

# The Effects of Co-Teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis

Mikkel Holding Vembye 

VIVE – The Danish Center for Social Science Research

Felix Weiss

Aarhus University

Bethany Hamilton Bhat 

University of Texas at Austin

*Co-teaching and related collaborative models of instruction are widely used in primary and secondary schools in many school systems. This systematic review and meta-analysis investigated the effects of such models on students' academic achievement and how these effects are moderated by factors of theoretical and practical relevance. Although previous research and reviews have asserted that the evidence base is scarce, we found 128 treatment and control group studies from 1984 to 2020. We excluded 52 studies due to critical risk of bias via Cochrane's risk of bias assessment tools and conducted a meta-analysis of 76 studies. This yielded 280 short-term effect sizes, of which 82% were pretest-adjusted. We found a moderate, positive, and statistically significant mean effect of  $\bar{g} = .11$ , 95% confidence interval [.035, .184] of collaborative instruction compared to single-taught controls, using the correlated-hierarchical effects (CHE-RVE) model. From moderator analyses, we found that collaborative instruction yields effects of mostly the same size, whether the interventions involved trained teachers or assistants with no teaching qualifications. This implies a potential for the expansion of such interventions at lower costs than otherwise expected. Moreover, factors that are highlighted in the co-teaching literature as preconditions for the effectiveness of collaborative instruction did not explain variations in effect sizes. Finally, we found no clear evidence for publication bias or small study effects. Notably, a large number of the studies that we drew upon were non-randomized studies; and therefore, more rigorous experimental research is needed, especially on relevant co-teaching interventions.*

**KEYWORDS:** co-teaching, teacher assistants, student achievement, meta-analysis, CHE-RVE model

Collaborative models of instruction have been applied since the 1950s (Willett et al., 1983), but their popularity has increased in recent decades in school systems throughout many high-income countries (Andersen et al., 2018; Blatchford et al., 2011; Friend, 2008; Muijs & Reynolds, 2003). Such models comprise co-teaching involving collaboration between general and special education teachers, but also the use of teacher assistants and paraprofessionals. This development has been fueled by a range of legislation and declarations (e.g., Individuals with Disabilities Education Act [IDEA], 2022; No Child Left Behind [NCLB], 2002; UNESCO, 1994) that warrant the right of all students to receive high-quality general education, regardless of their different abilities, needs, and challenges. Furthermore, the popularity of collaborative models of instruction is linked to the greater flexibility they offer compared to alternative options for improving the student-teacher ratio, such as class size reduction (Filges et al., 2018) or increased instruction time (Andersen et al., 2016; Kidron & Lindsay, 2014). In addition, the seemingly intuitive appeal, assuming that two educators working together will outperform a single teacher working alone, might also have contributed to their popularity (Bacharach et al., 2010; Friend, 2008). However, very little is known about the effects of collaborative models of instruction on students' academic achievement, not to mention how the effects vary across contexts, such as different subjects, grade levels, and/or student groups. In particular, there is a lack of research and reviews investigating the differential effects of various two-teacher instruction models, such as differences between co-teaching and teacher assistant interventions. In order to overcome this knowledge gap, we conduct a systematic review that includes and concentrates on interventions involving various models of collaborative instruction, allowing us to understand and contrast differences in effects between the various collaborative models of instruction. Previous research syntheses often refer to a limited evidence base as the main reason for the lack of understanding of the effects of collaborative models of instruction on student achievement (B. G. Cook et al., 2017; Friend, 2008; Iacono et al., 2021; Murawski & Swanson, 2001; Reinhiller, 1996). Meanwhile, we challenge this view by demonstrating that there exists a large body of literature, including many with a design appropriate for drawing causal conclusions. Despite applying more restrictive inclusion criteria than prior reviews, we found 128 relevant intervention studies published between 1984 and 2020. Based on these publications, we investigated not only the overall mean effect of collaborative models of instruction on student achievement but also how these effects varied across focal moderators highlighted in the methodological and theoretical literature as important factors in explaining differential effects of collaborative models of instruction.

### **Definition of Terms and Mechanisms for the Effect of the Intervention**

The underlying definitions used in this review follow the common use of key terms in the literature observed during the literature screening process. Inspired by Welch et al. (1999, p. 38), we broadly define collaborative instruction as the simultaneous presence of two or more educators/adults working together and sharing responsibilities in instructional and/or behavioral interventions. This definition encompasses all of the included compositions of two-teacher instruction, specified later in this section. Importantly, it is restricted neither to certain types

of two-teacher compositions/teacher actions nor to specific groups of students. In this way, we seek to include personnel without formal teaching qualifications, such as paraprofessionals, pedagogues, and parent volunteers.

Through our literature search, we identified three categories of collaborative models of instruction that fall under the overall definition but that were largely studied separately within the literature, with studies focusing on either *co-teaching*, *teacher assistants/aides*, or *team teaching*. Below, we outline the specific definition of each collaborative instruction model, as well as the causal theory behind the specific model.

### Co-Teaching

The largest share of the studies that we located refers to *co-teaching* interventions. We define co-teaching in line with L. Cook and Friend (1995) as “two or more professionals delivering substantive instruction to a diverse, or blended, group of students in a single physical space” (p. 2). In this regard, the term *professionals* specifically refers to the collaboration between a formally educated general education teacher and a formally educated special education teacher, such as a speech-language therapist, reading specialist, second-language teacher, or occupational therapist. The theoretical foundation within the co-teaching literature often highlights that co-teaching is context-dependent and only works effectively under quite specific conditions. For example, Friend (2008) states that

co-teaching partnerships require more than a casual agreement to work together in the classroom. For co-teaching to be effective, logistics must be addressed so that teachers’ schedules permit co-planning, teachers’ working relationships and classroom roles must be addressed, and administrative support must be in place. (p. 17)

Hence, providing time for co-planning is assumed to facilitate clear teacher roles by allowing the general and special educators to coordinate how to organize and (equally) share instruction time. Common instruction and equally shared instruction time, in turn, are assumed to be vital components for improving student learning by making full and complementary use of the professional competencies of the general and the special educator. To exemplify, the team might combine the general teacher’s in-depth knowledge of the curriculum and the specialized knowledge of the special education teacher about adapting instruction to the needs of the individual student. For the same reason, the co-teaching literature frequently presumes that co-teaching following the models “one-teach-one-assist” or “one-teach-one-observe” are ineffective (Friend, 2008; Scruggs et al., 2007; Szumski et al., 2017) since they do not take full advantage of the competencies of both educators. Additionally, it is emphasized that co-teaching is most effective when using a variety of co-teaching models (see L. Cook & Friend, 1995, pp. 5–6, for an overview of co-teaching models and support for this hypothesis).

Theoretical discussions and qualitative research related to the co-teaching literature furthermore suggest that voluntary participation and sound working relationships between the collaborating teachers maintain the co-teachers’

commitment to effective co-teaching (L. Cook & Friend, 1995; Friend, 2008; Scruggs et al., 2007).

As a somewhat concrete guideline, it has been suggested that co-teaching works best when provided to students in two 60- to 90-minute sessions per week (Friend cited in Stanek, 2017). Along similar lines, some have hypothesized that co-teaching only works properly when provided for more than a year, since it is a heavy developmental type of program that takes co-teachers' time to learn (see Friend cited in Dafolo, 2019).

Besides the hypotheses mentioned above, the co-teaching literature also expects a positive effect from the improved student-teacher ratio (L. Cook & Friend, 1995, pp. 3–4)—like other interventions such as class-size reductions (see Filges et al., 2018, and Supplementary Figure S29 in the online version of the journal for the causal theory behind the models). One hypothesis as to why a reduced student-teacher ratio might increase student outcomes is that it can reduce the number of disciplinary problems, which consequently increases instruction time and thus improves learning conditions. Another hypothesis is that students receive more appropriate and differentiated/personalized instruction, which allows for more in-depth presentation of the content, as well as increased student engagement.

### **Teacher Assistants/Aides**

Another set of collaborative models of instruction identified through our literature search was interventions involving teacher assistants/aides. We define teacher assistant(s) (TA) interventions as an in-class collaboration between a general education teacher and adults/paraprofessional educators without formal teaching qualifications, such as pedagogues or (volunteer) parents (Blatchford et al., 2011). These models can, to a large extent, be seen as a special case of co-teaching in which the primary instruction models used are “one-teach-one-assist” and “one-teach-one-observe,” but premised upon personnel without formal teaching qualifications. The teacher roles are assumed to be clearer in these models since the support personnel always play an assisting and secondary role relative to the general teacher. The mechanisms for the impact of TAs overlap with the reasoning behind the impact of reducing student-teacher ratios, as also presented in the co-teaching literature. The greatest difference is that the TA literature assumes that shared instructional responsibility and the formal qualifications of the second teacher are not a prerequisite for the effectiveness of the intervention.

In the literature, TAs are argued to have both an indirect and a direct impact on student achievement (Blatchford et al., 2011). The indirect effect is that TAs free up the general teacher from routine and clerical tasks, thus increasing the net instruction time, which in the end might benefit student achievement. The direct effect of TAs is argued to work through multiple complementary mechanisms (Blatchford et al., 2011; Muijs & Reynolds, 2003, pp. 221–222). For example,

- TAs can function as role models, showing students that the content is valued by adults other than the teachers.
- TAs provide an opportunity for greater student interaction with adults, which can scaffold student learning, provide more in-depth learning, and ensure that students are more active during lessons.

- TAs can facilitate students spending more time focused on tasks by addressing behavioral issues and thus increasing students' learning time.
- TAs can facilitate greater levels of concentration during whole-class activities by improving classroom management.
- TAs can increase the amount of immediate feedback and praise given to all students, boosting their confidence and motivation, reinforcing positive behaviors, and increasing their willingness to complete tasks.

### Team Teaching

Lastly, we identified a set of studies containing compositions of two-teacher instruction that fall outside the categories presented above, involving two regular/general in-class teachers with formal teaching qualifications (e.g., two math teachers). We refer to this category as *team teaching*,<sup>1</sup> which we define as two or more general education teachers sharing instructional and/or behavioral responsibilities for students in the same physical space. Since this model of instruction is not widespread in the literature, a specific causal theory for it is poorly developed. However, we assume that the reasoning is similar to the two other models, meaning that team teaching can reduce disciplinary problems and thereby ensure more instructional time, as well as increase student-teacher interaction and thereby allow students to receive more personalized instruction. A possible advantage of this model over the other two, however, is that students might be introduced to broader and more in-depth content knowledge due to the potentially complementary knowledge of the two general education teachers. A further benefit might be that it is less likely that one of the teachers will be ascribed the assisting role due to a lack of content knowledge, which often seems to happen to special education teachers in co-teaching interventions (Scruggs et al., 2007). This benefit might be especially pronounced in later grades when more advanced content knowledge is required.

### Previous Reviews

The systematic reviews most closely related to our review are Murawski and Swanson (2001), Khoury (2014), Willet et al. (1983), and Szumski et al. (2017). Common for all of these reviews is that they investigated the effects of collaborative models of instruction on student achievement outcomes by conducting systematic reviews, including statistical meta-analyses.

Murawski and Swanson (2001) investigated the effects of co-teaching on various outcome measures for students with special needs, including academic achievement. They found a large<sup>2</sup> mean effect size of .40. However, this must be seen against the backdrop of a rather small sample of six studies, of which five were single-case or single-group repeated measures designs—less conservative designs than studies with controls, generally yielding larger effects (Cheung & Slavin, 2016). Murawski and Swanson did not allow studies to contribute with multiple effect sizes, which, among other things, excluded the possibility of further investigating the differential effects of co-teaching across covariates varying within studies, such as outcome measures.

Khoury (2014) investigated the effect of co-teaching on academic achievement outcomes among students with special needs. It found a relatively large mean

effect size of .28, and it found that the effect of co-teaching neither varied across school levels, subjects, nor the type of study; nor different types of comparison groups used to calculate effect sizes. Finally, the review suggested that the effect became stronger the longer students were exposed to co-teaching. However, these analyses were based on the assumption of independence among effect sizes, which is likely violated when studies report multiple outcomes. As a consequence, the weighting schemes applied might likely be error-prone and yield models that do not adequately calibrate the nominal Type I error rate (Becker, 2000; Hedges et al., 2010; Van den Noortgate et al., 2013; Vembye et al., 2023).

By using multilevel meta-analysis, Szumski et al. (2017) investigated the effect of inclusive education on academic achievement among general students without special educational needs compared to general students in segregated classrooms from 1980 to 2013. The majority of the included studies involved the use of co-teaching models to accommodate inclusive education. Based on 35 studies, Szumski et al. found that inclusive education can have a positive effect ( $d = .12$ ) on the academic achievement of students without special educational needs, but this effect can vary with  $d = .05$ , 95% confidence interval (CI)  $[-.09, .18]$  for general students with a part-time co-teacher and  $d = .19$ , 95% CI  $[-.03, .41]$  for those with a full-time co-teacher. Szumski et al. conducted a range of moderator analyses but did not compare effects across subgroups since they fitted separate models for each subgroup. Furthermore, Szumski et al. included studies without control groups, which might have increased the estimated effects.

Willet et al. (1983), alone of the previous reviews, included studies of all types of in-class two-teacher instruction conducted from 1950 to 1983 and found a moderate mean effect size of .06 on students' science achievement. However, Willet et al. neither investigated moderating effects of collaborative teaching nor allowed studies to contribute with multiple outcomes, excluding knowledge about the differential effects of collaborative models of instruction.

All previous meta-analyses included quasi-experimental (QES) and observational (OBS) comparison studies but without ensuring or investigating the comparability of the intervention groups at baseline. This can potentially have jeopardized the accuracy of the estimation of the overall mean effect size.

### **Narrative Reviews and Syntheses**

Most commonly, previous reviews of research on collaborative models of instruction (see Table S1 in the online version of the journal for a list with an overview) have narratively synthesized studies, applying qualitative, mixed, and quantitative methods. These reviews have both included a mix of studies with different research designs, such as studies with single-case or single-group repeated measures designs and treatment and control group designs, as well as a mix of different outcomes, such as behavioral, social, emotional, and learning outcomes. A widespread conclusion across previous reviews is that research on the effects of collaborative instruction on student achievement is limited but that, based on what is known from the few existing studies, in-class collaboration seems to have a positive impact on learning. Prior reviews have typically concentrated on just one set of two teacher interventions and one sample of students, for example, only exploring the effects of co-teaching on outcomes related to special needs students.



One review, authored by Scruggs et al. (2007), is based purely on qualitative research. From interviews with and observations of co-teachers, they found that special education teachers frequently report that they are given subordinate assistant roles and that they consider administrative support vital in facilitating relevant training and the time needed for substantial co-planning and co-teaching. Finally, Sollis et al. (2012) conducted a review of reviews broadly focusing on collaborative models of instruction and inclusion interventions. They found mixed evidence of the effectiveness of co-teaching. However, this meta-review is primarily dominated by other types of interventions that are beyond the scope of our review.

### **Contribution of the Review**

This review goes beyond previous review studies in various ways. First, we aimed to fill the long gap since the first and hitherto only comprehensive review synthesizing different collaborative models of instruction was conducted in 1983. Second, we applied more clear-cut inclusion criteria for including study designs to ensure that we draw on the most reliable research for causal inference by only concentrating on quantitative studies with a treatment and control group design and only studies that measure students' academic achievement. In contrast to previous reviews, we conducted comprehensive risk of bias assessments, and we only included QES and OBS if they either reported pretest scores or had reliably ensured baseline equivalence (alternatively tested or adjusted for baseline differences) among treatment and control groups—for instance, by using matching techniques or controlling for focal covariates (see the full list of focal covariates in our protocol). Third, we conducted a more comprehensive review that combined and tested theoretically and empirically similar concepts. To this end, we included studies with different compositions of two-teacher instruction, as well as different samples of both general education and special needs students to understand the differential effects of collaborative models of instruction. To overcome limitations encountered in most previous reviews investigating differential effects of co-teaching, we used state-of-the-art meta-analysis methods to handle dependent effect sizes that stem from studies contributing multiple outcomes (Joshi et al., 2022; Pustejovsky & Tipton, 2021; Rodgers & Pustejovsky, 2021). Fourth, we sought to improve on previous meta-analyses by also accommodating common critiques leveled against effect size calculation in meta-analysis of being opaque (Maassen et al., 2020) by ensuring unprecedented transparency. In concrete terms, this means that all parts of the review, including effect size calculations and statistical analyses, are accessible at <https://osf.io/fby7w/>.

### **Methods**

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al., 2009; Page et al., 2021) reporting guidelines and the recommendations put forward by Pigott and Polanin (2019). Find completed PRISMA checklists at <https://osf.io/fby7w/>. The review has been pre-registered at the Open Science Framework (OSF): see <https://osf.io/ur2bs>.

## Inclusion and Exclusion Criteria

### *Study Designs*

Our review includes quantitative studies only. To draw on the most reliable research for general causal inferences, we only included studies with a treatment and control group design. Hence, single-case and single-group repeated measures designs were excluded. Included were (cluster and/or blocked) randomized controlled trials (RCTs), QES, and OBS. We characterize RCTs as studies in which researchers control the random assignment of students into treatment and control groups (either individually or in clusters, e.g., classrooms or schools) and initiate the implementation of the intervention. QES refers to studies in which researchers initiate the implementation of the treatment but do not randomly assign students to the intervention groups. Finally, OBS are studies where researchers neither influence the process of implementing the intervention nor control randomized assignment. However, such studies might still draw on randomization if, for example, participating schools randomly assigned students to classrooms before the treatment.

To avoid introducing more bias than we prevent by including study designs of varying quality (Egger et al., 2003), we applied strict rules for the inclusion of non-randomized studies. As such, we only allowed the inclusion of posttest effect sizes from QES and OBS if baseline equivalence was assured or if they either provided baseline/pretest achievement measures or covariate-adjusted<sup>3</sup> measures from which we could compute pretest- and/or covariate-adjusted effect sizes (Taylor et al., 2021). Otherwise, we considered nonequivalent studies that only reported posttest scores to be at critical risk of bias due to confounding factors. These studies were excluded via the ROBINS-I (Risk of Bias in Non-randomized Studies of Interventions) tool (Sterne et al., 2016). For further details, see the Risk of Bias Assessment subsection, below.

### *Intervention and Control Groups*

Collaborative models of instruction of all types were included as long as both teachers were at least 18 years old and the teaching took place in class with the educators sharing the same physical space (cf. Definition of Terms, above). Thus, we did not include collaborative models of instruction based on peer teaching or tutoring. Studies with more than two educators were allowed in this review, but none were found in the literature. We limited the included interventions to those with at least 2 weeks of treatment, which corresponds to at least 10 school days. Studies in which students had two teachers but the instruction was executed among a group of students *outside* the main classroom were excluded, as were studies where students were divided into distinct groups that received instruction in two different classrooms (e.g., see Jang, 2006a, 2006b). For an overview of teacher assistant interventions provided outside the main classroom (excluded here), see Farrell et al. (2010, p. 440).

Eligible control groups for this review can be divided into three categories: (a) noninclusive classrooms with a single general education teacher, that is, general education students only, compared to general students from co-taught classrooms (similar to those in Szumski et al., 2017); (b) inclusive classrooms with a general



education teacher, that is, a blended student composition, either compared to general and/or special needs students in co-taught classrooms; and (c) special education classrooms, such as resource rooms and pull-out classrooms, compared to students with special needs in co-taught classrooms (similar to those in Kråmer et al., 2021). We did not include two-teacher interventions conducted in special education school settings since the main focus of the review is on the general education setting. Moreover, to reduce confounding factors, we did not include studies comparing collaborative instruction to single-taught classrooms using reduced class sizes, as in Project STAR (Finn & Achilles, 1990).

### *Student Populations*

The eligible population sample for this review was students in Grades 1 to 12 attending primary or secondary schools, including public and private day schools as well as boarding schools. Special education schools were also included since they functioned as control schools. We did not find any studies that only included private school settings. Studies based on students in kindergarten, vocational, or postsecondary education were excluded. Overall, we included three types of student samples: (a) students with special educational needs and/or disabilities, (b) general education students, and (c) aggregated samples in which achievement outcomes were measured on a blended group of general and special needs students. If studies reported disaggregated measures for “at-risk” or “low-SES (socioeconomic status)” students but these were not formally characterized as special needs students, we amalgamated these results with the results for general students or, if that was not possible, interpreted the results as belonging to the general education population.

### *Country Context and Language*

To ensure a certain amount of comparability among the included population samples, the students had to come from high-income countries as defined by the 2020 World Bank Classification (World Bank, 2022). For instance, we excluded an Iranian study by Aliakbari (2013). Furthermore, we only included documents and studies written in English, Danish, Swedish, Norwegian, or German. All studies in the final sample used for data extraction and effect size calculation were written in English.

### *Outcomes*

Due to their important role in the policy debate and their high correlation with future academic and labor market success (Dietrichson et al., 2020; OECD, 2016), we concentrated on academic performance. Academic achievement tests of all types were included as eligible outcomes, including state- or nationwide standardized tests, norm-referenced commercial tests, grades, school-leaving examinations, marks for the year’s work, large-scale assessment tests, teacher-developed tests, researcher-developed tests, and textbook tests.

Only subjects within the areas of arts, social science, and STEM (science, technology, engineering, and math or mathematics) were considered eligible, such as language arts, social studies, history, science, biology, and mathematics. In subsequent analyses, we roughly divided effect sizes into “arts and social science”

versus STEM categories in order to make optimal use of all relevant outcomes and information. We excluded all practical and creative subjects, such as music, sports, home economics, or woodwork.

When studies were baseline-adjusted for test scores, we included them in the category with all other studies that were pretest-/baseline-adjusted. This included two smaller studies (Beam, 2005; Rea et al., 2002) that called the test applied an “IQ” test. The documentation of these tests varied, but in one case we would see the verbal test score used was similar to other test scores that were not called “IQ.” Thus, due to the vague and shifting definitions of IQ and other scores, we opted to include these two studies in the baseline-adjusted group of studies instead of the covariate-adjusted group.

We divided analyses between posttest and follow-up measures. The latter was characterized as effects measured 3 months or more after the end of the intervention. If studies reported effects across various time points and/or outcomes, we included them all.

### *Search Procedures*

The search string that we developed for our electronic searches was inspired by the previous review studies as well as several recent empirical studies. It covered the different types of interventions equally well (i.e., co-teaching, TAs, and team teaching). The search string is too extensive to be included in the main text but is documented within our preregistered protocol at <https://osf.io/ur2bs>. We conducted an electronic search from 1984 to June 2020 in the databases Scopus, Web of Science, APA PsycArticles, APA PsycInfo, Australian Education Index, Ebook Central, EconLit, Education Database, ERIC, Periodicals Archive Online, and ProQuest Dissertations and Theses Global. The main source for gray literature was the database ProQuest Dissertations and Theses Global, which identified a large number of eligible dissertations. Beyond this systematic search, we performed a less systematic search using Google Scholar and used snowball sampling for all previous reviews and for all journal articles that were included in the final dataset.

### *Expert and Author Solicitation*

We did not contact any primary authors or experts for further study detection, although stated in the first protocol attached to this review. Since previous research showed that only 12% of the primary study authors replied to solicitations and only 0.5% of these replies contained the requested information (Polanin et al., 2020; Schauer et al., 2020), we opted to change this initial plan due to the seemingly low chances of it adding to our data.

### *Screening Procedures*

The first and second review authors conducted independent abstract and full-text screening of all references found during the literature search. Disagreements were resolved via discussion and consensus among the authors. All screening was conducted and documented alongside reasons for excluding references in Covidence. The Covidence repository is accessible upon request.

### *Data Extraction*

Data extraction was conducted by the first author only. For quality assurance, the data extraction was conducted twice for each study. As a further quality check (suggested by Campbell Collaboration, 2019; Hofner et al., 2016), the third author inspected 12 of the most complex effect size calculations for coding errors and possible improvements.

We specifically extracted information regarding the study, sample, context, participants, design, treatment and control group, outcome, and estimation characteristics. Whenever data extraction or effect size calculation issues arose, these were resolved via discussion and in consensus among the authors. Most result data from studies reporting more than four outcome results were extracted by a student assistant.

To strengthen the theoretical relevance of the review, the data extraction scheme was developed in line with the co-teaching literature and theory (L. Cook & Friend, 1995; Friend, 2008, 2017). Thus, we were able to test the hypotheses discussed in this literature empirically. We pilot-tested the scheme on eight studies (these were Adams, 2014; Allen, 2008; Almon & Feng, 2012; Andersen et al., 2018; Andrews-Tobo, 2009; Fontana, 2005; Muijs & Reynolds, 2003; Murawski, 2006). We then optimized the data extraction scheme by reducing the number of extraction characteristics whenever certain characteristics were not retrievable from the pilot studies. This led us to exclude, for example, the variable that measures the quality of the collaboration between the collaborating teachers. All background information and covariates were extracted using MS Excel, while information related to the effect size calculation was extracted and managed using RStudio. To accommodate the one-coder-only practice, all extraction schemes, effect size calculations, and the final/complete dataset is available for critical inspection and future updates at <https://osf.io/fby7w/>.

### *Risk of Bias (RoB) Assessment*

To further ensure that the accuracy of the review was not compromised by including study designs of varying quality, we conducted comprehensive RoB assessments for all effect sizes individually. Studies contributing with multiple effect sizes underwent multiple and potentially different RoB assessments—for example, if a study reported results across different types of outcomes or student samples.

Since we amalgamated results across randomized and nonrandomized studies, we applied the RoB 2 tool for RCTs (Sterne et al., 2019), the RoB 2 cluster-randomized control trial (CRCT) tool for cluster RCTs (Eldridge et al., 2021), and the ROBINS-I tool for nonrandomized studies (Sterne et al., 2016). To ensure comparability between the three RoB assessment tools, we required that nonrandomized studies either provided raw data or a preregistered protocol in order to receive a low RoB assessment due to reporting. Moreover, to align the RoB 2 tools to social science standards, we did not consider questions regarding blinding and double blinding to have any significant impact on the overall RoB assessment.

The RoB assessment was conducted by the first author only. However, the RoB assessment was also used to exclude studies with a critical risk of bias, and

exclusion of these studies was always based on consensus between the first and the second author. In this regard, we excluded studies from the review as soon as they received the first critical RoB judgment for any domain in the ROBINS-I scheme. All conducted RoB assessments are available at <https://osf.io/fby7w/> for critical inspection and future updates. For further details about the RoB assessment procedure, see Section S6 in the online version of the journal.

### **Statistical Methods**

Effect size calculation and statistical data analyses were conducted using R 4.1.2 (R Core Team, 2022) in RStudio (RStudio Team, 2015). For the main analyses, we used the packages *metafor* (version 3.0-2; Viechtbauer, 2010), *clubSandwich* (version 0.5.5; Pustejovsky, 2020b), and *wildmeta* (version 0.0.0.9000; Joshi & Pustejovsky, 2022). For figure illustrations, we used *ggplot2* (version 3.3.3; Wickham, 2016). Find replication material for all statistical analyses of this review at <https://osf.io/fby7w/>.

#### *Effect Size Calculation*

Standardized mean differences are the effect size metric used in this review and were calculated via Hedges's  $g$  estimator (Hedges, 1981). We coded effect sizes so that positive values indicated a positive effect of collaborative instruction. We applied a broad range of techniques for obtaining effect sizes across the diverse set of research designs and estimation methods used in the primary studies (Borenstein, 2009; Hedges, 2007; Higgins et al., 2019; Pustejovsky, 2016; What Works Clearinghouse [WWC], 2020, 2021; Wilson, 2016). The majority of effect sizes were based on either pretest or covariate-adjusted computation techniques (Morris, 2008; Morris & DeShon, 2002; Pustejovsky, 2016; Taylor et al., 2021). In most cases, calculating covariate- and/or pretest-adjusted effect sizes requires information about the correlation between the covariate(s) and the outcome measures,  $\rho_{cor}$ , which are infrequently reported in primary studies but can be obtained from other measures that are usually provided (Pustejovsky, 2020a; Wilson, 2016). Whenever it was impossible to derive  $\rho_{cor}$  from reported results, we imputed  $\rho_{cor}$  following the guidelines proposed by the WWC (2020).

All calculated effect sizes were standardized by the *total variance*, here denoted as  $g_T$ . This means that all effect sizes encompass *variance* on the student level as well as cluster levels such as the classroom and/or school levels (Taylor et al., 2021). Thus, studies reporting *means* and variability measures on only one of these levels were converted to ensure that they represent the same unit of analysis:  $g_T$  (Hedges, 2007). This also entailed conducting approximate cluster bias corrections on all effect sizes coming from multisite studies (i.e., studies containing multiple treatment and control classrooms) not accounting for the nesting of students in classes and/or schools (Higgins et al., 2019; WWC, 2021). All conversions were premised upon intraclass correlation (*ICC*) values, which are rarely reported in educational research. We, therefore, imputed *ICC* values from Hedges and Hedberg (2007), as suggested by Hedges (2007), to conduct these two-level conversions. To further ensure a common unit of analysis across effect sizes and reduce unnecessary amounts of within-study variability, we aggregated results across subgroups and subtests if these were irrelevant to our moderator analyses.

For a detailed description of the full effect size calculation procedure, see Section S1 in the online version of the journal.

### *Dependent Effect Sizes*

The final dataset contains various dependency structures among effect sizes that necessitate the use of advanced meta-analytical techniques. First, 45 studies have what we define as a correlated effects dependency structure. This means that these studies reported multiple outcome results from the same sample of students, which produces correlated sampling errors among effect sizes and therefore breaks the assumption of independence among effect sizes. Second, six studies reported results from multiple nonoverlapping samples, which we define as a hierarchical effects dependency structure. What characterizes this dependency structure is that individual effect sizes are nested within samples that are nested within studies. Although results are from nonoverlapping samples, the fact that researchers applied the same measurement procedure, recruitment strategy, or other study procedures might create a dependency among the mean effects coming from the same study. Consequently, the assumption of independent results is violated. Third, four studies contained both of the above-mentioned dependency structures, which means that they reported multiple outcomes from multiple non-overlapping samples. The remaining studies contributed one effect size only.

A challenge when synthesizing dependent effect sizes is that the true/exact dependency among effect sizes is unknown, and only a few studies reported the information needed to assess the true dependency among the dependent effect sizes. To tackle this challenge, we applied robust variance estimation (*RVE*; Hedges et al., 2010; Pustejovsky & Tipton, 2021; Tipton & Pustejovsky, 2015), which has shown to be the most accurate method for meta-analyzing dependent effect sizes (Fernández-Castilla, Aloe, et al., 2020; Vembye et al., 2023). *RVE* implies the use of working models that tentatively aim to resemble the true dependency structures among effect size estimates coming from the same study. This is done by making various assumptions about the dependency structures, including the sample correlation,  $\rho$ , between within-study outcomes. These working models ensure more appropriate weighting schemes of effect sizes relative to univariate models that assume independence among effect sizes or use study-mean effect sizes. The most beneficial feature of using *RVE* is that it yields valid estimates even if the assumed working model is misspecified. To further ensure valid Type I error calibration even when analyses are predicated on a small number of studies, which is an issue especially common in subgroup analyses, we applied the “CR2” small-sample corrector (Joshi et al., 2022; Tipton, 2015; Tipton & Pustejovsky, 2015).

### *Mean Effect Size Estimation*

To derive the overall mean effect size across all effect size estimates, we applied the correlated-hierarchical effects (*CHE-RVE*) model (Pustejovsky & Tipton, 2021; Vembye et al., 2023). This model takes into account the multilevel structure of the effect size data with effect sizes nested in studies (Van den Noortgate et al., 2013, 2014) and guards against any misspecification of the model via *RVE* (Hedges et al., 2010; Tipton & Pustejovsky, 2015). At the same time, it

accounts for the correlated effects structure by imputing a constant sample correlation,  $\rho$ , between effect size estimates coming from the same study. Commonly, *CHE* models entail making a fair guess of the constant sample correlation,  $\rho$ , between effect size estimates coming from the same study. However, we obtained  $\rho$  by estimating Pearson's correlation from studies that both reported math and language arts scores, as suggested by Kirkham et al. (2012). We estimated  $\rho = .706$ . We considered this value to be plausible since it closely resembles the sample correlations obtainable from the Project STAR data (Achilles et al., 2008) across first-, second- and third-grade students assigned to either the teacher aides or single-taught arm, which were  $\rho_{\text{grd1}} = .718$ ,  $\rho_{\text{grd2}} = .722$ , and  $\rho_{\text{grd3}} = .735$ . Using restricted maximum likelihood techniques (Viechtbauer, 2005), we estimated two sources of heterogeneity: the standard deviations (*SD*) at the effect size level (also known as the within-study *SD*,  $\omega$ ) and at the study level (also known as the between-study *SD*,  $\tau$ ). Larger standard deviations indicate greater variability among effect sizes than would be expected from sampling error alone. See Section S2 in the online version of the journal for a detailed statistical description of the used *CHE-RVE* model.

### *Sensitivity Analyses*

Although the *CHE-RVE* model is expected to be valid even when the working model is misspecified, we conducted a sensitivity analysis in which we investigated the impact of changing the assumed sample correlation from  $\rho = 0$  to  $\rho = .95$ . The main reason for conducting this analysis was that the individual random variance components from *CHE* models can be substantially affected by the assumed magnitude of  $\rho$  (Pustejovsky & Tipton, 2021). Yet, the magnitude of the *total* variance component estimate is usually stable. Moreover, we conducted leaving-one-study-out analyses to investigate whether any specific study had a substantial impact on the mean effect size and heterogeneity estimations.

As further robustness checks, we conducted a range of sensitivity analyses in which we changed the inclusion criteria and the assumptions underlying the effect size calculation. Specifically, we conducted a range of sensitivity analyses in which we changed the *ICC* values used for the approximate cluster bias corrections and the pre-posttest correlation for difference-in-differences studies for which these correlation estimates were unobtainable. We also tested the impact of using neither cluster bias nor small sample corrections (i.e., calculating Cohen's *d* only). For studies from which we were able to obtain the same effect size using different calculation techniques (e.g., difference-in-differences and adjusted means), we applied all available approaches to probe potential discrepancies. We then conducted a sensitivity analysis in which we reestimated the mean effect size by using the most extreme alternative effect size estimate from these studies. Finally, we conducted a range of sensitivity analyses in which we repeatedly reestimated the mean effect size model while changing inclusion criteria by blockwise excluding the following categories of studies or effect sizes: observational studies, all nonrandomized studies, single-site studies (i.e., only one treatment and one control class), large-scale studies with sample sizes above 1,000 students, gray literature (i.e., studies not published in scientific peer-reviewed journals), serious risk of bias assessed effect sizes, non-U.S. studies, and outlier effect sizes. Outliers



were defined as effect size estimates falling more than three times the interquartile range below the first quartile or above the third quartile (Tukey, 1977; Winters et al., 2022). Under this definition, only one effect size calculated from math achievement in Dwyer (2018) was considered an outlier.

### *Publication Bias Testing*

We conducted five complementary publication bias and/or small-study effects tests, as suggested by Hedges and Vevea (2005). This included trim-and-fill tests based both on all the individual effect sizes and on effect sizes aggregated to the study level, Egger's regression tests accounting for dependent effect sizes using the *CHE-RVE* model (Egger et al., 1997; Rodgers & Pustejovsky, 2021), and step-function selection model tests. For the latter, we used three cutpoints at  $p = .05$ ,  $p = .10$ , and  $p = .50$ , as well as cutpoints at  $p = .025$  and  $p = 1$  (Hedges & Vevea, 2005), with effect sizes aggregated to the study level. The latter test functioned as a sensitivity analysis. We also conducted a sensitivity analysis for publication bias, which included the estimation of the worst-case meta-analysis, removing all positive and statistically significant effect sizes based on the assumption they were all false positives (Mathur & VanderWeele, 2020). Finally, we estimated the weighted average of the adequately powered (*WAAP*) studies (Stanley et al., 2017). For all tests, we either used a modified estimate of the standard error or sampling variance by removing the part of the variability estimation capturing the precision of the standard deviation used as the standardizer for the given effect size calculation (Hedges & Olkin, 1985; Pustejovsky & Rodgers, 2019). If not removed, it would have created an artificial correlation between the standardized mean differences and their variability measures, which would have risked yielding flawed evidence for publication bias and/or small-study effects. Moreover, we applied contour-enhanced funnel plots to illustrate potential publication bias/small-study effects (Peters et al., 2008). As a sensitivity analysis, we also applied contour-enhanced funnel plots based on transformed measures, which represents an alternative method for handling the artificial correlation between standardized mean differences and their variability measures (Pustejovsky & Rodgers, 2019). See Section S11 in the online version of the journal for an elaboration of the conducted publication bias tests.

### *Moderator Analyses*

To investigate whether focal moderators of methodological and theoretical relevance could explain differences in outcomes across studies, we conducted a comprehensive range of moderator analyses using three different working models from the *CHE* model family (Pustejovsky & Tipton, 2021).

Our meta-regression analyses fall into three categories: (a) subgroup analyses based on categorical variables without missing values (i.e., fully reported information across all studies), (b) subgroup analyses based on categorical moderators with missing values, and (c) meta-regression models including continuous moderators with missing values. For the first set of models, we investigated whether differential outcomes can be explained by methodological differences between research designs, publication status, the overall RoB assessment, and the type of effect size (i.e., covariate-adjusted vs. posttest-only effect sizes). We also

examined whether outcomes substantially differed across the following study characteristics: type of intervention, subjects, test modes, grade levels, the type of control group, and the type of control group used to calculate effect sizes for samples of special needs students only.<sup>4</sup> Across these models, we switched between fitting subgroup correlated effects plus (SCE+) or correlated multivariate effects plus (CMVE+) working models (find detailed information about the use and embedded assumptions of these models and the reasons for shifting across models in Sections S2–S4 in the online version of the journal). As with the overall mean effect size model, these models included heterogeneity at both the effect size and study level (indicated by the + sign). For each of these subgroup models, we investigated mean differences across subgroups using HTZ Wald tests (Tipton & Pustejovsky, 2015), as well as Wald tests based on cluster wild bootstrapping (CWB) with 1999 replications (Joshi et al., 2022). Using both these Wald tests allowed us to check for consistency, but our findings are primarily based on interpretations of the CWB values.

For the second set of models, we investigated whether effect size differences could be explained by factors highlighted as focal moderators in the co-teaching literature. First, we tested differences between studies in which time was provided for co-planning lessons against studies reporting no provision of common planning time. Second, we tested differences between studies in which training was provided in co-teaching methods against studies in which no training was provided. The SCE+ working model was the only model used for these analyses.

For the final set of models, we investigated whether the duration and intensity of the intervention, as well as the percentage of males in the sample, could explain true variation in effects across studies. All of these predictors were centered: The duration was centered around 40 weeks of treatment, which amounts to one school year; the intensity was centered around five sessions per week, amounting to one session per school day; and the percentage of males in the sample was centered around 50% males in sample. All models used the same *CHE* working model as in the summary model for the overall mean effect size.

Across all moderator analyses, we fitted models with and without adjusting for grade level, student sample, and subject differences—for some models, these variables were the independent variable of interest. Then, we adjusted for the remaining two control variables. We did not add further moderator factors to the models because we detected strong multicollinearity among the moderators. Therefore, we only focused on controlling for factors of substantial content importance. See the covariate correlation matrix in Supplementary Table S14 (online only). A detailed elaboration of the statistical conduct and model selection procedure can be found in Sections S3 and S4 in the online version of the journal.

### *Dealing With Missing Data*

To handle missing values on moderator variables, we used multiple imputation with 50 imputations and 50 iterations (Pigott, 2019; Van Buuren, 2018). We applied exploratory missingness analysis techniques (Schauer et al., 2021) to assess whether a covariate should be included in an analysis based on multiple imputation techniques. We excluded all variables with more than 50% missing

values or if the missingness structure of the variables was correlated with the effect sizes and their variance. Find our exploratory missingness analysis at <https://osf.io/fby7w/>.

Since there have yet to be developed reliable methods for pooling multicontrast Wald tests (i.e., HTZ Wald tests) across multiple imputed datasets, we applied a different procedure—relative to tests based on covariates without missing values—to obtain  $p$  values for the aggregated Wald test pooled across the 50 imputed datasets. First, we averaged coefficient estimates and variance-covariance matrices using Rubin's rule (Rubin, 2004), then we calculated the  $Q$ -statistics from Equation 10 in Tipton and Pustejovsky (2015) and obtained  $p$  values from  $F$ -tests with  $q$  and  $J - 1$  degrees of freedom, where  $q$  is number of coefficients in the model minus one and  $J$  is the number of studies. We used this approach because simply averaging Satterthwaite degrees of freedom across the imputations would yield rather conservative results with no fair chance of finding true mean differences between moderator categories.<sup>5</sup>

### Deviations From the Preregistration Protocol

The final review diverges in several ways from our preregistered protocol. The initial plan was to also use the EBSCO database to conduct literature searches. However, we chose to exclude this database after experiencing problems using our rather extensive search string.

We did not calculate variance-covariance matrices from studies where they were obtainable since this proved to be more problematic than anticipated based on the information given in the different publications. Moreover, we did not apply multilevel multiple imputation—the nesting structure of the missing values did not allow us to conduct this type of imputation since the covariates rarely varied within studies. Due to constraints of time and space, we did not conduct a subgroup analysis comparing the effects across urban and rural settings. However, this information can be estimated from our open-source data.

We also added several exploratory analyses that were not mentioned in the protocol but were able to strengthen the review. We added sensitivity analyses, including an analysis using  $ICC$  values from the pretest covariate models (instead of the unconditional models) for the population representing all schools from Hedges and Hedberg (2007) to cluster bias adjust effect sizes. Furthermore, we added an analysis investigating the impact of large studies (i.e., sample sizes above 1,000 individual students), an analysis omitting single-site studies (based on the assumption that they likely mislead more than they inform), an analysis examining U.S. studies only, and an analysis in which outliers were removed. We also conducted an analysis investigating the difference between effect sizes for special needs students based on general and special education control groups—partly to investigate if one of the alternative service delivery models outperformed the other and partly to add to the inclusion literature regarding special needs students (Krämer et al., 2021). To further link our study to related reviews regarding inclusion (Szumski et al., 2017), we also tested if differential effects could be identified when comparing two-teacher interventions among general education students to interventions in either inclusive or segregated settings. Lastly, we conducted a sensitivity analysis of the subgroup analyses based on studies that only

analyze the effect of co-teaching and hence limit the sample to collaborations between formally educated general and special education teachers.

## **Results**

Figure 1 presents a PRISMA flowchart documenting the search process and the criteria for exclusion of references. We identified 9,969 potentially relevant references from database searches and snowball sampling. After removing 1,962 duplicates, the first and second authors independently screened titles and abstracts of 8,007 references. The proportionate agreement between authors was 93.74% with Cohen's  $\kappa = .448$ , 95% CI [.447, .449], which indicates weak agreement according to commonly referenced guidelines (cf. McHugh, 2012; Orwin & Vevea, 2009). However, we are not especially concerned about this value since  $\kappa$  was mainly driven by initial disagreements where one of the authors applied a more inclusive screening strategy than the other. Subsequently, we independently screened 373 full texts. We excluded 245 studies for various reasons. The most common reasons for exclusion at this stage were ineligible study designs (65, most of which were studies with single-case or single-group repeated measures designs), ineligible interventions (32), and further duplicates (38). We do not report the interrater reliability for the full-text screening because many studies were excluded for multiple reasons, which artificially reduced the agreement rate between authors. We then assessed RoB for the remaining 128 studies, excluding 52 studies due to a critical RoB assessment for at least one domain in the ROBINS-I tool. The most common reason (25 studies) for a critical ROBINS-I assessment was studies with posttest designs that failed to ensure baseline equivalence between intervention and control groups or did not report relevant covariate/pretest/baseline measures. Following this process, the final meta-analytic dataset included 76 studies (find full reference list of the included studies at <https://osf.io/gnqde>).

## **Descriptive Statistics**

Figure 2 presents the included studies by year of publication and type of intervention. It appears that the number of studies investigating the academic effects of collaborative models of instruction has increased since 2000, with 61 studies (80%) published in this period. One factor that may have contributed to this significant increase was the need for more data highlighted in the previous review by Murawski and Swanson (2001). However, we were also able to identify seven additional studies in the period from 1990 to 2000 concerning the effects of co-teaching on special needs students' achievement not included by Murawski and Swanson, albeit we applied more narrow inclusion criteria. This might suggest there have been significant improvements in the quality of databases and search engines since they conducted their review.

## *Study and Sample-Level Characteristics*

Tables 1 and 2 and Tables S4 through S8 in the online version of the journal respectively present descriptive statistics, concerning the study, sample, and effect size levels. The final meta-analysis was based on 96 nonoverlapping samples from 76 studies, including 37,620 individual students. These studies based their

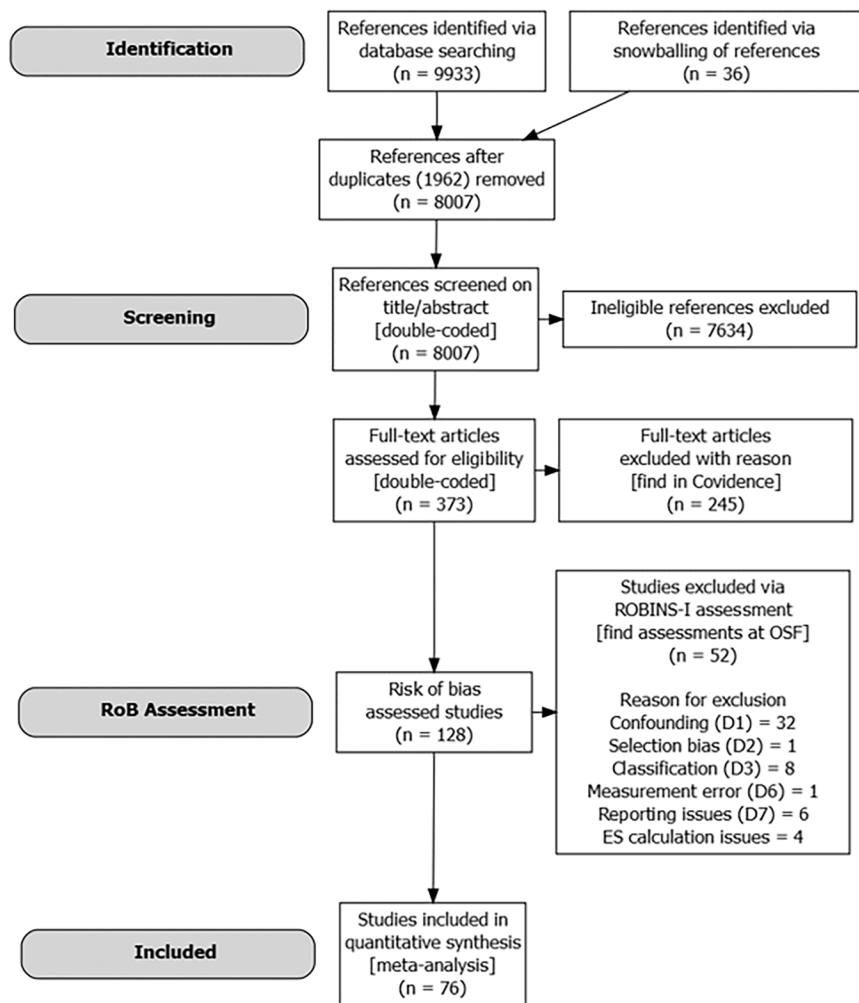


FIGURE 1. PRISMA flow chart showing the search, screening, and exclusion process.

results on 1 to 5 samples (mean = 1.263 samples per study). Most studies used U.S. data (69), with the remaining conducted in Belgium (1), Canada (1), Denmark (1), England (1), Hong Kong (1), and Taiwan (2). Across the included studies, 36 (47%) focused on elementary school students, 23 (30%) focused on middle school students, and 18 (23%) focused on high school students. Samples included students ranging from 1st to 11th grade (mean = 5.49). Consequently, studies solely focusing on 12th-grade students were absent from this review (see Figure S1 in the online version of the journal). Studies predominantly evaluated co-teaching

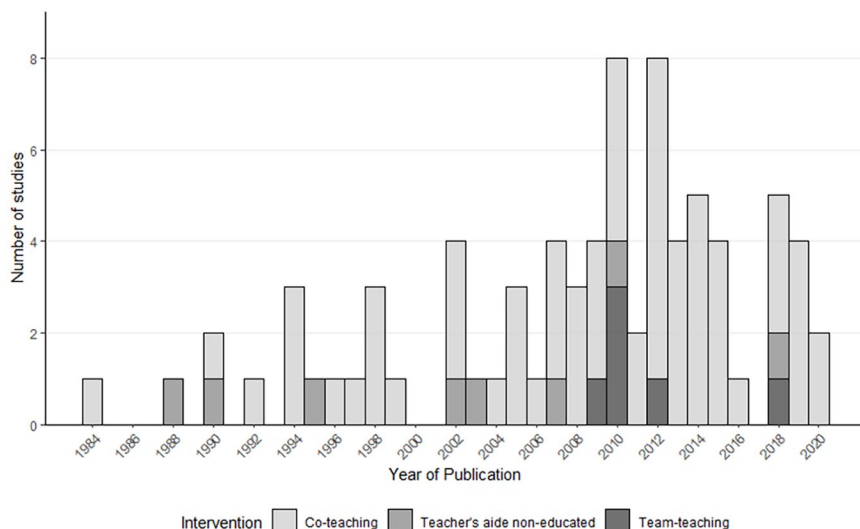


FIGURE 2. Number of studies included in the meta-analysis by year and intervention. Three studies (Andersen et al., 2018; LaFever, 2012; Powell, 2007) have examined more than one intervention. Therefore, 79 “studies” (intervention observations) are represented in the figure. Thereby, it illustrates the publication trends in the three subareas of collaborative models of instruction.

interventions, with co-teaching studies (65) far outnumbering studies of interventions involving teacher assistants (8) and team teaching (6). The size of the studies varied considerably. There were many small-sample studies within the co-teaching literature ( $M = 198$ ,  $SD = 515$ ); while studies involving teacher assistants, by contrast, were often based on large samples ( $M = 1915$ ,  $SD = 3264$ ), primarily driven by three large-scale cluster-randomized trials (these are Andersen et al., 2018; Finn & Achilles, 1990; Lapsley et al., 2002). Find further details about treatment and control group sample size distributions in Figure S2 and Tables S4 to S7 in the online version of the journal. The mean duration of interventions was 37.34 weeks ( $SD = 23.77$ ), which is close to 1 year of schooling. However, duration varied substantially between studies, ranging from 3 to 160 weeks. The average number of (45-minute) sessions per week was 11.4. However, this characteristic was seldom reported, which might have impacted this estimate.

Approximately half of the included studies reported whether co-teaching training was provided before the treatment, with 27 studies reporting the use of training and 10 studies reporting no training. Aligned with theoretical recommendations for the practice of co-teaching, it was common for studies (59) to document whether time was allocated for the co-planning of lessons, which was the case in all but 6 of these studies.

Regarding research designs, only nine studies were RCTs. Thus, the vast majority of studies (67) were quasi-experimental or observational studies. A



**TABLE 1**  
*Descriptive percentages for the included studies*

Study-level characteristics	Studies (J)	Samples (I)	Effect sizes (K)	Percentage <sub>J</sub>
Study context				
U.S. studies	69	88	253	0.908
Study design				
(C)RCT	9	9	59	0.118
QES	21	27	102	0.276
Observational studies	46	60	129	0.605
Study outlet				
Dissertation/thesis	55	67	158	0.724
Journal article	17	25	116	0.224
Others, incl. conf. abstracts	4	4	16	0.053
Interventions				
Co-teaching	65	80	236	0.855
Teacher assistants	8	11	36	0.105
Team-teaching	6	8	18	0.079
Student characteristics				
Elementary school (Grades 1–5)	36	52	148	0.474
Middle school (Grades 6–8)	23	25	84	0.303
High school (Grades 9–12)	18	19	54	0.237
Intervention characteristics				
Co-teaching training not provided	10	10	34	0.132
Co-teaching training	27	37	135	0.355
Common planning time not provided	6	6	33	0.079
Common planning time	53	71	193	0.697
Methodological features				
% no cluster treatment	67	87	241	0.882
Effect size–level characteristics	J	I	K	Percentage <sub>K</sub>
Outcome characteristics				
Language arts tests	53	71	165	0.569
Math tests	44	50	104	0.359
Science tests	8	8	13	0.045
Social science tests	3	3	4	0.014
History tests	1	1	1	0.003
Combi tests	3	3	3	0.01
Standardized tests	70	86	250	0.862
Follow-up test (3 months<)	2	2	8	0.028
Controls				
General education control group	53	61	190	0.697
Special education control group	33	44	100	0.434
Effect size characteristics				
Special education sample	43	54	137	0.472
General education sample	29	33	84	0.29
Blended sample	19	23	69	0.238
Pre-test adjusted	64	84	238	0.821
Covariates adjusted	69	89	255	0.879
Serious/high RoB	49	62	145	0.5

**TABLE 2**  
*Descriptive means of included studies*

Characteristics	J	I	K	Mean <sub>i</sub>	SD	Range
Sample characteristics						
Number of students	76	96	290	391	1286	10–10,781
Effective sample size <sup>a</sup>	76	96	290	18	20	5–113
Intervention group	76	96	290	157.96	494.648	5–4,016
Control group	76	96	290	232.835	812.151	5–6,765
Sample size (co-teaching)	65	80	236	198	514.973	10–4,368
Sample size (teacher assistant)	8	11	36	1915	3,264.341	54–10,781
Sample size (team-teaching)	6	8	18	518	1,186.489	46–3,450
Grade	76	96	290	5.492	2.845	1–11
Duration in weeks	70	90	273	37.348	23.772	3–160
Sessions per week	27	35	160	11.404	8.221	1–25
% Males in sample	57	66	239	55.971	9.632	31.8–77.5
Number of samples per study	76	96	290	1.263	0.772	1–5
Methodological features						
Effect sizes per study	76	96	290	3.816 <sup>b</sup>	4.21	1–27
Mean obtainable pre-posttest $\rho$	24	29	90	0.611 <sup>c</sup>	0.168	–0.036 – 0.92

<sup>a</sup>Calculated via  $4/\sigma^2_i$ , where  $\sigma^2_i$  is the effect size sampling variance both containing individual and cluster level heterogeneity.  
<sup>b</sup>This mean was aggregated to the study level.  
<sup>c</sup>This mean was calculated at the effect size level.

distinct feature of this review is that 59 (76%) studies were characterized as gray literature, including 55 dissertations (71%) and 4 conference papers (5%).

### *Effect Size Level Characteristics*

Several characteristics varied between the different effect sizes extracted from the same study. We calculated 290 effect sizes distributed across 96 samples from 76 studies. Most effect sizes (269) were calculated from achievement tests in either language arts (165 effect sizes from 53 studies) or mathematics (104 effect sizes from 44 studies). The mean percentage of male respondents in the sample was ~56%, ranging from 31.8% to 77.5%. There were 137 effect sizes from 43 studies on special needs students relative to 84 effect sizes from 29 studies and 69 effect sizes from 19 studies on general education and blended samples of students, respectively. In the majority (86.2%) of studies, effect sizes were obtained from standardized achievement test measures. Only eight effect sizes represented follow-up effect size estimates (i.e., effects measured more than 3 months after the end of the intervention). Further, only one study (Andersen et al., 2018) reported intention-to-treat effect sizes. As a consequence, we concentrated exclusively on treatment-on-the-treated effects.

We applied two-level cluster design adjustments for 67 studies (~88%) because these did not adequately account for school- and/or class-level nesting of students.

The mean number of effect sizes per study was 3.8 (ranging from 1 to 27, median = 2). Of the calculated effect sizes, 82% were adjusted for pretest differences among students, and another 6% were adjusted for other focal covariate differences (find further univariate descriptive information in Section S5 in the online version of the journal).

### **Risk of Bias**

Figures 3 and 4 depict weighted summary plot results of the RoB assessment for nonrandomized and randomized studies, respectively. Specifically, 67 studies were assessed via the ROBINS-I tool, while 9 were assessed using either the RoB 2 or RoB 2 CRCT tools. The plots are weighted by CHE model weights (Pustejovsky, 2020c). Therefore, the plots show “the proportion of information rather than the proportion of studies that is at a particular risk of bias” (McGuinness, 2021). See Section S6 in the online version of the journal for further details about the RoB assessment, including unweighted plots and separate plots for quasi-experimental and observational studies.

As illustrated in Figure 3, ROBINS-I assessed effect sizes were most frequently rated as having a moderate risk of bias due to confounding and/or reporting issues, mainly because pretest-adjusted effect sizes were rated to be of moderate risk of bias due to confounding. This assessment was based on the consideration that it might be unrealistic to expect the pretest adjustment (or focal covariate adjustment) to control out all imbalances between intervention groups under all circumstances. We only judged ROBINS-I assessed studies in which randomization was used but not controlled or initiated by the researchers to ensure a low risk of confounding. By contrast, most RoB 2 assessed effect sizes were rated as having a low risk of bias due to randomization, as shown in Figure 4.

The majority of the included studies across all research designs were assessed as having a moderate risk of bias due to (selective) reporting because none of the included studies were preregistered. We only considered studies that provided the raw data to be of low risk of bias due to reporting. For RCTs, the main reason for not being rated as of low overall risk of bias was the lack of preregistration.

Notably, 45 of 67 nonrandomized studies contained at least one effect size with an overall rating as having serious risk of bias. In total, 135 ROBINS-I assessed effect sizes received an overall rating of serious risk of bias. The most common reason (i.e., 24.4% of the ROBINS-I assessed effect sizes; see Table S12 in the online version of the journal) for a serious RoB assessment was limited descriptions of the implementation process (D4). Likewise, ~10% of the ROBINS-I assessed effect sizes received a serious RoB assessment due to classification issues, often because the control group was vaguely defined as “treatment as usual.” Generally, measurement of outcomes did not lead to many serious RoB ratings as 86% of the included studies applied standardized testing.

In sum, serious consideration must be given to the outcomes of risk of bias assessments for the sample of studies when conducting the subsequent meta-analysis. Half of the included effect size estimates were assessed to have a serious overall risk of bias. This result was mainly driven by the fact that most of the included studies applied nonrandomized research designs. In our first set of subgroup analyses, we contrast the differences between serious and nonserious risk

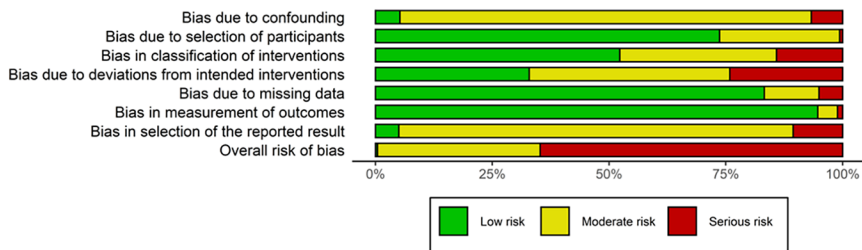


FIGURE 3. *ROBINS-I weighted summary plot. This plot contains information related to 225 effect sizes coming from 67 nonrandomized studies, of which 98 effect sizes come from 21 quasi-experimental studies, and 127 effect sizes come from 46 observational studies.*

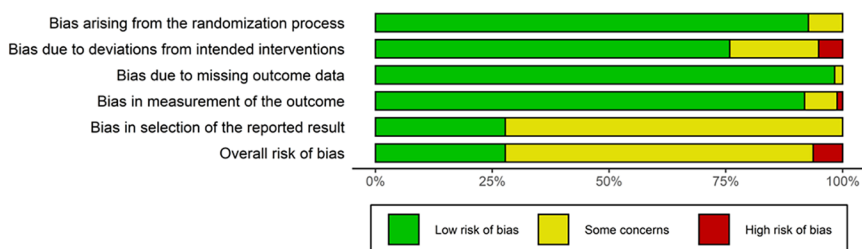


FIGURE 4. *RoB 2 and RoB 2 CRCT weighted summary plot. This plot contains information related to 55 effect sizes coming from nine randomized studies, of which 26 effect sizes come from four RCTs, and 29 effect sizes come from five cluster RCTs.*

of bias effect sizes to investigate the impact of including studies and effect sizes considered to be at serious risk of bias on the main findings.

## Meta-Analysis

### Mean Effect Size Estimation

The overall mean effect size from our meta-analysis summarizes a total of 280 effect sizes across 96 samples from 76 studies. For the mean effect size analysis, we excluded the eight follow-up effect size estimates and two effect sizes from the special needs student sample in Schaef (2014) since this sample overlapped with the blended student sample from which the rest of the effect sizes were estimated. The forest plot in Figure 5 depicts the distribution of dependent effect sizes from each study around the estimated overall mean effect size. Furthermore, the specific weight attributed to each single effect size can also be found in Figure 5.

We found a positive, statistically significant overall standardized mean difference of 0.11 standard deviations (*SD*),  $t(40.3) = 2.97$ ,  $p = .005$ , 95% *CI*[.035, .184]. In line with Kraft's (2020) benchmarks for interpreting education interventions with standardized achievement outcomes, we consider this to be a moderate effect size. Using Cohen's  $U_3$ , this result indicates that, on

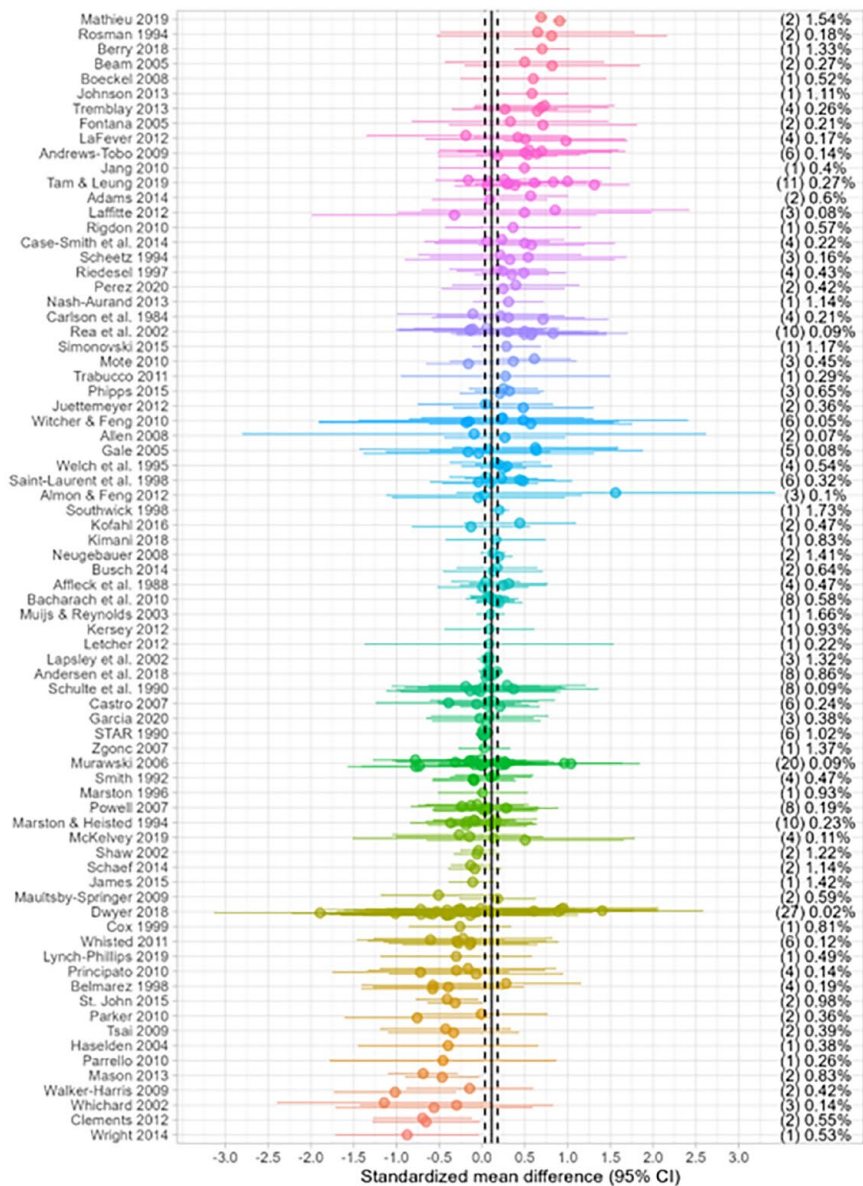


FIGURE 5. Mean effect size forest plot across dependent effect sizes. Number of effect sizes per study in parentheses. Percentages indicate the weight given to each point within the given study. Studies are ordered by the study mean effect size obtained from fitting the within-study effect sizes to a univariate meta-analysis model, as suggested by Fernández-Castilla et al. (2020). The bold line indicates the overall average effect size ( $\bar{g} = .11$ ), and the dashed lines indicate the 95% confidence interval from the fitted CHE-RVE model with  $\rho = .706$ .

average, co-taught students had better achievement scores than 54.4% of the control students (Baird & Pane, 2019; Valentine et al., 2019); or, put differently, there is a 54.4% chance that a randomly sampled score from the intervention group lies above the mean of the control group. At the student level, this translates to an expectation that a typical student from the control group would have had a percentile gain of 4.4% had they instead been exposed to a collaborative model of instruction (WWC, 2020). We found substantial heterogeneity among effect sizes,  $Q(279) = 1,164.2$ ,  $p < .0001$ ,  $I^2 = 91.73$ , with variance components (reported as *SDs*) of 0.255 *SD* at the effect size level, 0.102 *SD* at the study level, and a total *SD* of 0.274.<sup>6</sup> This suggests that both the study and effect level covariates might be able to explain differences in effect size estimates across studies, which in turn statistically justified all of our planned moderator analyses (cf. Pustejovsky & Tipton, 2021).

### *Sensitivity Analyses*

The overall mean effect size ( $\bar{g}$ ) was insensitive to changing assumptions about the sampling correlation,  $\rho$ , among effect sizes from the same study, with estimates varying from .107, 95% *CI*[.0332, .180] assuming  $\rho = .0$  to maximum .111, 95% *CI*[.034, .187] when  $\rho = .6$ . The total variance component estimate was largely insensitive to changing  $\rho$ . Nonetheless, some true variation was detected, with the total *SD* ranging from 0.232 to 0.347. By contrast, the individual variance components were highly sensitive to the magnitude of  $\rho$  (see Figure S10 in the online version of the journal). Therefore, the relative magnitude of the estimates of individual variance components should be interpreted with caution. The same patterns emerged from the leave-one-study-out analyses, in which  $\bar{g}$  was not substantially influenced by any single study.  $\bar{g}$  ranged from .088, 95% *CI*[.031, .144] when omitting Mathieu (2019) to .122, 95% *CI*[.05, .194] when omitting Mason (2013). The total variance component estimate was generally insensitive to the omission of any single study, ranging from .243 to .285. Yet between-study variation was heavily impacted by omitting Mathieu (2019). In fact, the study level *SD* reduced to .0 when omitting this study, indicating that the between-study variance estimation was fragile (see Figures S11 and S12 in the online version of the journal).

Figure 6 shows how the mean effect size and variance estimation respectively were influenced by changing the assumptions related to the effect size calculation and the applied inclusion criteria. Generally,  $\bar{g}$  was agnostic to assumptions related to the effect size calculation procedure, with  $\bar{g}$  ranging from .094, 95% *CI*[.028, .159] when using the most extreme alternative effect sizes from studies reporting multiple results eligible for different effect size calculations to .113, 95% *CI*[.044, .183] for calculations based on an imputed constant pre-posttest correlation of .8 for studies reporting difference-in-differences results without providing the opportunity to obtain the pre-posttest correlation. The total *SD* estimate varied slightly when changing assumptions related to effect size calculation, ranging from 0.248 to 0.377.

Overall,  $\bar{g}$  was not substantially influenced by changes to any of the inclusion criteria. Under all changed conditions,  $\bar{g}$  remained in the moderate effect size



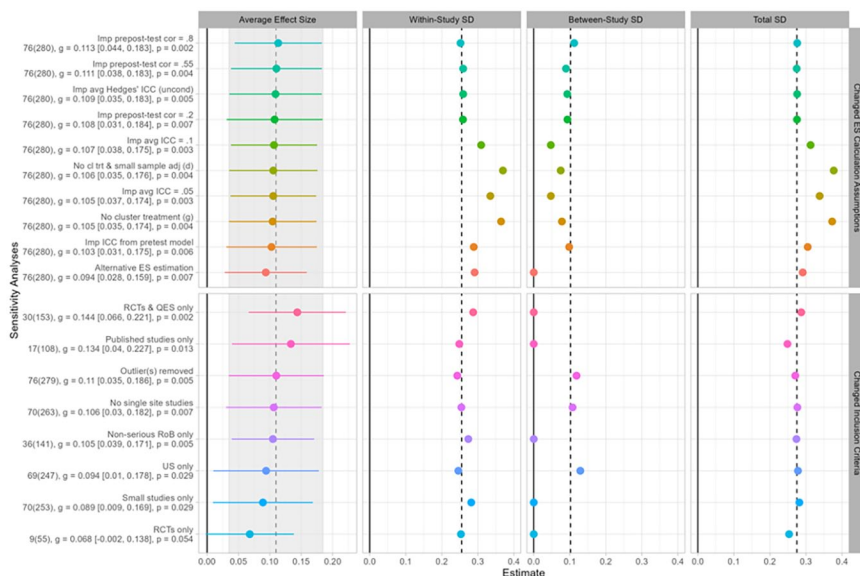


FIGURE 6. Sensitivity analyses changing effect size calculation assumptions and inclusion criteria. Dashed lines and shades indicate the estimated values and the confidence interval from the overall average effect size of the CHE-RVE model with  $\rho = .706$ .

interval (cf. Kraft, 2020), with  $\bar{g}$  ranging from .068, 95% CI[−.002, .138] when reestimated for only RCT studies to .144, 95% CI[.066, .221] when observational studies were excluded. As indicated by the confidence interval,  $\bar{g}$  only became statistically insignificant when the reestimation was based only on RCT studies. However, this sensitivity analysis was based on the smallest number of studies and effect sizes (nine studies and 55 effect sizes) relative to the rest of the analysis, substantially reducing the power of the conducted test.

In line with theoretical expectations regarding publication bias (Cheung & Slavin, 2016; Rothstein et al., 2005), we found that  $\bar{g}$  slightly increased when omitting gray literature. Interestingly, we found that  $\bar{g}$  is not impacted by the omission of all effect sizes with an overall rating as having serious risk of bias, despite this reducing the sample by 40 studies and 139 effect size estimates. Contrary to our expectations, omitting all large-scale studies with sample sizes greater than 1,000 students slightly reduced  $\bar{g}$ . We can therefore conclude that  $\bar{g}$  was not primarily driven by the included large-scale studies, such as the three large-cluster RCTs, although these were given more weight relative to the smaller studies. The total *SD* and the effect size level *SD* estimations were largely agnostic to the changed inclusion criteria, whereas the changes had more impact on the study-level *SD*; see the third column in the bottom row in Figure 6.

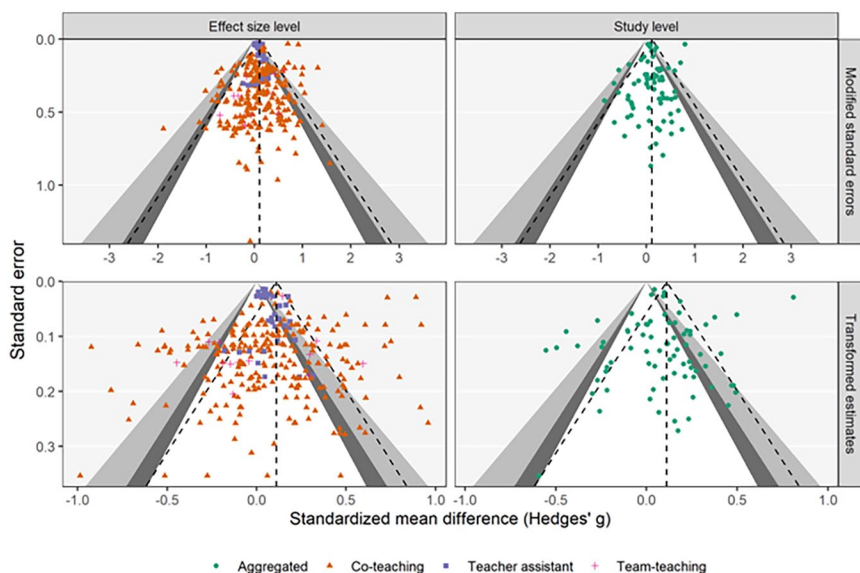


FIGURE 7. Contour-enhanced funnel plots. The contour-enhanced funnel plots present estimates at the effect size and study level using modified standard error and transformed estimates, respectively. The white region indicates  $p > .10$ , the dark gray region corresponds to  $p$  values from .05 to .1, and the gray region corresponds to  $p$  values from .01 to .05. The light gray region outside the funnel plot shows  $p$  values  $< .01$ . Dashed lines mark the distribution around the estimated mean effect size.

### Publication Bias

We conducted a range of complementary publication bias and/or small study effect tests without finding any systematic evidence for publication bias or small-study effects. Figure 7 depicts contour-enhanced funnel plots conducted at the effect size and the study level using modified standard errors and transformed effect size estimates, respectively. These plots indicate no small-study effects/publication bias. Generally, we found no systematic indication of publication bias or small-study effects based on the applied tests. This applies to the trim-and-fill test, cluster-robust Egger's regression tests, selection models test, and the WAAP test (see Section S11 in the online version of the journal for the results). We also conducted a sensitivity test for publication bias in which we estimated the worst-case meta-analysis by removing all affirmative effect sizes (i.e., assuming that all positive and statistically significant effect sizes are false positives) using the CHE-RVE model, which yielded  $\bar{g} = .043$ , 95%  $CI[-.049, .091]$ . In this case, the point estimate fell just below the threshold of a moderate effect size and became statistically insignificant. However, these analyses further indicated that it is unreasonable to expect publication bias to reduce the overall effect size to null and that affirmative effect sizes should be at least 11-fold more

likely to be published than negative or nonsignificant results for the confidence interval to include null. Overall, we do not expect publication bias to have substantially influenced estimations of the mean effect size. This is of little surprise given that more than 77% (57) of the included studies represent gray literature (Pigott et al., 2013).

### *Subgroup Analyses*

While the previously presented CHE-RVE model summarized the overall mean difference between co- and single-taught classrooms across all  $K = 280$  effect sizes, accounting for dependent effect sizes, it did not take into account potential differences in effect sizes across interventions, outcomes, participants, research designs, risk of bias assessment, and publication characteristics. In order to study such potential heterogeneities, we conducted a comprehensive range of meta-regression analyses. Table 3 reports the results of subgroup analyses of focal moderator variables that are categorical and did not contain any missing values (see Section S9 in the online version of the journal for relevant subgroup forest plots). All analyses reported in Table 3 were based on 275 effect sizes across 94 samples from 74 studies. In total, we excluded 15 effect sizes from six studies, of which two studies were fully excluded from the subgroup analysis dataset. As with the mean effect size model, we excluded the eight follow-up effect sizes due to the small sample preventing reliable estimation of the mean effect size for this subgroup dimension. We further excluded a number of effect sizes and studies since their research design did not allow us to explore moderating effects. This included excluding all four effect sizes calculated from Carlson et al. (1984) because the sample represented a mix of students across Grades 1 to 12. Furthermore, we excluded three effect sizes from three studies because they used achievement tests that were based on tests averaging results across language arts and math test measures. Contrary to the mean effect size model, we included all effect sizes from Schaefer (2014) since this allowed us to explore differences across student samples.

As can be seen from Table 3, we generally found quite robust effects of collaborative instruction across most of the conducted subgroup analyses. Most of the individual effects are not statistically different from zero, as might be expected due to the rather small samples and effects across subgroups. The empirically estimated average group means across most subgroup dimensions of collaborative instruction fell within the interval of moderate effects from .05 up to .20. We only found two statistically significant average group differences, both for the unconditional models (i.e., those without controls) and the covariate-adjusted models. These were between covariate-adjusted and posttest-only effect sizes, with  $F(1, 7.8) = 10.2, p = .013$  (CWB  $p = .001$ ); and between OBS, QES, and RCTs, with  $F(2, 12) = 4.86, p = .028$  (CWB  $p = .004$ ). The former test showed, contra to previous research (cf. Cheung & Slavin, 2016; Lipsey & Wilson, 2001), that covariate-adjusted effect sizes yielded substantially larger effect sizes than posttest effect sizes, while the latter test indicated that QES yielded larger positive effect sizes than RCTs and OBS. The results were similar when controlling for differences across subject, grade level, and student sample characteristics.

Table 3

*Subgroup analyses for focal moderators without missingness*

Subgroup-analyses		Unadjusted effects				Covariate-adjusted effects <sup>a</sup>			
Coefficient		Est. [95% CI]		HTZ $p$ value (CWB) <sup>b</sup>		Est. [95% CI]		HTZ $p$ value (CWB)	
Intervention characteristics	Studies	ES	HTZ $p$ value (CWB) <sup>b</sup>	Satt. d.f.	$SD$ ( $\tau + \omega$ )	HTZ $p$ value (CWB)	Satt. d.f.	$SD$ ( $\tau + \omega$ )	
Co-teaching <sup>c</sup>	63	226	.12* [.02, .22]	46.1	.332	.112 [-.012, .236]	31.3	.332	
Teacher assistant	8	33	.067 [-.012, .147]	<b>2.6</b>	.025	.08 [-.051, .212]	8.4	.017	
Team-teaching	6	16	.087 [-.022, .196]	<b>1.1</b>	.033	.082 [-.053, .216]	7.5	.045	
Wald test $p$ values <sup>d</sup>			.343 (.361)			.612 (.592)			
Outcome characteristics									
Arts and social science <sup>e</sup>	54	162	.137** [.056, .217]	34.6	.253	.177** [.059, .296]	21.9	.254	
STEM	48	113	.076 [-.021, .173]	34.4	.324	.125 [-.002, .252]	28.4	.335	
Wald test $p$ values			.179 (.201)			.234 (.266)			
Posttest ES <sup>c</sup>	11	35	-.264 [-.564, .035]	6.5	.236	-.262 [-.59, .067]	9.5	.258	
Covariate adjusted ES	67	240	.15*** [.075, .224]	26.4	.273	.155* [.04, .269]	23.9	.274	
Wald test $p$ values			.013* (.001**)			.014* (.003**)			
Nonstandardized test <sup>d</sup>	14	39	.107 [-.085, .298]	7.4	.096	.082 [-.15, .313]	9.7	.099	
Standardized test	68	236	.111** [.033, .189]	34	.286	.104 [-.032, .239]	28.6	.287	
Wald test $p$ values			0.961 (.961)			.827 (.812)			
Participants characteristics									
Blended sample <sup>e</sup>	19	61	.066 [-.005, .137]	<b>2.8</b>	.024	.043 [-.05, .136]	<b>2.4</b>	.021	
General education sample	26	79	.038 [-.081, .157]	18.8	.324	.026 [-.104, .156]	20.4	.329	
Special needs sample	42	135	.143* [.016, .269]	33.5	.344	.119 [-.007, .244]	33.7	.341	
Wald test $p$ values <sup>f</sup>			.213 (.220)			.279 (.269)			
Elementary school (1–5) <sup>g</sup>	35	142	.122** [.047, .198]	10.7	.078	.137* [.015, .258]	12.2	.046	
Middle school (6–8)	23	79	.08 [-.106, .267]	18.1	.372	.072 [-.121, .264]	19.7	.373	
High school (9–12)	18	54	.046 [-.133, .226]	12.9	.349	.033 [-.154, .219]	14.4	.346	
Wald test $p$ values			.676 (.697)			.502 (.515)			

(continued)

TABLE 3. (continued)

Subgroup-analyses		Unadjusted effects				Covariate-adjusted effects <sup>a</sup>			
Coefficient	Studies	ES	Est. [95% CI]		Satt. d.f.	SD ( $\tau + \omega$ )	Est. [95% CI]		Satt. d.f.
			HTZ <i>p</i> value (CWB) <sup>b</sup>	HTZ <i>p</i> value (CWB)			HTZ <i>p</i> value (CWB)	HTZ <i>p</i> value (CWB)	
Intervention characteristics									
Special edu. control group <sup>c</sup>	32	96	.173* [.023, .323]		25.8	.314	.137 [−.028, .301]		29.3
General edu. control group	51	179	.076* [.012, .141]		16.4	.26	.048 [−.032, .128]		13.4
Wald test <i>p</i> values			.225 (.249)				.225 (.297)		
Study characteristics									
Observational <sup>e</sup>	46	129	.064 [−.059, .187]		37.4	.304	.026 [−.132, .185]		25.4
QES	20	95	.245** [.126, .365]		8.8	.321	.225** [.103, .346]		17.4
(C)RCT	8	51	.063 [−.012, .137]		<b>2.5</b>	.254	.051 [−.079, .181]		13.4
Wald test <i>p</i> values			.028* (.004**)				.011* (.002**)		
Gray literature <sup>c</sup>	57	168	.081 [−.027, .189]		42.3	.308	.082 [−.059, .224]		28.1
Published literature	17	107	.136* [.04, .231]		6.1	.25	.156* [.016, .296]		17.9
Wald test <i>p</i> values			.421 (.416)				.292 (.299)		
RoB low/moderate <sup>c</sup>	34	136	.11** [.042, .177]		9.6	.275	.143 [−.026, .313]		22.7
RoB serious	46	139	.08 [−.038, .197]		35.4	.309	.075 [−.062, .213]		28
Wald test <i>p</i> values			.641 (.631)				.384 (.376)		

*Note.* Bold degrees of freedom (d.f.) estimates indicate low d.f. values, which in turn indicate that the given variance estimation was fragile. Find more detailed analysis in Section S12 in the online version of the journal. The table is based on 275 effect sizes across 94 samples from 74 studies.

<sup>a</sup>Adjusted for the grade level, student sample, and subject.

<sup>b</sup>CWB represents adjusted (CR2) cluster wild bootstrapping *p* values using 1999 replications.

<sup>c</sup>SCE+ model.

<sup>d</sup>Comparison was made between co-teaching and teacher assistant interventions only.

<sup>e</sup>CMVE+ model.

<sup>f</sup>Comparison were made between general and special needs students only.

\**p* < .05, \*\**p* < .01, \*\*\**p* < .001.

Notably, substantial heterogeneity remained across the majority of the conducted subgroup analyses.

The difference between the various categories of collaborative instruction was below practical relevance and not statistically significant. The average subgroup effect size ranges from .067, 95% *CI*[-.012, .147] for teacher assistant interventions to .12, 95% *CI*[-.020, .22] for co-teaching interventions in the unconditional model. We found neither statistical nor substantial important differences between subject categories, with the average group effect sizes ranging from .076, 95% *CI*[-.021, .173] for STEM; to .137, 95% *CI*[-.056, .217] for arts and social science outcomes in the unconditional model, with both results considered to be moderate in size. The unconditional model did indicate a small effect (cf. Kraft, 2020) not statistically distinct from zero, for effect sizes based on samples of general education students of .038, 95% *CI*[-.081, .157]; and a moderate effect statistically distinct from zero for effect sizes premised upon samples of special needs students of .143, 95% *CI*[-.016, .269]. Meanwhile, we did not find a statistically significant difference between the two means,  $F(1, 33.6) = 1.61, p = .213$  (CWB  $p = .220$ ), which suggests that general education and special needs students might benefit equally from collective instruction. Nevertheless, the effects for general students are substantially smaller than for special needs students, which might be of practical relevance if this estimate represents the true effect of collaborative instruction on the general student population. The models we used did not have enough statistical power to draw a firm statistical conclusion on this matter, but a previous meta-analysis suggested that the effect might be higher and moderate in size when general students receive full-time co-teaching (Szumski et al., 2017).

Results differed across grade levels, with  $\bar{g} = .122$ , 95% *CI*[-.047, .198],  $p = .004$  for elementary school outcomes;  $\bar{g} = .08$ , 95% *CI*[-.106, .267] for middle school outcomes; and  $\bar{g} = .046$ , 95% *CI*[-.133, .226] for high school outcomes. However, the results did not reveal any statistically significant differences across grade levels for the unconditional model,  $F(2, 25.6) = 0.398, p = .676$  (CWB  $p = .697$ ). Although the mean effect size for high school students fell within the small effect interval, it can be considered a substantial effect compared to the annual gain usually experienced in later grades (Lipsey et al., 2012). Similarly, the declining trend that we found for effects from earlier to later grade levels also confirms the tendency in terms of annual gains across subjects found in nationally normed tests in the United States (Lipsey et al., 2012).

Furthermore, we did not find any practically relevant mean differences between subgroups for risk of bias, the study outlet, and the type of test categories,<sup>7</sup> with mean values for all subgroups distributed closely around the overall average effect sizes, ranging from 0.08 *SD* to 0.136 *SD* across the unconditional models. Nor did we find any statistically significant difference between effect sizes based on general or special education control groups. As an exploratory analysis, we conducted the same test on a subsample only comprising special needs students in order to investigate if either of the service delivery models (inclusive/general education single-taught vs. special education classrooms) could be considered superior to the other. Once again, we did not find any statistically or practically significant difference between the mean effect sizes of different subgroups (see Table S16 in the online version of the journal).



The results from the unconditional models were largely equivalent to models controlling for subject, grade-level, and student sample differences, with no substantial differences across all HTZ and CWB Wald tests. In addition, we did not find any inferential discrepancies between HTZ and CWB  $p$  values across all types of models. Moreover, we conducted tests correcting for multiplicity<sup>8</sup> by using the false discovery rate method (Benjamini & Hochberg, 1995; Laird et al., 2005; Polanin, 2013), which downgraded our  $p$  value threshold to .01. This did not change our inferences when based on CWB  $p$  values. As a sensitivity analysis, we reestimated this set of subgroup analyses based only on co-teaching effect sizes, which are the effect sizes from interventions comprised of collaboration among general and special education teachers (see Table S15 in the online version of the journal). We did not find any noteworthy difference in results between the two sets of subgroup analyses.

### *Moderator Analyses With Missing Values*

We conducted two sets of moderator analyses for covariates/predictors of theoretical concern based on multiple imputed values for missing values on these variables (see Tables S18 and S19 in the online version of the journal). The first set of analyses included categorical moderators concerning the comparisons between studies allocating versus studies not allocating time for co-planning of lessons and between studies providing training in co-teaching versus providing no such training. The second set concerned continuous variables, including the effects of the duration and intensity of collaborative instruction as well as the average percentage of males in the sample. These analyses did not find any statistically significant effects, suggesting that none of these moderators explained differences in effectiveness, despite clear indications that this is the case in the co-teaching literature.

## **Discussion**

Over the last four decades, the volume of quantitative research on the effectiveness of collaborative models of instruction in improving students' academic achievement has increased considerably. We have demonstrated that this increase has been greater than often suggested in primary research and previous reviews on the topic by evaluating 128 studies with treatment-control designs. Notably, our search procedure resulted in the identification of a greater number of relevant studies across the entirety of this period than in previous reviews. We applied state-of-the-art techniques to meta-analyze the results across 96 samples of students from 76 studies that we assessed as not being of critical risk of bias. This yielded 280 effect sizes, of which most were based on standardized achievement outcomes from language arts and math tests and pretest-adjusted measures.

Across the studies included for meta-analysis, varying in terms of intervention, location, implementation, outcomes, research design, and participant characteristics, we found that collaborative models of instruction significantly increase student achievement compared to either single-taught or special education instruction models. The effect was moderate in size compared to the results of previous causal research on educational interventions with standardized achievement outcomes (Kraft, 2020). It remained moderate in size across almost all conducted

sensitivity analyses and publication bias tests, with the great majority of these tests supporting the conclusion that the mean effect size is statistically distinct from zero. Importantly, the overall mean effect was not altered by the inclusion of a large number of effect sizes assessed to be of serious risk of bias. In contrast to previous discussions (Achilles et al., 2000), this review provides unambiguous evidence for the effectiveness of collaborative models of instruction on student achievement.

In order to explain the heterogeneity among the effects, we identified moderators that are considered as theoretically or methodologically important in the literature, where to we fitted a range of meta-regression models. To our surprise, we found that the effects of collaborative instruction were generally robust across the assessed moderators. The vast majority of subgroup effects fell within the interval of a moderate effect from .05 up to .20. This applied to the unconditional as well as the covariate-adjusted meta-regression models, controlling for student, grade, and subject differences. Interestingly, we found neither statistical nor practical differences between interventions involving special education co-teachers and those with teacher assistants. Therefore, in contrast to much co-teaching literature—as portrayed by L. Cook and Friend (1995)—our results suggest that the effectiveness of collaborative models of instruction does not necessarily hinge on specific co-teacher compositions, the educational background of the second teacher, and/or an equal share of teaching responsibilities between co-teachers. As such, our study indicates that the mechanisms through which collaborative models of instruction work might be less complicated than often assumed in co-teaching literature. In addition, we did not find any notable differential effects across subjects. The mean difference between arts and social science versus STEM subjects was practically small and not statistically significant. Across grade levels, we found that the average effect size slightly declines for higher grade levels, which confirms trends found in previous evaluations and benchmarks of annual gains across grade levels in the United States (Lipsey et al., 2012). However, we did not find any statistically significant difference between the mean effect sizes for elementary, middle, and high school outcomes, suggesting that collaborative models of instruction can potentially be effective across all grade levels. Although we did find a small effect for general education students, the difference between the mean effect for general and special education students remained statistically insignificant. This might suggest that collaborative instruction can benefit general education students as well (as also suggested by Szumski et al., 2017). However, our results are somewhat uncertain in this regard as we found substantively relevant differences, although they were not statistically significant. Thus, these results should be interpreted with caution.

Furthermore, we tested a range of factors that in the co-teaching literature are highlighted as practically relevant preconditions for effective co-teaching. These factors included the allocation of time for co-planning lessons, training in co-teaching methods and strategies, the duration and intensity of the intervention, and the number of male students in the sample. We found that none of these factors were able to explain the difference in effects across studies or effect sizes. However, all of these analyses were based on variables with a large share of missing values. Although we used multiple imputation techniques to remedy this issue,

we suggest caution in drawing any conclusions, seeing the results as preliminary. Identifying moderators and conditions for a successful implementation of co-teaching is thus an area that calls for further experimental investigation. In this regard, qualitative studies might also add to our understanding of the mechanisms underpinning potential moderating effects (Shadish et al., 2002).

In a similar vein, our results indicate that the observed study characteristics of this review do not fully explain true differences across outcomes between and within studies since considerable heterogeneity remained at both the effect size and study level for the majority of the moderator analyses. This heterogeneity only disappeared for subgroups in which the total number of studies and effect sizes were limited. Consequently, there is still a need for further investigation of differences in the effects of different collaborative models of instruction and different settings.

### *Limitations*

Although we have performed a comprehensive literature search and attempted to offer in-depth analyses, this review has several clear limitations. One major limitation is that we only concentrated on students' academic achievement. This essentially circumscribes the general conclusion regarding the potential efficacy of collaborative models of instruction beyond academic achievement. From an educational perspective, academic achievement might not be the only reason for implementing collaborative models of instruction. Future research, reviews, and meta-analyses should certainly complement our results by studying the effects of collaborative models of instruction on other outcomes, such as student well-being, as well as social, behavioral, and teacher satisfaction measures. Additionally, we did not explore differential effects across subtypes of subjects, such as the differences between reading, writing, and spelling in terms of achievement outcomes, which might be essential in developing a more fine-grained and profound understanding of the effectiveness of collaborative instruction.

Several caveats should also be mentioned with regard to the included literature. In many cases, we experienced difficulties in obtaining information regarding the actual number of special needs students included in the specific general classroom (similar to Szumski et al., 2017). Moreover, it was often uncertain whether the number of special needs students remained constant across co-taught and single-taught classrooms. For special education control groups, it was in some cases quite difficult to decipher the exact number of adults present during the instruction. We assumed that the special education instruction was single-taught if not otherwise mentioned. Moreover, we were not always able to ensure that the class sizes across the treatment and control groups remained constant. It was likewise rare for studies to adjust for teacher differences across the treatment and control groups. In other words, only a few studies involved the same teachers across the treatment and control groups. Since we included a large number of observational studies, it was often difficult to assess the fidelity of the implementation of a given intervention. Altogether, these factors might potentially have induced some degree of error in our estimations and thus reduce the generalizability of the review.

Another limitation in terms of the generalizability of the review is the dominance of U.S. studies and its focus on education systems in high-income countries (according to the World Bank's definition). While we demonstrate that the interventions also proved effective in non-U.S. countries, there is a clear need for research into the generalizability to middle- and low-income countries in particular.

The review is also limited by the content of the included body of literature. Therefore, we were unable to investigate a range of questions of theoretical and practical importance (similar to Szumski et al., 2017), such as the impact of the number of included students with special needs in co-taught classes, the socioeconomic status of the students, the exact co-teaching model used, and the relationship between and teamwork skills in co-teacher teams. Such factors might be relevant topics for future research on differential effects of collaborative models of instruction. Furthermore, due to the focus on co-teaching in the vast majority of previous studies, our results are less conclusive with regard to the other collaborative models of instruction. While our results indicate that the effects are of similar size for all three instruction models, there are also small differences. Based on the imprecise estimates for "teacher assistants" and "team teaching," we cannot entirely rule out the possibility that there are differences between the effectiveness of the different methods of instruction.

Although we employed state-of-the-art review methods, several limitations also remain in this regard. For example, most of the risk of bias assessment and data extraction comprised single-coder and single-rater procedures, which may have induced some degree of error. However, all assessments and data extractions are available at <https://osf.io/fby7w/> for critical inspection and future updates. Finally, it is important to note that all publication bias tests that we applied have inherent deficiencies. Both the trim-and-fill test and cluster-robust Egger regression methods have limited power to detect small-study effects, especially when these effects are small and dependent effect sizes are present. Moreover, selection models based on dependent effect sizes aggregated to the study level do not fully calibrate the nominal Type-I error rate (Rodgers & Pustejovsky, 2021), and the WAAP test used was based on just three adequately powered studies. Therefore, while the results of all publication bias tests indicate the absence of reporting biases, the possibility of such bias cannot be ruled out entirely. Nevertheless, we do not consider publication bias to be a serious issue since this review included a large amount of gray literature.

### *Implications for Practice and Research*

Our results suggest that schools and teachers can improve academic learning for all students by implementing collaborative models of instruction and that the potential of such models is independent of the specific type of within-class collaboration. Moreover, it may be less complicated to increase the effectiveness of collaborative instruction than sometimes asserted in the co-teaching literature. As a consequence, we conclude that school leaders and educators can implement collaborative instruction across all arts and social science and STEM subjects, as well as all grade levels, even when no formally trained teachers are available or where there are scant resources allocated in terms of time for co-planning lessons and training in co-teaching methods and strategies. That is not to say that formal

qualifications have no impact on the efficacy of collaborative instruction, but our results suggest that the differences are too small to be of practical significance. Having an additional adult in the classroom thus seems to be a more relevant factor than the specific two-teacher composition or the qualifications of the second educator. Nonetheless, the effectiveness of collaborative models of instruction certainly might benefit from careful consideration of the local context and the concrete educators involved in the implementation and execution of the intervention. This also means that we cannot make a conclusion about the effects of support to co-teaching, such as kick-off workshops. Nor do we study the role of trust-building exercises between the two educators, mentoring by the organization, or other forms of support. All these should thus still be considered relevant.

Our results have several implications for educational research. The majority of the included studies only report short-term outcomes that were either measured during or immediately after the intervention. Thus, future research needs to concentrate more on assessing the long-term effects of collaborative instruction. Since there is a complete absence of cost-benefit analyses from the present body of literature, research is needed that focuses on exploring the costs and benefits of collaborative models of instruction, including comparisons with the costs and benefits of other related interventions, such as increasing instruction time and reducing class sizes.

As with all reviews and meta-analyses, the reliability, validity, and credibility of this review hinge on the quality of the included studies, which are predominantly nonrandomized studies. Most of the (cluster) randomized trials came from the teacher assistant literature and were large-scale trials. By contrast, co-teaching studies often had small sample sizes and were based on nonrandomized research designs. Thus, future co-teaching research must employ large-scale randomized controlled trials or high-quality matched-groups designs to assess the true effect of co-teaching. Overall, we argue that larger and more rigorously conducted studies are required, especially with regard to co-teaching interventions where there is a particular need for primary research studying variations across focal moderators and/or preconditions for effective co-teaching interventions (Hedges, 2018). There is also a need to learn more about the differences between the effects of collaborative instruction on general and special needs students. In contrast to other meta-analyses (Szumski et al., 2017), this review does not provide conclusive evidence of the effectiveness of such models in increasing the academic achievement of general education students.

### *Implications for Educational Policy*

Although we find a moderate and practically significant effect size, it is important to emphasize that introducing collaborative models of instruction can be costly. In contexts where resources are scarce, policymakers and local stakeholders could profitably start searching for cheaper and more efficient interventions. That said, our results suggest that interventions involving collaborative models of instruction have great potential in terms of their scalability and applicability since the effectiveness of these models did not appear to hinge on any specific composition of two-teacher instruction. This opens the possibility of relying on the comparatively inexpensive option of employing paraprofessional educators.

In contrast to previously discussed alternative policy options for improving student-teacher ratios, such as class size reduction (Achilles et al., 2000), an

advantage of collaborative instruction is that it can be implemented easily—also on a day-to-day basis. Notably, collaborative instruction is just as effective in increasing student achievement as other structural interventions, including increased instruction time (Kidron & Lindsay, 2014, p. 5, with  $\bar{g}$  ranging from  $-.04$  to  $.16$  across literacy and math subjects) and class reduction (Filges et al., 2018, p. 10,  $\bar{g} = .11$ , 95%  $CI[.05, .16]$ ,  $p = .0003$ ).


Although collaborative instruction cannot close the achievement gap between general education and special needs students (Dietrichson et al., 2017), our results suggest that it can function as a vital and significant tool for schools and school systems to accommodate the inclusion of students with special educational needs and/or disabilities in general education. We believe that collaborative instruction can, indeed, function as a contributing factor, further improving the educational achievements of special needs students in combination with other relevant interventions aimed at increasing this group's academic achievement (Dietrichson et al., 2020, 2021).

### *Conclusion*

The findings of this systematic review and meta-analysis provide evidence of the effectiveness of collaborative models of instruction in strengthening students' academic achievement. This is independent of the specific model of in-class collaboration between educators, the subject taught, and the grade level. Although the main results of this review were generally robust across all of the conducted sensitivity, publication bias, and moderator analyses, there is still plenty of room for further investigation within this field of study. A range of potentially relevant moderators could not be analyzed, for example, due to inadequate documentation in the existing research literature. Consequently, future studies should assign more weight to studying such moderators, and policymakers should bear in mind this gap in the current evidence base.

### **ORCID iDs**

Mikkel Holding Vembye  <https://orcid.org/0000-0001-9071-0724>

Bethany Hamilton Bhat  <https://orcid.org/0000-0002-9021-041X>

### **Notes**

Find the pre-registered protocols at <https://osf.io/ur2bs>. Find R codes to reproduce all parts and analyses of this paper at <https://osf.io/fby7w/>.

Thanks to James E. Pustejovsky for significant help on the effect size calculation, as well as for offering guidance on modeling. In this regard also thanks to Megha Joshi for guidance on how to implement cluster wild bootstrapping. Also thanks to Jens Dietrichson, Trine Filges, Terri D. Pigott, Wim van den Noortgate, and Monica Lervåg-Melby for their readings and comments.

<sup>1</sup> This definition should not be confused with Cock and Friend's (1995) "team teaching" model.

<sup>2</sup> When we interpret effect sizes throughout the article, we use Kraft's (2020) empirical guidelines and benchmarks for interpreting effect sizes related to causal research on education interventions with standardized achievement outcomes.



<sup>3</sup> Find the list of focal covariates/confounding factors and the reason why we included the given covariates in our preregistered protocol at <https://osf.io/ur2bs>.

<sup>4</sup> We expand on the theoretical and empirical reasons why we considered these covariates especially important in our protocol at <https://osf.io/ur2bs>.

<sup>5</sup> We sought statistical advice on this matter.

<sup>6</sup> The total *SD* is calculated by the square root of the sum of the within- and between-study variance.

<sup>7</sup> Recent research (Wolf & Harbatkin, 2022) has recommended not to amalgamate independent and nonindependent (e.g., standardized vs. researcher developed assessed) measures since the latter tend to overestimate the true effect. However, our meta-analytical data does not confirm this pattern. Meanwhile, we found that QES using standardized measures tend to systematically report more large, positive, and statistically significant effect sizes, suggesting a small-study effect for that group of studies. See Figure S13 in the online version of the journal for detailed analysis.

<sup>8</sup> Defined as the increased probability of committing a Type I error by conducting multiple statistical significance tests.

## References

- Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J. D., Folger, J., Johnston, J. M., & Word, E. (2008). *Project STAR Dataverse*. <https://dataverse.harvard.edu/dataverse/star>
- Achilles, C. M., Finn, J. D., Gerber, S., & Zaharias, J. B. (2000). *It's time to drop the other shoe: The evidence on teacher aides*. <https://eric.ed.gov/?id=ED447142>
- Adams, S. S. (2014). Coteaching in secondary special and general education classrooms and student mathematics achievement [Walden University]. *ProQuest Dissertations and Theses*.
- Aliakbari, M., & Nejad, A. M. (2013). On the effectiveness of team teaching in promoting learners' grammatical proficiency. *Canadian Journal of Education*, 36(3), 5–22. <https://www.jstor.org/stable/canajeducrevucan.36.3.5>
- Allen, J. L. (2008). The impact of speech-language pathologist service delivery models for concept imagery formation instruction on second grade students' language achievement outcomes [University of Nebraska at Omaha]. *ProQuest Dissertations and Theses*.
- Almon, S., & Feng, J. (2012). *Co-teaching vs. solo-teaching: Effect on fourth graders' math achievement*. <https://eric.ed.gov/?id=ED536927>
- Andersen, S. C., Beuchert-Pedersen, L. V., Nielsen, H. S., Thomsen, M. K., Beuchert, L., Nielsen, H. S., & Thomsen, M. K. (2018). The effect of teacher's aides in the classroom: Evidence from a randomized trial. *SSRN*, 18(1), 469–505. <https://doi.org/10.2139/ssrn.2626677>
- Andersen, S. C., Humlum, M. K., Nandrup, A. B., Knøth, H. M., & Brink, N. A. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences*, 113(27), 7481–7484. <https://doi.org/10.1073/pnas.1516686113>
- Andrews-Tobo, R. A. (2009). Coteaching in the urban middle school classrooms: Impact for students with disabilities in reading, math, and English/Language Arts classrooms [Capella University]. *ProQuest Dissertations and Theses*.
- Bacharach, N., Heck, T. W., & Dahlberg, K. (2010). Changing the face of student teaching through coteaching. *Action in Teacher Education*, 32(1), 3–14. <https://doi.org/10.1080/01626620.2010.10463538>

- Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, 48(4), 217–228. <https://doi.org/10.3102/0013189x19848729>
- Beam, A. P. (2005). The analysis of inclusion versus pullout at the elementary level as determined by selected variables [The George Washington University]. *ProQuest Dissertations and Theses*.
- Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Academic Press.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://www.jstor.org/stable/2346101>
- Blatchford, P., Bassett, P., Brown, P., Martin, C., Russell, A., & Webster, R. (2011). The impact of support staff on pupils' "positive approaches to learning" and their academic progress. *British Educational Research Journal*, 37(3), 443–464. <https://doi.org/10.1080/01411921003734645>
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–236). Russell Sage Foundation. <https://doi.org/10.7758/9781610441384>
- Campbell Collaboration. (2019). *Campbell systematic reviews: Policies and guidelines. 1.4*. <https://onlinelibrary.wiley.com/pb-assets/assets/18911803/CampbellPoliciesandGuidelinesv4-1559660867160.pdf>
- Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292. <https://doi.org/10.3102/0013189X16656615>
- Cook, B. G., McDuffie-Landrum, K. A., Oshita, L., & Cook, S. C. (2017). Co-teaching for students with disabilities: A critical and updated analysis of the empirical literature. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education* (2nd ed., pp. 233–248). Routledge. <https://doi.org/10.4324/9781315517698>
- Cook, L., & Friend, M. (1995). Co-teaching: Guidelines for creating effective practices. *Focus on Exceptional Children*, 28(3), 1–17. <https://doi.org/10.17161/foec.v28i3.6852>
- Dafolo. (2019). *Marilyn Friend om co-teaching*. <https://www.youtube.com/watch?v=4UUdXUJQ4PU>
- Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, 87(2), 243–282. <https://doi.org/10.3102/0034654316687036>
- Dietrichson, J., Filges, T., Klokke, 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), e1081. <https://doi.org/10.1002/cl2.1081>
- Dietrichson, J., Filges, T., Seerup, J. K., Klokke, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K–6: A systematic review. *Campbell Systematic Reviews*, 17(2), e1152. <https://doi.org/10.1002/cl2.1152>

- Dwyer, E. E. (2018). Co-teaching: The effects of co-teaching on reading and mathematics achievement for general education students in intermediate grade levels (grades 3-5) [University of St. Francis]. *ProQuest Dissertations and Theses*.
- Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. A. C. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment*, 7(1), 1–82. <https://doi.org/10.3310/hta7010>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Eldridge, S., Campbell, M. K., Campbell, M. J., Drahota, A. K., Giraudeau, B., Reeves, B. C., Siegfried, N., & Higgins, J. P. (2021). *Revised Cochrane risk of bias tool for randomized trials (RoB 2): Additional considerations for cluster-randomized trials (RoB 2 CRT)*. Cochrane Bias Methods Group. [https://drive.google.com/file/d/1yDQtDkrp68\\_8kJiUdbongK99sx7RFI/view](https://drive.google.com/file/d/1yDQtDkrp68_8kJiUdbongK99sx7RFI/view)
- Farrell, P., Alborz, A., Howes, A., & Pearson, D. (2010). The impact of teaching assistants on improving pupils' academic achievement in mainstream schools: A review of the literature. *Educational Review*, 62(4), 435–448. <https://doi.org/10.1080/00131911.2010.486476>
- Fernández-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods*, 53(1), 702–717. <https://doi.org/10.3758/s13428-020-01459-4>
- Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, N., Onghena, P., & Van den Noortgate, W. (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology*, 16(4), 299–315. <https://doi.org/10.5964/meth.4013>
- Filges, T., Sonne-Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, 14(1), 1–107. <https://doi.org/10.4073/csr.2018.10>
- Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A state-wide experiment. *American Educational Research Journal*, 27, 557–577. <https://doi.org/10.3102/00028312027003557>
- Fontana, K. C. (2005). The effects of co-teaching on the achievement of eighth grade students with learning disabilities. *Journal of At-Risk Issues*, 11(2), 17–23.
- Friend, M. (2008). Co-teaching: A simple solution that isn't simple after all. *Journal of Curriculum and Instruction*, 2(2), 9–19. <https://doi.org/10.3776/JOCI.%Y.V2I2P9-19>
- Friend, M. (2017). *Co-teaching i praksis: Samarbejde om inkluderende læringsfællesskaber* (1. udgave.). Dafolo.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. <https://doi.org/10.2307/1164588>
- Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32(4), 341–370. <https://doi.org/10.3102/1076998606298043>
- Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, 11(1), 1–21. <https://doi.org/10.1080/19345747.2017.1375583>

- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. <https://doi.org/10.3102/0162373707299706>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- Hedges, L. V., & Vevea, J. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145–174). Wiley Online Library. <https://doi.org/10.1002/0470870168.ch9>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M. S., Li, T., Page, M., & Welch, V. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley Online Library. <https://doi.org/10.1002/9781119536604>
- Hofner, B., Schmid, M., & Edler, L. (2016). Reproducible research in statistics: A review and guidelines for the Biometrical Journal. *Biometrical Journal*, 58(2), 416–427. <https://doi.org/10.1002/bimj.201500156>
- 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(1), 1–15. <https://doi.org/10.1080/13603116.2021.1900423>
- IDEA. (2022). *About IDEA*. <https://sites.ed.gov/idea/about-idea/#IDEA-History>
- Jang, S.-J. (2006a). The effects of incorporating web-assisted learning with team teaching in seventh-grade science classes. *International Journal of Science Education*, 28(6), 615–632. <https://doi.org/10.1080/09500690500339753>
- Jang, S.-J. (2006b). Research on the effects of team teaching upon two secondary school teachers. *Educational Research*, 48(2), 177–194. <https://doi.org/10.1080/00131880600732272>
- Joshi, M., & Pustejovsky, J. E. (2022). *wildmeta: Cluster wild bootstrapping for meta-analysis*. <https://github.com/meghaphsimatrix/wildmeta>
- Joshi, M., Pustejovsky, J. E., & Beretvas, S. N. (2022). Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with a small number of studies. *Research Synthesis Methods*, 13(4), 1–21. <https://doi.org/10.1002/jrsm.1554>
- Khoury, C. (2014). The effect of co-teaching on the academic achievement outcomes of students with disabilities: A meta-analytic synthesis [University of North Texas]. *ProQuest Information & Learning (US)*.
- Kidron, Y., & Lindsay, J. (2014). *The effects of increased learning time on student academic and nonacademic outcomes: Findings from a meta-analytic review*. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia. <https://ies.ed.gov/ncee/rel/Products/Publication/3603>
- Kirkham, J. J., Riley, R. D., & Williamson, P. R. (2012). A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine*, 31(20), 2179–2195. <https://doi.org/10.1002/sim.5356>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Krämer, S., Möller, J., & Zimmermann, F. (2021). Inclusive education of students with general learning difficulties: A meta-analysis. *Review of Educational Research*, 91(3), 432–478. <https://doi.org/10.3102/0034654321998072>

- LaFever, K. M. (2012). The effect of co-teaching on student achievement in ninth grade physical science classrooms [University of Missouri–St. Louis]. *ProQuest Dissertations and Theses*.
- Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., Turkeltaub, P. E., Kochunov, P., & Fox, P. T. (2005). ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1), 155–164. <https://doi.org/10.1002/hbm.20136>
- Lapsley, D. K., Daytner, K. M., Kelly, K., & Maxwell, S. E. (2002). *Teacher aides, class size and academic achievement: A preliminary evaluation of Indiana's Prime Time*.
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage.
- Maassen, E., van Assen, M., Nuijten, M., Olsson Collentine, A., & Wicherts, J. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PloS One*, 15(5), e0233107. <https://doi.org/10.1371/journal.pone.0233107>
- Mason, P. L. (2013). Comparing types of student placement and the effect on achievement for students with disabilities [Liberty University]. *ProQuest Information & Learning (US)*.
- Mathieu, L. (2019). An examination of special education instructional programs for English learners in New York City schools [Teachers College, Columbia University]. *ProQuest Information & Learning (US)*.
- Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5), 1091–1119. <https://doi.org/10.1111/rssc.12440>
- McGuinness, L. A. (2021). Risk of bias plots. In M. Harrer, P. Cuijpers, T. A. Furukawa, & D. D. Ebert (Eds.), *Doing meta-analysis in R: A hands-on guide*. PROTECT Lab. [https://bookdown.org/MathiasHarrer/Doing\\_Meta\\_Analysis\\_in\\_R/rob-plots.html](https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/rob-plots.html)
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11(2), 364–386. <https://doi.org/10.1177/1094428106291059>
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1), 105–125. <https://doi.org/10.1037/1082-989X.7.1.105>
- Muijs, D., & Reynolds, D. (2003). The effectiveness of the use of learning support assistants in improving the mathematics achievement of low achieving pupils in primary school. *Educational Research*, 45(3), 219–230. <https://doi.org/10.1080/0013188032000137229>
- Murawski, W. W. (2006). Student outcomes in co-taught secondary english classes: How can we improve? *Reading & Writing Quarterly*, 22(3), 227–247. <https://doi.org/10.1080/10573560500455703>



- Murawski, W. W., & Swanson, H. L. (2001). A meta-analysis of co-teaching research: Where are the data? *Remedial and Special Education*, 22(2), 258. <https://doi.org/10.1177/074193250102200501>
- No Child Left Behind (NCLB) Act of 2001. (2002). Pub. L. No. 107-110, § 101, Stat. 1425. <https://libguides.uww.edu/c.php?g=548489&p=4386220>
- OECD. (2016). *Skills matter: Further results from the survey of adult skills*. <https://doi.org/10.1787/9789264258051-en>
- Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., Vol. 2, pp. 177–203). Russell Sage Foundation. <https://doi.org/10.7758/9781610441384>
- Page, M. J., McKenzie, J. E., 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., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61(10), 991–996. <https://doi.org/10.1016/j.jclinepi.2007.11.010>
- Pigott, T. D. (2019). Missing data in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 367–382). Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>
- Pigott, T. D., & Polanin, J. R. (2019). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, 90(1), 24–46. <https://doi.org/10.3102/0034654319877153>
- Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, 42(8), 424–432. <https://doi.org/10.3102/0013189X13507104>
- Polanin, J. R. (2013). *Addressing the issue of meta-analysis multiplicity in education and psychology* [Loyola University Chicago]. [https://ecommons.luc.edu/luc\\_diss/539](https://ecommons.luc.edu/luc_diss/539)
- Powell, J. E. (2007). A comparison of learning outcomes for students with disabilities taught in three dissimilar classroom settings: Support services, team/collaborative and departmental/pullout [Auburn University]. *ProQuest Dissertations and Theses*.
- Pustejovsky, J. E. (2016). *Alternative formulas for the standardized mean difference*. <https://www.jepusto.com/alternative-formulas-for-the-smd/>
- Pustejovsky, J. E. (2020a). *An ANCOVA puzzler*. <https://www.jepusto.com/files/ancova-puzzle-solution.html>
- Pustejovsky, J. E. (2020b). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (0.5.5). [cran.r-project.org](https://cran.r-project.org/web/packages/clubSandwich/index.html). <https://cran.r-project.org/web/packages/clubSandwich/index.html>
- Pustejovsky, J. E. (2020c). *Weighting in multivariate meta-analysis*. <https://www.jepusto.com/weighting-in-multivariate-meta-analysis/>
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71. <https://doi.org/10.1002/jrsm.1332>



- Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(1), 425–438. <https://doi.org/10.1007/s11121-021-01246-3>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Rea, P. J., McLaughlin, V. L., & Walther-Thomas, C. (2002). Outcomes for students with learning disabilities in inclusive and pullout programs. *Exceptional Children*, 68(2), 203–222. <https://doi.org/10.1177/001440290206800204>
- Reinhiller, N. (1996). Coteaching: New variations on a not-so-new practice. *Teacher Education and Special Education*, 19(1), 34–48. <https://doi.org/10.1177/088840649601900104>
- Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, 26(2), 141. <https://doi.org/10.1037/met0000300>
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley Online Library.
- RStudio Team. (2015). *RStudio: Integrated development for R*. RStudio, Inc. <https://www.rstudio.com/>
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Schaef, R. J. (2014). Exploration of co-teaching in inclusive fourth-grade classrooms as a viable option for school districts [Indiana University of Pennsylvania]. *ProQuest Dissertations and Theses*.
- Schauer, J. M., Diaz, K., Pigott, T. D., & Lee, J. (2021). Exploratory analyses for missing data in meta-analyses and meta-regression: A tutorial. *Alcohol and Alcoholism*, 57(1), 35–46. <https://doi.org/10.1093/alcalc/agaal144>
- 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., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Cengage Learning, Inc.
- 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>
- Stanek, H. (2017). *Amerikansk ekspert: Co-teaching skal bruges varieret og med omtanke*. Folkeskolen.Dk. <https://www.folkeskolen.dk/604638/amerikansk-ekspert-co-teaching-skal-bruges-varieret-og-med-omtanke>
- Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36(10), 1580–1598. <https://doi.org/10.1002/sim.7228>
- Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., . . . Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919. <https://doi.org/10.1136/bmj.i4919>
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., & Eldridge, S. M. (2019). RoB 2: A revised

- tool for assessing risk of bias in randomised trials. *BMJ*, 366, 14898. <https://doi.org/10.1136/bmj.l4898>
- 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>
- Taylor, J. A., Pigott, T. D., & Williams, R. (2021). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher*, 51(1), 72–80. <https://doi.org/10.3102/0013189X211051319>
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3), 375–393. <https://doi.org/10.1037/met0000011>
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6), 604–634. <https://doi.org/10.3102/1076998615606099>
- Tukey, J. W. (1977). *Exploratory data analysis*. Pearson Modern Classic.
- UNESCO. (1994, June 7–10). *The Salamanca statement and framework for action on special needs education: adopted by the world conference on special needs education: Access and quality, Salamanca, Spain*. <https://unesdoc.unesco.org/ark:/48223/pf0000098427>
- Valentine, J. C., Aloe, A. M., & Wilson, S. J. (2019). Interpretation effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 433–452). Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press. <https://stefvanbuuren.name/fimd/>
- Van den Noortgate, W., López-López, J., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2), 576–594. <https://doi.org/10.3758/s13428-012-0261-6>
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2014). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, 47(4), 1274–1294. <https://doi.org/10.3758/s13428-014-0527-2>
- Vembye, M. H., Pustejovsky, J. E., & Pigott, T. D. (2023). Power approximations for overall average effects in meta-analysis with dependent effect sizes. *Journal of Educational and Behavioral Statistics*, 48(1), 70–102. <https://doi.org/10769986221127379>
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3), 261–293. <https://doi.org/10.3102/10769986030003261>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Welch, M., Brownell, K., & Sheridan, S. M. (1999). What's the score and game plan on teaming in schools? A review of the literature on team teaching and school-based problem-solving teams. *Remedial and Special Education*, 20(1), 36–49. <https://doi.org/10.1177/074193259902000107>
- What Works Clearinghouse (WWC). (2020). *WWC procedures and standards handbook* (4.1). Institute of Education Sciences. <https://ies.ed.gov/ncee/wwc/Handbooks>

- What Works Clearinghouse (WWC). (2021). *Supplement document for Appendix E and the What Works Clearinghouse procedures handbook, version 4.1*. Institute of Education Sciences. [https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-41-Supplement-508\\_09212020.pdf](https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-41-Supplement-508_09212020.pdf)
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. <https://cran.r-project.org/web/packages/ggplot2/index.html>
- Willett, J. B., Yamashita, J. J. M., & Anderson, R. D. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching*, 20(5), 405–417. <https://doi.org/10.1002/tea.3660200505>
- Wilson, D. B. (2016). *Formulas used by the “Practical meta-analysis effect size calculator.”* <https://mason.gmu.edu/~dwilsonb/downloads/esformulas.pdf>
- Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). *Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis*. MetaArXiv. <https://doi.org/10.31234/osf.io/bcky8>
- Wolf, B., & Harbatkin, E. (2022). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, 16(1), 1–28. <https://doi.org/10.1080/19345747.2022.2071364>
- World Bank. (2022). *World Bank country and lending groups*. <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>

## Authors

MIKKEL HELDING VEMBYE is a researcher in the Department of Quantitative Methods at VIVE – The Danish Center for Social Science Research, Søren Frichs Vej 36 G, 8230 Aabyhoej; e-mail: [mihv@vive.dk](mailto:mihv@vive.dk). His interests include questions of effectiveness in education and methods for conducting state-of-the-art systematic reviews and meta-analyses in education, especially methods for meta-analysis of dependent effect sizes.

FELIX WEISS is an associate professor in the Department of Educational Sociology at the Danish School of Education, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C.; e-mail: [fewe@edu.au.dk](mailto:fewe@edu.au.dk). His work focuses on social, ethnic and gender inequalities in the education system and on the transition from school to work.

BETHANY HAMILTON BHAT is a PhD student in the Department of Educational Psychology at the University of Texas at Austin, 1912 Speedway, Austin, Texas 78712-1289; e-mail: [bethanyhamilton@utexas.edu](mailto:bethanyhamilton@utexas.edu). Her research interests include statistical methods for meta-analysis, including methods for correcting effect sizes for clustering, methods for handling dependence structures, and methods for synthesizing using Bayesian inference.