see Supplementary Table 3).

With regard to course media (i.e., face-to-face, online asynchronous [students engage with the course content at different times and from different locations], online synchronous [the instructor and the students engage with the course content and each other at the same time, but from different locations], or blended, the majority of undergraduate-level studies were conducted in face-to-face (77.4%) and online asynchronous (3.6%) courses. For studies conducted in graduate-level courses, the most common medium was face-to-face (8%), followed by online asynchronous (5.8%). Online synchronous was the least common medium (2.2% at the graduate-level and no undergraduate-level studies; see Supplementary Fig. 3).

The majority of studies (71.5%) were conducted in courses covering social science topics (e.g., psychology of learning, social welfare; Arntzen & Hoium, 2010; Gayman et al., 2018; Truelove et al., 2013), followed by a small percentage of courses covering natural science topics (4.4%, e.g., anatomy and human physiology; Byrne & Guy, 2016; Mercer, 2014) and engineering (2.9%, e.g., biomedical engineering; Cezeaux & Keyser, 2018). Thus far, interteaching-related research at the graduate level has been conducted primarily in social science courses (see Supplementary Fig. 3).

Research Designs

An analysis of the research designs reported across the identified records showed that SCMs and group designs were implemented with similar frequency (49.6% and 48.2%, respectively; see Supplementary Table 4). Case study was the least frequently reported research method (2.2%). Among the SCM studies, only multielement/alternating treatments designs were implemented. For the studies that implemented group-comparison analyses, quasi-experimental without control and nonequivalent-groups designs were the most frequently reported.

Independent and Dependent Variables

The most frequently reported independent variables were manipulations of interteaching and traditional lectures (e.g., presence or absence of interteaching, comparison across or with other teaching strategies; see Supplementary Table 5 for details), and analyses of outcomes based on student grade point averages (GPA). Manipulations of specific components of interteaching were identified throughout the eligible studies (e.g., contents of the prep guides, points contingent on completion of the prep guides, number of examinations/quizzes/assessments, presence or absence of discussion component); however, each type of manipulation had few associated records (less than 10%). The most frequently reported dependent variables were quiz or examination scores, and assessments of social validity (significance of intervention goals, acceptability of intervention procedures, and social importance of effects). Although various other measures of student performance were also reported (e.g., laboratory reports, article reviews, projects, and assignments), the number of associated records for each of these other measures was low (below 5%). It is worth noting that studies frequently manipulated and/or measured multiple independent and dependent variables.

Components of Interteaching

An analysis of the studies that explicitly reported using each of the seven main components of interteaching, as defined by Boyce and Hineline (2002), was conducted. Studies in which it was not possible to determine the use of a given component were excluded from this analysis (i.e., were considered unclear). Prep guides and discussions were the most consistently implemented components (92.1%), followed by clarifying lectures (89.5%), use of record sheets (86.8%), and frequent assessment (81.6%; e.g., quizzes, probes, examinations). Among the less frequently implemented components were scheduling contingencies for the completion of prep guides or discussions (60.5%; e.g., points contingent on submission of the prep guides) and quality points (28.9%). Twenty-four percent of all the identified studies reported implementing at least five of the seven main components of interteaching (see details in Supplementary Table 6).

Prep Guides Different aspects of the prep guide component were extracted from the identified records, namely, number of questions on each prep guide, availability (beginning of the course, during each session, or prior to each session) and completion contingencies (reinforcement or no contingency). The most frequently reported number of questions per prep guide was 10–20, and prep guides were typically available prior to each session for both undergraduate- and graduate-level classes (see Supplementary Table 7).

The majority of records (70.4%) did not clearly specify contingencies related to prep guide completion and/or submission (Supplementary Table 8). Of the 30% that provided information, the majority reported using a positive reinforcement contingency (25%; e.g., Soldner et al., 2015).

Discussion The following aspects related to the discussion component of interteaching were analyzed: contingency scheduled on the quality or completion of the discussion, size of the discussion groups, assignment of the students to the discussion groups (different members every session or members randomly assigned), discussion facilitator (e.g., instructor, teaching assistant), and discussion length.

Few studies (6.5%) provided explicit information on whether contingencies were applied based on the discussion quality (see Supplementary Table 9). In all relevant records, positive reinforcement was implemented. Similarly, all records that provided information on a contingency for discussion completion (43.5%) reported using positive reinforcement (e.g., Felderman, 2016; Saville et al., 2012a, 2012b; Soldner et al., 2015).

The most commonly reported discussion-group size in undergraduate and graduate courses was student pairs (57.4% of all eligible records; see Supplementary Table 10). Of the remaining records, 20.3% reported between three and six students per group, and in a few cases, discussion groups exceeded six students (10.2%).

The most common method used to assign discussion groups was establishing a different partner or group each session for both undergraduate (33.3%) and graduate courses (4.6%). Random assignment of students to groups was also used in some undergraduate courses (24.1%). The remaining 38% of the records did not provide clear information on how students were assigned to discussion groups (see Supplementary Table 11).

Different combinations of discussion facilitators (instructor, teaching assistant, or tutor) were identified in the records that reported related information (64.8%). In only 2.8% of all the records, researchers reported that facilitation was not implemented. The most frequent facilitator in undergraduate courses was the instructor alone (30.6%) or a combination of instructor and a teaching assistant/tutor (23.2%). For graduate courses, facilitation by only the teaching assistant or tutor (6.5%) was the most frequent strategy reported (see Supplementary Table 12).

Almost half of the records (42.6%) did not provide clear information on discussion length. Among the remaining 57.4% of the records, the most frequent discussion length was 20–30 min (20.3%), followed by less than 20 min (15.8%) and 30–40 min (10.2%) sessions (see details in Supplementary Table 13).

Quality Points The majority of studies (84.3%) either did not implement or did not report clear details on the delivery of quality points. In 8.3% of records, quality points were delivered if performance on quizzes or tests was equal to or above 80%. In a few records (7.4%), researchers reported delivery of quality points but did not provide specific performance criteria (see Supplementary Table 14).

Scheduling of Interteaching and Frequency of assessment (Probes/Quizzes/Tests)

Biweekly interteaching sessions were the most frequently reported (30.6%), followed by weekly sessions (27.8%). Notably, 32.4% of the records did not report the frequency of interteaching sessions (see Supplementary Table 15). Regarding frequency of assessment, only 31.5% of all records provided relevant information. The most commonly reported schedule was every 1 to 3 class meetings (29.6%). Regarding the type of evaluation, a combination of short-essay and multiple-choice questions, or multiple-choice questions only, were the most frequently reported methods of assessment (see details in Supplementary Table 16).

Record Sheets/Forms An analysis of the information that students were asked to include in the record sheets was conducted (e.g., difficult topics or prep guide questions, quality of the discussions). The information most frequently requested from students was a list of difficult topics and an assessment of discussion quality (50.9%; see details in Supplementary Table 17), followed by a list of difficult topics only (22.2%). In some studies, researchers reported using the record sheet or similar form but did not provide details about the specific information requested from students (14.8% of records). Lastly, 10.2% of the records did not provide clear information about record sheets.

Students' Perception of Interteaching

Only 49.1% of the eligible records provided information on students' reported preferences for interteaching compared to other teaching approaches (e.g., traditional lecture or related manipulations). In 34.3% of these records (30.6% undergraduate and 3.7% graduate), students indicated a preference for interteaching. In the remaining records (14.8%), all of which included undergraduate participants, students reported a preference for methods other than interteaching (see Supplementary Table 18).

Other measures of the students' perceptions of interteaching included the extent to which students reported that they (a) acquired more knowledge with interteaching, (b) learned most with interteaching, (c) better understood the materials with interteaching, and (d) perceived the discussions of interteaching positively. The most salient aspect across these measures is the lack of consistency of implementation and limited information provided across the eligible studies. (Detailed information is provided in Supplementary Table 19.) These factors resulted in large percentages of unclear for each measure (80% or higher) and a wide range of percentages of students across the categories. Ultimately, it is not possible to characterize the students' perceptions of interteaching across these measures with the limited data available.

Meta-Statistical Findings

Twenty-four out of 38 studies reported data suitable for further effect size computations. (These studies are highlighted with asterisks (*) in Supplementary Table 1 and Supplementary References.)

Meta-Analysis Results

Interteaching to Traditional Lecture The majority of studies ($n = 14$) compared traditional lecture to interteaching across all research designs. The outcome measures most commonly used to assess the effectiveness of interteaching were examination and quiz results, assignments (i.e., written essays and American Psychological Association assignments), and assessment of improvement (e.g., level of understanding of the subject taught). Figure 1 shows the details of these 14 studies, a forest plot, and effect size statistics for all the comparisons between traditional lecture (TL) and interteaching (IT). An overall large and significant effect size of interteaching when compared to traditional lecture or other control or experimental groups (e.g., reading) was found, $g = 0.814$, $p = 0.000$, 95% CI (0.491–1.137).

Study name		Outcome	Statistics for each study				
			Hedges's g	Standard error	Lower limit	Upper limit	p-Value
Cezeaux & Keyser (2018)	TL to IT	Exam results	-0,733	0,327	-1,373	-0,092	0,025
Felderman (2015)	TL to IT	Exam results	0,451	0,214	0,031	0,871	0,035
Filipiak et al. (2010)	TL to IT prep guides	Quiz results	0,662	0,298	0,078	1,246	0,026
Filipiak et al. (2010)	TL to IT prep guides R	Quiz results	0,837	0,303	0,244	1,430	0,006
Rieken et al. (2018)	TL to IT	Quiz results	0,373	0,152	0,076	0,670	0,014
Saville et al. (2014)	TL/Control to IT	Quiz results	2,558	0,242	2,083	3,033	0,000
Saville et al. (2005)	Control to IT	Quiz results	1,550	0,347	0,869	2,230	0,000
Saville et al. (2005)	Reading to IT	Quiz results	1,186	0,319	0,562	1,810	0,000
Saville et al. (2005)	TL to IT	Quiz results	0,893	0,323	0,261	1,526	0,006
Saville et al. (2006)	TL to IT	Quiz results	1,650	0,274	1,112	2,188	0,000
Scoboria & Pascual-Leone (2009)	TL to IT	Written assignment	0,365	0,159	0,053	0,676	0,022
Slezak & Faas (2017)	TL to IT	APA assessment	1,137	0,385	0,383	1,892	0,003
Zayac & Paulk (2014)	TL to IT	Exam results	-0,054	0,303	-0,648	0,539	0,857
Gayman et al. (2018)	TL to IT	Exam results	0,651	0,191	0,277	1,025	0,001
Soldner et al. (2015)	TL to IT	Quiz results	1,148	0,245	0,667	1,630	0,000
Anbro (2012)	TL to IT	Quiz results	0,763	0,294	0,186	1,340	0,010
Edwards (2005)	TL to IT	Improvements	0,504	0,194	0,124	0,884	0,009
			0,814	0,165	0,491	1,137	0,000

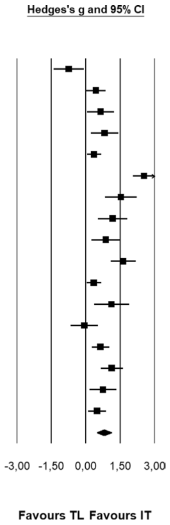
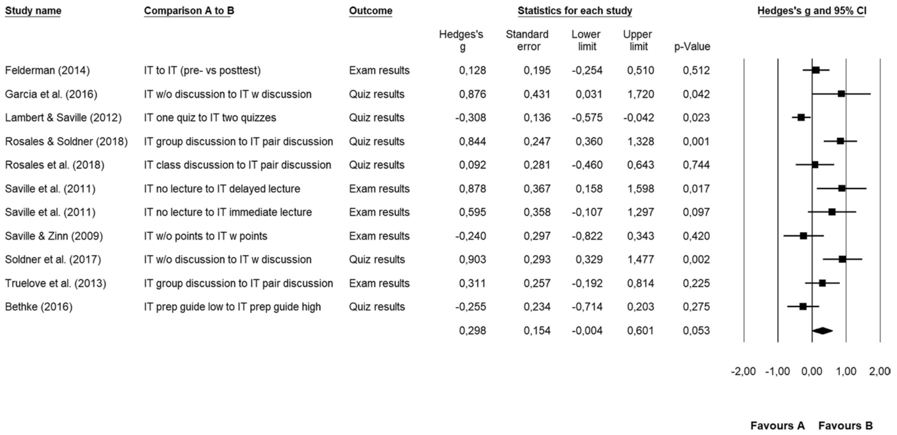


Fig. 1 Forest plot and effect size statistics for traditional lecture (TL)—interteaching (IT) comparisons. The diamond-shaped data point at the bottom of the plot represents the summary effect size across all studies

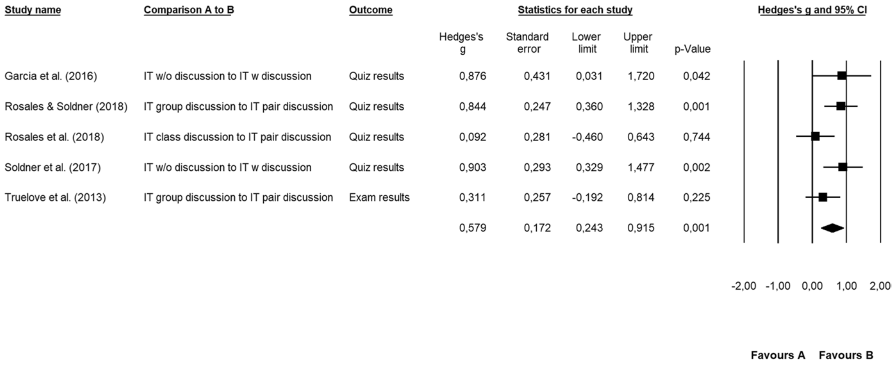
Three out of these 14 studies compared different variations of control and treatment groups. Filipiak et al. (2010) compared a traditional lecture format to interteaching with a reinforcement contingency for completion of the prep guides and interteaching without that completion contingency. When compared to traditional lecture, interteaching with the reinforcement contingency led to higher quiz scores and yielded a large effect size ($g=0.837$, $p=.006$, 95% CI [0.244–1.430]), whereas interteaching without the reinforcement contingency resulted in a medium effect ($g=0.662$, $p=.026$, 95% CI [0.078–1.246]). Saville et al. (2005) compared a control group (no intervention) to three experimental groups (i.e., traditional lecture, reading, and interteaching) in a simulated classroom. They found that interteaching lead to higher overall quiz scores, and comparisons of interteaching to control, reading, and traditional-lecture groups in all cases yielded large and significant effects (range of g's from 0.893 to 1.558, all p's $<.05$, range of 95% CI [0.261–2.230]). Finally, Saville et al. (2014) compared a control group (no intervention) and traditional lecture to interteaching during introductory psychology courses on quiz scores. The comparison resulted in a very large and significant effect size ($g=2.558$, $p=.00$, 95% CI [2.08–3.03]), favoring interteaching over traditional lecture and control conditions.

Interteaching to Interteaching Ten studies tested six different variations of interteaching (level of difficulty of prep guides, number of students in the discussion groups, number of quizzes, presence or absence of discussions, presence or absence of lectures, and presence or absence of quality points; Bethke, 2016; Garcia et al., 2016; Lambert & Saville, 2012; Rosales & Soldner, 2018; Saville et al., 2011a; Saville & Zinn, 2009) using two different outcome measures (i.e., results of examinations and quizzes). Figure 2 displays a forest plot and the effect-size analyses for these studies (Hedge's g, standard error, CI, and p-values). Overall, these different forms of interteaching-to-interteaching comparisons yielded a small and marginally signifi-



Study name	Comparison A to B	Outcome	Statistics for each study				
			Hedges's g	Standard error	Lower limit	Upper limit	p-Value
Felderman (2014)	IT to IT (pre- vs posttest)	Exam results	0,128	0,195	-0,254	0,510	0,512
Garcia et al. (2016)	IT w/o discussion to IT w discussion	Quiz results	0,876	0,431	0,031	1,720	0,042
Lambert & Saville (2012)	IT one quiz to IT two quizzes	Quiz results	-0,308	0,136	-0,575	-0,042	0,023
Rosales & Soldner (2018)	IT group discussion to IT pair discussion	Quiz results	0,844	0,247	0,360	1,328	0,001
Rosales et al. (2018)	IT class discussion to IT pair discussion	Quiz results	0,092	0,281	-0,460	0,643	0,744
Saville et al. (2011)	IT no lecture to IT delayed lecture	Exam results	0,878	0,367	0,158	1,598	0,017
Saville et al. (2011)	IT no lecture to IT immediate lecture	Exam results	0,595	0,358	-0,107	1,297	0,097
Saville & Zinn (2009)	IT w/o points to IT w points	Exam results	-0,240	0,297	-0,822	0,343	0,420
Soldner et al. (2017)	IT w/o discussion to IT w discussion	Quiz results	0,903	0,293	0,329	1,477	0,002
Truelove et al. (2013)	IT group discussion to IT pair discussion	Exam results	0,311	0,257	-0,192	0,814	0,225
Bethke (2016)	IT prep guide low to IT prep guide high	Quiz results	-0,255	0,234	-0,714	0,203	0,275
			0,298	0,154	-0,004	0,601	0,053

Fig. 2 Forest plot and effect size statistics for all interteaching-to-interteaching comparisons. The diamond-shaped data point at the bottom of the plot represents the summary effect size across all studies



Study name	Comparison A to B	Outcome	Statistics for each study				
			Hedges's g	Standard error	Lower limit	Upper limit	p-Value
Garcia et al. (2016)	IT w/o discussion to IT w discussion	Quiz results	0,876	0,431	0,031	1,720	0,042
Rosales & Soldner (2018)	IT group discussion to IT pair discussion	Quiz results	0,844	0,247	0,360	1,328	0,001
Rosales et al. (2018)	IT class discussion to IT pair discussion	Quiz results	0,092	0,281	-0,460	0,643	0,744
Soldner et al. (2017)	IT w/o discussion to IT w discussion	Quiz results	0,903	0,293	0,329	1,477	0,002
Truelove et al. (2013)	IT group discussion to IT pair discussion	Exam results	0,311	0,257	-0,192	0,814	0,225
			0,579	0,172	0,243	0,915	0,001

Fig. 3 Forest plot and effect size statistics for interteaching-to-interteaching comparisons featuring various types of discussions. The diamond-shaped data point at the bottom of the plot represents the summary effect size across all studies

cant summary effect size ($g = 0.298$, $p = .053$, 95% CI [-0.004 to 0.601]). However, this overall effect size should be interpreted with caution, as these studies differ considerably from each other (i.e., they investigated various and specific questions on the efficacy of different interteaching components).

Discussion Effects of interteaching with or without the discussion component and different discussion-group sizes (e.g., pairs or two to four participants per group) were investigated in more than three studies, which allowed for further analysis. Figure 3 shows a forest plot and effect size statistics for these interteaching-to-interteaching comparisons featuring various types of discussions. Variations of the discussion-group size yielded a moderate and significant summary effect size ($g = 0.579$, $p = 0.001$, 95% CI [0.243–0.915]), favoring pair discussions to produce better test scores. The two studies that compared the absence or presence of discussions yielded the largest effects ($g = 0.903$, $p = 0.002$, 95% CI [0.293–1.477]; ($g = 0.876$, $p = 0.042$, 95% CI [0.031–1.720]; Soldner et al., 2017; Garcia et al., 2016, respectively), favoring inclusion of the discussion component of interteaching to produce higher test scores.

Taken together, these results indicate that interteaching, regardless of different variations or configurations, is more effective than traditional lecture or other alternative control conditions. Furthermore, it appears that variations in the configuration of the different interteaching components do not substantially limit its effectiveness, as long as the discussion component is included.

Heterogeneity Assessment

Assessment of the heterogeneity of the 17 studies that compared traditional lecture to interteaching indicates that it was statistically significant ($Q = 121.01$ [$df = 16$], $p = .00$, $T^2 = 0.39$). The calculation of the proportion of heterogeneity (I^2) yielded

87%, confirming the initial assessment. Heterogeneity analysis of the 11 studies that compared variations of interteaching also indicates it was beyond expected by sampling error ($Q = 39.35$, [$df = 10$], $p = .00$, $T^2 = 0.18$), with I^2 yielding 75%. To summarize, heterogeneity across all eligible studies was pronounced, and studies did not seem to share a common effect size. Put differently, this seems to indicate that the effects found in the studies were not simply due to sampling error.

Table 1 Meta-regression statistics across 30 moderator variables

Moderator variables	Number of covariates	Q	p-value	R^2 (%)
Assignment of discussion groups	3	0.52	0.77	0
Availability of prep guides	3	6.52	0.038*	14
Clarifying lectures	2	0.01	0.94	0
Class size	6	8.22	0.145	14
Content of probes/examinations/tests	3	5.77	0.056	3
Contingency on discussion/prep guide	2	4.51	0.033*	7
Details of quality points	3	3.76	0.15	11
Details of record sheets/forms	5	2.5	0.645	0
Facilitation	5	4.93	0.29	0
Frequency of probes	2	5.21	0.022*	9
Instructor type	5	1.2	0.878	0
Introductory lectures	3	0.28	0.871	0
IT session frequency	4	9.71	0.021*	18
Key components of interteaching	2	0.03	0.87	0
Lecture details	4	1.31	0.727	0
Length of discussions	6	4.24	0.51	4
Media	3	0.47	0.79	0
Number of examinations/probes/quizzes	4	11.14	0.011*	27
Number of instructors	3	0.02	0.99	0
Preparation guides	2	3.15	0.076	0
Quality of discussions	2	0.51	0.47	0
Quality of prep guide questions	3	2.27	0.321	0
Quality points	2	0.25	0.613	0
Record sheets	2	0.8	0.37	0
Research designs	2	0.09	0.76	0
Several examinations/probes/quizzes	2	5.82	0.016*	16
Size of discussion groups	3	3.12	0.21	0
Student level	2	0.12	0.726	0
Students' preference for IT	2	2.3	0.13	32
Type of questions in probes/quizzes	3	7.58	0.023*	14

See Supplementary Fig. 4 for individual scatterplots of each of these significant regressions

*Indicates a statistically significant result ($p < .05$)

Moderator Analysis

Meta-regressions were calculated for traditional lecture to interteaching comparisons only, as the interteaching-to-interteaching comparisons already addressed research questions about the effects of specific interteaching components at the individual study level (e.g., Felderman, 2014; Querol et al., 2015; Truelove et al., 2013; Fig. 3). Table 1 shows the thirty variables (e.g., types of discussion, preparation guides, or quality points) that were tested as potential moderators for the efficacy of interteaching when compared to traditional lecture or control groups. The number of covariates, and *Q, p,* and R^2 values for each regression are also included in Table 1.

Seven statistically significant regressions were identified, as indicated by asterisks in the *p*-value column of Table 1 (see Supplementary Fig. 4 for individual scatterplots of each of these significant regressions). The variable *availability of prep guides* (students had access to the assigned prep guides prior or during the corresponding class meeting/session) was related to the effectiveness of interteaching. Specifically, making prep guides first accessible to students during the corresponding session resulted in a higher effect size ($p < .05$). *Interteaching session frequency* moderated the effectiveness of interteaching, with weekly interteaching meetings having a higher effect size ($p < .05$). The variable *type of question* during the examinations/probes (short essay, multiple choice, etc.) was also associated with a larger effect, with multiple-choice and/or true/false questions having higher effects ($p < .05$). Similarly, variables related to number and frequency of assessments (i.e., *several examinations/probes/quizzes*, *frequency of examinations/probes/quizzes,* and *number of examinations/probes/quizzes*) also seem to influence effectiveness. Conducting multiple examinations, probes, or quizzes and administering them at every or during every third meeting, at least, seem to increase the effectiveness of interteaching ($p < .05$).

Lastly, *contingency on discussion or prep guide completion* was associated with the effectiveness of interteaching ($p < .05$). Studies in which no contingencies were implemented (e.g., Edwards, 2005; Saville et al., 2014; Saville et al., 2005; Scoboria & Pascual-Leone, 2009; Slezak & Faas, 2017) had larger effect sizes than those that had contingencies in place (Anbro, 2015; Cezeaux & Keyser, 2018; Felderman, 2016; Rieken et al., 2018; Soldner et al., 2015; Zayac & Paulk, 2014). However, these findings should be interpreted carefully because a closer inspection of the records related to this regression showed that one of them reported a somewhat disproportionate effect size (Saville et al., 2014). When the regression was recalculated excluding this potential outlier ($g = 2.56$, $p = .00$, 95% CI [2.08–3.03]), the result of the meta-regression was not significant (number of covariates = 2, $Q = 3.81$, $p = .051$, $R^2 = 3\%$). This issue aligns with previous reports that very large effects sizes tend to excessively impact meta-regressions and moderator analyses (Lipsey & Wilson, 2001).

Publication Bias

A funnel plot was created to visually analyze and, if present, detect publication bias among all eligible studies (see Fig. 4). Inspection of the funnel plot found an

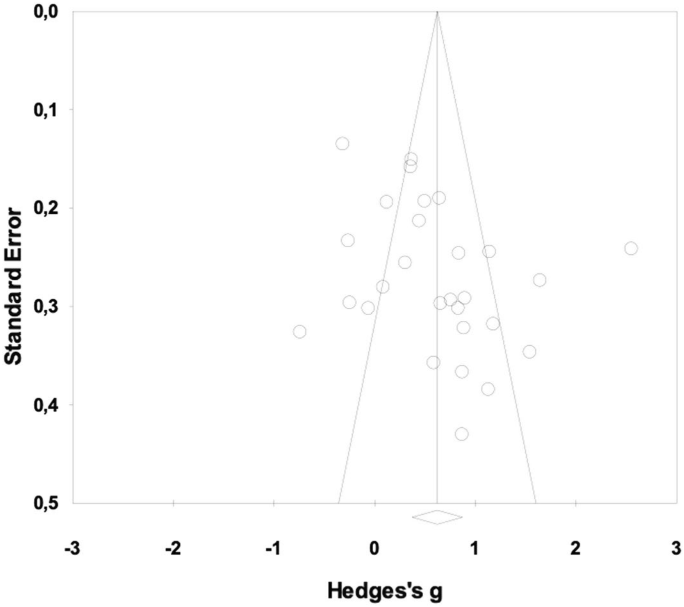


Fig. 4 Funnel plot depicting the distribution of all eligible studies (open circles)

asymmetrical pattern (i.e., no data points at the left corner of the bottom of the plot), which indicated the presence of publication bias. This particular pattern of the plot indicates an underrepresentation of small sample-size studies that yielded small effects. In contrast, large sample studies tend to be published more frequently due to increased statistical power (Lipsey & Wilson, 2001). To test this interpretation, an Egger's regression (i.e., a linear regression between the standard error of all eligible studies and Hedges' *g*; Egger et al., 1997) was computed. This regression was statistically significant ($p < .05$), confirming the presence of publication bias.

Only 24 of the identified studies reported data suitable for meta-statistical analyses. The majority compared traditional lecture to interteaching using different outcome measures (e.g., examination or quiz results, written assignments). These studies showed an overall large and significant effect size of interteaching.

Ten studies tested six different variations of interteaching (e.g., number of students in the discussion groups, number of quizzes, presence, or absence of discussions). Overall, these comparisons yielded a small and marginally significant summary effect size; however, this finding should be interpreted with caution, as these studies differ considerably from each other.

Analyses focused on the discussion component showed that variations of the discussion-group size yielded a moderate and significant summary effect size. Furthermore, a comparison between absence and presence of discussions yielded a large significant effect size.

Seven of the different variables that were tested as potential moderators for the efficacy of interteaching were statistically significant. Making the prep guides first accessible to students during the corresponding session, scheduling weekly interteaching meetings, and implementing multiple-choice and/or true/false questions were associated with higher effects. Conducting multiple examinations, probes or quizzes and administering them at every or during every third meeting, at most, was also related to higher interteaching effectiveness. Although lack of a contingency for completion of discussions or prep guides was related to larger effect sizes, this finding should be interpreted with caution because of the disproportionate effect size of one of the records included in the analysis (potential outlier).

Discussion

The purpose of the present study was twofold: (a) update and expand previous reviews on interteaching and (b) conduct a meta-analytic review of its effectiveness. The systematic search yielded 38 records that met the inclusion criteria. Twenty-four of these records reported sufficient data for inclusion in the meta-analysis. The systematic review identified empirical evaluations of interteaching published between 2005 and 2018. The majority of these studies were published in the latter four years (2014–2018) of this range. It seems possible that this uptick in research was related, among others, to the publication of systematic reviews on interteaching (Querol et al., 2015; Saville et al., 2011b; Sturmey et al., 2015).

All identified studies were conducted in higher-education settings, and most took place at North American undergraduate-level social science classes that involved at least some face-to-face instruction. Although this pattern seems somewhat expected considering the origins of interteaching (e.g., behavior analysis and psychology) and its recent inception (less than 20 years), promising recent changes include implementation in more diverse areas of knowledge (e.g., natural sciences and engineering; Byrne & Guy, 2016; Cezeaux & Keyser, 2018) and evaluation of interteaching in asynchronous (i.e., Gayman et al., 2018; Rieken et al., 2018) or synchronous (i.e., Soldner et al., 2017) online settings.

Evaluations of the efficacy of interteaching have most often compared interteaching to traditional lecture, and several studies also considered student GPAs in their analyses. Quiz or examination scores and assessments of social validity (e.g., students' preferences for interteaching compared to traditional lecture) were common dependent measures. Nearly half of the studies reported social validity measures, in which the majority reported that students preferred interteaching to other experimental conditions (e.g., traditional lecture). Lastly, less than a quarter of studies included five or more of the seven main interteaching components, which suggests few studies have evaluated the approach as it was originally outlined by Boyce and Hineline (2002). The most commonly implemented components across studies were prep guides, discussions, clarifying lectures, record sheets, and frequent probes.

The majority of the studies included in the meta-analysis focused primarily on the comparative effects of interteaching and traditional lecture. Collectively, these evaluations showed that interteaching resulted in better outcomes (e.g., examination and quiz scores) than traditional lecture and other comparison conditions (e.g., reading-only group). Compared to traditional lecture, interteaching that included a reinforcement contingency for completing prep guides led to a slightly higher effect size than interteaching without this contingency (e.g., Filipiak et al., 2010; see Fig. 1). The other ten studies included in the analysis compared one variation of interteaching to another (e.g., interteaching with pair vs. group discussions) or used a pre-test/post-test design to evaluate manipulations of interteaching (i.e., Felderman, 2014). Overall, these interteaching-to-interteaching comparisons produced minimal differential effects on outcome measures (i.e., examination and quiz scores; see Fig. 2). However, we identified some notable findings among studies that examined the discussion component of interteaching (see Fig. 3). Our analysis provides moderate support for having students work in pairs during discussions, rather than in larger groups (e.g., Boyce & Hineline, 2002; Rosales & Soldner, 2018). Furthermore, the data suggest that removing the discussion component of interteaching is likely to limit the effectiveness of the approach substantially (e.g., Garcia et al., 2016). This notion is in line with recent studies on the effectiveness of active learning components across in-person and online settings (Gayman et al., 2018; Müller & Wulf, 2020; Pollock et al., 2011). Overall, discussions seem to improve academic performance via enabling comparison and knowledge exchange between students, revision of previous content, and promoting critical thinking and higher-order learning (Gayman et al., 2018; Müller & Wulf, 2020; Pollock et al., 2011).

In addition to our analysis of the experimental comparisons summarized above, we also conducted a moderator analysis of 30 participant- and study-related characteristics. Several variables significantly increased the effectiveness of interteaching compared to traditional lecture or control conditions. Making prep guides first available during the corresponding interteaching session, scheduling weekly interteaching sessions, using multiple-choice and/or true/false questions, conducting multiple probes, and administering them every one to three meetings seem to be factors that improve the effects of interteaching. The analysis also indicated that *not* scheduling a contingency for completing discussions or prep guides seems to increase effectiveness. Although previous studies have found a similar effect, namely, that interteaching continues to be effective without the application of cooperative contingencies (e.g., Saville & Zinn, 2009), this finding is limited by the presence of outliers and the small sample size used in the analysis.

Future Research Directions

Interteaching is an empirically supported teaching method with a growing body of literature supporting its efficacy. However, many empirical questions have yet to be answered in the existing literature. Most interteaching studies have been conducted with undergraduate students in a face-to-face higher-education setting, usually in social science classes. Thus, further research is needed to investigate the efficacy of interteaching in additional academic areas such as classes on liberal arts, political science, history, literature courses, legal studies, and in graduate-level courses outside of the social sciences.

Only a few studies investigated interteaching in an online environment. Accordingly, additional research should be conducted to determine whether effects found in face-to-face settings generalize to the online environment and to evaluate components specific to online settings. This seems of especial relevance considering the dramatic increase in online instruction that resulted from the COVID-19 pandemic (Dhawan, 2020; Mahmood, 2020). For example, no study has yet compared synchronous to asynchronous online discussions, and interteaching has rarely been evaluated in a blended course where some parts of the class are taught face-to-face, and other parts are taught in an online format. Interteaching leads to better learning outcomes and is preferred by students over more passive teaching strategies such as lectures. However, interteaching has never been evaluated against other active learning techniques such as flipped classroom (Lage et al., 2000), discussion-based learning (see review by Aloni & Harrington, 2018), or more hands-on workshop-style learning.

Future research should also evaluate whether results generalize to settings outside of higher education, such as middle/high school classrooms, workplace training, caregiver training (including animal companion), rehabilitation instruction, or continuing education (Querol et al., 2015; Sturmey et al., 2015). Interteaching has been an effective teaching method even when implemented in a single session (e.g., Saville et al., 2005), so it stands to reason that the method could be used in a diverse range of settings where didactic style teaching is currently used.

It is unclear which components of interteaching are necessary or sufficient, as the small number of records in the meta-analysis did not allow for many strong conclusions to be made about individual components of interteaching. There is a need for more component analyses to evaluate what effect each component adds to the efficacy of the overall method, and how different components might be manipulated to improve outcomes (Querol et al., 2015; Sturmey et al., 2015). For example, the effect on learning outcomes of each component of interteaching could be evaluated separately, or in combination with each other to determine which components are necessary and sufficient (see Ward-Horner & Sturmey, 2010 for methods to complete component analyses using single-subject experimental designs). The majority of studies in the existing literature have used test scores and student preference as the primary dependent variables measured. Future studies could investigate additional outcome measures (e.g., measures of long-term retention, open-ended examination questions or assignments, generalization probes), including testing generalization. There is also a need to assess social validity from the instructor's point of view. As noted in the introduction, interteaching was designed as a method to address limitations that previously hindered the implementation of behavioral teaching methods in classrooms (i.e., methods that were time-consuming to prepare and did not fit well within the confines of the typical academic structure of higher-education settings). Some researchers have indicated that interteaching takes less time to prepare after the first-class section taught with the method (Sturmey et al., 2015); however, empirical evidence regarding instructor ratings of interteaching is needed.

Sturmey et al. (2015) noted a lack of specificity in how components of interteaching were reported in the literature and called for an increase in procedural integrity measures. The current investigation found a wide range of procedural differences in the implementation of interteaching procedures, and many studies were unclear about the components implemented. Only a small number of records included clear procedural integrity data. For example, the majority of studies did not clearly report details on whether there was a contingency in place for submission or completion of prep guides (70.4%), or quality (93.5%), or completion (56.5%) of the discussion. Size of discussion groups, how students were assigned to them, type of facilitator, length, and frequency of discussions all varied across studies and was not clearly reported in many of them. Frequency of assessment and type of evaluation often were not described in records. This review did not find any record of quality points implemented in the manner originally outlined by Boyce and Hineline (2002). This component entails an explicit collaborative contingency where points are contingent on everyone in a discussion group answering certain quiz questions covering material discussed by the group accurately. Instead, nine records (8.3%) indicated that quality points were earned based on an overall performance of 80% or better on quizzes, and eight records (7.4%) indicated that quality points were used but did not specify performance criteria. The remaining records (84.3%) did not report clear details on delivery of quality points. This disparity in the implementation of interteaching and the lack of clarity on how included components were executed makes replication of findings challenging. The field of interteaching could benefit from additional studies with strong procedural integrity measures.

Pedagogical Implications

Our findings may inform pedagogy in several ways. These results indicate that interteaching, regardless of different variations or configurations, is more effective than traditional lecture or other alternative control conditions. Furthermore, it appears that variations in the configuration of the different interteaching components do not limit its effectiveness significantly, as long as the discussion component is included. The discussion component thus seems to be crucial to the effectiveness of interteaching. One of the strengths of interteaching seems to lie in having discussions that focus on reviewing and clarifying difficult concepts instead of presenting introductory information on a topic. An interesting finding of the present review is that making prep guides first available during the interteaching session seems to increase its effectiveness. This finding seems counterintuitive, as one may expect that students' engagement with the material previous to the corresponding session may be more effective. However, it seems possible that this approach works better because it prevents students from copying other students' answers without having a real opportunity to analyze the readings, and/or it promotes more meaningful and direct interaction with the material in the context of the discussion between peers. Further research is needed to explore this effect systematically, as the sample of related studies identified in the present review was small.

Scheduling weekly interteaching sessions, using multiple-choice and/or true/false questions, conducting multiple examinations, probes, or quizzes, and administering them at every or during every third meeting seem to be additional factors that moderate interteaching effectiveness. Though it seems that not scheduling a contingency for completing discussions or prep guides may increase effectiveness, this finding seems limited by the presence of outliers and overall small sample size of studies used for the analysis. A promising area of research in interteaching relates to testing the reliability of the effects identified here, and investigating how to further maximize the effective components.

Conclusion

Lecture-based methods continue to be the predominant college pedagogy, notwithstanding mounting evidence that has demonstrated their limited efficacy (Stains et al., 2018). Interteaching was introduced almost two decades ago (Boyce & Hineline, 2002) as a behavior analytic alternative built upon previous behavioral teaching methods (e.g., Programmed Instruction Holland & Skinner, 1961; Personalized System of Instruction Keller, 1968). The growing interest in interteaching since its inception resulted in several studies aimed at describing its application and testing its effectiveness. Earlier reviews of this literature (Querol et al., 2015; Saville et al., 2011b; Sturmey et al., 2015) overall indicated that interteaching was more effective and accepted by students, when compared with traditional lecture-based instruction across a wide range of academic disciplines and settings. Here we updated and expanded those previous reviews. Our findings overall indicated that interteaching was importantly more effective than traditional lecture-based methods. The fact that variations in the configuration of the interteaching components did not seem to substantially limit its effectiveness, as long as the discussion component was included, suggested that the discussion component is crucial to the effectiveness of interteaching. However, future systematic component analyses are needed to test the necessity and sufficiency of the discussion component and its potential interaction with other components (Ward-Horner & Sturmey, 2010). Other promising efforts to extend our knowledge about the efficacy and versatility of interteaching include investigating it across other academic areas, online environments, and workplace and continuing education settings. Hopefully, these efforts will continue promoting its dissemination as an evidence-based instruction method.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10864-021-09452-3.

Acknowledgements This project was partially supported with funding provided by the Faculty Development Committee (FDC) of Troy University.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

List of references of studies included in the systematic review and meta-analysis are available in Supplementary References.

Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology, 71*(2), 103–112. https://doi.org/10.4097/kjae.2018.71.2.103

Aloni, M., & Harrington, C. (2018). Research based practices for improving the effectiveness of asynchronous online discussion boards. *Scholarship of Teaching and Learning in Psychology, 4*(4), 271–289. https://doi.org/10.1037/stl0000121

Anbro, S. J. (2015). *An evaluation of the efficacy of interteaching in an undergraduate classroom*. Carbondale, Illinois: Southern Illinois University.

Arntzen, E., & Hoium, K. (2010). On the effectiveness of interteaching. *The Behavior Analyst Today*, *11*(3), 155–160. https://doi.org/10.1037/h0100698

Bethke, V. S. (2016). *Interteaching: Types of prep guide questions and their effect on student quiz performance*. Harrisonburg, Virginia: James Madison University.

Biostat, I. (2019). *The comprehensive meta-analysis©* (Version 3.0) [Computer software]. Englewood, NJ, USA: Biostat, Inc. Retrieved from https://www.meta-analysis.com/index.php?cart=BMNJ1648138.

Borenstein, M. (2005). Software for publication bias. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 193–220). New York: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Borenstein, M., Hedges, L.V., Higgins, J. P., & Rothstein, H. R. (2015). *Regression in meta-analysis*. Retrieved from https://www.meta-analysis.com/downloads/MRManual.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Boyce, T. E., & Hineline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst, 25*(2), 215–226.

Byrne, B., & Guy, R. (2016). Interteaching within a human physiology course: A comparison of first- and second-year students' learning skills and perceptions. *Advances in Physiology Education*, *40*(3), 349–353. https://doi.org/10.1152/advan.00141.2015

Cannella-Malone, H. I., Axe, J. B., & Parker, E. D. (2009). Interteach preparation: A comparison of the effects of answering versus generating study guide questions on quiz scores. *Journal of the Scholarship of Teaching and Learning*, *9*(2), 22–35. https://files.eric.ed.gov/fulltext/EJ854891.pdf

Cezeaux, J. L., & Keyser, T. K. (2018). *Introducing active learning strategies into an undergraduate engineering physiology course*. Paper presented at the ASEE Annual Conference and Exposition on Salt Palace Convention Center, Salt Lake City, UT. Retrieved from https://scholarworks.iu.edu/journals/index.php/josotl/article/download/1723/1721

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129. https://doi.org/10.2307/3001666

Deeks, J.J., Higgins, J.P.T., Alman, D.G. (2019). Chapter 10: Analysing data and unterdaking meta-analyses. In: Higgins, J.P.T, Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., & Welch, V.A. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.0 (updated July 2019). Retrieved 28th Sept 2020. from: http://www.training.cochrane.org/handbook.

Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems*. https://doi.org/10.1177/0047239520934018

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions.

Journal of Epidemiology and Community Health, 52(6), 377–384. https://doi.org/10.1136/jech.52.6.377

Dreier, M. (2013). Quality assessment in meta-analysis. In S. A. Doi & G. M. Williams (Eds.), *Methods of clinical epidemiology* (pp. 213–228). Springer.

Edwards, J. R. (2005). Effect of various teaching approaches on business ethics instruction. Nacogdoches, Texas: Stephen F. Austin State University.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Felderman, T. A. (2014). Preliminary analysis of interteachings frequent examinations component in the community college classroom. *Journal of College Teaching & Learning (TLC)*, *11*(4), 149–156. https://doi.org/10.19030/tlc.v11i4.8851

Felderman, T. A. (2016). A systematic replication comparing interteaching and lecture in the community college classroom. *Community College Journal of Research and Practice*, *40*(9), 739–749. https://doi.org/10.1080/10668926.2015.1075445

Filipiak, S., Anne Rehfeldt, R., Heal, N. A., & Baker, J. C. (2010). The effects of points for preparation guides in interteaching procedures. *European Journal of Behavior Analysis*, *11*(2), 115–132. https://doi.org/10.1080/15021149.2010.11434338

Garcia, Y. A., Orozco, L., & Martin, G. (2016). Comparación de dos procedimientos de enseñanza universitaria: Un ejemplo de interteaching. *Psicologia Escolar e Educacional*, *20*(3), 493–501. https://doi.org/10.1590/2175-3539/2015/02031029

Gayman, C. M., Hammonds, F., & Rost, K. A. (2018). Interteaching in an asynchronous online class. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 231–242. https://doi.org/10.1037/stl0000126

Germain, S. M., Wilkie, K. D., Milbourne, V. M. K., & Theule, J. (2018). Animal-assisted psychotherapy and trauma: A meta-analysis. *Anthrozoos, 31*(2), 141–164. https://doi.org/10.1080/08927936.2018.1434044

Goto, K., & Schneider, J. (2010). Learning through teaching: Challenges and opportunities in facilitating student learning in food science and nutrition by using the interteaching approach: Classroom techniques. *Journal of Food Science Education*, *9*(1), 31–35. https://doi.org/10.1111/j.1541-4329.2009.00087.x

Gutierrez, M. (2017). *Interteaching: The effects of discussion group size on undergraduate student performance and preference Michael Gutie*. Seattle, Washington: University of Washington.

Higgins, J.P., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. Retrieved from http://handbook-5-1.cochrane.org.

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston M, Li, T., Page, M.J., & Welch, V.A. (2020). *Cochrane handbook for systematic reviews of interventions* version 6.1 (updated September 2020). Cochrane, 2020. Retrieved from https://www.training.cochrane.org/handbook.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior*. McGraw-Hill.

Keenan, C. (2016). *Universal preschool- and school-based education programmes for reducing ethnic prejudice and promoting respect for diversity among children aged 3–11: A systematic review and meta-analysis*. Queen's University Belfast. https://doi.org/10.1002/cl2.164

Keller, F. S. (1968). Good-bye, teacher. *Journal of Applied Behavior Analysis, 1*(1), 79–89. https://doi.org/10.1901/jaba.1968.1-79

Kratochwill, T. R., & Levin, J. R. (2014). Introduction: An overview of single-case intervention research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: methodological and statistical advances* (pp. 3–23). American Psychological Association.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education, 31*, 30–43.

Lambert, T., & Saville, B. K. (2012). Interteaching and the testing effect: A preliminary analysis. *Teaching of Psychology*, *39*(3), 194–198. https://doi.org/10.1177/0098628312450435

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Lipsey, M. W. (2019). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (1st ed.). SAGE Publications Ltd.

Littell, J. H., & Corcoran, J. (2010). Systematic reviews. In B. A. Thyer (Ed.), *The handbook of social work: Research methods* (2nd ed., pp. 313–338). SAGE Publications Ltd.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press.

Logan, L. R., Hickman, R. R., Harris, S. R., & Heriza, C. B. (2008). Single-subject research design: Recommendations for levels of evidence and quality rating. *Developmental Medicine and Child Neurology, 50*(2), 99–103. https://doi.org/10.1111/j.1469-8749.2007.02005.x

Mahmood, S. (2020). Instructional Strategies for online teaching in COVID-19 pandemic. *Human Behavior and Emerging Technologies*. https://doi.org/10.1002/hbe2.218

Mason, L. L. (2012). Interteaching to increase active student responding and differentiate instruction. *Behavioral Technology Today*, *15*(7), 1–15. Retrieved from https://www.behavior.org/resources/661.pdf

Mercer, D. E. (2014). Interteaching: success and frustrations in implementing active learning methodology in anatomy and physiology. *Human Anatomy & Physiology Society Educator*, *18*(1), 18–23.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000097

Moran, D. J., & Malott, R. W. (Eds.). (2004). *Evidence-based educational methods: Advances from the behavioral sciences*. Academic Press.

Müller, F. A., & Wulf, T. (2020). Technology-supported management education: A systematic review of antecedents of learning effectiveness. *International Journal of Educational Technology in Higher Education, 17*, 47. https://doi.org/10.1186/s41239-020-00226-x

Pear, J. J., Schnerch, G. J., Silva, K. M., Svenningsen, L., & Lambert, J. (2011). Web-based computer-aided Personalized System of Instruction. *New Directions for Teaching and Learning, 128*, 85–94.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing Ltd.

Pollock, P. H., Hamann, K., & Wilson, B. M. (2011). Learning through discussions: Comparing the benefits of small-group and large-class settings. *Journal of Political Science Education, 7*(1), 48–64. https://doi.org/10.1080/15512169.2011.539913

Querol, B. I. D., Rosales, R., & Soldner, J. L. (2015). A comprehensive review of interteaching and its impact on student learning and satisfaction. *Scholarship of Teaching and Learning in Psychology, 1*(4), 390–411. https://doi.org/10.1037/stl0000048

Rehfeldt, R. A., Walker, B., Garcia, Y., Lovett, S., & Filipiak, S. (2010). A point contingency for homework submission in the graduate school classroom. *Journal of Applied Behavior Analysis*, *43*(3), 499–502. https://doi.org/10.1901/jaba.2010.4

Rieken, C. J., Dotson, W. H., Carter, S. L., & Griffith, A. K. (2018). An evaluation of interteaching in an asynchronous online graduate-level behavior analysis course. *Teaching of Psychology*, *45*(3), 264–269. https://doi.org/10.1177/0098628318779275

Rosales, R., & Soldner, J. L. (2018). An assessment of group size in interteaching. *Journal of the Scholarship of Teaching and Learning*, *18*(2), 105–117. https://doi.org/10.14434/josotl.v18i2.22539

Rosales, R., Soldner, J. L., & Zhang, L. (2018). An evaluation of the pair discussion component of interteaching. *The Psychological Record*, *68*(1), 71–79. https://doi.org/10.1007/s40732-018-0269-0

Saville, B. K., Bureau, A., Eckenrode, C., Fullerton, A., Herbert, R., Maley, M., Porter, A., & Zombakis, J. (2014). Interteaching and lecture: A comparison of long-term recognition memory. *Teaching of Psychology*, *41*(4), 325–329. https://doi.org/10.1177/0098628314549704

Saville, B. K., Cox, T., O'Brien, S., & Vanderveldt, A. (2011a). Interteaching: The impact of lectures on student performance. *Journal of Applied Behavior Analysis*, *44*(4), 937–941. https://doi.org/10.1901/jaba.2011.44-937

Saville, B. K., Lambert, T., & Robertson, M. S. (2011b). Interteaching: Bringing behavioral education into the 21st century. *The Psychological Record, 61*, 153–165. https://doi.org/10.1007/BF03395752

Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012). Interteaching and the testing effect: A systematic replication. *Teaching of Psychology*, *39*(4), 280–283. https://doi.org/10.1177/0098628312456628

Saville, B. K., Pope, D., Truelove, J., & Williams, J. (2012). The relation between GPA and exam performance during interteaching and lecture. *The Behavior Analyst Today*, *13*(3–4), 27–31. https://doi.org/10.1037/h0100728

Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis*, *42*(2), 369–374. https://doi.org/10.1901/jaba.2009.42-369

Saville, B. K., & Zinn, T. E. (2011). Interteaching. *New Directions for Teaching and Learning, 2011*(128), 53–61. https://doi.org/10.1002/tl.468

Saville, B. K., Zinn, T. E., & Elliott, M. P. (2005). Interteaching versus traditional methods of instruction: A preliminary analysis. *Teaching of Psychology*, *32*(3), 161–163. https://doi.org/10.1207/s15328023top3203_6

Scoboria, A., & Pascual-Leone, A. (2009). An "interteaching" informed approach to instructing large undergraduate education. *Journal of the Scholarship of Teaching and Learning*, *9*(3), 29–37. Retrieved from https://scholarworks.iu.edu/journals/index.php/josotl/article/view/2140

Slezak, J. M., & Faas, C. (2017). Effects of an interteaching probe on learning and generalization of American Psychological Association (APA) Style. *Teaching of Psychology*, *44*(2), 150–154. https://doi.org/10.1177/0098628317692619

Soldner, J. L., Rosales, R., & Crimando, W. (2015). A comparison of interteaching and classroom lecture in rehabilitation education. *Rehabilitation Counselors and Educators Journal*, *8*(1), 91–100.

Soldner, J. L., Rosales, R., Crimando, W., & Schultz, J. C. (2017). Interteaching: Application of an empirically supported behavioral teaching Method in Distance Rehabilitation Education. *Rehabilitation Research, Policy, and Education*, *31*(4), 372–386. https://doi.org/10.1891/2168-6653.31.4.372

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science, 359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119–1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Sturmey, P., Dalfen, S., & Fienup, D. M. (2015). Inter-teaching: A systematic review. *European Journal of Behavior Analysis, 16*(1), 121–130. https://doi.org/10.1080/15021149.2015.1069655

Truelove, J., Saville, B., & Van Patten, R. (2013). Interteaching: Discussion group size and course performance. *Journal of the Scholarship of Teaching and Learning*, *13*(2), 23–30.

Ward-Horner, J., & Sturmey, P. (2010). Component analyses using single-subject experimental designs: A review. *Journal of Applied Behavior Analysis, 43*(4), 685–704.

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*(2), 235–268. https://doi.org/10.1353/etc.2012.0010

White, H. D. (2019). *The Handbook of Research Synthesis and Meta-Analysis.* In J. C. Cooper, H., Hedges, L.V., & Valentine (Eds., 2nd ed.). Russell Sage Foundation.

Wright, R. A., & Wright, W. C. (2011). *The use of interteaching to evaluate short - and long-term concept retention*. West Point, New York: Center for Teaching Excellence, United Stated Military Academy.

Zayac, R., & Paulk, A. L. (2014). Interteaching: Its effects on exam scores in a compressed-schedule format. *Journal of the Scholarship of Teaching and Learning*, *14*(1), 1–12. https://doi.org/10.14434/josotl.v14i1.3649

Zelinsky, N. A. M., & Shadish, W. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation, 21*(4), 266–278. https://doi.org/10.3109/17518423.2015.1100690

Authors and Affiliations

Camilo Hurtado-Parrado[1] · **Nicole Pfaller-Sadovsky**[2] · **Lucia Medina**[3] · **Catherine M. Gayman**[4] · **Kristen A. Rost**[4] · **Derek Schofill**[4]

[1] School of Psychological and Behavioral Sciences, Southern Illinois University, 1125 Lincoln Drive, Carbondale, IL 62901, USA

[2] School of Biological Sciences, Queen's University Belfast, Belfast, Northern Ireland

[3] Faculty of Psychology, Fundacion Universitaria Konrad Lorenz, Bogotá, Colombia

[4] Department of Psychology, Troy University, Troy, AL, USA