Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

Studying the effectiveness of team teaching : a systematic review on the conceptual and methodological credibility of experimental studies

Reference:
De Weerdt Dries, Simons Mathea, Struyf Elke, Tack Hanne.- Studying the effectiveness of team teaching : a systematic review on the conceptual and methodological credibility of experimental studies
Review of educational research - ISSN 0034-6543 - (2024), p. 1-46
Full text (Publisher's DOI): https://doi.org/10.3102/00346543241262807
To cite this reference: https://hdl.handle.net/10067/2071370151162165141

uantwerpen.be
Institutional repository IRUA

Studying the Effectiveness of Team Teaching: A Systematic Review on the Conceptual and Methodological Credibility of Experimental Studies

Dries De Weerdt[1], Mathea Simons[1], Elke Struyf[1], and Hanne Tack[2,3]

[1] Antwerp School of Education, Faculty of Social Sciences, University of Antwerp (Belgium)

[2] Department of Training and Education Sciences, Faculty of Social sciences, University of Antwerp (Belgium)

[3] Department of Educational Studies, Faculty of Psychology and Educational Sciences, Ghent University (Belgium)

Author note

Dries De Weerdt https://orcid.org/0000-0001-6903-6587

Mathea Simons https://orcid.org/0000-0002-9239-9324

Elke Struyf https://orcid.org/0000-0003-1067-6357

Hanne Tack https://orcid.org/0000-0002-6199-4411

At the time of writing this study, there was a change in author affiliation: Hanne Tack is now at the Flemish Inspectorate of Education.

We have no known conflict of interest to disclose. We received funding from Fonds Wetenschappelijk Onderzoek (FWO), the Research Foundation Flanders, as this study is part of the larger ESTAFETT project (www.teamteaching-estafett.be/english).

Correspondence concerning this article should be addressed to Dries De Weerdt, University of Antwerp, Sint-Jacobstraat 2, 2000 Antwerp, Belgium. Email: dries.deweerdt@uantwerpen.be

Abstract

The aim of this systematic review is to provide insight into the conceptual and methodological credibility of experimental research designs on the effectiveness of team teaching—a promising instructional strategy wherein two or more professionals collaboratively provide education for a group of students. A total of 31 studies were included. These studies were conceptually and methodologically examined according to two actualized quality-appraisal frameworks. The findings reveal that it remains a challenge to design rigorous experimental studies with clear conceptualizations of key variables related to team teaching. To make convincing claims on the effectiveness of team teaching, there is an urgent need for better-defined quality experimental research. Therefore, we conclude with recommendations for future research, specifically how experimental studies on the effectiveness of team teaching should be conceptually and methodologically implemented to provide policymakers and stakeholders with information for evidence-informed decision-making on educational practices.

Keywords: team teaching, co-teaching, experimental research, effectiveness, systematic review

Studying the Effectiveness of Team Teaching: A Systematic Review on the Conceptual and Methodological Credibility of Experimental Studies

Up until now, the nature of the teaching profession has been highly individualistic, with little or no cooperation between teachers and their colleagues (OECD, 2020). However, evidence suggests that teachers may have a greater impact on their students' outcomes and develop stronger teaching practices when they pool their individual expertise and collaborate with other teachers (Esterhazy et al., 2021; Goddard et al., 2015; Vangrieken et al., 2015). For this reason, many researchers and stakeholders have turned to teacher collaboration as a strategy to break down teacher isolation (Ostovar-Nameghi & Sheikhahmadi, 2016) and expand professional identities in ways that foster sustained teacher professional development and overall education quality (Gast et al., 2017; Hargreaves, 2019).

A growing body of literature recognizes the pivotal role of team teaching as a collaboration-focused educational strategy (Walsh, 2020). Team teaching is a teaching model in which "two or more teachers work together in some level of collaboration in the planning, delivery and/or evaluation of a course" (Baeten & Simons, 2014, p. 93). Because of its promising character, team teaching is receiving increasing attention from policymakers and is starting to become an integral part of many school and educational formats. Though this is especially so in countries like Finland and Japan where teacher collaboration serves a central function in instructional improvement (Honigsfeld & Dove, 2019), its merits are also being recognized in other countries in North America and Europe (Iacono et al., 2021; King-Sears et al., 2021).

However, to implement team teaching in a sustainable way, stakeholders must be informed about its effectiveness through clear evidence. To deliver this evidence and to attain an understanding of what really works in the field of education, there is a need for rigorous causal inference research (Gopalan et al., 2020). In line with best practices from other

research fields, experimental research designs are considered the most appropriate strategy for convincingly claiming causal evidence about the effectiveness of interventions or educational strategies (Connolly et al., 2018; Reynolds et al., 2014). In this respect, Gopalan et al. (2020) have already noted an increasing trend since the beginning of 2009 in the use of experimental studies in the broad field of education research.

Despite the undeniable value of experimental research, multiple authors in the field of team teaching have argued that far too little attention has been paid to the use—and quality—of experimental studies to provide an understanding of the effectiveness of integrating team teaching into current educational practices (e.g., Baeten & Simons, 2014; Friend et al., 2010; King-Sears et al., 2021; Murawski & Swanson, 2001; Rexroat-Frazier & Chamberlin, 2019). In other words, the call for high-quality experimental studies that produce usable knowledge, curate research evidence, and utilize scientific knowledge in daily practices echoes through the field of team teaching. Recently, Vembye et al. (2023) recognized this statement in their meta-analysis on the effects of collaborative teaching models on students' learning achievement, in which they argue the need for more quality experimental research before the true effect of team-taught practices can be convincingly claimed. Drawing upon this need, the aim of our study is to further elaborate on the use of experimental studies in team teaching research and especially how it can be conducted in a qualitative way.

Establishing quality research is a challenging endeavor. Based on a research quality framework of Ming and Goldenberg (2021), we present two central components for describing how new scientific knowledge can be defined, constructed, and validated through experimental research. Ming and Goldenberg (2021) highlight the importance of conceptual and methodological credibility to providing high-quality research. Conceptual credibility refers to clarity in the definition, operationalization, and coherence of studied variables. Methodological credibility ensures transparency in study design and execution, with attention

to differences that affect to whom and under what conditions research results are meaningful. These basic arguments on research quality serve as the basics for this review study as we elaborate on how experimental research into the effectiveness of team teaching can be conducted in a way it can be considered as conceptually and methodologically credible.

While previous research confirms that experimental studies have been carried out within the field of team teaching, to date there has been no overview of the extent to which these studies have contributed to the examination of the effectiveness of team teaching across populations, outcomes, settings, and different modes of implementation by making use of quality research designs. An overview study that presents clear insight into conceptual and methodological credibility can enhance the knowledge base of the research field and guide future studies to examine the effectiveness of team teaching in a convincing way. In our review study, we expand on previous review studies by providing more in-depth insight into the use of experimental studies, as these are a main concern in team teaching research (e.g., Baeten & Simons, 2014; Vembye et al., 2023). Furthermore, and in contrary to other review studies, we did not limit our literature overview to particular units (e.g., students, teachers) or outcomes (e.g., learning achievement, engagement, teaching behavior). This broader view can be of interest as the choice for implementing team teaching in education should not rely solely on its effectiveness examined from a single perspective. Our paper addresses the following research questions:

- RQ1: To what extent do experimental studies on the effectiveness of team teaching take into account conceptualizations of key variables in team teaching research?
- RQ2: To what extent do experimental studies on the effectiveness of team teaching comply with methodological quality requirements?
- RQ3: To what extent do current experimental studies inform us about the effectiveness of team teaching?

Theoretical Foundation

In this section, we first outline the concept of team teaching. Second, we elaborate on the value of experimental research designs for examining the effectiveness of team teaching. Third, and to answer the research questions, we describe how conceptual and methodological credibility can be established in experimental research on team teaching. In doing so, we present (1) a conceptual framework regarding key variables in team teaching research and (2) methodological quality characteristics of experimental research designs.

Team Teaching as a Collaborative Educational Strategy

Teacher collaboration in teaching practices is gaining attention in contemporary education research (Hargreaves, 2019). Collaboration refers to an interactive group process in which group or team members work together for task-related purposes (Kelchtermans, 2006; Vangrieken et al., 2015). Research focuses on various forms of collaborative practices with different levels of intensity and entitativity, such as collegial visits, lesson study, professional learning communities, and team teaching (Gast et al., 2017; Muckenthaler et al., 2020; Vangrieken et al., 2015). Most of these teacher collaborations are often restricted to discussing teaching didactics and daily teaching problems, or practical issues related to teaching activities and materials (Vangrieken et al., 2015). In general, little attention goes to collaborative instruction.

What differentiates team teaching from other collaborative activities is that it can be considered as more than just an interactive group process, but also as an instructional strategy where two or more teachers collaboratively teach a group of students (Walsh, 2020). Through this teaching format, team teaching creates situations where teachers learn from colleagues and contribute to shared knowledge construction through exchanging ideas, discussing alternative perspectives on teaching and learning, and providing feedback to each other, with a direct focus on improving students' outcomes during the teaching practice (Hargreaves,

2019; Walsh, 2020). This makes team teaching perhaps the most intense form of teacher collaboration (de Jong et al., 2019; Muckenthaler et al., 2020).

However, recent studies in team teaching research highlight the complexity of defining team teaching as multiple definitions and synonyms exist, such as co-teaching, collaborative teaching, and cooperative teaching (Simons, Coetzee, et al., 2020; Walsh, 2020). The existence of different terms and the fact that these terms are often used interchangeably may cause some confusion among researchers and stakeholders in education (Friend et al., 2010; Walsh, 2020). Especially, contradiction occurs between the term 'co-teaching' and 'team teaching'. Co-teaching refers to the collaboration between a (general) teacher and a special-education teacher to help students with special educational needs (Cook & Friend, 1995). By this definition, co-teaching refers narrowly to a teaching team that combines specific and different expertise from each educational professional, and team teaching is often considered a specific model of co-teaching with the highest amount of classroom interaction between teachers (Cook & Friend, 1995; Cook & McDuffie-Landrum, 2020).

From a historical point of view, however, the roots of co-teaching are found in the practice of team teaching (Friend et al., 2010). In the 1950s, team teaching was already included in US educational practices. During the second half of the 20th century, team teaching gained in popularity, and several variations of team teaching emerged. One of these variations stemmed from the movement toward inclusive education, where special education and related services were being offered more frequently in general education settings by combining the expertise of different educational professionals (Friend et al., 2010; Fluijt et al., 2016). The practice of co-teaching arose as a result of this idea (Cook & Friend, 1995). Thus, co-teaching can be considered more specific than team teaching because of the difference in expertise between the educational professionals within the team composition. Over time,

however, the two terms “team teaching” and “co-teaching” have come to be used interchangeably.

Considering this development, in this review study, we argue that team teaching should be regarded as the broader concept, referring to all possible variations in collaboration between two or more educational professionals, in all possible subjects, across all different kinds of educational settings (Simons, Coetzee, et al., 2020).

Experimental Research Designs for Effectiveness Studies in Education

To convince policymakers and stakeholders of the added value of team teaching, compelling scientific evidence is needed. An important feature of scientific research is the identification of cause-effect relationships among variables (Pearl, 2009). Causation provides insight into the effectiveness of educational practices as we deem the implementation of certain practices to be effective when it leads to improved student, teacher, or school outcomes (Connolly et al., 2018; Reynolds et al., 2014; Shadish et al., 2002). These causal insights are then used as a key source of evidence to guide policymakers and stakeholders in deciding what strategies are worth spending time and money on (Levin, 2004; Tipton & Olsen, 2018).

Establishing causal insights requires that three fundamental conditions be met: (1) the cause (e.g., the educational intervention) must be associated with the effect, (2) the cause must precede the effect, and (3) there must be no other reasonable explanation for the effect (Shadish et al., 2002). A main strategy for showing that these conditions are satisfied is to employ a valid counterfactual inference (Morgan & Winship, 2014), which involves comparing the outcome of participants’ receiving a particular treatment/intervention with the outcome of a situation in which they do not receive it (Schanzenbach, 2012).

In view of the above statements, rigorous research designs can make it possible to establish causal claims. Experimental methods are considered the best-suited methodology for

studying causal relationships (Fraenkel et al., 2019; Schanzenbach, 2012; Shadish et al., 2002). Experimental designs have the potential to establish causation by introducing one or more treatments into a situation and observing the outcome(s) of the treatment(s). In doing so, a variety of techniques (e.g., randomization, control group) are utilized during the experiment to diminish the likelihood that the (in)effectiveness of the treatment could have other explanations besides the purported cause (Cohen et al., 2018; Shadish et al., 2002). This way, experiments can provide unbiased estimates of treatment effects. In contrast, other research designs (e.g., observational studies) are limited in their capability to isolate the effects of a particular educational intervention from other, extraneous influences in the study setting.

However, experimental studies are no simple endeavor in education research. Despite the fact that experiments can provide evidence of "what works," they are often criticized for ignoring the complexity and changing dynamics of social contexts (Connolly et al., 2018). In this regard, several scholars have argued that although rigorous experimental studies have their place in education research, attention must be given not only to methodological quality but also to the context, participant characteristics, and multiple elements of the intervention (Cohen et al., 2018; Shadish et al., 2002; Sullivan, 2011). All this to ensure that contextual factors do not trump the findings from the experiment and to foster opportunities for causal generalization that moves beyond the localized nature of the original experimental setting.

In essence, this focus on methodology and conceptual meaning reflects the two central concepts of our review, namely the conceptual and methodological credibility of an experimental study. Therefore, in the following sections, we elaborate on both perspectives in the research field of team teaching. By this means, we promote the use of experimental studies in team teaching research by capitalizing on the inherent methodological strengths of experimental designs without ignoring the social context in which they are conducted.

Conceptual Framework

In practice, team teaching has various meanings and interpretations as well as a variety of aspects that can influence how it is implemented. From a conceptual point of view, this implies that studies on team teaching—and particularly experimental studies—are subject to a large number of variables. Although previous research states that, based on preliminary insights, mediating and moderating processes for explaining the (in)effectiveness of team teaching may be less intricate than expected, the mechanisms that underpin effective team teaching practices are still not fully understood (Vembye et al., 2023). We developed a conceptual framework that summarizes the most highlighted factors, retrieved from previous review studies in the field (e.g., Baeten & Simons, 2014; Iacono et al., 2021), that may influence the implementation of team teaching in practice and its effectiveness for enhancing education outcomes.

Our conceptual framework is based on work of Cronbach (1982), who proposes that each experiment should include a conceptual meaning for (1) units, such as people or groups; (2) a treatment that (some of) the units receive; (3) observations, such as outcome measures; and (4) a setting, which refers to the larger context in which the study takes place (Albright & Malloy, 2000; Shadish, 2011). Based on these four elements, a framework can be developed that includes the most important conceptual aspects when examining the effectiveness of team teaching (see Figure 1).

Units

Units in experimental research on team teaching, defined as persons or groups on which data are collected (Shadish et al., 2002), may be chosen to focus on central educational actors such as students (who are taught by a team teaching approach) or teachers (who teach in a team-taught arrangement). It is also possible to consider other units, such as mentors observing teams of student teachers. The choice of a particular unit may affect the results of the experimental study. Previous review studies by Altstaedter et al. (2016) and Baeten and

Simons (2014) reported different kinds of advantages and disadvantages depending on the position (i.e., student, teacher, or mentor) in the team-taught context that was chosen as a unit. Although different units exist, most research in the field of team teaching is conducted with teachers as the research unit (e.g., Walsh, 2020).

Team Teaching as a Treatment

Three differentiating aspects influence the meaning of team teaching as a treatment. These aspects are (1) the composition of the teaching team, (2) the appearance of the team teaching in practice (i.e., the model(s) applied), and (3) the phases included in the team teaching practice.

First, regarding composition, team-teaching teams are composed of multiple educational professionals. Previous review studies have already described a variety of groups. Most often, these studies focused on co-teaching, where a general-education teacher collaborates with a special-education teacher to support students with disabilities (Solis et al., 2012). Research can also be found regarding teams consisting of two general-education teachers (Brojčin et al., 2012), teams of student teachers (e.g., Weinberg et al., 2020), collaborations between teacher educators (e.g., Nevin et al., 2009), teams pairing a student teacher with a mentor (e.g., Baeten & Simons, 2016a), teams of two subject experts with different knowledge bases (e.g., Dehnad et al., 2021), and teams pairing an educator with a paraprofessional (e.g., Heisler & Thousand, 2021). As noted by Friend et al. (2015), different compositions with different objectives may have different impacts on certain units.

Second, differences can be noticed in how team teaching is organized with respect to teaching practices (Iacono et al., 2021). Different team teaching models have been extensively described (e.g., Baeten & Simons, 2014; Cook & Friend, 1995; Devecchi & Nevin, 2010; Honigsfeld & Dove, 2019). Most of these configurations differ in the way they are categorized or in how the separate models are defined. In this review study, the models of

Baeten and Simons (2014) are highlighted as an overarching framework. Table 1 presents these models and corresponding explanations (Simons, Baeten, et al., 2020).

Third, in most definitions of team teaching, three main phases of teaching practice are repeatedly mentioned: planning, teaching, and evaluation/reflection. Previous research recognizes the importance of all three phases (Veteska et al., 2022), especially the planning and evaluation/reflection phases, which are related to opportunities for teacher professionalization and teaching practice optimization (Cook & McDuffie-Landrum, 2020; Fluijt et al., 2016; Kim, 2019). Despite the critical role of these phases, they are not always acknowledged in current teaching practices and in corresponding research on team teaching (Fluijt et al., 2016; Pratt et al., 2017).

Outcomes

The dependent variable in an experimental study is fulfilled by its outcomes. Generally, in education effectiveness research, a variety of outcome variables at different levels and units are deemed useful to judge effectiveness. Most often, student achievement is used as the central criterion in effectiveness studies (Reynolds et al., 2014). However, increased interest is going to noncognitive outcomes such as well-being, engagement, and motivation (Creemers & Kyriakides, 2007; Reynolds et al., 2014). Outcomes can be investigated on the classroom and school levels, where teacher behavior and teacher development are predominant factors (Muijs et al., 2014).

Current team teaching research follows this trend, as different outcome variables are represented for students, teachers, and others (e.g., student teachers' mentors). For example, in a meta-analysis by King-Sears et al. (2021), effects were summarized on student achievement in language, mathematics, and science. Besides these cognitive outcomes, non-cognitive student outcomes have also been found interesting. Scruggs et al. (2007) noted increased student engagement during team-taught lessons. Johnson and King-Sears (2020)

referred to student perceptions and experiences in team-taught lessons as valuable outcomes. For teachers, outcomes may focus on perceptions of increased emotional and professional support as well as professional and personal growth (Baeten & Simons, 2014). Furthermore, previous studies have indicated more effective instruction in team-taught classes (Johnson & King-Sears, 2020; King-Sears et al., 2014). As for participants in other roles, Simons and Baeten (2016) found advantages for mentors such as decreased workload, learning gains, and increased collaboration at school. In this context, Iacono et al. (2021) warn that focusing on different outcome variables may result in finding different effects of team teaching.

Setting

Setting characteristics have a significant influence on the way units experience a treatment and the way that treatment is implemented. These characteristics may even explain perceived heterogeneity in outcome measures (Gollwitzer & Schwabe, 2022). Setting variables in experimental studies on team teaching can be grouped into three categories: team teacher factors, interventional factors, and contextual factors. Each category involves multiple influencing variables.

Team Teacher Factors. The first category involves variables regarding the team teachers. In accordance with team-based research, Gast et al. (2017) drew a distinction between the individual level and the team level. Individual teacher characteristics concern teacher competencies, defined as the knowledge, skills, and attitudes required to cooperate in an effective teaching team. For this type of factor, researchers highlight the impacts of the epistemological background and teaching experience of individual teachers on the level of the collaboration implemented (i.e., the models used), the parity in roles, and the division of responsibilities (Gately & Gately, 2001; Kim, 2019). Also required for successful collaboration are interpersonal and communication skills (Cook & Friend, 1995; Jortveit & Kovač, 2021). Furthermore, attitudes such as a commitment to the concept of team teaching

and the belief that students will benefit from team-taught lessons may influence teachers' motivation and thus the quality of their teaching practices (Bešić et al., 2017; Härkki et al., 2021; Solis et al., 2012). These attitudes reflect the personal readiness and willingness to share responsibility, modify teaching styles and preferences, and work closely with another adult, all traits that better equip teachers to succeed in team-teaching situations (Cook & Friend, 1995; Jortveit & Kovač, 2021; Rexroat-Frazier & Chamberlin, 2019).

When individual teachers team up, compatibility, mutual recognition, and a collective mindset influence the effectiveness of the partnership. Team compatibility mainly concerns a theoretical match between teachers, in which overlapping competencies, personalities, and experiences create a complementary partnership (Baeten & Simons, 2014; Pratt, 2014). Additionally, there is a need for mutual recognition and a collective mindset to establish a relational connection. In team-teaching relationships, it is critical for teachers to recognize each other's expertise and personal preferences and to find common ground on which to build (Van Garderen et al., 2012). This common ground can be created by establishing a collective mindset (Magiera et al., 2005). A collective mindset can be established in partnerships that are rooted in feelings of openness, respect, trust, and parity (Baeten & Simons, 2016a; Heisler & Thousand, 2021). Openness, trust, and respect are conditions for effective collaboration and communication in partnerships (Cook & McDuffie-Landrum, 2020). Parity means both teachers' taking an active role in the classroom and sharing responsibility (Kim, 2019; Leatherman, 2009). Also, shared visions and beliefs with regard to what constitutes good education build a mindset supported by the (whole) team (Jortveit & Kovač, 2021) and reduce difficulties during team teaching practice (Magiera et al., 2005).

Interventional Factors. The second category deals with interventional factors, which are related to the design of experimental studies in the field of team teaching. In this category, five variables arise. First, the way the partnership is established can play an important role in

successful team teaching and creating an environment for student success (Rexroat-Frazier & Chamberlin, 2019). Voluntary engagement and choices in group composition generally result in better partnership development (Friend et al., 2010; Van Garderen et al., 2012).

Second, training or professional development initiatives are often considered a key factor to prepare teachers for team teaching practices (Baeten & Simons, 2016a; Sweigart & Landrum, 2015). Altstaedter et al. (2016) distinguished two major components that teacher preparation programs should contain: foundations training and pairs training. In foundations training, teachers slated to teach on the same team learn about the same conceptual framework and create a common language about team teaching strategies. In pairs training, participants engage in activities for collaboration and relationship building.

Third, prior to and during the intervention, shared time for planning and evaluation/reflection must be provided (Dehnad et al., 2021; Friend et al., 2010). In previous research, the most frequently cited challenge to collaboration is a lack of time for planning and conducting evaluation/reflection in a structured way (i.e., using established procedures and resources) (Honigsfeld & Dove, 2019; Pratt et al., 2017; Scruggs et al., 2007).

Fourth, and to make the above kinds of efforts effective, communication and collaboration are essential. Effective team teaching requires open discussion among team members based on shared experiences with the goals of modifying and enhancing teaching and learning (Baeten & Simons, 2014; Gurgur & Uzuner, 2011). Topics that teachers should communicate about include teaching approaches, objectives, expectations, areas of responsibility, and role clarity (Bouck, 2007; Härkki et al., 2021; Zach, 2020).

Fifth, it is important to take note of teaching intensity during the intervention. It takes time to build a constructive relationship among team teachers (Friend et al., 2010; Pratt, 2014). Teachers must be given ample opportunities to grow together as a team and to establish an appropriate team-teaching relationship. For experimental research, this implies

that, when new teaching teams are established, teachers need time to get to know each other; alternatively, researchers may prefer to focus on experienced team-teaching teams that have already established a strong relationship. Dietrichson et al. (2020) differentiated three variables to capture intensity: the duration of the intervention period, the frequency of team-teaching sessions, and the duration of team-teaching sessions.

Contextual Factors. The third category covers contextual factors, which can be grouped under the class, school, and system levels. On the class level, the class composition and size may influence team-teaching effectiveness. In addition to the number of students in a class, student characteristics and the variation in student needs are possible confounders (Pearl et al., 2012). Educational effectiveness research shows that various student characteristics (e.g., IQ, gender, social background, language deficiency) may influence the intended outcome measures (Reynolds et al., 2014).

On the school level, school culture and climate can play important roles in successful collaborative practices (Kyndt et al., 2016). Team teaching is an educational approach that has to be accepted and supported by more than the individuals directly involved in the teaching partnership (Rexroat-Frazier & Chamberlin, 2019). Schools with a safe climate where teachers can raise concerns or express conflicting views without fear of being ignored may be more suitable for effective team teaching practices (Härkki et al., 2021; Vesikivi et al., 2019). In establishing this climate, the school administration plays a pivotal role. School administrators should encourage teacher collaboration and assist team teachers with program scheduling as well as provide incentives and resources that allow team teachers to design and reflect on desired changes to the ways they provide services (Heisler & Thousand, 2021). Also, the school infrastructure should have classrooms large enough to accommodate multiple teachers (Baeten & Simons, 2014).

Finally, previous research refers to possible differences between studies on the effectiveness of team teaching that may depend on the education system in which the studies take place (Szumski et al., 2017). These differences may result from the way team teaching is understood and the purpose for which it is deployed, the duration of experience with team teaching implementation, and the consistency of the educational policy in favoring team teaching in schools.

Methodological Quality Characteristics

Besides embodying conceptual clarity, experimental studies into the effectiveness of team teaching must meet requirements for methodological quality. Different quality-appraisal tools for experimental research in education exist, including the Checklist for the Rigor of Education-Experiment Designs (CREED). This critical-appraisal tool was developed by Sung et al. (2019) and focuses on the assessment of three types of validity: internal validity (are alternative explanations for the effect ruled out?); construct validity (is it possible to generalize from the research operations to underlying constructs?); and statistical conclusion validity (is there an association between cause and effect?). These domains are evaluated according to six criteria, respectively addressing (1) the type of experimental design, (2) the methods for establishing baseline equivalence, (3) the number of participants in each group, (4) the reliability and validity of the measurements, (5) the fulfillment of statistical assumptions, and (6) the reporting of effect sizes. In addition, the level of rigor and the statistical power of experimental studies can be determined.

Despite the fact that CREED is a comprehensive and easy-to-use quality-appraisal tool, in this review study, some adjustments were made to the original meaning Sung et al. (2019) gave to the criteria. Next, we clarify each criterion.

Experimental Design Type

Three variants of experimental research designs are distinguished (i.e., randomized experiments, quasi-experiments, and pre-experiments), based on two design elements (i.e., random assignment and group comparisons) (Shadish et al., 2002; Sung et al., 2019). A randomized experiment assigns units to experimental and control group(s) based on chance (i.e., random assignment). A quasi-experiment uses a control group but lacks random assignment. A pre-experiment has no control or comparison group and no randomization.

In general, experimental studies with a comparative factor and random assignment are more convincing and rigorous than those without (Hammersley, 2008). The comparative factor delivers information about the counterfactual inference, which means that a situation is created from what would have happened to the same or a similar group of people in the absence of the treatment (Shadish et al., 2002). The difference between these situations provides evidence from which causal effects can be determined (Gopalan et al., 2020). A randomization process assures that variables, known and not known, that may theoretically affect the treatment outcomes are equally distributed between the experimental and control groups (Sullivan, 2011).

Methods for Baseline Equivalence

Experiments should start with the aim of creating equal groups to eliminate the threat of extraneous variables. In this regard, a key factor in experimental research is the principle of baseline equivalence, in which methods are used to determine whether the intervention group and the control group had similar enough characteristics at the beginning of the study (What Works Clearinghouse, 2020). If the two groups differ at the start of the study in crucial traits that could influence outcomes, the effect observed at the end could be due to the disparities that already existed (Steeger et al., 2021). A wide range of methods for ascertaining baseline equivalence can be used (Anderson & Maxwell, 2021). As already mentioned, in randomized experimental studies, randomization procedures ensure equivalence. Quasi-experimental

studies can use, for example, the principle of matching. In this case, participants are matched according to a certain variable of interest, and one member of each matching pair is subsequently assigned to the control group while the other is assigned to the experimental group. Another option is to use counterbalancing to compare the same participants under different conditions. Alternatively, a pretest with equating measures can ensure equality between groups (e.g., a t-test can confirm that no differences exist between groups on pre-test scores; analysis of covariance can adjust post-test scores according to pre-test scores or according to other possible confounding variables).

Number of Participants

To derive reliable inferences from study findings, an adequate sample size must be determined (Memon et al., 2020). Although there is no clear-cut answer to the question of whether the size of a sample is sufficient (e.g., for random assignment to be operable), it was determined that a minimum of 30 participants per group is required based on recommendations by Chang et al. (2006) and Fraenkel et al. (2019).

Apart from these guidelines, power analyses can calculate a more accurate sample size to find a certain effect size, for a certain amount of statistical power, with a suggested statistical significance level and the statistical test used. More details on power analyses are provided below.

Reliable and Valid Measurements

To claim valid and reliable measurements, relevant evidence should be collected and assessed to determine the degree to which that evidence supports the intended meaning of a measurement's scores and inferences about the characteristics it was designed to measure (Cook et al., 2015; Hess & Kvern, 2021). It should be recognized that reliability and validity characterize not the test itself, but rather how a test is used in a certain situation (Kane, 2013). The psychometric properties of a test can vary substantially depending on the purpose, setting,

and population, meaning that the same instrument that led to reliable and valid results when used in one situation can be misused in another, leading to unreliable or invalid results (Cook, 2014).

To identify the most important pieces of evidence, Kane (2013) refers to four types of inferences: scoring, generalization, extrapolation, and decision-making. The scoring inference requires evidence of the extent a score adequately represents key components of the observed performance. Generalization entails drawing conclusions from observed scores to universal scores of all possible observations across tasks, conditions, and/or raters. Extrapolation starts from the assumption that interpretations of scores in the test setting can be extrapolated to performances in other contexts (i.e., in a larger domain). Finally, decision rules ensure that the interpretation of scores is an appropriate source for meaningful decisions (high-stakes or low-stakes).

Fulfillment of Statistical Assumptions

To draw valid inferences from the results of statistical tests and provide solid underpinning for causal claims, the fulfillment of statistical assumptions should be addressed (Hoekstra et al., 2012). If statistical assumptions are violated, the probability of a test statistic may be inaccurate, distorting Type I or Type II error rates (Erceg-Hurn & Mirosevich, 2008). In turn, this results in incorrect *p* values and effect sizes (Hu & Plonsky, 2021; Osbourne & Waters, 2002). However, multiple studies in the field of educational research (e.g., Hoekstra et al., 2012; Lindstromberg, 2016) have noted that researchers often do not check for violations of assumptions and that the scientific community tends to tolerate this.

In response, Nimon (2012) offered a checklist to support researchers in reviewing key statistical assumptions of commonly-used statistical tests. In compliance with Nimon's checklist, Table S1 (online only) presents, for each of the assumptions, when it should be

checked, how it can be examined, and how researchers can respond appropriately to any violations (see Nimon, 2012, for more details).

Reporting of Effect Sizes

To provide stakeholders with results that are practically significant, effect sizes should be reported (Schafer & Schwartz, 2019; Harrison et al., 2009). The effect size measures the effectiveness of a treatment by providing the magnitude of the effect as it unfolds (Ellis, 2010). Often, outcome scores are judged based on their statistical significance, where a significance level indicates whether a result occurred by chance. However, this kind of significance does not demonstrate the impact of a treatment in the real world. Especially in educational research, which has a broader audience that includes teachers and policymakers, there is a demand for meaningful research results. Research is practically significant when the magnitude of the effect found is noteworthy according to the researcher and research field (Mohajeri et al., 2020). To determine whether an effect is meaningful or not, two principal approaches exist. First, the comparative approach compares the effect of a study with previous effects in the same area of research. However, publication bias should be taken into consideration, as published effects can be much larger than what holds true for the population (Bakker et al., 2012). Second, when there is a lack of previous research, Cohen's conventional approach can serve as an alternative (Cohen, 1988). In this approach, effect sizes are categorized as small, medium, or large (see Table S2, online only). However, such thresholds have been criticized for their relativity to each other, the specific content, and the research method (Correll et al., 2020).

Level of Rigor

The level of rigor is a crucial factor in defining research quality (Ming & Goldenberg, 2021). The CREED flowchart distinguishes five levels of experimental rigor based on four of the criteria described above (see Figure 2). First, and following the experimental design type

of a study, the CREED flowchart sorts studies based on whether they make use of a control group and random assignment procedures. Second, if studies use random assignment, they are divided depending on whether their sample size was sufficiently large to make random assignment operable. Third, when studies have insufficient participants for a meaningful random assignment procedure, or when there is no random assignment at all, baseline equivalence serves as the dividing factor. Fourth, in all cases with at least a control group, reliable and valid measurements result in the final rating of a study.

Statistical Power

The power of a test is the probability of not making a Type II error. In other words, when a study has sufficient power, it ensures that there is a reasonable chance of finding effects, assuming those effects exist (Bartlett et al., 2022). The most commonly accepted minimum level of power is 0.80 (Cohen, 1992). If a test has 80% power, then it means that the test has an 80% chance of finding a difference of a given effect size if such a difference exists. Underpowered studies are problematic because they lead to biased conclusions (Anderson et al., 2017; Christley, 2010). First, underpowered studies have a relatively low probability of finding a statistically significant effect compared to a study with sufficient power (Crutzen & Peters, 2017). Second, underpowered studies that find significant results yield excessively wide sampling distributions for the sample estimates (Ioannidis, 2008). This means that all the parameters computed from the sample (e.g., effect sizes) can differ considerably from the population value. Furthermore, running an underpowered study might raise ethical concerns because it requires investments from participants in a study whose ability to generate insights might be limited. Also, overpowered studies might bring some issues as, for example, discovered associations can contain deflated effect sizes compared with the true ones (Ioannidis, 2008).

Previous power surveys in the social and behavioral sciences have indicated that samples generally do not provide adequate power for finding small and medium-size effects (Aberson, 2019; Ellis, 2010). As power appears low and publication bias remains an issue, the published literature may provide a skewed view of true population effects.

Method

To answer the research questions, a thorough review of the literature on team teaching was conducted. To support this process, the Preferred-Reporting of Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (Page et al., 2021) was used. PRISMA provides a four-phase diagram and corresponding checklist to systematically search the literature and include studies that meet the suggested criteria. All selected studies were coded and categorized according to a coding scheme. In the process, they were all subjected to a conceptual and methodological evaluation.

Search and Inclusion

Figure 3 summarizes the whole search and inclusion process. The following paragraphs provide a more detailed description of each phase.

Identification

The first phase was a search for relevant studies. This search was executed in January 2022 in three databases, namely Web of Science, Scopus, and ProQuest (ERIC, Publicly Available Content Database). To cover as many eligible articles as possible, a series of search terms was used. The first set of search terms focused on the selection of studies on team teaching and entailed the following keywords: “team teaching,” “co-teaching,” “collaborative teaching,” “cooperative teaching,” and “paired placement.” The second set of search terms was intended to restrict the results to studies that were carried out in a school or educational environment. The queries used were “school,” “education,” “classroom,” “instruction,” and “teaching.” The third and final set of search terms aimed to narrow the remaining group down

to studies that were labeled as experimental research. Search terms related to experiments were added: “experiment,” “experimental,” “quasi-experiment,” “quasi-experimental,” “randomized trial,” “randomised trial,” “control group,” and “control condition.”

Besides the search terms, the following criteria for inclusion were applied:

- To yield an overview of relatively recent literature, studies had to be included in the database between January 2000 and January 2022. Also, more recent review studies in the field of team teaching indicate that there may be an increase in the use of experimental studies during the 21st century (e.g., King-Sears et al., 2021).
- To keep the search process manageable, and because our primary concern was to offer insight into the credibility of published research of high academic quality, we restricted the search to “peer-reviewed journals” and “academic journals.” This implies that unpublished manuscripts, online reports, book chapters, master’s theses, doctoral dissertations, conference abstracts, and other gray literature were not considered for inclusion. According to Alexander (2020), the inclusion of nontraditional research sources has ‘pros’ and ‘cons’. For example, one of the counterarguments is that there may be variable (overall) research quality, making it difficult to assess these nontraditional works. On the other hand, we admit that there may be publication bias, indicating a potential overemphasis on positive significant effects in our included studies.
- Manuscripts had to be written in English, and their full text had to be available.

Screening

The second phase entailed a preliminary review of the identified studies based on title and abstract. At first, all identified articles were uploaded into Endnote 20 software. In the Endnote library, duplicates were removed. In screening titles and abstracts, several criteria for exclusion were applied. Studies were excluded in instances that had (1) no mention of collaboration between teachers, (2) no direct link to a school environment, or (3) no reference

to any sort of experimental research design. Regarding the population, no exclusion criteria were used for educational level (pre-primary, primary, secondary, higher, or teacher education), nor were there any exclusion criteria for personal or contextual characteristics of students and teachers. Also, all types of outcomes were eligible for inclusion. A double-screening of all abstracts resulted in a 97% agreement rate.

Eligibility

In the third phase, the remaining studies' full texts were thoroughly screened to further evaluate eligibility. Beyond the screening-phase criteria, three extra exclusion criteria were applied. First, studies had to describe team teaching in conformity with the definition cited in the introduction of this paper (i.e., collaboration between two or more teachers in the planning, delivery, and/or evaluation of a course; Baeten & Simons, 2014). Concepts that deviated too much from the original meaning of team teaching were excluded. Second, the objective of studying the effects of team teaching had to be explicitly stated. Studies that investigated the impact of a larger program that happened to contain team teaching were not included. These studies often evaluate the effect of the program as a whole and not necessarily the effect of team teaching. Third, there had to be a clear link to the use of an experimental research design.

The screening and eligibility procedures were performed by the first author. When there was uncertainty about the inclusion or exclusion of a given article, a discussion with the co-authors followed. For example, in the beginning, it was uncertain whether to include so-called "design experiments" (Bradley & Reinking, 2011). There are contradictory statements concerning whether or not this type of study is experimental (Cohen et al., 2018). After a discussion, all studies that were originally labeled as an experiment were deemed eligible for inclusion. Appraisal tools like CREED lend support to this choice by providing a subsection to accommodate studies of this type (i.e., pre-experiments).

In total, 66% of the studies that remained after the screening phase were excluded. First, 15% of these studies were excluded because of an incompatible definition (e.g., Zhang et al., 2016). Second, in 25% of these studies, it was not clear whether they examined the effects of team teaching (e.g., Jang, 2010). Finally, 60% of these studies were excluded for being typical examples of case studies or action research (e.g., Gurgur & Uzuner, 2011).

Inclusion

In the fourth and final phase, the studies that survived the eligibility process were listed and retained to answer the research questions. All included studies can be found in the reference list of this paper, where they are marked with an asterisk. In addition, Table S3 (online only) provides an overview of all included studies with their main demographic characteristics (e.g., the country in which they were conducted).

Coding Scheme

To categorize all included studies and assess them from a conceptual and methodological perspective, a coding process was performed in NVivo 12 software. First, the conceptual and methodological elements of each article were coded by the first author. Next, the coding process was collaboratively discussed with the co-authors, and a consensus was reached. Conceptual elements were coded according to the aspects of the conceptual framework (see Figure 1). For methodological quality, all studies were coded according to the CREED criteria and corresponding flowchart (see Figure 2). Some elements and criteria require more explanation according to the coding process than is provided in previous sections. Therefore, a special note follows for the coding of (1) the use of reliable and valid measurements, (2) fulfillment of statistical assumptions, and (3) statistical power.

Reliable and Valid Measurements

Included studies were categorized according to the extent to which they described a validation process. In line with ideas from the validity framework of Kane (2013) and criteria

presented by Sung et al. (2019), this review focused on three pieces of evidence for quality use and interpretation of measurements. First, studies had to define the construct and describe the extent to which items were representative and relevant to capturing that construct. Often, evidence is provided by a panel of experts who check the relevance and representativeness of items. Second, the alignment between the construct and the chosen measurement tool had to be described, for example, examining whether items correlated as expected, related to a single factor, and were unrelated to a different set of items. For this purpose, information about factor analysis or item response theory models could serve as evidence. Third, the reproducibility of the scores had to be described by providing examination evidence of reliability coefficients. For this purpose, minimum standards from the What Works Clearinghouse standards handbook, Version 4.1 (What Works Clearinghouse, 2020), were used. These standards suggest that measurements can be classified as reliable when studies report an internal consistency of measures with a Cronbach's alpha of 0.50 or higher, a test–retest reliability score of at least 0.40, or an inter-rater reliability (such as percentage agreement, correlation, or kappa) of 0.50 or higher.

Existing evidence from previous research was considered acceptable when it was gathered in a similar setting, with a similar group of participants, and for the same purpose.

Fulfilling Statistical Assumptions

First, the statistical test(s) a study used and the described information on statistical assumptions were coded. Subsequently, a study was categorized as fulfilling the statistical assumptions after checking the necessary assumptions according to Nimon's framework and verifying that any violations had been appropriately addressed.

Statistical Power

To retrospectively calculate statistical power, the procedure of Cohen (1962, as cited in Ellis, 2010) was followed. First, the sample size and statistical test were recorded for each

included study. Statistical tests were assumed to be nondirectional (two-tailed). Second, given the above information and assuming alpha levels of 0.05, a calculation was carried out on the minimum statistical power of each study relevant to the observation of hypothetical effects. Due to a lack of previous quantitative studies on the effectiveness of team teaching, Cohen's conventional approach was used. Third, the results were averaged across all studies to get the mean power scores for observing small, medium, and large effects. Because of possible inflations of the average power levels due to outliers, median scores were also calculated.

The power analysis was carried out in R software, with the aid of the 'pwr' package (Champely et al., 2020). Furthermore, as some studies may have chosen a clustered randomized protocol, the 'WebPower' package (Zhang & Yuan, 2018) was applied. The clustered randomized experiment can be of value in team teaching research because it considers team teaching as a class-level intervention and, therefore, the entire cluster can be randomly assigned to an experimental condition (Liu, 2013; Raudenbush, 1997). Finally, power was not calculated for single-case designs because of the probability that results would be close to zero (Bouwmeester & Jongerling, 2020).

Findings

To provide context for the conceptual and methodological findings, Figure 4 reveals the evolution of experimental studies on the effectiveness of team teaching. A significant increase in the use of these types of studies can be noticed since 2015. From the period between 2015 and 2022, 25 experimental studies were found, in contrast to the period between 2000 and 2015, when only six experimental studies were found. However, there are observable differences according to geographical area. Asian countries have contributed the most to experimental research on the effectiveness of team teaching. Of the studies conducted between 2015 and 2022, 64% were conducted in Asia (n = 16). Next were studies from European countries with a proportion of 20% (n = 5), followed by studies from the US and

Australia with 8% each (n = 2). Of the studies published before 2015, 50% were conducted in Asian countries (n = 3), 33% in the US (n = 2), and one study (17%) in Europe (viz., in Belgium).

Conceptual Evaluation (RQ1)

Units

Across the 31 included studies, units were defined 35 times (due to studies that examined effects on both students and teachers). Table 2 provides the numbers of studies respectively examining the effects of team teaching on students, teachers, and others, for each educational level.

Students were chosen as units in the largest percentage of cases (80%). Within this group, students in secondary education (e.g., Migdadi & Baniabdelrahman, 2016; Shein & Tsai, 2015) received slightly more attention than students in primary (e.g., Ali et al., 2021; Bardaglio et al., 2015) or higher education (e.g., Una, 2016; Wang et al., 2019). Other educational settings examined were a summer school and a language institute (Chandler-Olcott, 2017; Yeganehpour & Zarfsaz, 2020). Studies that used teachers as units were limited and mostly embedded in higher education (e.g., McKenzie et al., 2022; Sharma et al., 2021). There was also one study examining the effects of team teaching on mentors of student teachers (Simons & Baeten, 2016).

Team Teaching as a Treatment

Regarding team composition, 90% of the studies described a specific collaboration between two or more teachers. Table 3 shows that experimental studies examined the effects of two sorts of team-teaching compositions. The first is teams consisting of teachers of the same type. Within this group, a collaboration of general education teachers was the most prevalent form (e.g., Jang, 2006; Migdadi & Baniabdelrahman, 2016). The second category covers mixed teams of educational professionals. Usually, this comprises a general-education

teacher in a leading role combined with a special-education teacher supporting students with disabilities (e.g., Lehane & Senior, 2020; Tremblay, 2013; Welch, 2000).

To give an idea of the versions of team teaching in practice, Table 4 provides an overview of the cases in which team-teaching models were defined. More than half of the studies did not refer to any model (e.g., Besharati & Mazdayasna, 2017; Maharani et al., 2019a). When studies did describe the model(s) used, they often used a mix of models. In these cases, teachers could generally choose or experiment with different models during the intervention period (e.g., Boland et al., 2019; Muhammad & Jahan, 2019). But some studies explicitly examined team teaching in the guise of one or more specific models. The parallel teaching model was mentioned most frequently (e.g., Aliakbari & Bazyar, 2012; Una, 2016), followed respectively by the sequential teaching model and the assistant teaching model (e.g., Ali et al., 2021; Ansari & Wahyu, 2017). Station teaching and teaming were each mentioned only once (Aliakbari & Nejad, 2013; Jang, 2006). The observation model and the coaching model were not mentioned.

With respect to phases, as shown in Table 5, most of the studies considered at least the planning and teaching phases when defining team teaching as a treatment (e.g., Al-Saaideh & Al-Zyoud, 2015; Boland et al., 2019). Conversely, barely one-quarter of the studies mentioned the evaluation/reflection phase (e.g., Baeten & Simons, 2016b; Sharma et al., 2021). When the evaluation/reflection phase was described, the purpose was mostly to reflect or deliver feedback on team teaching practices (n = 6; e.g., Hadley et al., 2000; Welch, 2000) or to evaluate student performances and/or learning (n = 2; Aliakbari & Nejad, 2013; Stapleton et al., 2021).

Regarding the control group, a key characteristic in experimental research, generally a team-teaching context was compared with a solo-taught context (e.g., Rao & Yu, 2021; Yeganehpour & Zarfsaz, 2020). Some cases, however, compared two different models as well

(e.g., sequential teaching versus parallel teaching; Baeten & Simons, 2016b) or different team-teaching compositions (e.g., qualified partner versus unqualified partner; Andersen et al., 2018).

Outcomes

Among students, outcome measures were related to language proficiency (n = 13; e.g., Aliakbari & Nejad, 2013; Besharati & Mazdayasna, 2017), mathematical (n = 9) or scientific (n = 1) skills (e.g., Ansari & Wahyu, 2017; Shein & Tsai, 2015), or coordinative motor skills (n = 1; Bardaglio et al., 2015). In addition, students' experiences of a given lesson (period) and student attendance (n = 4) were studied (e.g., Al-Saaideh & Al-Zyoud, 2015; Saeed et al., 2018). For (student) teachers and mentors, observations of outcomes mostly focused on perceptions or attitudes regarding the team-teaching practices performed (n = 7; e.g., Simons & Baeten, 2016; Welch, 2000). In cases of student teachers, these outcomes were supplemented with measurements of teaching efficacy or readiness to teach (n = 2; Sharma et al., 2021; Stapleton et al., 2021).

Setting

In terms of team-teacher factors, individual teacher characteristics were only described in a limited way. Teacher competencies as reflected in qualifications obtained were reported in 16 studies (52%; e.g., Bardaglio et al., 2015; Lehane & Senior, 2020), and general teaching experience was described in 14 studies (45%; e.g., Andersen et al., 2018; Tremblay, 2013). Only two studies (6%) reported experience in team teaching (Hadley et al., 2000; Lehane & Senior, 2020). In one of these studies, teachers had fewer than three years of team teaching experience. Meanwhile, 23 studies (74%) explicitly reported information about the assembling of a new team (e.g., Rao & Yu, 2021; Shein & Tsai, 2015). One study (3%) reported personality characteristics (Jang, 2006). This description referred to mildness, temper, and forthrightness in communication. On the team level, six studies (19%) reported

information about relationship-building between teachers prior to the intervention but did not provide any details (e.g., Ghanaat Pisheh et al., 2017; Stapleton et al., 2021).

In terms of the intervention period, five studies (16%) explicitly reported that teachers volunteered to participate in a teaching team (e.g., Hadley et al., 2000; Lehane & Senior, 2020). Thirteen studies (42%) reported some sort of training prior to the intervention. These training sessions were mostly organized to inform team teachers about basic concepts relevant to implementing team teaching in teaching practices (n = 12, 92%; e.g., Ali et al., 2021; Jang, 2006; Rao & Yu, 2021). In addition, information about the lesson content(s) and specific research goals was presented during training sessions (n = 6, 46%; e.g., Bardaglio et al., 2015; Muhammad & Jahan, 2019). Other opportunities provided during training sessions were relationship-building activities (n = 4, 31%; e.g., Ghanaat Pisheh et al., 2017; Stapleton et al., 2021) and observations of others' team-teaching practices (n = 3, 23%; e.g., Lehane & Senior, 2020; Welch, 2000). In two studies, extra guidance was provided for team teachers during the intervention period (Tremblay, 2013; Welch, 2000). Seven of the 13 studies (54%) that mentioned a training period also described the time intensity of the program (e.g., Lehane & Senior, 2020; Muhammad & Jahan, 2019). In most cases, three or fewer training sessions lasting a few hours each were organized.

Seventeen studies (57%) reported information on the collaboration and communication between teachers during the intervention, referring to discussion about the teaching objectives, teaching materials, forms of team teaching activities, and classroom arrangement (e.g., Aliakbari & Nejad, 2013; Boland et al., 2019). Also mentioned were reflections on teaching experience, exchange of ideas on teaching methods, comparison of cultural and educational differences to reach a mutual understanding, and conversations about responsibility and division of roles (e.g., Baeten & Simons, 2016b; Sharma et al., 2021). However, in 23 studies (74%), no planning or reflection time was reported. Of the remaining

eight studies, three reported weekly planning sessions (Besharati & Mazdayasna, 2017; Hadley et al., 2000; Rao & Yu, 2021). Another three studies reported that planning time was provided before each lesson (Boland et al., 2019; Chandler-Olcott, 2017; Muhammad & Jahan, 2019). The final two studies reported monthly planning (Al-Saaideh & Al-Zyoud, 2015; Welch, 2000).

In addition, 21 studies (68%) mentioned the duration of the intervention (e.g., Jang, 2006; McKenzie et al., 2022). Table 6 presents the durations of the respective interventions and the intensity of team teaching practices, expressed in hours per week. Usually, an intervention period lasted approximately one semester, with multiple team-teaching hours per week.

Lastly, class variables regarding student characteristics were often mentioned in studies on team teaching. Information on students (e.g., IQ) was also collected when matching procedures were used to allocate students into experimental and control group(s). In addition, 22 studies (71%) described class size. Eleven of these studies involved 20–30 students each (e.g., Andersen et al., 2018; Boland et al., 2019), seven studies involved more than 30 students each (e.g., Al-Saaideh & Al-Zyoud, 2015; Shein & Tsai, 2015), and four studies had fewer than 20 students each (e.g., Aliakbari & Bazyar, 2012; Tremblay, 2013). Variables concerning the school level were almost never described. Regarding the class infrastructure, only one study explicitly stated that the school did not have sufficient accommodations for team teaching practices (Jang, 2006). Concerning school administration support, descriptions were limited to receiving permission from the school leader to execute the research (n = 5, 16%; e.g., Muhammad & Jahan, 2019; Saeed et al., 2018).

Methodological Evaluation (RQ2)

Experimental Design Type

Figure 5 shows that quasi-experimental studies make up the largest group of included studies (68%). Randomized experiments and pre-experiments each account for a smaller proportion (23% and 9%, respectively). This means that more than 90% of the identified experimental studies on the effectiveness of team teaching made use of either a control group or a comparison group (e.g., Andersen et al., 2018; Besharati & Mazdayasna, 2017). However, almost 80% of the studies did not make use of random assignment procedures. Two of those studies did describe the reasons why random assignment was not valuable or convenient (Lehane & Senior, 2020; Yeganehpour & Zarfsaz, 2020). Reasons cited included ethical or practical issues (e.g., not disturbing the school's academic programs).

Some studies used a mixed-method approach to enumerate reasons for the success or failure of the experimental intervention (e.g., Aliakbari & Bazyar, 2012; Jang, 2006). In these cases, a variety of data collection approaches were used, such as interviews with teachers, logbooks about teachers' perceptions and feelings, and videotaped records of teaching implementation in practice. Supplementary questionnaires on students' perceptions of team teaching were also used.

Methods for Baseline Equivalence

Table 7 shows how many studies used each of several methods for baseline equivalence. Almost three-quarters of the included experimental studies indicated baseline equivalence. In most of these studies, a t-test on pre-test scores was used to confirm that groups were equal (e.g., Hadley et al., 2000; Rao & Yu, 2021). These t-test measures were often preceded by some sort of matching process. In almost one-fifth of the studies, the comparison was between a certain group of participants in a non-treatment situation and the same group of participants in a situation with the team teaching treatment (e.g., Saeed et al., 2018; Wang et al., 2019). In fewer than 10% of the studies (Baeten & Simons, 2016b; Shein & Tsai, 2015), statistical tests (such as ANCOVA) were used to control for confounding

variables (e.g., class size, teaching style, gender, academic ability, and socio-economic status).

Number of Participants in Each Group

Across the 31 included studies, the number of participants was defined 33 times (due to studies that examined effects on both students and teachers). One included study did not describe the number of participants involved (Maharani et al., 2019b). Table 8 presents the number of participants studied for each group of units. Studies that used students as participants (n = 20, 69%) were mostly limited to fewer than 50 participants apiece in the control and experimental groups (e.g., Lehane & Senior, 2020; Migdadi & Baniabdelrahman, 2016). In situations where teachers or mentors served as participants, all included studies contained fewer than 30 participants in each group (e.g., Al-Saaideh & Al-Zyoud, 2015; Chandler-Olcott, 2017).

Reliable and Valid Measurements

In general, 18 out of 31 studies (58%) did not provide sufficient information about the validation process for their measurements. Almost half of the studies (n = 16, 52%) did not provide reliability information. The 15 studies that did provide information on reliability usually calculated inter-rater reliability (n = 6; e.g., Bardaglio et al., 2015; Ghanaat Pisheh et al., 2017) or Cronbach's alpha (n = 5; e.g., Baeten & Simons, 2016b; Wang et al., 2019). Other studies (n = 4) referred to previous research where reliability was mentioned for the same group of participants, purpose, and setting (e.g., Hadley et al., 2000; Lehane & Senior, 2020; Muhammad & Jahan, 2019). Twenty-five studies (81%) reported content evidence by describing the extent to which items adequately represented the content of the construct in question (e.g., Ali et al., 2021; Migdadi & Baniabdelrahman, 2016; Shein & Tsai, 2015). Eight studies (26%) provided evidence about the alignment between the construct and the chosen measurement tool. Most often, a factor analysis was performed (n = 6; e.g., Shein &

Tsai, 2015; Wang et al., 2019), followed by models of the IRT (*n* = 2; Ali et al., 2021; Andersen et al., 2018).

Fulfillment of Statistical Assumptions

A total of 26 studies reported use of a statistical test. One study reported two tests, leading to a total of 27 tests requiring evidence of the fulfillment of statistical assumptions. Table 9 summarizes all tests reported and whether corresponding assumptions were fulfilled. T-test variations were most common (*n* = 20, 74%), followed by the OVA variations of tests (*n* = 4, 15%) and regression analyses (*n* = 3, 11%). Regarding the assumptions, only 30% of the tests (*n* = 8) that were used in the studies fulfilled the necessary assumptions (e.g., Besharati & Mazdayasna, 2017; Yeganehpour & Zarfsaz, 2020).

Reporting an Effect Size

Eight out of 31 studies (26%) reported an effect size (e.g., Bardaglio et al., 2015; Boland et al., 2019). Of the 23 studies (74%) that did not report an effect size directly, most (*n* = 14, 61%) did provide information for manual calculation (e.g., means and standard deviations).

Level of Rigor

As Figure 6 shows, most of the studies had a medium-low level of rigor (35%). These were all quasi-experimental studies without reporting of reliable and valid measurements (e.g., Jang, 2006; Rao & Yu, 2021). The studies with a low level of rigor (26%) included three pre-experiments and five quasi-experiments without methods for baseline equivalence (e.g., Maharani et al., 2019a; Welch, 2000). Fewer than half of the studies had a rigor rating equal to or above medium (39%). Within these higher-categorized groups of studies, the studies with a high level of rigor were all randomized experiments with a quality use of measurements and a sufficient sample size (e.g., Andersen et al., 2018; Stapleton et al., 2021). One randomized experiment did not fulfill standards for reliable and valid measurements and

was therefore categorized as medium-high (Una, 2016). Another randomized experiment did not have a sufficient number of participants but provided baseline equivalence and had quality measures and was thus categorized as medium (Migdadi & Baniabdelrahman, 2016). Five other quasi-experiments with quality use of measurements were also classified as medium (e.g., Lehane & Senior, 2020; Wang et al., 2019).

Statistical Power

Power analyses could be conducted for 25 out of the 31 studies. One of these studies mentioned an a priori power analysis to justify the sample size (Stapleton et al., 2021). The mean statistical power values for finding small, medium, and large effects were 0.25, 0.67, and 0.91, respectively. Median power levels were even lower for small and medium effects, while being slightly higher for large effects: 0.16 for finding small effects, 0.61 for medium effects, and 0.96 for large effects.

Effectiveness Variability (RQ3)

Taking into account the reported conclusions of all studies in relation to the reported conceptual and methodological characteristics, we found some variability in the concluded effectiveness of team teaching. In general, 90% (n = 28) of our included studies provided a conclusion supporting the effectiveness of team teaching. The other 10% (n = 3) did not conclude with a recommendation to choose team teaching over the more traditional solo teaching approach (Aliakbari & Bazyar, 2012; Aliakbari & Nejad, 2013; Stapleton et al., 2021). The studies that did not find evidence to support the effectiveness of team teaching had either a medium-low level of rigor (n = 2) or a high level of rigor (n = 1). None of these studies provided information about fulfilling statistical assumptions. The two studies with a lower level of rigor lacked statistical power to find medium effects, whereas the study with a high level of rigor did not have enough power to find small effects.

Most studies that concluded with a positive statement on the effectiveness of team teaching had a low or medium-low level of rigor (n = 17, 55%), without reporting effect sizes (n = 21, 75%), and fulfilling statistical assumptions (n = 20, 65%). Four of the included studies that mentioned a positive impact of team teaching found that all team-teaching models can be beneficial in teaching practices (e.g., Boland et al., 2019; Rao & Yu, 2021). Nevertheless, there were differences in the benefits reported when different models were used (e.g., students indicated that the parallel model was more beneficial than the sequential model; Baeten & Simons, 2016b). Another study mentioned a positive impact of team teaching, regardless of the composition of the team-teaching team (i.e., qualified partner versus unqualified partner; Andersen et al., 2018).

Among the studies with a high level of experimental rigor (n = 5), four reported a positive conclusion regarding the implementation of team teaching in teaching practices. However, only one of these most rigorous studies reported on all aspects of team teaching as a treatment (i.e., team composition, team-teaching models, and phases of the teaching practice; Muhammad & Jahan, 2019). None of the studies with a rigorous methodological design managed to report all influencing variables within the larger study setting (i.e., team teacher factors, interventional factors, and contextual factors).

Discussion and Conclusion

The aim of this study was to review the conceptual and methodological credibility of experimental research designs in research on the effectiveness of team teaching. A total of 31 experimental studies were (1) examined according to a conceptual framework and (2) critically assessed using an adapted methodological quality appraisal tool. In the course of this examination and assessment, we transferred two existing frameworks, one due to Cronbach (1982) and one due to Sung et al. (2019), to the team-teaching context and actualized them based on recent and widely-cited scientific knowledge. The combination of the two

frameworks makes an important contribution to the field of team teaching (and even beyond) because of its originality and its relative simplicity for evaluating and designing experimental studies from a conceptual and methodological perspective, both of which perspectives are inherently associated with high-quality research (Ming & Goldenberg, 2021).

The first general finding of this review concerns a noticeable increase in the use of experimental studies in the field of team teaching since 2015. Gopalan et al. (2020) discovered a similar trend in the broad field of educational research. However, the latter trend started more than half a decade earlier, in 2009. This indicates that research on team teaching lags behind the broad field of educational research in the evolution of experimental designs. This result may be explained by the fact that experimental research on team teaching is difficult to conduct because of various influencing variables affecting study results.

Conceptual Credibility

From a conceptual point of view, experimental researchers must consider many variables. Based on four elements (units, outcomes, treatment, and setting) presented by Cronbach (1982), a framework (see Figure 1) was developed to group these variables in order to assess the conceptual credibility of experimental studies on the effectiveness of team teaching.

Results show that students were the most studied units in experimental studies on team teaching. This finding contrasts with other research performed in the field of team teaching. Generally, teachers' perceptions are collected using a self-reporting measurement tool (Walsh, 2020). When students are examined in experimental studies, test scores on language or mathematical proficiency are usually used. Similar outcome measures were found in a meta-analysis on co-teaching by King-Sears et al. (2021). Cognitive outcomes are represented far more than non-cognitive outcomes, such as student attendance. For teachers, meanwhile,

previous research indicates multiple advantages, such as an increase in personal and professional growth, but there is no experimental research to support these claims.

Besides the impact of different units on outcomes, the way team teaching is implemented as an intervention may have an impact on the perceived results. Therefore, it is important to describe team compositions, team teaching models, and phases of teaching practice. First, with respect to team composition, equal teams are distinguished from mixed teams. Equal teams consisted mostly of two subject-specific teachers, and mixed teams were mostly related to co-teaching practices where a general-education teacher collaborates with a special-education teacher (Solis et al., 2012). This is an accurate reflection of current research within team teaching, where co-teaching has the upper hand. Second, regarding models, fewer than half of the studies referred to a team-teaching model in describing teaching practices. In line with Iacono et al. (2021), not describing teaching models results in a limited understanding of how team teaching appears in practice and makes it impossible to draw unambiguous conclusions from research findings. When models were described, teachers most often applied a mix of models. A smaller number of studies focused on one specific model. Among these cases, equal-status models were most common. Third, the planning and teaching phases were often described, and the evaluation/reflection phase was often ignored. The lack of attention to evaluation/reflection is in line with findings from previous review studies (Fluijt et al., 2016). However, all three phases, especially in long-lasting intervention periods, are critical for effective team teaching practices (Cook & McDuffie-Landrum, 2020).

Furthermore, three different sorts of setting factors may have an impact on how units perceive a treatment, how that treatment is administered, and even how reported variation in outcome measures is explained. Due to the lack of previous knowledge on the mechanisms underlying the (in)effectiveness of team teaching, experimental studies that provide insight into these aspects are of special interest for identifying why and how causal relationships hold

(Shadish et al., 2002). First, regarding team-teacher factors, only half of the studies described teacher competencies or teacher experience, which makes it difficult to gain insight into the compatibility of teaching teams and other relational factors (Pratt, 2014). Also, experimental studies do not satisfy requirements for describing interpersonal factors such as compatibility, mutual recognition, and collective mindset, although these are essential for effective collaboration and team teaching partnerships (Baeten & Simons, 2014; Pratt, 2014; Van Garderen et al., 2012). As team teaching practices can have various meanings and interpretations, extensive information on the characteristics of the teaching team is of special interest when researchers aim to generalize the results of their experiment to a broader population. Furthermore, the included studies were conducted in different continents worldwide, with the majority of studies conducted in Asia. This variety in geographical region and differences in education system may cause different interpretations on how team teaching is understood and the purpose for which it is deployed. Therefore, experimental findings must be contextualized within cultural properties and characteristics of the education system.

Second, as to interventional variables, training was mentioned in fewer than half of the studies, despite its importance in compensating for low levels of experience in team teaching. During the intervention period, more than half of the studies reported collaboration and communication about teaching practices and visions of teaching and learning. Interventional periods lasted for approximately one semester and, even though studies reported the importance of planning in team teaching, most of the studies did not mention planning time during this period. Third, concerning contextual factors, class composition was described mostly in co-teaching settings. Most classes consisted of 20–30 students. Finally, limited information was reported on school infrastructure and administration support.

From a conceptual perspective, the conclusion we can draw is that, so far, it remains a challenge to design experimental research that (1) provides sufficient information about how

team teaching practices occurred, (2) recognizes setting factors, and (3) deliberately delineates units and outcome measures. Corresponding facts have already been noted regarding other types of research designs, but the finding that this trend continues in experimental studies causes uncertainty about making statements on the effectiveness of team teaching. Also, definitive claims on moderating or mediating processes that may explain how effective team teaching practices occur, cannot be made based on the current experimental knowledge base. In this regard, it is strongly recommended to consider the presented influencing variables in the experimental process.

Methodological Credibility

From a methodological perspective, the quality of an experimental study stands or falls with its level of rigor (Ming & Goldenberg, 2021). Using an adapted version of the Checklist for the Rigor of Education-Experiment Designs (CREED; Sung et al., 2019), we assessed the level of rigor of experimental studies on the effectiveness of team teaching and found that 61% of all included studies had a medium-low or low level of rigor. This finding illustrates the limited progress in methodological quality of experimental studies on the effectiveness of team teaching, which was already indicated by previous researchers (e.g., Baeten & Simons, 2014; Friend et al., 2010; King-Sears et al., 2021; Murawski & Swanson, 2001; Rexroat-Frazier & Chamberlin, 2019). Two causes explain why these rigor ratings were so low even though most of the included experimental studies used a control group (90%) and methods for baseline equivalence (74%).

First, 69% of the studies were limited to fewer than 50 participants in the control or experimental group(s). When teachers served as units, the number was even lower, not even exceeding 30 participants in each group. Such a limited number of participants results in insufficient statistical power. The findings show that published research on the effectiveness of team teaching is underpowered, meaning that average statistical power levels are below the

recommended level of 0.80. This study found that the average statistical power for finding medium effects was 0.67. This means that an average experimental study on the effectiveness of team teaching ran the risk of overlooking medium effects 33% of the time. Furthermore, when published experimental research did find an effect, the size of this effect probably differed from the population effect size due to wide sampling distributions for sample estimates (Ioannidis, 2008). This lack of sufficient power is in line with previous power surveys in social and behavioral science, which found that samples generally do not provide adequate power for finding small and medium size effects (Aberson, 2019; Ellis, 2010).

Despite the strong emphasis on an adequate sample size, in many research situations it can be difficult to collect enough data to test the proposed (complex) hypotheses. As a result, researchers may have to work with data sets that are too small for the complexity of the statistical model they aim to use (van de Schoot & Miočević, 2020). For the sake of convenience, researchers could simplify the model they want to apply such that a smaller sample size suffices for the analysis. However, this may have consequences for testing the hypotheses of interest. van de Schoot and Miočević (2020) discuss guidelines and different solutions that allow researchers to apply an appropriate statistical model for their research situation and hypotheses when the sample is small (e.g., Bayesian estimation, unilevel design-based analysis for analyzing a single-case experimental design, item parceling; also see van de Schoot & Miočević, 2020, for more details).

The second reason for the lower rigor rating is insufficient information about the validation process to claim reliable and valid use of measurements. The experimental research included in this study frequently refers to existing measurement tools but lacks information about the validation process. This can be explained by researchers' perceptions that existing evidence is enough to justify claims about the quality of measurements. However, Cook et al. (2015) state that evidence applies only to the purpose, context, and population group in which

it was collected. Previous research might guide the choice for an assessment approach but does not support its future interpretations and use. Therefore, studies must examine how situational differences might influence the relevance of evidence.

Furthermore, fewer than half of the studies fulfilled statistical assumptions. This may have led to inaccurate *p* values and effect sizes (Hu & Plonsky, 2021). In this respect, research on team teaching does not deviate from educational research in general, where statistical assumptions are evaluated only to a limited extent (Hoekstra et al., 2012). However, to provide high-quality research, checking the fulfillment of statistical assumptions should become more common.

Finally, effect sizes were not reported directly 74% of the time, despite the fact that it has become more common to supplement reporting of statistical significance with effect sizes (Schäfer & Schwarz, 2019). A possible explanation for this might be that researchers continue to value statistical significance more than practical significance. However, extrapolating the meaning of effects to the real world is of absolute interest for policymakers and teachers who seek to implement team teaching in a sustainable way (Ellis, 2010).

Due to this lack of methodological rigor, it is recommended to design studies with an appropriate sample size (and apply the appropriate statistical techniques; van de Schoot & Miočević, 2020) and an instrumentation of sufficient psychometric quality to provide valid counterfactual inferences by using, for example, a well-established control group to ensure baseline equivalence. Only by enhancing their quality, experimental studies have the potential of informing stakeholders about the effectiveness of educational strategies, such as team teaching (Gopalan et al., 2020). If not, experiments with little methodological credibility due to, for example, problems in isolating the effects of a particular educational intervention from extraneous influences in the study setting (e.g., invalid use of control groups, randomization), can provide biased estimates of treatment effects.

Limitations and Recommendations for Future Research

This systematic review contributes to the literature on team teaching by providing insight into experimental research that examines the effectiveness of team teaching. Despite its important contributions to the field, some limitations need to be acknowledged. First, and as with all systematic reviews, the selection of databases and search terms can produce biased results. However, our search string contained commonly-used terms gathered from previous review studies on team teaching and experimental research. Second, by restricting our search to peer-reviewed journals, this review may have omitted relevant literature. However, gray literature was deliberately excluded for the purpose of focusing on studies of high academic quality, and to keep the search process and quality appraisal manageable. Nevertheless, as publication bias can be considered a problem in education literature (Polanin et al., 2016), future studies may also check the conceptual and methodological credibility of unpublished experimental research on the effectiveness of team teaching. This could potentially provide a more balanced view of the research quality of studies into the effectiveness of team teaching. Third, we did not address treatment fidelity directly, despite the fact that fidelity is a key factor in the experimental process because it exposes the extent to which the treatment was implemented as intended. Therefore, we recommend that researchers include fidelity assessment in their experiments. For example, Nelson et al. (2012) offered a five-step procedure for systematically assessing treatment fidelity. This procedure can be used in conjunction with our conceptual framework and the methodological quality criteria presented in this review study to design rigorous experimental studies. Fourth, by means of our conceptual framework, we present an overview of potential influencing factors that may moderate the effects of team teaching on certain outcomes. However, most of these factors are theoretical constructs, which have not been tested in experimental research. Therefore, it is an interesting avenue for future research to test whether the factors presented in our conceptual

framework are actual moderators or conditions for a successful implementation of team teaching. This way, future studies may gain more insight into the true causal effect of team teaching on educational outcomes.

Concluding Remarks

Planning and conducting education research is a complex process in which many different views of social reality feature (Cohen et al., 2018). In this regard, several paradigms exist, all of which have their own contribution in providing insight into educational practices. Considering education research from a more holistic point of view, Cohen et al. (2018) argue that education research will thrive most strongly when considering methodological, paradigmatic, and theoretical pluralism, in which strengths of alternative views are combined. To some extent, in this review study, we contribute to this approach by focusing not only on methodological requirements when conducting experimental studies to examine cause-and-effect relationships, but also by linking these to the conceptual understanding of the broader study context.

However, and although we solely focus on the potential of applying experimental research, comprehensive scientific knowledge should be the product of the dialogue between alternative research paradigms (Moss & Haertel, 2016). By this means, experimental studies are well-suited for testing hypotheses and causality, by mainly relying on positivist approaches (Creswell, 2014). Alternatively, qualitative research designs such as phenomenological or ethnographical research, and case studies may gain better understanding into how interventions work, what individuals experience during these interventions, and in constructing testable theories, through constructivist perspectives.

In addition, the objective of our review study was to enhance the knowledge base on the use of experimental studies in the research field of team teaching. However, the main underlying principles of both our conceptual and methodological framework can be applied in

a variety of contexts. Researchers from other fields may adapt our frameworks to their own needs. By this means, experimental research can be conducted that include the most important conceptual aspects when examining the effectiveness of a particular educational phenomenon (Cronbach, 1982; Shadish, 2011), and comply with requirements for methodological quality (Sung et al., 2019).

In conclusion, it is found to be a challenge to design experimental studies on the effectiveness of team teaching for all sorts of units and all sorts of outcome measures while clearly defining team teaching as a treatment and considering situational variables. Therefore, future experimental research should be explicit about conceptualizations and operationalizations of important variables related to team-teaching practices. Also, quality research designs must be used to present trustworthy and meaningful findings that either support or disconfirm the hypothesis that team teaching has a positive effect on students, teachers, and/or other stakeholders. Taken together, these points indicate that a potential next aim in research on team teaching may be to produce experimental studies with quality methodological designs and clear conceptualizations of units, team teaching as a treatment, outcome measures, and the larger study setting. The implementation of such studies in the field of team teaching could provide policymakers and teachers with the additional information to make evidence-informed decisions about team teaching practices. The conceptual framework and methodological quality criteria presented in this study can serve as a guideline for the development of these future experimental studies and thus be of interest to all experimental researchers in the field of team teaching. We therefore recommend that future researchers build on both frameworks to create better opportunities for cross-study comparability and cumulative science with a high utilization factor for educational practices.

References

References marked with an asterisk indicate studies included in the review. The full list of included studies is added as an online supplementary file.

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9781315171500

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research*, 90(1), 6–23. https://doi.org/10.3102/0034654319854352

*Al-Saaideh, M., & Al-Zyoud, N.-A. (2015). Benefits of teaching interdisciplinary subjects collaboratively in Jordanian pre-vocational education. *Educational Research and Reviews*, *10*, 2702–2712. https://doi.org/10.5897/ERR2015.2461

Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of General Psychology*, *4*(4), 337–353. https://doi.org/10.1037/1089-2680.4.4.337

*Ali, A., Ahmad, N., & Hussain, S. (2021). An experimental study of collaborative instructional strategy (CIS) for teaching mathematics at primary level in Pakistan. *Mathematics Teaching Research Journal*, *13*(1), 94–105.

*Aliakbari, M., & Bazyar, A. (2012). Exploring the impact of parallel teaching on general language proficiency of EFL learners. *Pan-Pacific Association of Applied Linguistics*, *16*, 55–71.

*Aliakbari, M., & Nejad, A. (2013). On the effectiveness of team teaching in promoting learners' grammatical proficiency. *Canadian Journal of Education*, *36*, 5–22.

Altstaedter, L., Smith, J., & Fogarty, E. (2016). Co-teaching: Towards a new model for teacher preparation in foreign language teacher education. *Hispania*, *99*, 635–649. http://doi.org/10.1353/hpn.2016.0108

*Andersen, S. C., Beuchert, L., Nielsen, H. S., & Thomsen, M. K. (2018). The effect of teacher's aides in the classroom: Evidence from a randomized trial. *Journal of the European Economic Association*, *18*(1), 469–505. https://doi.org/10.1093/jeea/jvy048

Anderson, M., & Maxwell, N. (2021). Baseline equivalence: What it is and why it is needed. *Mathematica*.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

*Ansari, B., & Wahyu, N. (2017). Mathematics understanding and anxiety in collaborative teaching. *Journal of Physics: Conference Series*, *943*. https://doi.org/10.1088/1742-6596/943/1/012040

Baeten, M., & Simons, M. (2014). Student teachers' team teaching: Models, effects, and conditions for implementation. *Teaching and Teacher Education*, *41*, 92–110. https://doi.org/10.1016/j.tate.2014.03.010

Baeten, M., & Simons, M. (2016a). Innovative field experiences in teacher education: Student-teachers and mentors as partners in teaching. *The International Journal of Teaching and Learning in Higher Education*, *28*, 38–51.

*Baeten, M., & Simons, M. (2016b). Student teachers' team teaching: How do learners in the classroom experience team-taught lessons by student teachers? *Journal of Education for Teaching*, *42*(1), 93–105. https://doi.org/10.1080/02607476.2015.1135226

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

*Bardaglio, G., Marasso, D., Magno, F., Rabaglietti, E., & Ciairano, S. (2015). Team-teaching in physical education for promoting coordinative motor skills in children: The more you invest the more you get. *Physical Education and Sport Pedagogy*, *20*(3), 268–282. https://doi.org/10.1080/17408989.2013.837434

Bartlett, M. E., Edmunds, C. E. R., Belpaeme, T., & Thill, S. (2022). Have I got the power? Analysing and reporting statistical power in HRI. *ACM Transactions on Human-Robot Interaction*, *11*(2), 1–16. https://doi.org/10.1145/3495246

*Besharati, M., & Mazdayasna, G. (2017). Investigating the effect of team-teaching approach on ESP students' English proficiency: Evidence from students' attitudes. *International Journal of Applied Linguistics and English Literature*, *6*(5), 41–50. https://doi.org/10.7575/aiac.ijalel.v.6n.5p.41

Bešić, E., Paleczek, L., Krammer, M., & Gasteiger-Klicpera, B. (2017). Inclusive practices at the teacher and class level: The experts' view. *European Journal of Special Needs Education*, *32*(3), 329–345. https://doi.org/10.1080/08856257.2016.1240339

*Boland, D. E., Alkhalifa, K. B., & Al-Mutairi, M. A. (2019). Co-teaching in EFL classroom: The promising model. *English Language Teaching*, *12*, 95–98. https://doi.org/10.5539/elt.v12n12p95

Bouck, E. C. (2007). Co-teaching … Not just a textbook term: Implications for practice. *Preventing School Failure: Alternative Education for Children and Youth*, *51*(2), 46–51. https://doi.org/10.3200/PSFL.51.2.46-51

Bouwmeester, S., & Jongerling, J. (2020). Power of a randomization test in a single case multiple baseline AB design. *PLOS ONE*, *15*(2). https://doi.org/10.1371/journal.pone.0228355

Bradley, B. A., & Reinking, D. (2011). Enhancing research and practice in early childhood through formative and design experiments. *Early Child Development and Care*, *181*(3), 305–319. https://doi.org/10.1080/03004430903357894

Brojčin, B., Bankovič, S., Glumbič, N., & Weiss, S. (2012). Effects of cooperative teaching in inclusive education. *Didactica Slovenica - Pedagoska Obzorja*, *27*(5), 66–79.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & de Rosario, H. (2020). *Pwr: basic functions for power analysis*. (Version 1.3-0). https://cran.r-project.org/package=pwr

*Chandler-Olcott, K. (2017). Co-teaching to support early adolescents' writing development in an inclusive summer enrichment program. *Middle School Journal*, *48*(1), 3–12. https://doi.org/10.1080/00940771.2017.1243916

Christley, R. M. (2010). Power and error: Increased risk of false positive results in underpowered studies. *The Open Epidemiology Journal*, *3*(1), 16–19. http://doi.org/10.2174/1874297101003010016

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Cook, D. A. (2014). When I say… validity. *Medical Education*, *48*(10), 948–949. https://doi.org/10.1111/medu.12401

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Cook, L., & Friend, M. (1995). Co-teaching: Guidelines for creating effective practices. *Focus on Exceptional Children*, *28*(3). https://doi.org/10.17161/foec.v28i3.6852

Cook, S. C., & McDuffie-Landrum, K. (2020). Integrating effective practices into co-teaching: Increasing outcomes for students with disabilities. *Intervention in School and Clinic*, *55*(4), 221–229. https://doi.org/10.1177/1053451219855739

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (4th ed.). Thousand Oaks, CA: Sage.

Creemers, B., & Kyriakides, L. (2007). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Crutzen, R., & Peters, G.-J. Y. (2017). Targeting next generations to change the common practice of underpowered research [Opinion]. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01184

de Jong, L., Meirink, J., & Admiraal, W. (2019). School-based teacher collaboration: Different learning opportunities across various contexts. *Teaching and Teacher Education*, *86*, 102925. https://doi.org/10.1016/j.tate.2019.102925

Dehnad, A., Jalali, M., Shahabi, S., Mojgani, P., & Bigdeli, S. (2021). Students' view on supportive co-teaching in medical sciences: A systematic review. *BMC Medical Education*, *21*(1). https://doi.org/10.1186/s12909-021-02958-4

Devecchi, C., & Nevin, A. (2010). Leadership for inclusive schools and inclusive school leadership. In A. H. Normore (Ed.), *Global perspectives on educational leadership reform: The development and preparation of leaders of learning and learners of leadership* (Vol. 11, pp. 211–241). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-3660(2010)0000011014

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7–12: A systematic review. *Campbell Systematic Reviews*, *16*(2). https://doi.org/10.1002/cl2.1081

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Esterhazy, R., de Lange, T., Bastiansen, S., & Wittek, A. L. (2021). Moving beyond peer review of teaching: A conceptual framework for collegial faculty development. *Review of Educational Research*, *91*(2), 237–271. https://doi.org/10.3102/0034654321990721

Fluijt, D., Bakker, C., & Struyf, E. (2016). Team-reflection: The missing link in co-teaching teams. *European Journal of Special Needs Education*, *31*(2), 187–201. https://doi.org/10.1080/08856257.2015.1125690

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed.). New York: McGraw-Hill.

Friend, M., Cook, L., Hurley-Chamberlain, D., & Shamberger, C. (2010). Co-teaching: An illustration of the complexity of collaboration in special education. *Journal of Educational and Psychological Consultation*, *20*(1), 9–27. https://doi.org/10.1080/10474410903535380

Friend, M., Embury, D. C., & Clarke, L. (2015). Co-teaching versus apprentice teaching. *Teacher Education and Special Education*, *38*(2), 79–87. https://doi.org/10.1177/0888406414529308

Gast, I., Schildkamp, K., & Van Der Veen, J. T. (2017). Team-based professional development interventions in higher education: A systematic review. *Review of Educational Research*, *87*(4), 736–767. https://doi.org/10.3102/0034654317704306

Gately, S. E., & Gately, F. J. (2001). Understanding coteaching components. *Teaching Exceptional Children*, *33*(4), 40–47. https://doi.org/10.1177/004005990103300406

*Ghanaat Pisheh, E. A., Sadeghpour, N., Nejatyjahromy, Y., & Mir Nasab, M. M. (2017). The effect of cooperative teaching on the development of reading skills among students with reading disorders. *Support for Learning*, *32*(3), 245–266. https://doi.org/10.1111/1467-9604.12168

Goddard, R., Goddard, Y., Sook Kim, E., & Miller, R. (2015). A theoretical and empirical analysis of the roles of instructional leadership, teacher collaboration, and collective efficacy beliefs in support of student learning. *American Journal of Education*, *121*(4), 501–530. https://doi.org/10.1086/681925

Gollwitzer, M., & Schwabe, J. (2022). Context dependency as a predictor of replicability. *Review of General Psychology*, *26*(2), 241–249. https://doi.org/10.1177/10892680211015635

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243. https://doi.org/10.3102/0091732X20903302

Gurgur, H., & Uzuner, Y. (2011). Examining the implementation of two co-teaching models: Team teaching and station teaching. *International Journal of Inclusive Education*, *15*(6), 589–610. https://doi.org/10.1080/13603110903265032

*Hadley, P. A., Simmerman, A., Long, M., & Luna, M. (2000). Facilitating language development for inner-city children: Experimental evaluation of a collaborative, classroom-based intervention. *Language, Speech, and Hearing Services in Schools*, *31*(3), 280–295. https://doi.org/10.1044/0161-1461.3103.280

Hammersley, M. (2008). Paradigm war revived? On the diagnosis of resistance to randomized controlled trials and systematic review in education. *International Journal of Research & Method in Education*, *31*(1), 3–10. https://doi.org/10.1080/17437270801919826

Hargreaves, A. (2019). Teacher collaboration: 30 years of research on its nature, forms, limitations and effects. *Teachers and Teaching*, *25*(5), 603–621. https://doi.org/10.1080/13540602.2019.1639499

Härkki, T., Vartiainen, H., Seitamaa-Hakkarainen, P., & Hakkarainen, K. (2021). Co-teaching in non-linear projects: A contextualised model of co-teaching to support educational change. *Teaching and Teacher Education*, *97*, 103–188. https://doi.org/10.1016/j.tate.2020.103188

Heisler, L. A., & Thousand, J. S. (2021). A guide to co-teaching for the SLP: A tutorial. *Communication Disorders Quarterly*, *42*(2), 122–127. https://doi.org/10.1177/1525740119886310

Hess, B. J., & Kvern, B. (2021). Using Kane's framework to build a validity argument supporting (or not) virtual OSCEs. *Medical Teacher*, *43*(9), 999–1004. https://doi.org/10.1080/0142159X.2021.1910641

Hoekstra, R., Kiers, H. A., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*, Article 137. https://doi.org/10.3389/fpsyg.2012.00137

Honigsfeld, A., & Dove, M. G. (2019). Preparing teachers for co-teaching and collaboration. In L. C. de Oliveira (Ed.), *The Handbook of TESOL in K-12* (pp. 405–421). Wiley. https://doi.org/10.1002/9781119421702.ch26

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, *37*(1), 171–184. https://doi.org/10.1177/0267658319877433

Iacono, T., Landry, O., Garcia-Melgar, A., Spong, J., Hyett, N., Bagley, K., & McKinstry, C. (2021). A systematized review of co-teaching efficacy in enhancing inclusive education for students with disability. *International Journal of Inclusive Education*, 1–15. https://doi.org/10.1080/13603116.2021.1900423

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

*Jang, S.-J. (2006). Research on the effects of team teaching upon two secondary school teachers. *Educational Research*, *48*(2), 177–194. https://doi.org/10.1080/00131880600732272

Jang, S.-J. (2010). The impact on incorporating collaborative concept mapping with coteaching techniques in elementary science classes. *School Science and Mathematics*, *110*(2), 86–97. https://doi.org/10.1111/j.1949-8594.2009.00012.x

Johnson, T. M., & King-Sears, M. E. (2020). Eliciting students' perspectives about their co-teaching experiences. *Intervention in School and Clinic*, *56*(1), 51–55. https://doi.org/10.1177/1053451220910732

Jortveit, M., & Kovač, V. B. (2021). Co-teaching that works: Special and general educators' perspectives on collaboration. *Teaching Education*, 1–15. https://doi.org/10.1080/10476210.2021.1895105

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kelchtermans, G. (2006). Teacher collaboration and collegiality as workplace conditions: A review. *Zeitschrift für Pädagogik*, *52*(2), 220-237. https://doi.org/10.25656/01:4454

Kim, J. (2019). Implementing a co-teaching model in music student teaching: A literature review. *Update: Applications of Research in Music Education*, *38*(1), 18–24. https://doi.org/10.1177/8755123319843169

King-Sears, M. E., Brawand, A. E., Jenkins, M. C., & Preston-Smith, S. (2014). Co-teaching perspectives from secondary science co-teachers and their students with disabilities. *Journal of Science Teacher Education*, *25*(6), 651–680. https://doi.org/10.1007/s10972-014-9391-2

King-Sears, M. E., Stefanidis, A., Berkeley, S., & Strogilos, V. (2021). Does co-teaching improve academic achievement for students with disabilities? A meta-analysis. *Educational Research Review*, *34*, 100405. https://doi.org/10.1016/j.edurev.2021.100405

Kyndt, E., Gijbels, D., Grosemans, I., & Donche, V. (2016). Teachers' everyday professional development: Mapping informal learning activities, antecedents, and learning outcomes. *Review of Educational Research*, *86*(4), 1111–1150. https://doi.org/10.3102/0034654315627864

Leatherman, J. (2009). Teachers' voices concerning collaborative teams within an inclusive elementary school. *Teaching Education*, *20*(2), 189–202. https://doi.org/10.1080/10476210902718104

*Lehane, P., & Senior, J. (2020). Collaborative teaching: Exploring the impact of co-teaching practices on the numeracy attainment of pupils with and without special educational needs. *European Journal of Special Needs Education*, *35*(3), 303–317. https://doi.org/10.1080/08856257.2019.1652439

Levin, B. (2004). Making research matter more. *Education policy analysis archives*, *12*, 56. https://doi.org/10.14507/epaa.v12n56.2004

Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research*, *20*(6), 741–768. https://doi.org/10.1177/1362168816649979

Liu, X. S. (2013). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203127698

Magiera, K., Smith, C., Zigmond, N., & Gebauer, K. (2005). Benefits of co-teaching in secondary mathematics classes. *Teaching Exceptional Children*, *37*(3), 20–24. https://doi.org/10.1177/004005990503700303

*Maharani, A., Darhim, A., Sabandar, J., & Herman, T. (2019a). Pbl-team teaching: Supporting vocational students logical thinking and creative disposition. *Journal of Physics: Conference Series*, *1188*. https://doi.org/10.1088/1742-6596/1188/1/012026

*Maharani, A., Darhim, A., Sabandar, J., & Herman, T. (2019b). Problem based learning-team teaching to improve vocational school students' mathematical disposition. *Journal of Physics: Conference Series*, *1157*(4). https://doi.org/10.1088/1742-6596/1157/4/042078

*McKenzie, S., Hains-Wesson, R., Bangay, S., & Bowtell, G. (2022). A team-teaching approach for blended learning: An experiment. *Studies in Higher Education*, *47*(4), 860–874. https://doi.org/10.1080/03075079.2020.1817887

Memon, M. A., Ting, H., Cheah, J.-H., Thurasamy, R., Chuah, F., & Cham, T. H. (2020). Sample size for survey research: Review and recommendations. *Journal of Applied Structural Equation Modeling 4*(2), i–xx. https://doi.org/10.47263/JASEM.4(2)01

*Migdadi, A., & Baniabdelrahman, A. (2016). The effect of using team teaching on Jordanian EFL eleventh grade students' reading comprehension and their attitudes towards this strategy. *Journal of Education and e-Learning Research*, *3*, 38–50. https://doi.org/10.20448/journal.509/2016.3.2/509.2.38.50

Ming, N. C., & Goldenberg, L. B. (2021). Research worth using: (Re)framing research evidence quality for educational policymaking and practice. *Review of Research in Education*, *45*(1), 129–169. https://doi.org/10.3102/0091732X21990620

Mohajeri, K., Mesgari, M., & Lee, A. S. (2020). When statistical significance is not enough: Investigating relevance, practical significance, and statistical significance. *MIS Quarterly*, *44*(2), 525–559. https://doi.org/10.25300/MISQ/2020/13932

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107587991

Moss, P. A. & Haertel, E. H. (2016). Engaging methodological pluralism. In D. Gitomer and C. Bell (Eds), *Handbook of research on teaching* (5th Ed), (pp. 127---247). Washington, DC: AERA.

Muckenthaler, M., Tillmann, T., Weiß, S., & Kiel, E. (2020). Teacher collaboration as a core objective of school development. *School Effectiveness and School Improvement*, *31*(3), 486–504. https://doi.org/10.1080/09243453.2020.1747501

*Muhammad, Z., & Jahan, A. (2019). Co-teaching effectiveness: Students' achievement in mathematical proficiencies and content strands. *Pakistan Journal of Education*, *35*. http://doi.org/10.30971/pje.v35i3.772

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451

Murawski, W. W., & Swanson, H. (2001). A meta-analysis of co-teaching research. *Remedial and Special Education*, *22*(5), 258–267. https://doi.org/10.1177/074193250102200501

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*(4), 374–396. https://doi.org/10.1007/s11414-012-9295-x

Nevin, A., Thousand, J., & Villa, R. (2009). Collaborative teaching for teacher educators—What does the research say? *Teaching and Teacher Education*, *25*, 569–574. https://doi.org/10.1016/j.tate.2009.02.009

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*, 322. https://doi.org/10.3389/fpsyg.2012.00322

OECD. (2020). *TALIS 2018 results (Volume II): Teachers and school leaders as valued professionals*. OECD Publishing. https://doi.org/10.1787/23129638

Osbourne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, *8*, Article 2. https://doi.org/10.7275/r222-hv23

Ostovar-Nameghi, S., & Sheikhahmadi, M. (2016). From teacher isolation to teacher collaboration: Theoretical perspectives and empirical findings. *English Language Teaching*, *9*(5), 197. https://doi.org/10.5539/elt.v9n5p197

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,… McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Pearl, C., Dieker, L. A., & Kirkpatrick, R. M. (2012). A five-year retrospective on the Arkansas Department of Education Co-teaching Project. *Professional Development in Education*, *38*(4), 571–587. https://doi.org/10.1080/19415257.2012.668858

Pearl, J. (2009). *Causal inference in statistics: An overview*. Cambridge University Press.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, 86, 207–236. https://doi.org/10.3102/0034654315582067

Pratt, S. (2014). Achieving symbiosis: Working through challenges found in co-teaching to achieve effective co-teaching relationships. *Teaching and Teacher Education*, *41*, 1–12. https://doi.org/10.1016/j.tate.2014.02.006

Pratt, S. M., Imbody, S. M., Wolf, L. D., & Patterson, A. L. (2017). Co-planning in co-teaching. *Intervention in School and Clinic*, *52*(4), 243–249. https://doi.org/10.1177/1053451216659474

*Rao, Z., & Yu, H. (2021). Enhancing students' English proficiency by co-teaching between native and non-native teachers in an EFL context. *Language Teaching Research*, *25*(5), 778–797. https://doi.org/10.1177/1362168819873937

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Rexroat-Frazier, N., & Chamberlin, S. (2019). Best practices in co-teaching mathematics with special needs students. *Journal of Research in Special Educational Needs*, *19*(3), 173–183. https://doi.org/10.1111/1471-3802.12439

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230. https://doi.org/10.1080/09243453.2014.885450

*Saeed, A. A., Mutashar, A. M., & Aldakheel, A. (2018). The impact of applying collaborative team teaching method on students' outcomes. *Indian Journal of Forensic Medicine & Toxicology*, *12*(4). https://doi.org/10.5958/0973-9130.2018.00212.8

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

Schanzenbach, D. W. (2012). Limitations of Experiments in Education Research. *Education Finance and Policy*, *7*(2), 219–232. https://doi.org/10.1162/EDFP_a_00063

Scruggs, T. E., Mastropieri, M. A., & McDuffie, K. A. (2007). Co-teaching in inclusive classrooms: A metasynthesis of qualitative research. *Exceptional Children*, *73*(4), 392–416. https://doi.org/10.1177/001440290707300401

Shadish, W. R. (2011). The truth about validity. *New Directions for Evaluation*, *2011*(130), 107–117. https://doi.org/10.1002/ev.369

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

*Sharma, U., Grové, C., Laletas, S., Rangarajan, R., & Finkelstein, S. (2021). Bridging gaps between theory and practice of inclusion through an innovative partnership between university academics and school educators in Australia. *International Journal of Inclusive Education*, 1–16. https://doi.org/10.1080/13603116.2021.1882052

*Shein, P. P., & Tsai, C.-Y. (2015). Impact of a scientist–teacher collaborative model on students, teachers, and scientists. *International Journal of Science Education*, *37*(13), 2147–2169. https://doi.org/10.1080/09500693.2015.1068465

*Simons, M., & Baeten, M. (2016). Student teachers' team teaching during field experiences: An evaluation by their mentors. *Mentoring & Tutoring: Partnership in Learning*, *24*(5), 415–440. https://doi.org/10.1080/13611267.2016.1271560

Simons, M., Baeten, M., & Vanhees, C. (2020). Team teaching during field experiences in teacher education: Investigating student teachers' experiences with parallel and sequential teaching. *Journal of Teacher Education*, *71*(1), 24–40. https://doi.org/10.1177/0022487118789064

Simons, M., Coetzee, S., Baeten, M., & Schmulian, A. (2020). Measuring learners' perceptions of a team-taught learning environment: Development and validation of the Learners' Team Teaching Perceptions Questionnaire (LTTPQ). *Learning Environments Research*, *23*(1), 45–58. https://doi.org/10.1007/s10984-019-09290-1

Solis, M., Vaughn, S., Swanson, E., & McCulley, L. (2012). Collaborative models of instruction: The empirical foundations of inclusion and co-teaching. *Psychology in the Schools*, *49*(5), 498–510. https://doi.org/10.1002/pits.21606

*Stapleton, J., Fogarty, E., Tschida, C., Cuthrell, K., & Chıttum, J. (2021). Impact of coaching, co-teaching, and student characteristics on teacher readiness *Journal of Teacher Education and Educators*, *10*, 131–155.

Steeger, C. M., Buckley, P. R., Pampel, F. C., Gust, C. J., & Hill, K. G. (2021). Common methodological problems in randomized controlled trials of preventive interventions. *Prevention Science*, *22*(8), 1159–1172. https://doi.org/10.1007/s11121-021-01263-2

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1

Sung, Y.-T., Lee, H.-Y., Yang, J.-M., & Chang, K.-E. (2019). The quality of experimental designs in mobile learning research: A systemic review and self-improvement tool. *Educational Research Review*, *28*, 100279. https://doi.org/10.1016/j.edurev.2019.05.001

Sweigart, C. A., & Landrum, T. J. (2015). The impact of number of adults on instruction: Implications for co-teaching. *Preventing School Failure: Alternative Education for Children and Youth*, *59*(1), 22–29. https://doi.org/10.1080/1045988X.2014.919139

Szumski, G., Smogorzewska, J., & Karwowski, M. (2017). Academic achievement of students without special educational needs in inclusive classrooms: A meta-analysis. *Educational Research Review*, *21*, 33–54. https://doi.org/10.1016/j.edurev.2017.02.004

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

*Tremblay, P. (2013). Comparative outcomes of two instructional models for students with learning disabilities: Inclusion with co-teaching and solo-taught special education.

Journal of Research in Special Educational Needs, *13*(4), 251–258. https://doi.org/10.1111/j.1471-3802.2012.01270.x

*Una, S. (2016). The use of parallel team-teaching: The case of teaching speaking for economics students in the Indonesian context. *Asian EFL Journal*, *2016*(89), 4–22.

van de Schoot, R., Miočević, M. (2020). *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners.* London: Routledge. https://doi.org/10.4324/9780429273872

Van Garderen, D., Stormont, M., & Goel, N. (2012). Collaboration between general and special educators and student outcomes: A need for more research. *Psychology in the Schools*, *49*(5), 483–497. https://doi.org/10.1002/pits.21610

Vangrieken, K., Dochy, F., Raes, E., & Kyndt, E. (2015). Teacher collaboration: A systematic review. *Educational Research Review*, *15*, 17–40. https://doi.org/10.1016/j.edurev.2015.04.002

Vembye, M. H., Weiss, F., & Hamilton Bhat, B. (2023). The Effects of Co-Teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis. *Review of Educational Research*, 0(0). https://doi.org/10.3102/00346543231186588

Vesikivi, P., Lakkala, M., Holvikivi, J., & Muukkonen, H. (2019). Team teaching implementation in engineering education: Teacher perceptions and experiences. *European Journal of Engineering Education*, *44*(4), 519–534. https://doi.org/10.1080/03043797.2018.1446910

Veteska, J., Kursch, M., Svobodova, Z., Tureckiova, M., & Paulovcakova, L. (2022). Longitudinal co-teaching projects: Scoping review. In D. Ifenthaler, P. Isaías, & D. G. Sampson (Eds.), *Orchestration of learning environments in the digital world* (pp. 35–53). Springer International Publishing. https://doi.org/10.1007/978-3-030-90944-4_3

Walsh, T. (2020). 'Promoted widely but not valued': Teachers' perceptions of team teaching as a form of professional development in post-primary schools in Ireland. *Professional Development in Education*, *48*(4), 1–17. https://doi.org/10.1080/19415257.2020.1725596

*Wang, S.-L., Chen, H.-P., Hu, S.-L., & Lee, C.-D. (2019). Analyzing student satisfaction in the technical and vocational education system through collaborative teaching. *Sustainability*, *11*(18), 48–56. https://doi.org/10.3390/su11184856

Weinberg, A. E., Sebald, A., Stevenson, C. A., & Wakefield, W. (2020). Toward conceptual clarity: A scoping review of coteaching in teacher education. *The Teacher Educator*, *55*(2), 190–213. https://doi.org/10.1080/08878730.2019.1657214

*Welch, M. (2000). Descriptive analysis of team teaching in two elementary classrooms: A formative experimental approach. *Remedial and Special Education*, *21*(6), 366–376. https://doi.org/10.1177/074193250002100606

What Works Clearinghouse. (2020). *What Works Clearinghouse procedures and standards handbook.* (Version 4.1). Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Educational Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/handbooks.

*Yeganehpour, P., & Zarfsaz, E. (2020). The effect of co-teaching on advanced EFL learners' writing ability. *Journal of Language and Linguistic Studies*, *16*, 1833–1853. https://doi.org/10.17263/jlls.851009

Zach, S. (2020). Co-teaching – An approach for enhancing teaching-learning collaboration in physical education teacher education (PETE). *Journal of Physical Education and Sport*, *20*(3), 1402–1407. https://doi.org/10.7752/jpes.2020.03193

Zhang, Y., Wang, J., & Lu, S. (2016). Research into the topic-based inter-curriculum cooperative teaching model of the higher school English major. *Journal of Interdisciplinary Mathematics*, *19*(3), 585–600. https://doi.org/10.1080/09720502.2016.1196051

Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using Webpower and R*. ISDSA Press.

Table 1

Team Teaching Models

Level of collaboration	Model	Role teacher 1	Role teacher 2
Low ↓	(1) Observation model	Full responsibility teaching	Observer
	(2) Coaching model	Full responsibility teaching	Coach
	(3) Assistant teaching model	Main responsibility teaching	Assistant
	(4) Equal status model	Identical status and responsibilities	
	(a) Parallel teaching model	(a) The class group is divided into subgroups. Each teacher teaches the same learning contents/activities to a subgroup.	
	(b) Sequential teaching model	(b) The learning contents or activities are divided. Each teacher is responsible for a different phase of the lesson.	
	(c) Station teaching model	(c) The class group and the learning contents/activities are divided. Each teacher teaches a specific content/activity to a subgroup.	
High	(5) Teaming model	Full collaboration in the planning, delivery, and evaluation of the lesson.	

Table 2

Number of Units According to Education Level

Units	Primary education	Secondary education	Higher education	Other educational setting
Students	8	10	8	2
Teachers	1	2	3	
Others		1		

Note. Blank spaces mean there is no data reported.

Table 3

Reported Team Composition

Type of team	*n*	%
Equal teams	15	48
Two or more qualified general education teachers in …		
English	6	
Mathematics	3	
Physical education	1	
Not subject-related	2	
Two student teachers	2	
Two higher education teachers/academics	1	
Mixed teams	13	42
General education teacher with …		
Special education teacher	5	
Teacher from another specific subject (math, languages)	2	
Higher education teacher/academic	2	
Paraprofessional (e.g., speech-language pathologist)	1	
Unqualified teaching assistant	1	
Student teacher with a mentor	1	
Higher education teacher/academic with an industry expert	1	
No composition specified	3	10

Note. Three studies did not provide a team composition as no information was specified about the type of educational professionals working together in a team teaching team.

Table 4

Reported Team Teaching Models

Type of model	*n*	%
Observation model		
Coaching model		
Assistant model	2	6
Equal status model		
Parallel teaching model	3	9
Sequential teaching model	2	6
Station teaching model	1	3
Teaming model	1	3
Mix of models	7	21
No models specified	17	52

Note. Blank spaces mean there is no data reported.

Table 5
Reported Phases of Teaching Practice

Phase of teaching practice	*n*	%
Planning	23	74
Teaching	23	74
Evaluation/reflection	8	26
No phases specified	8	26

Table 6
Reported Intensity of Team Teaching Practices

Duration of intervention period	*n*	Team teaching intensity (hours/week)	
		Min	Max
<1 week			
1-5 weeks	2	4	4
6-10 weeks	5	2	7
11-20 weeks	7	1	18
21-40 weeks	4	3	12
>40 weeks	3	1	Full time

Note. Blank spaces mean there is no data reported.

Table 7
Reported Methods Used for Baseline Equivalence

Baseline equivalence	*n*	%
Methods defined		
- Yes	23	74
- No	8	26
Type of method		
- T-test to confirm equivalence	10	44
- Statistical control for confounding variables	2	9
- Random assignment	3	13
- Random assignment and confirmation method	4	17
- Counterbalanced design	4	17

Table 8

Reported Number of Participants in Each Group

Group of units	<30		30-50		51-100		101-500		>500	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Pupils/students										
- Primary education	1	8	4	36	2	40	1	33	1	100
- Secondary education	4	31	4	36	1	20	1	33		
- Higher education	2	15	3	27	2	40	1	33		
- Other context	2	15								
Subtotal	9	69	11	100	5	100	3	100	1	100
Teachers										
- Primary education	1	8								
- Secondary education	1	8								
- Higher education	1	8								
Subtotal	3	24								
Other educational professionals										
- Mentors	1	8								
Total	13	100	11	100	5	100	3	100	1	100

Note. Blank spaces mean there is no data reported.

Table 9

Fulfillment of Statistical Assumptions According to the Test Used

Statistical test	Times reported		Assumptions fulfilled	
	n	%	*n*	%
T-test				
- Independent samples t-test	15	48	3	20
- Dependent samples t-test	4	13	3	75
- One-sample t-test	1	3	1	100
OVA test				
- Independent ANOVA	1	3	Not fulfilled	
- Mixed ANOVA	1	3	Not fulfilled	
- ANCOVA	2	6	Not fulfilled	
Regression				
- Multiple linear regression	1	3	1	100
- Multilevel modeling	2	6	Not fulfilled	
No test used	5	16		
Total	32	100		

Figure 1

Conceptual Framework for Examining the Effectiveness of Team Teaching

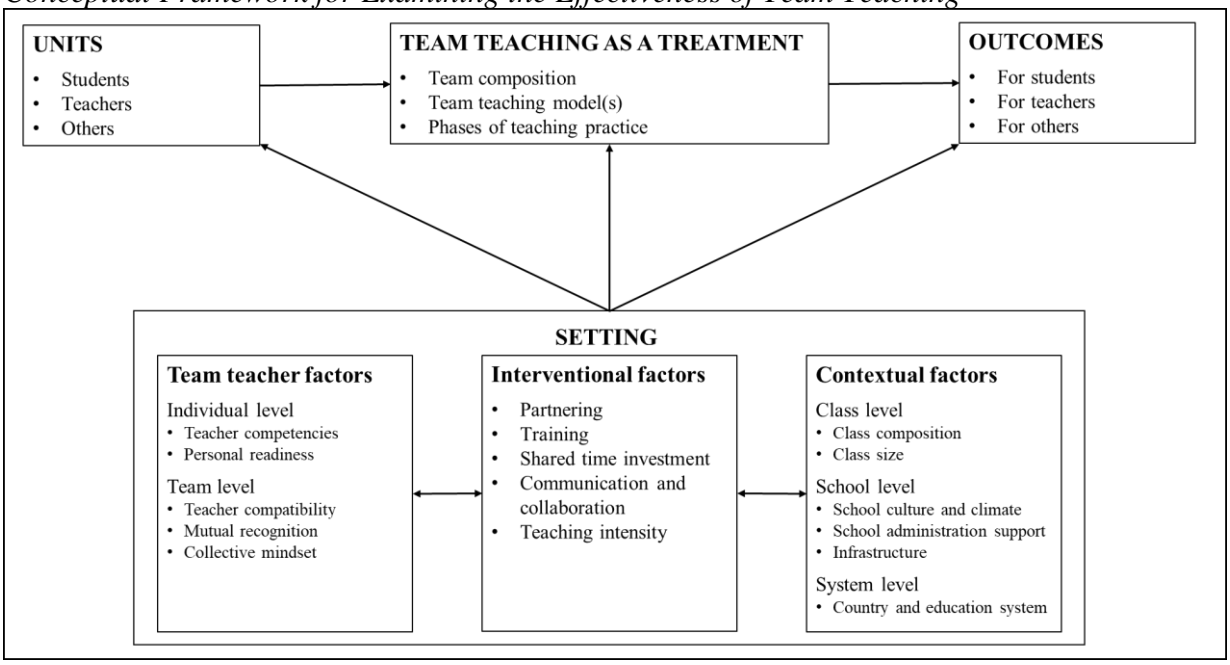

Figure 2

CREED Flowchart for Assessing the Experimental Level of Rigor

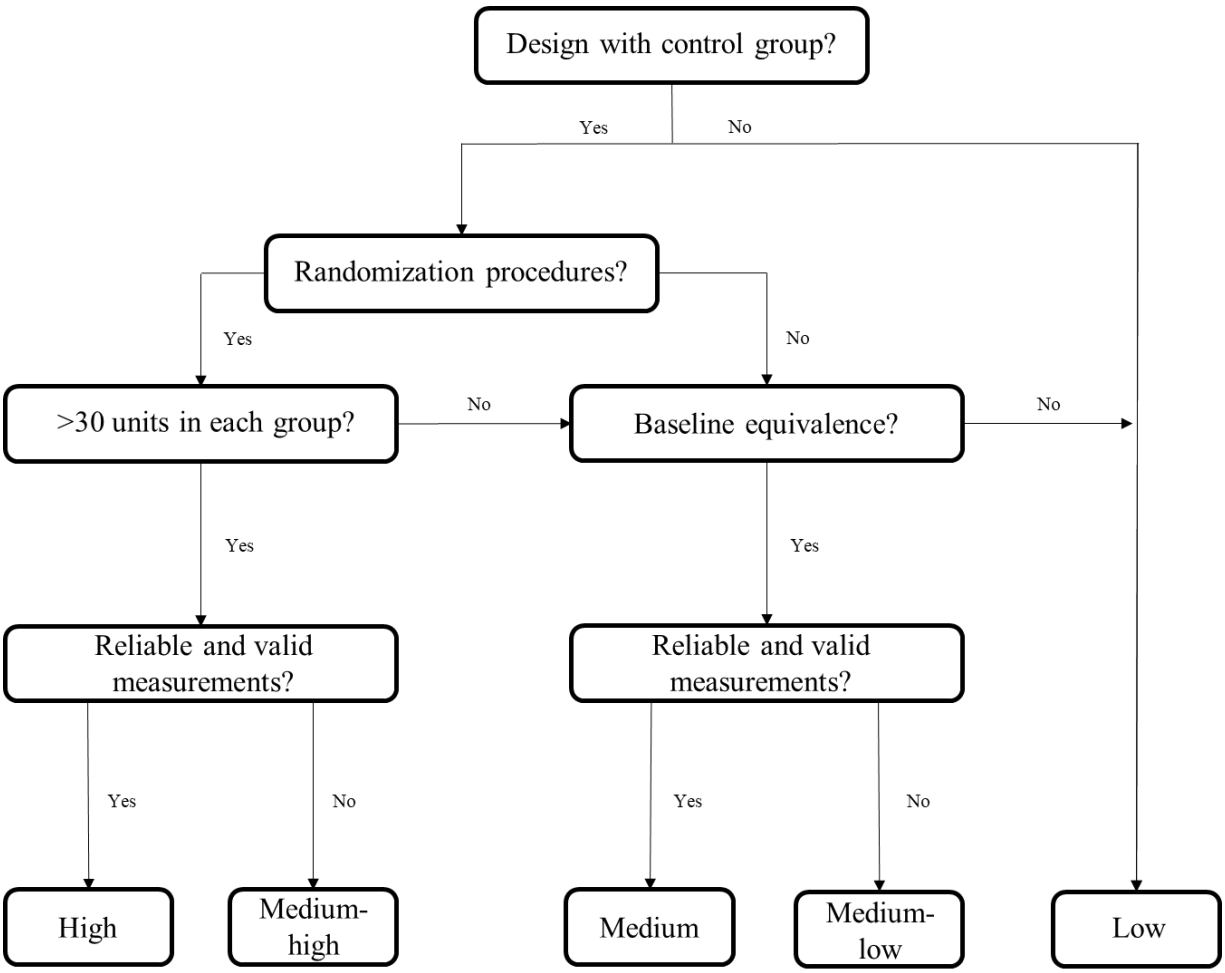

Figure 3

Search and Inclusion Process

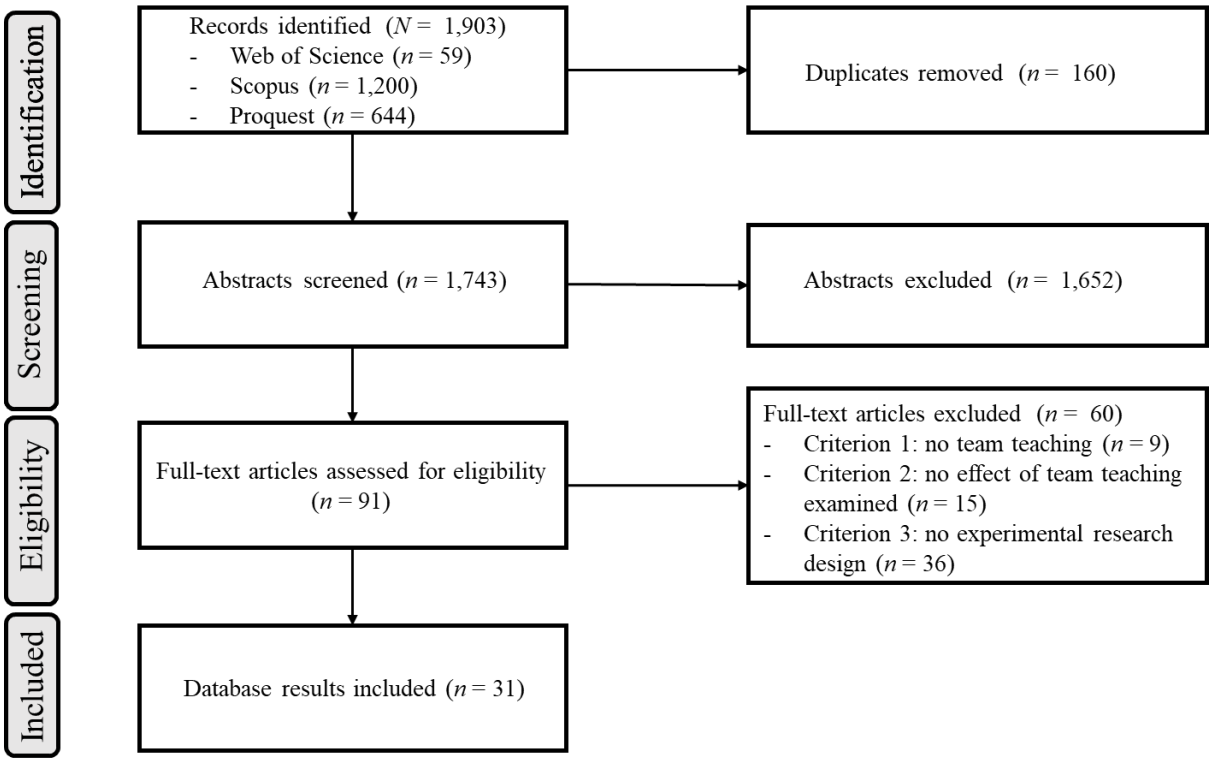

Figure 4

Evolution of Experimental Research on the Effectiveness of Team Teaching

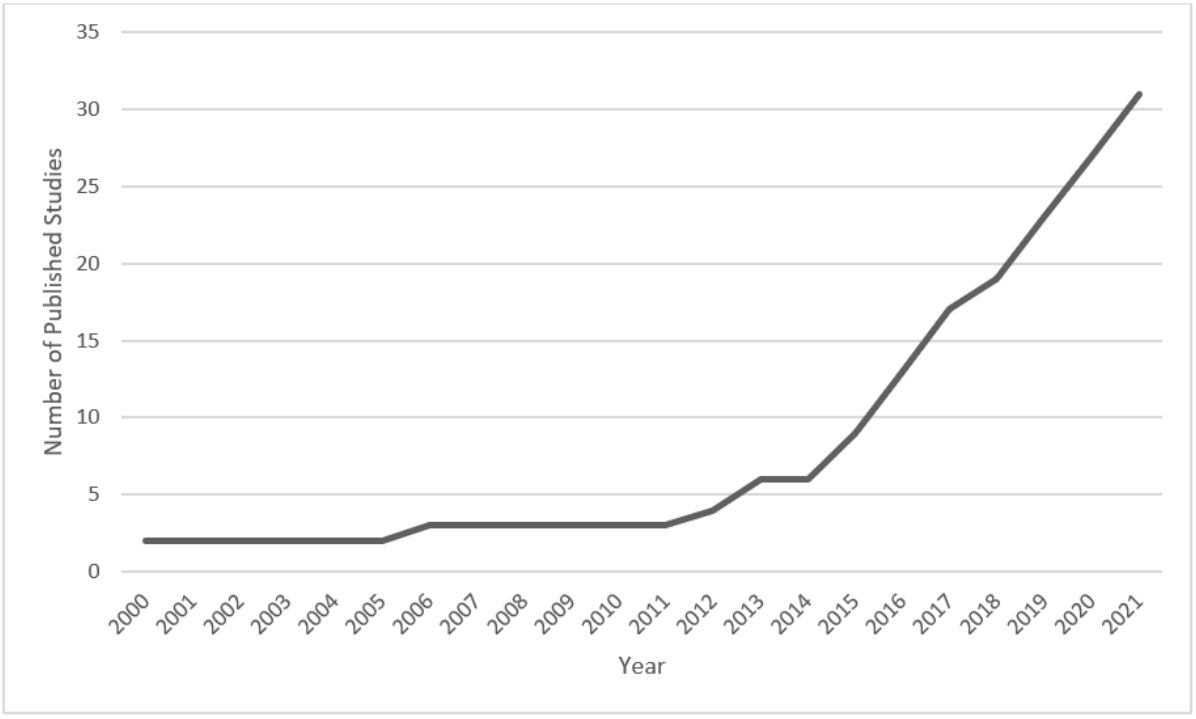

Figure 5

Experimental Design Types

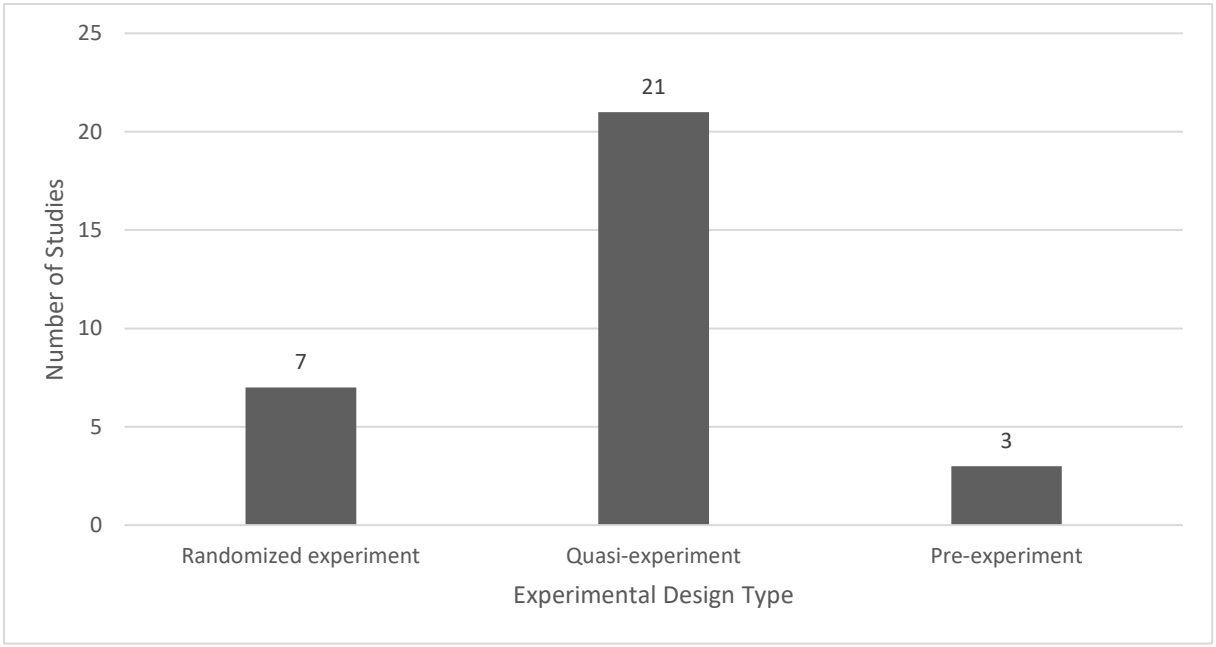

Figure 6

Obtained Levels of Experimental Rigor