

Methodological Innovation in Systematic Reviewing and Statistical Meta-Analysis in Education and Beyond

Mikkel Holding Vembye

Methodological Innovation in Systematic Reviewing and Statistical Meta-Analysis in Education and Beyond (Metodisk innovation af systematisk forskningskortlægning og statistisk meta-analyse inden for uddannelsesforskningen med videre [Danish Title])

Mikkel Holding Vembye

Ph.D. thesis handed in at the Danish School of Education (DPU), Aarhus University

Main supervisor: Felix Weiss

Co-supervisors: Christian Chrstrup Kjeldsen and Hans Siggaard Jensen

Founded by an Open Call Ph.D. scholarship from Aarhus University

Preface

The thesis is based on three articles, all of which address questions regarding systematic reviewing and meta-analysis. In all, the thesis is composed of four chapters, where to the three articles are enclosed in Chapters II to IV, making up the core of the thesis. The first chapter of the thesis aims to situate the three articles by introducing the overall Ph.D. project and by giving an overview of the main research questions that have driven the three articles and how these are related. Before the chapters, a brief summary of the three articles and their main findings and scientific contributions is provided both in English and Danish.

Acknowledgments

First and foremost, I wish to thank my main supervisor Felix Weiss. I simply could not have asked for any better. I am truly grateful for your generosity with your time and comments on the project, as well as the work you provided on our joint article. I truly appreciate that you have always been standing on my side and, as one of the few internals, valued my scientific work and endeavors. Furthermore, it has been a true pleasure to be in training as a researcher under your supervision since you have always trusted my ability to manage my own project and let me freely follow my scientific ideas.

Then, I owe the greatest thanks to James E. Pustejovsky and Terri D. Pigott, without whom this thesis would never have reached its sound quality. Without exaggeration, it has been the greatest experience of my career to get the opportunity to work together with my scientific idols. It has simply been a sincere pleasure to work with you, and I hope we can continue our collaboration in many years to come. It has been amazing to experience your kindness and learn that friendliness can easily be a part of academia. I will always bear this with me in future collaborations. James, thanks for always replying to all my emails no matter the subject and without ever complaining. I hope you know I will do whatever I can to somehow return your effort. Moreover, I have really enjoyed doing and learning R coding with and from you. Professionally, I must say that I have never learned so much from anybody as from you. Terri, thanks for taking me under your wings and letting me into the field of meta-analysis, and providing me the opportunity to be a part of the Society for Research Synthesis Methodology. I appreciate all of our discussions which have had ground-breaking impacts on my work and thinking. To paraphrase Ryan T. Williams, “while this chapter of my life ends, you will hopefully always be my mentor.” I am looking forward to working with you at Campbell Collaboration.

Next, I would like to thank Bethany H. Bhat for our collaboration and for never complaining about my untidy first-draft codes.

I would also like to thank my co-supervisors, Hans Siggaard Jensen, for shaping the basic idea of the project, and Christian Christrup Kjeldsen, for placing trust in me at an early career stage.

Among previous and current colleagues at Aarhus University, I wish to thank Morten T. Korsgaard for all our discussions about the current stage of Danish research in education, Jørn Bjerre for our discussions about causality and evidence-based research, Oliver Kauffmann for supporting me in the early stage of my project, and Savannah Schulz for creating the hex sticker for our R package. Also, a great thanks to Ida Gran Andersen for always caring about me and listening to my frustrations along the way. As a young scholar, it has been invaluable to feel the presence and encouragement of a more senior scholar.

Moreover, I wish to give a great thanks to Jens Dietrichson and Trine Filges for your ground-breaking feedback on the project. I am truly looking forward to coming closer to you. In this regard, also thanks to Hans Hummelgaard for the trust in me.

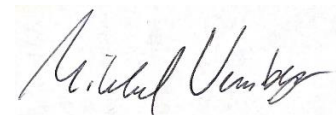
Generally, thanks to my family for always listening to me talking about meta-analysis, and a special thanks to Magnus Birkemose Nordam for all of our discussions regarding R, calculus, and matrix algebra, as well as for help with subtle coding problems. Finally, and absolutely most importantly, I owe the greatest of all thanks to Lotte, Dagmar, and Theodor. You mean everything to me. Lotte, my dear, thanks for your patience and forbearance. I could never have achieved anything close to the current result of the project without you. For this, I will forever be grateful to you.

Mikkel Holding Vembye, Nørresundby, April 2022

Updated Acknowledgments

I am very grateful to the members of the assessment committee Monica Melby-Lervåg, Wim Van Den Noortgate, and Rune Müller Kristensen (chair) for their careful reading and comments.

Mikkel Holding Vembye, Nørresundby, June 2022



Dedication

To Lotte, Dagmar, Theodor, James, Terri, and Anna Margrethe Vembye

Not every competent education researcher knows what is necessary to be a competent systematic reviewer

Larry Hedges

Table of Content

English Summary.....	8
Summary of results and contributions.....	9
Dansk resumé.....	11
Resumé af resultater og forskningsbidrag.....	12
 Chapter I.....	 15
<i>Overview Article</i>	
Abstract	16
1. Introduction	17
2. Overcoming Common Issues in Systematic Reviews and Meta-Analyses.....	23
3. Educational Theory	53
4. Philosophy of Science	57
5. Open Science – Preregistration, Open Material, and Open Data	60
6. Methodology	62
7. Summary and Discussion of Findings.....	71
8. References	78
 Chapter II	 96
<i>The Effects of Co-Teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis</i>	
Appendix 1: Studies Included in Meta-Analysis.....	155
Appendix 2: Supplementary Material (Chapter II)	164
Appendix 3: OSF Preregistered Protocol (Second Version).....	214
 Chapter III.....	 263
<i>Power Approximations for Meta-Analysis of Dependent Effect Sizes</i>	
Appendix 4: Supplementary Material (Chapter III).....	299
 Chapter IV.....	 309
<i>Conducting Power Analyses for Meta-Analysis of Dependent Effect Sizes: Common Guidelines and an Introduction to the POMADE R Package</i>	

English Summary

This thesis aims to conduct a state-of-the-art systematic review and contribute to the improvement of systematic reviewing and meta-analysis techniques in education and beyond. The thesis is composed of three articles that each makes methodological contributions to educational research as well as to systematic reviewing and statistical meta-analysis. The thesis has two overall aims. *First*, it seeks to remedy two frequently used and error-prone features of systematic reviews in education, i.e., the use of narrative synthesis of *quantitative research literature* and meta-analysis of studies contributing multiple effect sizes not sufficiently accounting for statistical dependencies among effect sizes coming from the same study. Among other things, the former issue has repeatedly been shown to produce conclusions driven by the preconceptions of the reviewers, while the latter prompts systematic reviews to yield too many false-positive results. To guard against these issues, the dissertation aims to provide a use case both for how to avoid narrative syntheses and tackle common reasons used to justify narrative synthesis and for how to adequately account for dependent effect sizes in meta-analysis. *Second*, it aims to expand the ballpark of statistical methods to handle dependent effect sizes by providing new power approximation formulas for the most common models used to handle dependency among effect sizes. These are the correlated hierarchical effects (CE), the multi-level meta-analysis (MLMA), and correlated-hierarchical effects (CHE) models. These new statistical power analyses can, for example, be utilized at the planning stage of systematic reviews in order to investigate if a review will be able to detect the smallest effect size of practical concern with a given certainty.

Throughout, the thesis complies with open science standards so that applied researchers, by re-using the available codes, can more easily implement accurate meta-analysis in future reviews and/or test our results. All background materials supporting the thesis can be found on Open Science Framework (OSF) and be accessed via <https://osf.io/fby7w/>, <https://osf.io/auj2e/>, and <https://bit.ly/3uuinTz>. Moreover, the *POMADE* (**P**ower for **M**eta-**A**nalysis of **D**ependent **E**ffects) R package is developed in this thesis to ease the use and accessibility of the newly developed and rather complex power approximation formulas. All material related to the package development can be found on GitHub at <https://github.com/MikkelVembye/POMADE>. In the following section, the main findings and contributions of each of the enclosed articles are described in more detail.

Summary of results and contributions

Article 1: The Effects of Co-Teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis (with Felix Weiss & Bethany H. Bhat, University of Texas at Austin)

This article [Chapter II] is a large-scale systematic review of the effects of collaborative models of instruction on students' academic achievement, which functions as a use case for how to apply cutting-edge meta-analysis techniques. Although the overall focus of the dissertation is on the statistical conduct and improvement of meta-analysis, the empirical work of this article is also substantially motivated by educational theories about collaborative models of instruction. In particular, the article wants to challenge common claims made in the co-teaching literature, saying that the literature is only mature for narrative synthesis and asserting that the evidence base supporting the effectiveness of co-teaching is scarce. Hereto, we found 128 treatment and control group designed studies in the period from 1984 to 2020 through databases- and snowballing searches. In fact, we located more eligible studies within all historical periods previously reviewed. From this pool of studies, 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, including 96 independent student samples and 280 short-term effect sizes, of which 82% were pretest-adjusted. We found a moderate statistical significant mean effect size equal to $\bar{g} = 0.11$, 95% CI[0.035, 0.184]. From moderator analyses, we found that collaborative instruction generally yields stable, moderate effects on academic achievement, and we show that the effect does not hinge on any specific two-teacher compositions, suggesting an increased potential for the scalability of these collaborative instruction interventions. All included models were based on the correlated-hierarchical effects (CHE-RVE) working models that combine multi-level meta-analytical modeling with robust variance estimation techniques while accounting for various dependency structures among effect sizes. We also found that important factors of effectiveness highlighted in the co-teaching literature did not explain any substantial variation in effect sizes. Finally, we did not find any clear evidence for publication bias or small study effects, which was not surprising since more than 80 percent of the included studies came from gray literature.

Article 2: *Power Approximations for Meta-Analysis of Dependent Effect Sizes (with James E. Pustejovsky, University of Wisconsin-Madison & Terri D. Pigott, Georgia State University)*

This article [Chapter III] introduces power approximations for tests of the overall average effect size from the most common models for handling dependent effect sizes mentioned above. These approximations aim to replace previous power approximation in meta-analysis, which was based on the assumption of independent effect sizes. In a Monte Carlo simulation, we show that the new power formulas can accurately approximate the true power of common meta-analytic models for dependent effect sizes when these approximations are based on reliable pilot data. We also show that the original method for approximating the power of the overall average effect size in meta-analysis performs inadequately in terms of predicting the power of models handling dependent effect sizes. Finally, the article investigates the Type I error rate and power for several common models. Findings show that tests using robust variance estimation provide better Type I error calibration than tests based on model-based variance estimation.

Article 3: *Conducting Power Analyses for Meta-Analysis of Dependent Effect Sizes: Common Guidelines and an Introduction to the POMADE R Package*

While the second article of the thesis concentrated on the statistical accuracy and quality assurance of the performance of the newly developed methods, it focuses less on the practical challenges encountered by researchers for obtaining the relevant quantities required to implement reliable power approximations for meta-analyses involving statistically dependent effect sizes. Therefore, this article [Chapter IV] aims to support applied reviewers by making these power approximation methods practically accessible. For this purpose, the article develops common guidelines for how I/we think power analysis for meta-analysis of dependent effect sizes can be conducted. Furthermore, it introduces the *POMADE* R package with the purpose of making these methods easily applicable in systematic reviews. Specifically, I/we provide R codes for how reviewers can investigate and illustrate *power*, *the number of studies required to detect a given effect size considered to be of practical concern*, and *the minimum detectable effect size* across various plausible data and model assumptions as well as with prespecified levels of statistical significance and power. Finally, we introduce the *traffic light power plot* for presenting power analyses across a range of plausible scenarios while clearly indicating the exact assumptions made by the reviewers.

Dansk resumé

Formålet med denne afhandling er at gennemføre en state-of-the-art systematisk forskningskortlægning samt at forbedre nuværende systematiske forskningskortlægnings- og statistiske meta-analytiske teknikker i pædagogikken og uddannelsesforskningen med videre. Afhandlingen består af tre artikler, som hver især skaber metodiske bidrag til pædagogikken og uddannelsesforskningen samt systematisk forskningskortlægning og statistisk meta-analyse. Overordnet set har denne afhandling to hovedformål. *For det første* søger afhandlingen at udbedre to ofte benyttede og fejlbehæftede metoder i systematisk forskningskortlægning inden for uddannelsesforskningen, dvs. brugen af narrativ syntese til at samle *kvantitativ forskningslitteratur* samt brugen af statistisk meta-analyse af studier, som bidrager med flere effektstørrelser, uden at der tages tilstrækkeligt højde for den afhængighed, der eksisterer mellem effektstørrelser, som kommer fra det samme studie. I forhold til den første problematik er det blandet andet gentagne gange blevet påvist, at dette kan føre til konklusioner, som hovedsageligt er drevet af forskeres forudindtagede overbevisninger, mens den anden problematik forårsager, at systematiske forskningskortlægninger afkaster for mange falske positive resultater. For at imødegå disse problematikker har afhandlingen til formål at skulle fungere som en use case delvis til at vise, hvordan man undgår narrative syntese og overkommer de normale argumenter, der bliver brugt som berettigelsesgrundlag for narrativ syntese og delvis til at illustrere hvordan afhængige effektstørrelser håndteres korrekt i meta-analyse. *For det andet* har afhandlingen til formål at bidrage med nye metoder til at håndtere afhængige effektstørrelser i meta-analyse. Til dette formål præsenterer afhandlingen nye styrke/power udregningsformler for de mest udbredte modeller til at håndtere afhængige effektstørrelser. Disse er correlated effects (CE), correlated-hierarchical effects (CHE) og multi-level meta-analyse (MLMA) modellerne. Disse nye statistiske styrkeanalyser kan eksempelvis benyttes i planlægningsfasen af systematiske forskningskortlægninger til at undersøge, hvorvidt kortlægningen vil være i stand til med en vis sikkerhed at kunne finde den mindste effektstørrelse vurderet til at være af praktisk relevans.

Afhandlingen følger gennem alle sine dele open science standarder, således at anvendte forskere ved at genbruge de tilgængelige koder med større lethed kan implementere adækvat meta-analyse i fremtidige forskningskortlægning og/eller teste vores resultater. Alle afhandlingens bagvedliggende materialer ligger på OSF (Open Science Framework) og kan tilgås via <https://osf.io/fby7w/>, <https://osf.io/auj2e/> og <https://bit.ly/3uuinTz>. Ydermere udvikles POMADE

(**P**ower for **M**eta-**A**nalysis of **D**ependent **E**ffects) R pakken i denne afhandling, som har til formål at øge adgangen til og brugen af de nyudviklede og relativt komplekse styrkeanalyseformler. Alt materiale relateret til pakkeudviklingen ligger på GitHub og kan tilgås via <https://github.com/MikkelVembye/POMADE>. I den følgende del, præsenteres mere detaljeret hovedresultaterne og -bidragene for hver af de indlagte artikler.

Resumé af resultater og forskningsbidrag

Artikel 1: Effekterne af co-teaching og relaterede tolærerordninger på elevers faglige præstationer: Et systematisk review og en meta-analyse (medforfattere Felix Weiss & Bethany H. Bhat, University of Texas at Austin)

Denne artikel [Kapitel II] repræsenterer et stor-skala review af effekter af tolærerordninger på elevers faglige evner. Derudover fungerer artiklen som en use case for, hvordan state-of-the-art meta-analyse teknikker benyttes. Selvom afhandlingens hovedfokus er på den statistiske udførelse og udvikling af meta-analyse, så er det empiriske arbejde i denne artikel ligeledes motiveret af substantielle pædagogiske tolærerordningsteorier. Artiklen ønsker specifikt at udfordre to udbredte antagelser om, at co-teaching litteraturen kun kan samles via narrativ syntese, og at vidensgrundlaget for de faglige effekter af tolærerordninger er spinkelt. Hertil fandt vi i alt 128 intervention og kontrolgruppe designede studier i perioden 1984-2020 baseret på database- og citationssøgninger. Det viste sig endvidere, at vi fandt flere relevante studier inden for alle tidsperioder, som har været benyttet i tidligere systematiske reviews angående tolærerordninger. Fra denne mængde af studier ekskluderede vi 52 studier, da disse ved hjælp af Cochrane risk of bias værktøjer blev vurderet til at indeholde en alvorlig risiko for bias. I alt gennemførte vi en meta-analyse af 76 studier, som bestod af 96 uafhængige grupper af elever og 280 effektstørrelser målt maksimalt tre måneder efter interventionens ophør. Hertil var 82% af disse effektstørrelser kontrolleret for elevers før-testscore. Vi fandt en moderat, statistisk signifikant gennemsnitlige effektstørrelse på $\bar{g} = 0.11$, 95% KI[0.035, 0.184]. Gennem moderatoranalyser fandt vi, at tolærer-undervisning har stabile moderate effekter på elevers faglige præstationer, og vi viser, at effekten ikke knytter sig til nogen specifik sammensætning af lærere/voksne, hvilket peger på et øget potentiale i forhold til at kunne udbrede disse undervisningsformer på en større skala. Alle anvendte modeller i artiklen baserer sig på correlated-hierarchical effects (CHE-RVE) arbejdsmodeller, som kombinerer multi-level meta-analyse modellering med robust varians estimeringsteknikker og samtidig tager højde for flere typer

af afhængighedsstrukturer mellem effektstørrelser. Vi fandt ingen tydelige tegn på publikations-bias eller small study effects, hvilket ikke er overraskende da mere end 80% af de inkluderede studier ikke har været formelt udgivet i et videnskabeligt peer-reviewed tidsskrift.

Artikel 2: Styrkeapproximationer for meta-analyse af afhængige effektstørrelser (medforfattere James E. Pustejovsky, University of Wisconsin-Madison & Terri D. Pigott, Georgia State University)

Denne artikel [Kapitel III] introducerer styrkeapproximationer for tests af den overordnede gennemsnitlige effektstørrelse for de mest normale modeller til at håndtere afhængige effektstørrelser som nævnt ovenfor. Disse approksimationer har til formål at erstatte tidligere styrkeapproximation for meta-analyse, som var baseret på antagelsen om uafhængighed mellem effektstørrelser. Gennem et Monte Carlo simuleringstudie viser vi, at de nye styrkeudregningsformler kan præcist approksimere den sande styrke af alle almene meta-analytiske modeller, der håndterer afhængige effektstørrelser, når disse approksimationer bygger på pålidelig pilot data. Vi viser ligeledes, at den oprindelige metode til at approksimere styrke for den overordnede gennemsnitlige effektstørrelse i meta-analyse præsterer statistisk utilstrækkeligt i forhold til at forudse styrken for modeller, der håndterer afhængige effektstørrelser. Til sidst undersøger artiklen Type I fejlraten og styrken for de mest normale modeller til at håndtere afhængige effektstørrelser. Resultaterne viser i den forbindelse, at tests som benytter robust varians estimering kontrollerer Type I fejlraten bedre sammenlignet med tests, som baserer sig på modelbaseret varians estimering.

Artikel 3: Udførelse af styrkeanalyser for meta-analyse af afhængige effektstørrelser: Almene guidelines og en introduktion til POMADE R pakken.

Mens den anden artikel i afhandlingen koncentrerer sig om den statistiske nøjagtighed og kvalitetssikring af de nyudviklede metoder, så fokuserer artiklen i mindre grad på de praktiske udfordringer, som forskere vil møde i forhold til at skulle skaffe de informationer og parametre, som er nødvendige for at kunne implementere pålidelige styrkeapproximationer i meta-analyse, der involverer statistisk afhængige effektstørrelser. Derfor har denne artikel [Kapitel IV] til formål at støtte anvendte reviewere ved at gøre styrkeapproximationerne mere tilgængelige i praksis. Til dette formål udvikles der i artiklen almene guidelines, for hvordan jeg/vi tænker, at styrkeanalyser

for meta-analyse af afhængige effektstørrelser kan blive udført. Ydermere introduceres *POMADE* R pakken, som har til formål at gøre disse metoder mere anvendelige i systematiske reviews. Specifikt fremsætter artiklen R koder, som understøtter reviewere i forhold til at undersøge og illustrere *styrke*, *antallet af studier krævet for at kunne finde en given effektstørrelse vurderet til at være af praktisk relevans* og *den mindst mulige detekterbare effektstørrelse* på tværs af plausible antagelser om ens data og model samt med præspecificerede niveauer for den statistiske signifikans og styrke. Til sidst introducerer vi *the traffic light power plot*, som har til formål at kunne præsentere styrkeanalyser på tværs af en række plausible scenarier, samtidig med at plottet tydeligt indikerer de eksakte antagelser fremført af reviewerne.

Chapter I

Overview Article

Mikkel H. Vembye

Abstract

This chapter gives an overview of the dissertation as required by the Graduate School of Arts, Aarhus University, to fulfill a Ph.D. degree from an article-based Ph.D. dissertation (cf. Aarhus University, 2010, p. 12). The chapter has several aims. First, it presents the overall project of the thesis and relates it to the scientific fields of systematic reviewing and meta-analysis. By doing so, it also demonstrates how the thesis aims to provide a means for overcoming a range of issues commonly encountered in systematic reviews and meta-analyses in education and the social sciences. Second, it explicates the overall research questions that have driven the Ph.D. project and shows the relationship between the research questions and the *three* enclosed research articles. Third, it presents the theory and methods that the Ph.D. project employed in order to make a scientific contribution. Forth, it shows how open science and open data policies have played a key role in ensuring the credibility of the work presented in this thesis. This should also provide codes to applied reviewers so that these can be re-used, or at least inspire, future reviews and meta-analyses in education and beyond. Fifth and finally, the article gives a brief summary and discussion of the achieved results and contributions of each of the enclosed research articles.

1. Introduction

Systematic reviews,^{1,2} i.e., “review[s] of existing research using explicit, accountable rigorous research methods” (Gough et al., 2017, p. 4), and *meta-analyses*,³ i.e., “the quantitative procedures that (...) statistically combine the results of studies” (Cooper & Hedges, 2019, p. 7), are critical tools for guiding developments of educational theory and educational decision-making for policy and practice (Campbell Collaboration, 2019; Gough et al., 2017; Karseth, Sivesind, & Gita, 2022; White, 2022; WWC, 2020). For that reason, systematic reviews and meta-analyses have substantially proliferated over the last three decades in the social sciences, including education (Ahn, Ames, & Myers, 2012; Cooper, Hedges, & Valentine, 2019; Pigott, 2012; Polanin, 2013; Williams, 2012). In education, in particular, systematic reviewing and meta-analysis have gained special attraction and trust since this type of research has been expected to link research, practice, and policy closer together (Hargreaves, 1996; Hattie, 2009; OECD, 2004). In fact, it has been the goal for many educational researchers that meta-analyses should become a standard practice in educational research (Campbell Collaboration, 2019; Hargreaves, 1996; KSU, 2021; WWC, 2022), moving toward the same systematic approach to the production of causal knowledge as in the field of medicine (Higgins et al., 2019). Given the increased dissemination of and the immense reliance on systematic reviews and meta-analyses for policy and practice decisions in education, it is all-important to understand and scrutinize the advantages, boundaries, and pitfalls of this type of research and, not least, ensure its validity (Polanin, 2013). Otherwise, researchers risk disseminating error-prone and potentially misleading guidelines for policy, practice, and research. Although numerous guidelines have been developed for how to conduct state-of-the-art reviews and meta-analyses in education and the social sciences (Campbell Collaboration, 2019; Cooper, 2015; Moher, Liberati, Tetzlaff, Altman, & Group, 2009; Pigott & Polanin, 2019; Siddaway, Wood, & Hedges, 2019; WWC, 2020), it is still widespread to find systematic reviews and meta-analyses

¹ Also known as *research syntheses* similarly defined as “integrating past research by drawing overall conclusions (generalizations) from many separate investigations that address identical or related hypotheses” (Cooper, 2015, p. 7).

² When I use the term review in this article, it refers to this definition of a systematic review.

³ The terms systematic review and meta-analysis are often used interchangeably. However, these conceptions are clearly distinguished in the present thesis since it is possible to conduct systematic reviews without using meta-analysis as the synthesis method and contrarily it is possible to conduct meta-analysis of a body of literature that has not been systematically gathered.

not following highest and best practice standards. This is also the case in prestigious journals⁴ for educational research, such as *Review of Educational Research* (Tipton et al., 2019b), *Educational Research Review* and *Review of Research in Education* (Ahn et al., 2012), as well as in evidence institutions such as Campbell Collaboration (Wang et al., 2021). Ultimately, this risks compromising the accuracy of the causal inferences and the credibility of this type of (high-stake) research, which in the end might promote error-prone decision-making in policy and practice.

On these grounds, one of the primary aims of the present thesis is to create knowledge that helps to overcome some of the most widespread and common issues encountered in educational and social science reviews. By conducting a systematic review including state-of-the-art meta-analysis techniques, the thesis aims to provide a use case that illustrates the strength of meta-analysis methods over *narrative synthesis*⁵ for amalgamating quantitative research literature and that practically demonstrates how to handle dependencies among effect sizes adequately when studies report multiple eligible results. One of the reasons why reviewers should avoid using narrative synthesis of quantitative literature is that it has repeatedly been shown to be subject to a number of fundamental deficits and biases. For example, that conclusions based on narrative synthesis are vulnerable to being driven by the researchers' preconception of the content area under review (Cooper, 2015). On the other hand, dependencies among effect sizes are frequently ignored in meta-analyses (Ahn et al., 2012; Tipton et al., 2019b), prompting reviews to yield too many false-positive results. I elaborate more thoroughly on these issues in the next section. Although dependence among effect sizes is often improperly treated in meta-analysis in education, authors of systematic reviews are not always the ones to blame simply because of lacking methods for handling dependent effect sizes in all parts of systematic reviews and meta-analyses. To exemplify, researchers have previously been compelled to use power approximation based on the assumption of independent effect sizes (Hedges & Pigott, 2001, 2004) if they wanted to understand the

⁴ Previously, it was also common to find systematic reviews and meta-analyses in *Psychological Bulletin* that treated dependent effect sizes inadequately (Tipton et al., 2019b). However, when looking through reviews concerning educational topics in issues published in 2021, it is *not* possible to find any reviews not using proper methods to handle dependencies among effect sizes, indicating a substantial improvement in this journal.

⁵ Defined as "a method to summarize [results] by using words" (Melendez-Torres et al., 2017, p. 109). Sometimes also defined as "narrative summaries" (Littell, 2008) or "thematic summaries" (Thomas et al., 2017). Narrative synthesis should not be confused with *narrative reviews*. A narrative review refers to a review based on a convenient sample of studies that is often based on the previous knowledge of the reviewer(s). However, narrative synthesis specifically refers to the method used to amalgamate results across studies independently of whether these are selected systematically or conveniently (Popay et al., 2006, p. 5).

approximate statistical power of their more complex model for handling dependent effect sizes at the planning stage of the review. Furthermore, until recently, researchers have only had limited software tools to support the conduct of power analysis for meta-analysis (Harrer et al., 2019). Thus, the thesis aims to remedy this apparent deficit of power analysis for meta-analysis by developing and quality assuring new power approximation formulas for the most common meta-analysis models that accurately handle dependence among effect sizes, i.e., the Multi-Level Meta-Analysis (MLMA; Van den Noortgate et al., 2013), the Correlated Effects (CE; Hedges et al., 2010b), and the Correlated-Hierarchical Effects (CHE; Pustejovsky & Tipton, 2021) models. Hereto, the thesis presents the first version of the *POMADE* (**P**ower for **M**eta-**A**nalysis of **D**ependent **E**ffects) R package that aims to ease the usability and accessibility of these rather complex power approximations.

Albeit the thesis might seem to make dispersed contributions to the fields of education, systematic reviewing, and meta-analysis alike, it has been driven by three overall and closely interrelated research questions. In the following sections, I will present the fine-grained connections between the three research articles of the thesis and the connections between their inherent research questions.

Main research questions

Throughout my dissertation work, a number of minor research questions have continuously sufficed, but three overarching and interrelated *methodological* research questions can be said to have been the main drivers of my work. These are:

- 1) *How to overcome common deficits and issues used to justify the use of narrative synthesis of quantitative research?*
- 2) *How to conduct state-of-the-art meta-analysis in education?*
- 3) *How to improve state-of-the-art methods to handle dependent effect sizes in education and beyond?*

This is not to say that *theoretical* educational research questions have not played a key role in the dissertation, but these are mainly subordinated to the second research question. Essentially, it is paramount to emphasize that it is not possible to conduct a high-quality systematic review, and

thereby answer the second research question, without a profound insight into the educational theories about the causal connections under review (Pigott, 2012). I elaborate more thoroughly on how educational theory has guided our review in the below Theory Section.

How the research questions evolved

The original idea leading to the first research question of the thesis evolved from my previous exploration and examination of the deficits related to the conduct of narrative synthesis of large bodies of literature based on quantitative analyses (Vembye & Jensen, 2022). We previously showed that narrative syntheses tend predominantly to produce simplistic and law-like causal statements about what works due to the narrow focus on statistical significance and lacking opportunities for rigorous testing of moderating effects across studies.

The second question appeared through the investigation of the first research question. From these investigations, it rapidly became apparent that the aim of overcoming the shortcomings of narrative synthesis was one of the main motivations for the development of meta-analysis techniques back in the 1970s as a response to the at the time common use of narrative reviews and syntheses (Cooper, 2015; Glass, 2000; Hedges & Olkin, 1985; Shadish & Lecy, 2015). Therefore, I started investigating how and if meta-analysis could provide a means to remedy common deficits and justifications of narrative syntheses.

Although it rapidly became clear that meta-analytical techniques could solve most of the shortcomings embedded in narrative reviews and syntheses, it was also clear from my exploration of meta-analytical methods that, as with every scientific method, they are not without deficits and pitfalls (Borenstein, 2019; Hedges et al., 2010b; Van den Noortgate et al., 2013). It was especially clear that it is still widespread to find meta-analyses that do not follow best practice methods for meta-analysis of social science research (Ahn et al., 2012; Littell, 2008). In particular, it is common to find lacking use of adequate methods for handling dependencies among effect sizes coming from the same study (Tipton et al., 2019b), which in turn compromises the accuracy of the derived inferences and results. On this basis, I got interested in understanding how to conduct state-of-the-art meta-analysis.

From my investigation of how to conduct a state-of-the-art meta-analysis in education, I found certain boundaries for methods to handle dependent effect sizes, leading to my interest in

understanding how new methods could be developed to adequately handle dependencies among effect sizes in meta-analysis. The boundaries of meta-analysis of dependent effect sizes appeared at first when I began to calculate *a priori* power for the complex models we intended to use in our planned review. At that point, these approximations were only available under the assumptions of independence among effect sizes. Hereto, it was obvious that these approximations were problematic to use for approximating the statistical power of the more complex models handling dependent effect sizes. Therefore, I began exploring how to develop new techniques for power approximation for meta-analysis of dependent effect sizes to improve the ballpark of methods to handle dependent effect sizes in meta-analysis.

Relations between dissertation chapters (articles) and research questions

The first article [Chapter II] of the thesis represents a systematic review and meta-analysis of the effects of co-teaching and related collaborative models of instruction on student achievement. Besides the aim of investigating the theoretical and substantial issues related to collaborative models of instruction, the underlining aim of the review is to function as a use case for answering the first and the second research question of the thesis. Particularly, the article aims to answer the first research question by replicating⁶ a previous systematic review conducted by the Danish Clearinghouse for Educational Research (henceforth DCER), in which study results of the effects of co-teaching and teacher assistants interventions were combined via narrative synthesis techniques (Dyssegaard & Larsen, 2013). The underlying intention of the article, therefore, is to show the difference between narrative synthesis and meta-analysis. In particular, it aims to show the advantage of doing meta-analysis relative to narrative synthesis of quantitative research in terms of demonstrating that meta-analysis can address and answer a larger number of substantial questions that are unreachable for narrative synthesis. Moreover, the article aims to show how to overcome common justifications for the use of narrative synthesis of quantitative literature (Campbell et al., 2019; Melendez-Torres et al., 2017; Petticrew & Roberts, 2008; Popay et al., 2006). To give a brief and typical example of why reviewers opt not to use meta-analysis of quantitative research (Ioannidis, Patsopoulos, & Rothstein, 2008), DCER argues (Dyssegaard & Larsen, 2013, p. 12)

⁶ Replication is defined as follows: “*Replicability* concerns whether another investigator can obtain the same results when they obtain their own (new) data by attempting to repeat the study that was carried out by the first investigator” (Hedges, 2019, p. 4).

that meta-analysis can only be conducted when studies are based on randomized controlled trials (RCTs). With the first article of the thesis, we implicitly want to show, among other things, that heterogeneity (in research designs) is not necessarily a valid reason for not conducting meta-analysis since this excludes the opportunity for investigating and creating an understanding of differential (design) effects. In this regard, we also demonstrate how biases of including non-randomized studies can be reduced through the use of thorough risk of bias assessments, pretest-adjusted effect size calculation, and statistical modeling techniques. I elaborate in more detail on these matters in Section Two of this overview article.

In order to answer the second research question of the thesis, the first article aims to provide a use case for how to deploy state-of-the-art methods to handle dependent effect sizes also in cases when various dependency structures are simultaneously present in the review data. This is the case when the data both contain studies reporting multiple outcomes from the same sample (also known as the *correlated effects dependency structure*) and studies reporting multiple outcomes from non-overlapping samples (also known as the *hierarchical effects dependency structure*). Until recently (Pustejovsky & Tipton, 2021), applied reviewers have been compelled to either model a hierarchical or a correlated dependence structure, but in the first article, we aim to demonstrate how to use the newly developed correlated-hierarchical effects models (CHE-RVE) that account for both types of dependence structures while combining MLMA (Van den Noortgate et al., 2013, 2014) and robust variance estimation (RVE; Hedges et al., 2010b; Tipton & Pustejovsky, 2015). At the current stage of the methodological meta-analysis literature, I/we considered this approach to represent a state-of-the-art of meta-analysis technique because the CHE model most adequately captures the true dependency structures encountered in reviews in education, including our review.

The second and the third article [Chapters III and IV] of the dissertation are both centered around answering the third research question by developing and quality testing new power approximation formulas for meta-analysis of dependent effect sizes. Through a Monte Carlo simulation, the second article also investigates which meta-analytical models are insufficient with regard to their Type I error calibration. In line with previous research on the topic (Moeyaert et al., 2017), the second article shows that averaging effect sizes from the same studies to avoid dependence among study-level effect sizes does not control the nominal Type I error rate when less than 60 studies are available for meta-analysis. Moreover, the article demonstrates that original suggested

power approximation for meta-analysis (Hedges & Pigott, 2001) is poorly suited for approximating the power of meta-analytical models based on study-mean effect sizes, albeit they are statistically independent.

A part of improving state-of-the-art meta-analytical methods is also to ensure that applied reviewers have sufficient guidelines and all necessary tools to apply these methods. Therefore, the third article of the thesis both revolves around developing common guidelines for how to conduct power analyses for meta-analysis of dependent effect sizes and around presenting the first version of the *POMADE* (**P**ower for **M**eta-**A**nalysis of **D**ependent **E**ffects) R package (R Core Team, 2022), which aims to ease the use of these rather complex power approximations. Furthermore, the third article of the thesis presents R functions for obtaining the *minimum detectable effect size* and the *number of studies required to detect a given effect size considered to be of smallest practical interest* under assumed conditions of the data and model as well as with prespecified levels of statistical power (β) and significance (α).

2. Overcoming Common Issues in Systematic Reviews and Meta-Analyses

In this section, I outline in more detail typical shortcomings encountered in educational and social science reviews. Hereto, I present how the thesis aims to contribute to overcoming these issues and, thereby, how the thesis aims to answer the three main research questions. Along with this presentation, I also reflect on the limitations of the suggested alternative solutions. It is obvious that no method is perfect (Borenstein, 2019), but I will carefully argue that some methods are indeed better suited to reach certain aims than others (Murnane & Willett, 2010).

Handling deficits of narrative synthesis of quantitative research

*Narrative synthesis*⁷ is seemingly still one of the most widespread methods used to combine results across studies included in systematic reviews (Thomas, O'Mara-Eves, Harden, & Newman, 2017; Valentine et al., 2017), and it is common to find narrative syntheses of large amounts of quantitative data (Campbell et al., 2019; Melendez-Torres et al., 2017), also in education (Dyssegaard & Larsen, 2013). In fact, multiple clearinghouses in education have been or are built entirely or partly

⁷ i.e., a “synthesis of findings from multiple studies that relies primarily on the use of words and text to summarise and explain the findings of the synthesis” (Popay et al., 2006, p. 5).

on this approach toward systematic reviewing (DCU, 2013; EPPI-Centre, 2010; KSU, 2021; Valentine et al., 2017). However, narrative synthesis techniques have ever since the development of statistical meta-analysis been shown to be suspect in producing error-prone results (Cooper, 2015; Hedges & Olkin, 1985; Lipsey, 2007; Littell, 2008). In the following section, I will discuss some of the most grievous problems of doing narrative synthesis of quantitative research. I am taking a critical stance towards narrative synthesis here because these critiques have played a vital role first in my understanding of common and practical challenges in systematic reviewing and second in my understanding of how meta-analytical techniques can overcome most of them. In this regard, it is all-important to emphasize that criticizing narrative synthesis is neither to say that narrative synthesis in all circumstances malfunctions nor that meta-analysis functions without narratives. My incoming critique is solely directed toward narrative synthesis of *quantitative* research studies (Campbell et al., 2019; Melendez-Torres et al., 2017; Petticrew & Roberts, 2008; Popay et al., 2006). Moreover, the critique mainly concerns systematic reviews that examine the effectiveness of given interventions and follow rigorous methods in all parts surrounding the narrative literature syntheses. Although the critique is certainly relevant for other types of reviews, such as narrative reviews, these are not of my concern because they contain other shortcomings that fall outside the scope of this thesis. To further support my argumentation, I would also like to stress that narrative synthesis can, obviously, provide a pivotal means for combining findings of qualitative studies to create qualitative understandings and theorizations of why given interventions work or fail to work properly. In other words, narrative syntheses can be important for theoretical developments in the given content area of interest (for a great example, see Scruggs et al., 2007). Needless to say, it still, however, requires rigorous procedures to avoid running into the below-presented shortcomings of narrative synthesis (Campbell et al., 2018). On the other hand, it is also important to emphasize, as highlighted by Thomas et al. (2017, p. 185), that it is to some extent artificial to make bold distinctions between narrative synthesis and meta-analysis since “all synthesis involve narrative[s] of some kind.” As will be shown in the below Theory Section, all high-quality meta-analyses hinge on deep theoretical narratives of why certain effects are expected to or not to appear and/or why they vary or not across scientifically relevant factors. That said, narrative syntheses of quantitative studies are subject to numerous pitfalls and deficits that meta-analysis is intended to remedy. I will go through three major deficits of narrative synthesis of quantitative literature and show how the thesis aims to come around these issues via statistical

meta-analysis. Along with these expositions, common critiques against meta-analysis are discussed as well.

Cognitive algebra and how to avoid it with meta-analysis

One of the most widespread and persistent critiques against narrative synthesis is that it depends on *cognitive algebra*, i.e., that the weighting schemes and rules used to combine findings “are rarely known as to anyone but the synthesists themselves” (Borenstein et al., 2009; Campbell et al., 2019; Cooper, 2015, p. 9). In effect, it has been shown that cognitive algebra in narrative synthesis frequently leads to *confirmation bias*, where researchers “unintentionally (...) seek information that supports a hypothesis, give preferential treatment to evidence that confirms existing beliefs, and dismiss evidence to the contrary” (Littell, 2008). This further means that narrative reviewers will often have the tendency to predominantly place more weight on studies conducted by themselves or their colleagues (Cooper, 2015; Glass, 1976, 2000). Consequently, it will be unlikely that other scholars will be able to reproduce the same conclusion even under the same circumstance (Valentine et al., 2010). Even the best educational researchers will likely fall short of keeping a valid overview of the overall synthesis conclusions not driven by arbitrarily weighting rules as soon as the number of studies and reported results increases. It can, therefore, be expected that the deficit of narrative synthesis increases as a function of the number of included studies and the number of reported results from each study. Thus, the critique is also to say that without statistical methods, “[t]he accumulated findings of (...) studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than hundreds of data points in a single study” (Glass in Cooper, 2015, p. 13).

As a justification for narrative syntheses of quantitative results, it is often claimed that they are valid when they include only a few studies (data points) (see Table 1 in Ioannidis, Patsopoulos, & Rothstein, 2008, p. 1413; Valentine et al., 2017). However, even with a small number of studies, it is easy to imagine in social science reviews that it might be an insurmountable task to reach accurate conclusions from narrative synthesis, particularly when studies report multiple eligible outcomes. To give an example from the review of the thesis, four studies yielded a total of 70 effect sizes. Even with only four studies, it would require an extraordinary cognitive capacity to make reliable conclusions about the connection among so many effect sizes, not to mention how the effects vary and how to adequately account for the dependencies among the effect sizes coming

from the same studies. In such cases, I will argue in line with Valentine et al. (2010, p. 241) that meta-analysis is more appropriate, “[n]ot because it is ideal but rather because given the needs for a conclusion, it is a better analysis strategy than the alternatives.” Alternatively, as we suggest in the third article of the thesis, researchers should consider if a synthesis is at all necessary when only having a few studies and almost no power to find the effects of practical concerns. Hereto, it might be better to ask for more evidence and use alternative techniques not involving any type of synthesis (see suggestions in McKenzie & Brennan, 2019 & Valentine et al., 2010).

Although weighting schemes might be used obscurely in narrative synthesis of quantitative studies, it is also common to find systematic reviews in education not using any weighting scheme across the *included* studies. Thus, all study results are taken at face value (Lipsey, 2007), independently of the study sample size and other relevant features, such as the fidelity of the intervention, etc. (see, e.g., systematic reviews conducted by DCU, 2013; DPU, 2022).⁸ This issue becomes particularly problematic when narrative reviewers bear their evidence on statistical significance, as is common in narrative synthesis (Vembye & Jensen, 2022). Say, for instance, that one study with a sample size of 500 students yields a statistically significant result, and one study with a sample size of 200 students yields a non-statistically significant result, everything else being equal. Taken at face value, this would be considered as evidence supporting the conclusion of a “failure to replicate” (Hedges & Olkin, 1985). Thereby narrative reviewers would likely conclude that the evidence for the effectiveness of the given intervention is ambiguous when they, in fact, just see an underpowered study to detect the true effect. As pointed out by Borenstein et al. (2010, p. 13), it gets “[e]ven worse, when the significant study was performed in one type of sample and the nonsignificant study was performed in another type of sample, researchers would sometimes interpret this difference as meaning that the effect existed in one population but not the other.” As these examples indicate, using statistical significance as the main evidence for assessing the effectiveness of educational interventions sets out unrealistic requirements for fair effectiveness judgments in education since “[e]ven under conditions of high power (.80), the probability of both studies rejecting a false null hypothesis is only .64. Both studies would have to be conducted under conditions of extraordinary statistical power (.975) for there to be a 95% chance that they would both correctly result in a rejection of a false null hypothesis” (Valentine et al., 2010, p. 240). The

⁸ Note that many previous reviews of collaborative models of instruction have been subject to these critiques as well (see Supplementary Table S1 in Chapter II).

same issues are embedded in the more systematic use of *vote counting* (Hedges & Olkin, 1980, 1985). Generally, methods based on amalgamating statistically significant results have been shown to produce overly pessimistic and conservative results since educational intervention studies are frequently underpowered. Most often, a majority of studies will yield insignificant results, which will lead many researchers to conclude that the evidence decisively shows that there is no effect of the given intervention. This is also why Borenstein et al. (2009, p. 14) stress that “*doing arithmetic with words (...) when the words are based on p-values (...) are the wrong words*. Hereto, it is also important to remember that *absence of evidence is not evidence of absence* (Altman & Bland, 1995; Senn, 2009), meaning that the absence of statistical significance does not necessarily imply that a study found no effect of practical importance.

Weighting scheme used in the thesis

To overcome the above-mentioned issues related to cognitive algebra and wrong narratives about patterns of p values when taken at face value, the first article of the thesis draws on meta-analytical techniques to synthesize results across and within studies. Among other things, an advantage of using meta-analysis is that it ensures transparent, albeit complex, weighting schemes that can be reproduced. To give an example of the weights used and attached to each study in our review, I will present the CHE model weights we used to estimate the overall average effect size. For an overview of the weighting schemes used for some of the more complex meta-regression models used in the thesis, see Pustejovsky (2020) and the Supplementary Material of Pustejovsky & Tipton (2021).

To concisely explain this procedure, assume that we have a collection of J studies, each reporting $k_j \geq 1$ effect size estimates. Then let T_{ij} be effect size estimate i from study j with a corresponding sample error σ_{ij} , for $i = 1, \dots, k_j$ and $j = 1, \dots, J$. In the model we used in the thesis, we assumed, as is common in meta-analysis, that T_{ij} is an unbiased estimate of the effect size parameter θ_{ij} and that σ_{ij} is fixed and known. These assumptions can be expressed as

$$T_{ij} = \theta_{ij} + e_{ij} \tag{1}$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = s_{ij}^2$. We assumed across all models that effect sizes coming from different studies were uncorrelated, so $\text{cor}(e_{hj}, e_{il}) = 0$ when $j \neq l$.

Since we had more than 55 studies reporting multiple results with various dependency structures⁹ and wanted to synthesize studies across various interventions and student populations, we used a model encompassing random effects, both capturing between-study and within-study heterogeneity to estimate the overall average effect size. Hierarchically, the model can be written as (Pustejovsky, 2020)

$$\begin{aligned} T_{ij} &= \theta_j + v_{ij} + e_{ij} \\ \theta_j &= \mu + u_j \end{aligned} \tag{2}$$

where θ_j represents the average effect size parameter of the j^{th} study, i.e., $\hat{\theta}_j = \frac{1}{k_j} \sum_{i=1}^{k_j} T_{ij}$. μ is the overall average effect size, u_j is the between-study error with $\text{Var}(u_j) = \tau^2$ and v_{ij} is the within-study error with $\text{Var}(v_{ij}) = \omega^2$. $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = \sigma_{ij}^2$. In this model, we made the simplifying assumption that sample variance estimates from the same study were equal so that $\sigma_j^2 = \frac{1}{k_j} \sum_{i=1}^{k_j} \sigma_{ij}^2$. We also assumed that there was a constant sample correlation, ρ , between effect size i and m for $i, m = 1, \dots, k_j$ and $j = 1, \dots, J$. Thus, $\text{Cov}(e_{ij}, e_{mj}) = \rho \sqrt{\sigma_{ij}^2 \sigma_{mj}^2}$. From this model, the overall average effect size can be estimated as

$$\hat{\mu} = \frac{1}{W} \sum_{j=1}^J w_j \hat{\theta}_j, \text{ where } W = \sum_{j=1}^J w_j$$

when the CHE model is correctly specified, then

$$\text{Var}(\hat{\mu}) \approx \frac{1}{W}$$

⁹ I will return to this issue in later sections.

The weight, w , attached to study j for this model is given by

$$w_j = \frac{1}{\tau^2 + \frac{1}{k_j}(\omega^2 + (k_j - 1)\rho\sigma_j^2 + \sigma_j^2)} \quad (3)$$

and the weight, w , attached to effect size i in study j is given by

$$w_{ij} = \frac{1}{k_j\hat{\tau}^2 + \hat{\omega}^2 + (k_j - 1)\rho\sigma_j^2 + \sigma_j^2} \quad (4)$$

Notably, we used Equation (4) to calculate the weight given to each effect size used to estimate μ in our review. The estimated weights can be found in Figure 5 in Chapter II. The above weights have certain characteristics that help to understand how studies and effect sizes are weighted (Pustejovsky, 2020). *First*, when the between- and within-study variance is zero, the weight given to a study depends on the number of reported effect sizes k_j , the assumed (often constant) sample correlation ρ , and the average sample variance of the study σ_j^2 . Under this scenario, if ρ is close to one when the weight given to the study will be close to the same as the weight given to one of the single effect size estimates from the study. In contrast, if ρ is near zero, a study will gain more weight as a function of k_j . Furthermore, under the assumption that the assumed ρ is reasonable close to the true sample correlation, these weights will properly take into account the number of effect sizes within each study and their precision. Notably, this is a feature that I consider to be impossible to properly account for in narrative synthesis of quantitative studies.

Second, when there is no between-study variation, i.e., $\tau^2 = 0$, but a substantial amount of within-study heterogeneity, ω^2 , studies reporting more effect sizes will prevalingly get more weight relative to studies reporting few results.

Third, as with most weighting schemes used in meta-analysis (Borenstein et al., 2010), larger studies will get more weight than smaller and noisier (i.e., imprecise) studies. However, as with simpler random-effects models (Borenstein et al., 2010), when the between-study variance τ^2 gets larger, μ will come closer to the simple average of the study mean effect sizes, θ_j s. Put differently, the relative weight difference between large and small studies reduces as τ^2 increases.

A statistical advantage of using CHE weights relative to the original weighting scheme for meta-analytical RVE models (Hedges et al., 2010b; Hedges, Tipton, & Johnson, 2010a) is that they represent inverse-variance weights under the working model, which consequently produces fully efficient estimation of μ when the working model¹⁰ is correctly specified, meaning that the weights yield the most precise variance estimate of μ (Pustejovsky & Tipton, 2021).

Potential critiques against weights used in meta-analysis

Despite the fact that the weights used in meta-analysis are *mathematically transparent*, it still fair to say that many meta-analyses lack *empirical transparency* since it has been shown that it is often impossible to reproduce the effect sizes and the corresponding variance components σ_j^2 s used for the weight construction and estimation in the meta-analysis (Maassen et al., 2020). To remedy this issue, all parts of our meta-analysis, including effect size and sample variance calculations, have been made public at <https://osf.io/fby7w/>. It should even be possible for readers of the review to determine the page numbers from which we obtained the results used for the effect size calculations from each study. In addition, to ensure that we did not use unreasonably biased estimates of σ_j^2 in Equation (3), we conducted *approximate cluster design adjustment* techniques (Hedges, 2007; Higgins et al., 2019) so that lower-quality studies (assuming that lower-quality studies rarely account for clustering of students) did not get disproportionately more weight relative to more rigorously conducted studies (see Supplementary Section S1 in Chapter II). In other words, without conducting cluster design corrections, the sample variance from studies not accounting for multi-level data structures would be too small, meaning that inappropriately more weight would be ascribed to these studies. However, as a sensitivity analysis, we tested the impact of not applying cluster bias corrections on the effect size estimation, showing that it only had a minor, non-substantial effect on our main conclusion (see Supplementary Figure S13 in Chapter II).

As with narrative synthesis, it could be argued that the CHE model weights are based on quantities that are strongly dependent on the idiosyncratic assumptions of the researchers and thus readily subject to manipulation. For example, researchers might arbitrarily choose and change the value of ρ , which has a strong impact on the relative magnitude of τ^2 and ω^2 (Pustejovsky & Tipton, 2021) and thereby the weights. To accommodate this concern, we conducted a range of

¹⁰ I elaborate on what is meant by a working model in a later section.

sensitivity analyses in which we changed our assumption about ρ to investigate how these assumptions impacted our main conclusion. Although, τ^2 and ω^2 often were substantially impacted by changed assumptions, the estimation of μ was more or less insensitive to these changes (see Supplementary Figure S10 in Chapter II).

Additionally, the weights might be criticized for placing more weight on large studies than on smaller studies, under the assumption that these are more rigorously conducted compared to smaller studies and thus more reliable. Some would argue that this assumption is invalid since large studies require more resources which might include involving less proficient personnel to carry out the intervention, which in the end might lead to measurement errors. Consequently, it is assumed that giving more weight to larger studies can induce wrong null findings, also defined as *a null bias* (find these arguments in Ioannidis, 2005). Therefore, to understand the impact of large studies, i.e., studies with a sample size of more than 1000, we conducted a sensitivity analysis where we re-estimated μ by excluding large studies. However, that did not change our results in any substantial way (see Supplementary Figure S13 in Chapter II).

To sum up, although meta-analytical weights might be subject to errors, I try to show in this thesis that the main advantage of meta-analysis compared to narrative synthesis is that it provides a means for reviewers to investigate and evaluate the impact of their proposed weighting schemes and model assumptions.

The deficit of not using effect sizes and why we need them

Effect sizes¹¹—in this thesis, narrowly understood as standardized mean differences¹² (Hedges, 1981)—are the main currency of meta-analysis that provide a vital means to “facilitate[] the comparability of results across studies, across programs and policies, and across outcome measures” (Baird & Pane, 2019, p. 217). Without effect size estimates, it is impossible for narrative reviewers

¹¹ Effect sizes “are quantitative indexes of relations among variables” and “refer to any index of relation between variables” (Hedges, 2008, p. 167).

¹² In the most simple case, standardized mean difference refers to a mean difference often scaled/divided by a pooled standard deviation, i.e., $d = (\bar{X}_1 - \bar{X}_2)/S$, where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the treatment and control group, respectively, and S is the pooled standard deviation given by $S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ with S_1 and S_2 being the standard deviations and n_1 and n_2 being the sample sizes of the two groups, respectively (Borenstein et al., 2009, p. 26). This metric is often defined as Cohen’s d .

to make fair comparisons about the magnitude of the effect of an intervention across studies using different scales of measurement, e.g., for student achievement tests. As Hedges (2008, p. 167) points out, it is hard to tell which effect is larger “[i]f Study A finds that the treatment effect of an intervention is 2.3 scale score points on the Woodcock–Johnson reading comprehension test, but Study B finds that the treatment effect is 7.5 points on the Terra Nova reading comprehension scale.” Again, as with cognitive algebra, interpreting magnitudes of effects across studies without using effect sizes only gets more devastating as the number of studies and effect sizes increases. It would require a profound knowledge of all measurement scales included in a review to make reliable judgments about the relative magnitude of the effect without the use of effect sizes. Most educational researchers will not have this knowledge, including myself. To exemplify, our review would have required content knowledge on more than 40 different achievement test scales to make a reliable judgment about the magnitude of the effect of collaborative models of instruction on student achievement. Effect sizes overcome this issue and allow the interpretation of results across studies using different scales with incommensurable units.

Moreover, the lacking use of effect sizes is perhaps also the main reason why narrative synthesis of quantitative research tends to place interpretations of statistical significance as evidence for the size of the effect. For example, it is common to find narrative reviewers using *p* values as evidence for a large effect of an intervention (Borenstein et al., 2009) or use wordings like “highly significant” to support the interpretation of a large effect. However, *p* values are strongly dependent on the sample size of the study (Hedges, 2008). This also means that it is possible to prove statistical significance for even the smallest effect with no practical relevance only by increasing the sample size of the studies. Therefore, it is important to remember “the fact that [if] one study reported a *p*-value of 0.001 and another reported a *p*-value of 0.50 [it] does not mean that the effect was larger in the former. The *p*-value of 0.001 *could* reflect a large effect size, but it *could* also reflect a moderate or small effect in a large study (...). The *p*-value of 0.50 could reflect a small (or nil) effect size but could also reflect a large effect in a small study (...). This point is often missed in narrative reviews “(Borenstein et al., 2009, p. 12). Furthermore, when systematic reviewers only concentrate on statistical significance, they only lend themselves to answer binary research questions about whether the intervention effect is statistically different from zero or not, but they have no means to talk about the size of the effect. Concretely, this means that without effect size estimates, narrative synthesis simply has no reliable means to answer “*how big*

is the effect?”. Moreover, it is almost impossible for narrative synthesis to compare the effectiveness of the given intervention(s) and student population(s) to other related interventions and student populations.

Unlike statistical significance testing, one of the benefits of using effect sizes is that they “depend[] only on the underlying population parameters [for standardized mean differences, i.e., $\delta = (\mu_1 - \mu_2)/\sigma$],¹³ not on sample size, which is particular to the study” (Hedges, 2008, p. 168). That said, *p* values can still provide valuable information about how reliable the mean difference might be, but they certainly do not provide an index for the magnitude of an effect (Hedges, 2008).

Besides the above-presented benefits, there are at least five additional advantages of using effect size estimates as the alternative to *narrative* interpretations of study results. *First*, the use of effect sizes makes it possible to correct error-prone reported results, e.g., when studies do not account for nesting of students in classes and school (Hedges, 2007, 2011). It is also possible to correct for explicit measurement errors when the reliability of the used scale is known (Hedges, 1981; Schmidt & Hunter, 2015). The former method was highly relevant for our review since we have 67 out of the 76 studies that did not adequately account for the nesting of students in classes or schools. As previously mentioned, we conducted cluster design adjustments for these studies to assure that the sampling variance σ_{ij}^2 of each effect size estimate was reasonably accurate. This was important to do in order to; 1) most reliably estimate between-study variance, 2) determine the weights used to estimate the overall average effect, μ , 3) assess the uncertainty of the estimation of μ , and 4) assess the extent of uncertainty in the between-study variance estimate. As we will show in a later section, RVE ensures reliable estimates of the variance of μ even when the sample variance components are completely wrong. However, the three other scenarios (1, 2, and 4) hinge on the accuracy of the sample variance estimation, which underpins the importance of conducting corrections for cluster bias. Hereto, it is important to emphasize that the cluster bias adjustments that we used do not yield the exact sample variance of each effect size from studies not accounting for clustering. However, we used these techniques based on the assumption that “making no correction for the effects of clustering at all corresponds to assuming that [the intraclass correlation]=0: The assumption that [the intraclass correlation]=0 is often very far from the case, and thus

¹³ μ_1 and μ_2 represent the true population mean treatment effect of the treatment and control group, respectively. σ is the true common standard deviation of the two populations.

it may introduce more serious biases in the computation of variances than using values of [the intraclass correlation] that are slightly in error” (Hedges, 2007, p. 359). We did not correct for measurement errors, but whenever reported in primary studies, we coded the reliability of the test scale used (Schmidt & Hunter, 2015). This information is retrievable from our coding/data extraction scheme (find at <https://osf.io/fby7w/>) so that other researchers can deploy this approach if they find the Hunter & Schmidt meta-analysis technique more relevant.

Second, it is often argued, as previously shown, that meta-analysis is only viable if all studies represent randomized controlled trials (Dyssegaard & Larsen, 2013; Ioannidis et al., 2008). However, by using either pretest- and/or covariate-adjusted effect size calculation techniques, the bias of including non-randomized studies can, in effect, be reduced since it “allows researchers to control for preexisting differences, allowing estimates of treatment effectiveness even when treatment and control groups are nonequivalent” (Morris, 2008, p. 365). Therefore, to avoid inducing more bias than we prevent by including research designs of varying quality (Egger et al., 2003), we required in our meta-analysis that non-randomized studies had ensured baseline equivalence (as suggested in Campbell Collaboration, 2019). If not ensured, the study should either provide baseline/pretest achievement or covariate-adjusted measures from which we could compute pretest- and/or covariate-adjusted effect sizes (Morris, 2008; Morris & DeShon, 2002; Pustejovsky, 2016; Taylor, Pigott, & Williams, 2021; WWC, 2021). By virtue of our inclusion rules, we excluded all non-randomized studies that did not ensure baseline equivalence between the intervention and the control groups or did not provide results from which we could calculate pretest- and/or covariate-adjusted effect sizes. This was done via the use of Cochran’s risk of bias assessment tools (Higgins et al., 2019). Independently of the research design, we always prioritized pretest- and covariate-adjusted effect sizes above post-test score effect sizes since a further advantage of this family of effect sizes is that they increase the estimated precision of the effect size estimation (Morris, 2008; Pustejovsky, 2020). For further details about our effect size calculation, see Supplementary Section S1 and the risk of bias assessment described in Chapter II.

Third, effect size estimates can be used to check for the consistency between reported study results. For example, if a study both reports difference-in-differences and ANCOVA (Analysis of Covariance) results from the same sample of students. In our meta-analysis, we calculated all

possible effect sizes whenever relevant and used these tests to inform our risk of bias assessment.

Forth, a clear benefit of using effect sizes relative to narrative interpretations and syntheses of study results is that it is possible to reliably quantify the uncertainty/precision attached to the given effect size/result. This properly acknowledges the fact that all research results always come with some uncertainty due to the fact that these always are derived from finite (student) samples.

Fifth and finally, effect sizes allow for the use of accurate statistical methods to conduct a range of tests investigating essential and potential sources and reasons for heterogeneity (Tipton et al., 2019a) and publication bias (Rothstein et al., 2005) that is not possible in narrative syntheses. I will elaborate more on this matter in the below subsection regarding the advantage of heterogeneity in meta-analysis. Before that, I will briefly turn to some of the common complications of using effect sizes.

Complications of using effect sizes

Although effect sizes have clear benefits, certain complications arise when they are used. Like narrative syntheses, effect size estimation has been strongly criticized for being easy to manipulate so that they can be fitted to the preconception of the researchers (Simpson, 2017; Stegenga, 2011). Along the same line, particularly, effect sizes based on standardized mean differences are criticized for their strong dependence on the standard deviation (SD). For example, when calculating effect sizes from more complex multi-level designs, such as two-level models with students nested in schools, effect sizes can be calculated in three different ways, either by using the within-school SD, the between-school SD, or the total SD that both accounts for the between and within-school variation. The choice of SD, in this case, will lead to very different effect size estimates (Hedges, 2007, 2008). Thereby, researchers can potentially select the effect size that are most in line with the hypothesis they want to prove. To accommodate this type of cherry-picking, we ensured in our systematic review to standardize the mean difference by the total SD across all studies, as recommended by Taylor et al. (2021). To exemplify, we found a couple of studies that only reported results at the class level (LaFever, 2012; Southwick, 1998). For these studies, we first standardized the classroom mean difference by the between-classroom SD by either using Equations (11) and (12) or (21) and (22) from Hedges (2007), depending on whether the exact class sizes were reported for all included classes. We then converted these measures into effect sizes scaled by the

total variance by using Equations (25) and (26) from Hedges (2007). For these conversions, we imputed intraclass correlations from Hedges & Hedberg (2007), as suggested by Hedges (2007). See Supplementary Section S1 in Chapter II for more details or the raw effect size calculation codes at <https://osf.io/fby7w/>.

Next, Cohen's d has been shown to be subject to upward biases, i.e., that it yields too large effect size and sampling variance estimates when computed from studies with small sample sizes. Therefore, we used Hedges (1981) proposed small-sample correction. Hence, the effect size metric used in our meta-analysis was the Hedges's g estimator.

Finally, one of the greatest complications of using effect size estimates based on standardized achievement tests is that it is complicated to interpret their practical importance because they are described in standard deviation units, i.e., units without a natural meaning (Baird & Pane, 2019; Kraft, 2020; Lipsey et al., 2012; Valentine et al., 2019). In this regard, the complication is further that it is difficult to re-scale standardized mean differences into natural units of analysis. Four strategies commonly used to overcome these issues and make interpretable translations of standard mean differences are; *a*) the number of years of learning to induce the effect, *b*) estimating the probability of scoring above a proficient threshold, *c*) benchmarking against other effect sizes, and *d*) converting to percentile growth (Baird & Pane, 2019, p. 217). However, it has been shown that options *a* and *b* have substantial flaws for accurate interpretations of effect sizes. For example, comparing effect sizes to years of learning is based on the assumption that “learning rates accumulate linearly over time (Baird & Pane, 2019, p. 225), which rarely holds hold in practice (Lipsey et al., 2012). On the other hand, “[o]ne problem with thresholds is that information is discarded by taking the continuous variable of the standardized score and converting it into a discrete variable” (Baird & Pane, 2019, p. 227). Therefore, we opted not to use these translation metrics to interpret the results of our meta-analysis. Next, option *c* has been criticized for being subject to cherry-picking (Baird & Pane, 2019), meaning that researchers can just find comparison effect sizes that are aligned with their belief of the right interpretation. To remedy this issue, we preregistered (see pre-registration protocol in Chapter II) all translation metrics that we used to benchmark the detected effects of our meta-analysis. Notably, we placed much weight on option *d* since this option “has several desirable properties and no strong weaknesses. (...) The only assumption made is that score distributions are normal, an assumption already made for standardized effect sizes” (Baird

& Pane, 2019, p. 226). In sum, we used three different translation options to interpret the practice importance of our meta-analysis results. *First*, we used Cohen's U_3 metric (Cohen, 1988; Valentine et al., 2019) and converted the overall mean effect size into a percentile growth score (Baird & Pane, 2019; WWC, 2020). *Second*, we compared all detected effects to Kraft's (2020) empirical benchmark scheme for interpreting causal research on educational interventions with standardized achievement outcomes. *Third*, we compared the overall mean effect size to two related interventions, i.e., class-size reduction (Filges et al., 2018) and increased instruction time (Kidron & Lindsay, 2014). We applied these alternative empirical benchmark effect sizes since they were particularly relevant to the intervention, target populations, and outcome measures included in our review (Hill et al., 2008) partly because these interventions represent true structural alternatives to collaborative instruction, and partly because the effect sizes were based on similar population and outcome characteristics as our review.

As I/we also recommend in the third article of the thesis, we deliberately chose not to compare our findings to any universal benchmark schemes, such as the ones suggested by Cohen (1988) and Hattie (2009), although widely used in education and the social sciences (Baird & Pane, 2019; Cheung & Slavin, 2016; Hedges, 2008; Kraft, 2020; Lortie-Forgues & Inglis, 2019). The main reason for this is that

[u]sing those categories to characterize effect sizes from education studies ... can be quite misleading. It is rather like characterizing a child's height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals (Lipsey et al., 2012, p. 4)

In this regard, it is also important to remember and repeat that Cohen (1988, p. 25) himself cautioned against characterizing effect sizes equal to 0.2 as small, 0.5 as medium, and 0.8 as large. Therefore, one of the underlining aims of the thesis is also to show how to avoid using universal schemes to interpret education research and make more relevant interpretations of effect size estimates in education.

Heterogeneity: An advantage, not a shortcoming

“All studies differ and the only interesting questions to ask about them concern how they vary across the factors we conceive of as important” – Gene Glass (2000)

One of the primary aims of the thesis is to show that heterogeneity is not a shortcoming—and thereby a justification of narrative synthesis in systematic review—but an advantage for meta-analysis. In effect, I strive to show with this thesis that heterogeneity is one of the main reasons for doing meta-analysis. As highlighted by Viechtbauer (2005, p. 264), “researchers are becoming increasingly aware of the fact that detection and estimation of moderator effects is often the most valuable contribution of meta-analysis to the research domains in which it is applied.” In essence, this has been known since the beginning of meta-analysis (Glass, 2000; Hedges & Olkin, 1985). In fact, it was one of the main reasons why meta-analysis was originally developed in education and psychology and *not* in medicine because there was a more pressing need to understand effects across diverse outcomes matrix (Shadish & Lecy, 2015, p. 258). As Tipton, Pustejovsky, & Ahmadi clarify, “[a]t the inception of meta-analysis as a field, understanding moderators of effect sizes was viewed as a central aim and unique strength of research synthesis” (2019a, p. 161) and since “meta-analysis (...) are growing in size and scope, with meta-analyses of 100 or more studies becoming increasingly common (...) the goals of meta-analysis have shifted from focusing predominantly on overall average effects towards understanding and explaining heterogeneity in effect sizes.” (2019b, p. 180). Nevertheless, heterogeneity is still the most common justification for conducting narrative synthesis and not doing meta-analysis of quantitative research (Campbell et al., 2019; Ioannidis et al., 2008; Melendez-Torres et al., 2017; Petticrew & Roberts, 2008; Valentine et al., 2010). This is especially clear when looking at the main reasons and justifications for not doing meta-analysis in systematic reviews in medicine, as presented in Table 1 (Ioannidis et al., 2008, p. 1413)

TABLE 1: Reasons for not conducting meta-analysis in 135 systematic reviews from the Cochrane Database of Systematic Reviews

Reason	No (%) of systematic reviews (<i>n</i> = 135)*
Statistical <i>heterogeneity</i> too high	32 (24)
<i>Different</i> interventions compared	41 (30)
<i>Different</i> metrics or outcomes evaluated	26 (19)
<i>Different</i> study designs	21 (16)
<i>Different</i> study participants, settings	21 (16)
<i>Data many counts per participants</i>	5 (4)
Data too limited	11 (8)
Clinical <i>heterogeneity</i> (not otherwise specified)	5 (4)
Synthesis considered inappropriate (not specified why)	3 (2)
Non-normality data	1 (1)
No reason given	10 (7)
Artefact	3(2)
Quantitative synthesis is given in text	7 (5)

Note: Percentages in parenthesis. * “Several reviews gave more than one reason without clarifying which was the most important. In these cases, all reasons are counted.” (Ioannidis et al., 2008, p. 1413). Italic text marks that the reason involves heterogeneity.

From Table 1, it is apparent that heterogeneity is clearly the most widespread reason for not conducting a meta-analysis. Although Table 1 is obtained from medicine, the exact same patterns might be expected to be found in education. As previously mentioned, DCER opted to use narrative synthesis instead of meta-analysis due to variation in the included study designs (Dyssegaard & Larsen, 2013). The same patterns underpinning heterogeneity as the main reason for selecting not to conduct meta-analysis are also found in other disciplines as well (see Littell, 2008; Melendez-Torres et al., 2017).

In the presence of heterogeneity, it is occasionally argued that narrative synthesis *should not* be seen as a second-best solution to meta-analysis (Thomas et al., 2017) since narrative synthesis “can render dense quantitative data intelligible and can increase the policy readiness of a systematic review” (Melendez-Torres et al., 2017, p. 110). However, by not using meta-analysis, including meta-regression techniques, to investigate differential effects across effect sizes, I strive to show with this thesis that narrative synthesis lacks the possibility of asking and answering key questions of *political*, *scientific*, and *practical* concerns. Therefore, we actually allowed heterogeneity on purpose in our meta-analysis in order to answer these key questions about the relative effects of collaborative models of instruction and in order to test differential effects across focal theoretical and methodological moderators. For example, we included various interventions, student populations, outcomes measures, and research designs in our review so that we were able to

answer questions such as: “*Is co-teaching substantially more effective than teacher assistant interventions?*”, “*Does collaborative models of instruction work equally well in general education and special education populations?*”, “*Are collaborative models more effective in Arts compared to STEM subjects?*”, and “*Does the effect vary across RCTs and non-randomized studies?*”. These are all questions that narrative syntheses have no good means to answer reliably (Borenstein et al., 2009; Cooper, 2015; Lipsey, 2007). The argument here is further that meta-regression techniques are probably one of the best methods we have in the toolbox to validly answer the key question of *what works for whom and under what circumstances*.

To ensure valid and reliable inference of our subgroup and meta-regression analyses, we used HTZ Wald tests (Pustejovsky & Tipton, 2021; Tipton & Pustejovsky, 2015) and Cluster Wild Bootstrapping (CWB) techniques (Joshi et al., 2022) to contrast differences across covariates of methodological and theoretical importance.

To these matters, it is occasionally argued that RCTs lack external validity (Cartwright & Munro, 2010; Reiss, 2018), i.e., that they do not provide knowledge that is relevant for understanding “whether the cause-effect relationship holds over variations in persons, settings, treatment variables, and measure measurement variables” (Cook, Campbell, & Shadish, 2002, p. 38). Hereto it is argued that “we do not have good explicit methodologies for how to establish tendency claims[/external validity]¹⁴” (Cartwright, 2011, p. 767). With the thesis and by exploring heterogeneity of the effectiveness of collaborative models of instruction, I strived to show that meta-analysis can actually provide a tool for supporting claims of the external validity/stable tendencies of the effectiveness of interventions. As I try to indicate with the above questions, meta-analysis comes exactly with the possibility to investigate the consistency of effects across variations in persons, settings, treatment variables, and measurement variables. The critique of not having any methods for making general claims might arise from the misunderstanding of meta-analysis as only concerning the estimation of the overall average effect size (see Cartwright & Hardie, 2012). This is also known as the *earth is flat critique* against meta-analysis (Glass, 2000), in which it is assumed that meta-analysts are only concerned with the overall average effect size and deliberately ignore more fine-grained understandings of the effectiveness of interventions.

¹⁴ Cartwright wants to talk about tendency claims instead of external validity (see Cartwright, 2011).

Another advantage of using meta-analysis is that the amount of true heterogeneity can be quantified and investigated in reliable ways instead of being based on arbitrary, *a priori* assumptions and judgments of the reviewers, as in narrative synthesis (Borenstein, Higgins, Hedges, & Rothstein, 2017; Ioannidis et al., 2008; Langan et al., 2019; Viechtbauer, 2005). Moreover, heterogeneity estimates in meta-analysis can provide pivotal diagnostic information with regard to what types of covariates might explain true variation across effect sizes (Pustejovsky & Tipton, 2021). For example, in our meta-analysis, we both found a substantial amount of between-study and within-study heterogeneity, suggesting that both moderator factors that vary at the within-study and between-study level might explain variation across effect sizes and should thereby be added to our subgroup analyses.

Finally, it is also paramount to emphasize that the problem of heterogeneity does not disappear by selecting to use narrative synthesis. To this matter, I will argue in line with Valentine et al. (2010, p. 239) that “if the assertion that the studies are ‘too heterogeneous to combine’ is taken seriously, it precludes both a quantitative and a qualitative summary in most circumstances.”

Complications of using subgroup and meta-regression models

Although subgroup and meta-regression analyses have clear advantages over narrative syntheses in terms of providing evidence for how effects might vary across covariates of theoretical and methodological relevance, they also have certain limitations. As with all post hoc analyses, meta-regression analysis can be prone to *p*-hacking (Deeks et al., 2019; Lakens et al., 2018), i.e., “non-principled decisions during data analysis that are aimed at reducing the *p*-value of a significance test and thus make the data look more robust than they actually are” (Friesse & Frankenbach, 2020). To overcome this issue, we preregistered all of our conducted subgroup and meta-regression analyses, and we only included and tested factors that we considered to be of scientific (i.e., theoretical and methodological) interest in order to restrict the number of included variables in our models. This was particularly important since “[t]he likelihood of a false-positive result among subgroup analyses and meta-regression increases with the number of characteristics investigated” (Deeks et al., 2019, p. 269). To mitigate this problem, also known as *multiplicity* (Polanin, 2013), we applied the *false discovery rate* method (Benjamini & Hochberg, 1995; Laird et al., 2005), as suggested by Polanin (2013). Concretely, this means that cluster wild bootstrapping (CWB) *p* values below 0.05 from our Wald testing were compared to a nominal Type I error rate threshold of 0.01 and

not the conventional value of 0.05 (find these analyses at <https://osf.io/fby7w/>). Yet, accounting for multiplicity, in our case, did not change any of the conclusions reached from the subgroup analyses.

Furthermore, a clear limitation of subgroup and meta-regression models is that they are often substantially underpowered, meaning that there are too few studies and effect sizes in each subgroup for detecting meaningful effects. Therefore, it is often not sensible to conduct subgroup and meta-regression analyses. This issue is even further enhanced when covariates are strongly imbalanced and/or the moderator variables contain a substantial number of missing observations (Deeks et al., 2019). Therefore, we excluded variables with more than 50 percent missing values from our subgroup and meta-regression analyses, and we downgraded our interpretations of the moderator analyses based on multiple imputation techniques to handle missing values (Van Buuren, 2018).

Moreover, a common critique against meta-regression, when based on continuous variables, is that it is subject to *ecological fallacies*, which means that patterns revealed at the study level do not represent the true patterns at the student level (Cooper, 2015; Cooper & Patall, 2009; Deeks et al., 2019). For the same reason, we did not spend much effort on coding factors such as the percentage of low socioeconomic students or minority students in the study sample. However, this critique should certainly be born in mind when interpreting our meta-regression model investigating if the percentage of males in the study sample had a moderating effect on collaborative models of instruction.

The ecological fallacy in meta-regression has been known for a long time. As Gene Glass highlights:

Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in abbreviated reports of research in journals and other printed sources (...) Meta-analysis needs to be replaced by archives of raw data that permit the construction of complex data landscapes that depict the relationships among independent, dependent, and mediating variables (Glass in Cooper & Patall, 2009).

Hence, future meta-analyses should ideally be based on the amalgamation of individual participant data (IPD) to overcome ecological biases. So far, much effort has been made in medical research to overcome this issue (Riley et al., 2021). However, it would require that educational researchers radically change their attitudes toward data sharing. Though, it is possible to use hybrid methods for amalgamating raw and aggregated study data when some studies provide raw data (Goldstein, Yang, Omar, Turner, & Thompson, 2000; Pigott, Williams, & Polanin, 2012). However, we did not use this hybrid approach because we only found five studies that made their raw data available. Yet, we strived to improve the original analysis made in primary studies whenever possible. For example, we fitted multi-level models (Raudenbush & Bryk, 2002), guarding against misspecification via robust variance estimation (Cameron & Miller, 2015) for the raw Project STAR data (Achilles et al., 2008), which was not originally conducted.

Since subgroup and meta-regression analyses are observational in nature, the most serious critique against these analyses is that they generally do not yield causal knowledge because studies are not randomly assigned to be in either subgroup characteristic. Thus, confounding factors might potentially distort the estimated relationships. Hereto, Cooper makes a distinction between *study-generated evidence* and *synthesis-generated evidence*, where the former can establish causal connections because it often draws on random assignment of students to the treatment and control groups, whereas the latter can only generate hypotheses about causal relations¹⁵ due to the lack of randomization. Hence, too strong causal interpretations of subgroup and meta-regression should be avoided, but they might produce invaluable knowledge about which causal connection future primary research should concentrate on. Therefore, it is important to emphasize that meta-analysis is clearly not without deficits, but it is also critical to note that the critique related to causality is not idiosyncratic to meta-analysis but a general critique against all types of systematic reviews and synthesis techniques, including narrative synthesis (Cooper, 2015). Thus, I still consider meta-analysis as the best tool we have to date for analyzing variation across educational interventions. Since no experiment will ever be able to investigate all relevant differential effects, I usually think of this critique against meta-analysis this way (to paraphrase Donald Rubin in Van Buuren, 2018, Preface): It is not that meta-analysis is so good; it's really that other methods for synthesizing quantitative research are so bad.

¹⁵ This is also described as the ability of meta-analysis to indicate causal signs (Cook, 1994).

Handling dependent effect sizes in meta-analysis

One of the main issues that required attention in order for the meta-analysis of the thesis to represent a state-of-the-art meta-analysis concerned how to handle dependencies between effect sizes coming from the same study most adequately. This is an omnipresent issue in educational reviews and meta-analyses, including our review. The following section describes the rationale behind using multi-level modeling and robust variance estimation (RVE) to handle dependent effect sizes in our meta-analysis, and it argues why we considered this method to represent a state-of-the-art technique. Furthermore, I discuss some of the inherent limitations of the method.

Dependence structures

Most research syntheses in the social sciences encounter that studies, authors, or research labs contribute with multiple effect sizes to the review, creating various statistical dependencies among the effect sizes. Dependence among effect sizes can be said to follow two broad dependence structures; *the correlated effects structure* and *the hierarchical effects structure*. The correlated effects structure is characterized by the dependence occurring in effect size estimate through the sample error terms, e_{ij} , from Equation (1) (Hedges, 2019; Joshi, 2021). Common reasons for the occurrence of correlated effect sizes are; 1) studies reporting multiple measures on the same individuals across different time points; 2) studies measuring multiple outcomes on the same sample of individuals (e.g., math and language arts scores, respectively); 3) studies comparing different treatment groups to the same control group or comparing the same treatment to multiple controls. On the other hand, the hierarchical effects structure is characterized by the dependence occurring through the true effect size estimates, θ_{ij} , from Equation (1) (Hedges, 2019; Joshi, 2021). Common reasons for the appearance of hierarchical dependent effect sizes are; 1) the same author(s) or research labs contribute with multiple studies of the same questions, or 2) the same study reporting multiple effect sizes across nonoverlapping samples (e.g., result for primary and secondary students, respectively). In our review, we found 45 studies having a correlated effects dependence structure, six studies having a hierarchical effects dependence structure, and four studies containing both dependency structures with multiple outcomes coming from multiple non-overlapping samples.

Common methods to handle dependencies among effect sizes

In previous co-teaching reviews (Khoury, 2014; Murawski & Swanson, 2001; Willett, Yamashita, & Anderson, 1983), dependencies among effect sizes have been ignored. This approach can perform well when very few studies contribute with multiple effect sizes (Becker, 2000; Hedges et al., 2010b). However, as we show in our meta-analysis, it is actually more common for studies regarding collaborative models of instruction to contribute with multiple effect sizes than just a single one. Consequently, ignoring dependence among effect sizes will lead to incorrect standard errors and incorrect inference from hypothesis tests.

A popular method for handling dependence among effect sizes in education is to create a synthetic effect size for each study by averaging all within-study effect sizes across different outcomes, e.g., across various kinds of achievement scores (Tipton et al., 2019b). For current examples of this approach, see Bredow et al. (2021), Furenes, Kucirkova, & Bus (2021) and Bethhäuser, Bach-Mortensen, & Engzell (2022). However, as we show in Figure 3 in Chapter III, this method does not adequately control the nominal Type I error rate when multiple dependency structures are present in the review data,¹⁶ as we experienced. Therefore, we did not consider this to be a viable method to use in our review. Furthermore, it would have restricted us from investigating differential effects across factors varying within studies, such as different outcome measures.

The most sophisticated method to handle dependent effect sizes is the so-called multivariate meta-analysis method (Becker, 2000; Raudenbush et al., 1988), in which the dependencies among effect sizes are explicitly modeled by the use of the true variance-covariance matrix from each study reporting multiple effect sizes. However, this method has rarely been used in the social sciences since it requires exact knowledge about the covariance or correlations among effect sizes, which is information that is rarely reported in primary studies (Gleser & Olkin, 2009). This was also the case for most studies included in our meta-analysis. Only a few studies provided the necessary information required to obtain or estimate the variance-covariance matrix of the dependent effect sizes within the given study. For those studies that reported relevant information to obtain a variance-covariance matrix, we coded this information in our data extraction scheme so that future updates of the review can locate and make use of this information. Due to time constraints, we did not estimate these matrices.

¹⁶ However, it is in some cases entirely reasonable to aggregate within-study effect sizes (Pustejovsky, 2019b).

Robust variance estimation

To overcome the challenge of not knowing the true dependencies among effect sizes, and thereby to overcome the shortcomings of the above-presented methods, we applied *robust variance estimation* (RVE; Hedges et al., 2010b; Pustejovsky & Tipton, 2021; Tipton, 2015; Tipton & Pustejovsky, 2015) since it has been shown to be the most accurate method for handling dependent effect sizes (Fernández-Castilla et al., 2020; Vembye et al., 2022 [Chapter III]).

To explain RVE used across all models in our review, I here extend the model presented in Equations (1) and (2) in order to allow for the sources of heterogeneity we found during our overall average effect size estimation. To investigate heterogeneity, we assumed the effect size estimates represent a sample from some underlying population of effects and that the average effect sizes can be explained set of covariates or predictors (as Pustejovsky & Tipton, 2021). To explain this procedure, let $\mathbf{x}_{ij}$ denote a row vector of p covariates and β denote a vector of p regression coefficients, so that the meta-regression model can be express as

$$T_{ij} = \mathbf{x}_{ij}\beta + u_{ij} + e_{ij}$$

where u_{ij} represent the variation not accounted for by the covariates.

RVE

To further explicate the underlying computational details of the meta-regression models used in the thesis, I will present RVE via matrix notation. The general meta-regression model can then be written as

$$\mathbf{T}_j = \mathbf{X}_j\beta + \mathbf{u}_j + \mathbf{e}_j$$

where $\mathbf{T}_j$ represent a vector of k_j effect size estimates, $\mathbf{X}_j$ denote the $k_j \times p$ design matrix of covariates (for the intercept-only model, i.e., the overall average effect size model, $\mathbf{X}_j$ denote a $k_j \times 1$ vector of 1's), and β denote a $p \times 1$ vector of regression coefficients, $\mathbf{u}_j$ represent a vector of random effects, and $\mathbf{e}_j$ represent a vector of sample error, all for studies $j = 1, \dots, J$.

To further describe the weighted least squares (WLS) estimator that we used in our models, let $\mathbf{W}_j$ be an $k_j \times k_j$ matrix of weights for the j^{th} study, and let $\mathbf{\Phi}_j = \text{Var}(\mathbf{u}_j + \mathbf{e}_j)$ be an $k_j \times k_j$ matrix that describes the true dependency structures of the effect size estimates in the j^{th} study (Pustejovsky & Tipton, 2021). The WLS estimator can then be written as

$$\hat{\beta} = \mathbf{M} \left(\sum_{j=1}^J \mathbf{X}_j' \mathbf{W}_j \mathbf{T}_j \right), \text{ where } \mathbf{M} = \left(\sum_{j=1}^J \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j \right)^{-1}$$

and the true sampling variance of $\hat{\beta}$ is given by

$$\text{Var}(\hat{\beta}) = \mathbf{M} \left(\sum_{j=1}^J \mathbf{X}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \mathbf{X}_j \right) \mathbf{M}$$

If the true dependence structure had been known, then we could have calculated weights that were fully inverse of the variance-covariance of each study, i.e., $\mathbf{W}_j = \hat{\mathbf{\Phi}}_j^{-1}$. The weights would then have been fully efficient and thereby produced the smallest possible sampling variance of $\hat{\beta}$. This also means that under this scenario, $\text{Var}(\hat{\beta}) = \mathbf{M}$. When this information is not available, as in our case, RVE can be used to roughly and empirically approximate the dependence structure of the effect sizes by using the observed residuals and a small-sample adjustment matrix as substitutes for $\mathbf{\Phi}_j$. Following this approach, the robust estimator of the sampling variance can be expressed as

$$\mathbf{V}^R = \mathbf{M} \left(\sum_{j=1}^J \mathbf{X}_j' \mathbf{W}_j \mathbf{A}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j' \mathbf{A}_j \mathbf{W}_j \mathbf{X}_j \right) \mathbf{M},$$

where $\hat{\mathbf{e}}_j = \mathbf{T}_j - \mathbf{X}_j \hat{\beta}$ is a vector of the residuals from the j^{th} study, and $\mathbf{A}_j$ denote a $k_j \times k_j$ small-sample matrix, given by

$$\mathbf{A}_j = \mathbf{W}_j^{-\frac{1}{2}} \left[\mathbf{W}_j^{-\frac{1}{2}} (\mathbf{W}_j^{-1} - \mathbf{X}_j \mathbf{M} \mathbf{X}_j) \mathbf{W}_j^{-\frac{1}{2}} \right] \mathbf{W}_j^{-\frac{1}{2}},$$

and $\mathbf{W}_j^{-\frac{1}{2}}$ is the inverse of the symmetric square root of the weight matrix $\mathbf{W}_j$. These are the CR2 adjustment matrices (Tipton, 2015; Tipton & Pustejovsky, 2015), which are constructed so that $E(\mathbf{V}^R) = \text{Var}(\hat{\beta})$ if $\mathbf{W}_j = \Phi_j^{-1}$ for all $j = 1, \dots, J$, i.e., when the working model is exactly correct and the weights are inverse-variance.¹⁷

Working models

RVE in meta-analysis implies the use of working models that tentatively aim to resemble the true dependency structures between effect sizes as close as possible. Until recently, methods to handle dependencies were limited to either making purely correlated (Hedges et al., 2010b) or purely hierarchical (Fernández-Castilla et al., 2020; Hedges et al., 2010b; Van den Noortgate et al., 2014, 2013) assumptions about the dependency structure, which, in turn, decreases the estimated precision of these models when the model diverges from the true dependency structures of the meta-analytical data. In order to remedy this issue, the correlated-hierarchical effect (CHE) model was developed in which multi-level modeling (Van den Noortgate et al., 2013, 2014) and RVE (Hedges et al., 2010b) are combined (also defined as the CHE-RVE model, see Chapter II) while simultaneously accounting for correlated and hierarchical dependency structures (Pustejovsky & Tipton, 2021). Because we expected at the planning stage of the review (see our protocols) both to find correlated and hierarchical dependence structures among effect sizes but also because we expected to find true random variation among effect sizes both at the between- and within-study levels—also in moderator analyses—all of our models draw on the CHE working models. Specifically, we applied three different working models from the CHE model family. I will describe these working models below and the reasons why we used these three models.

Correlated-Hierarchical Effects (CHE) model

We used the CHE(-RVE) working model for estimation of the overall average effect size and all meta-regression models with continuous moderator variables. The CHE model is given by

¹⁷ This passage is taken from the first draft of the second paper of the thesis.

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + v_{ij} + e_{ij} \quad (5)$$

where $\text{Var}(u_{ij}) = \tau^2$, $\text{Var}(v_{ij}) = \omega^2$, $\text{Var}(e_{ij}) = s_j^2$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. τ and ω represent the between-study and within-study SDs, respectively, and $s_j^2 = \frac{1}{k_j} \sum_{i=1}^{k_j} s_{ij}^2$ denote the average sampling variance of study J . Notably, this is one of the models that we developed power approximations for in Chapters II and III. As previously mentioned, the CHE model involves making the simplifying assumption that there is a single constant correlation among all dependent effect sizes, ρ . As suggested by (Kirkham et al., 2012, p. 2182), we estimated ρ “by calculating the Pearson correlation between the pairs of available treatment effect estimates in those studies that provide data on both outcomes,” in our case, this means mathematics and language arts effect sizes. We also used the STAR study data (Achilles et al., 2008) to obtain sample correlation to compare the difference between the two approaches for obtaining ρ . Yet, for both methods $\rho \approx .7$ (see Chapter II). Since our meta-analysis data contained 280 across 96 nonoverlapping samples of students from 76 studies, it could potentially have been beneficial to fit a CHE+ model, accounting for true random variation at the sample level as well. However, we conducted a *likelihood ratio test* (Viechtbauer, 2022), showing that nesting effect sizes in samples did not improve our model in any way. In fact, it just moved all variation from the study level to the sample level. Therefore, we only used the simpler and clearer version of the CHE model.

Subgroup Correlated Effects (SCE+) model

For subgroup analyses based on categorical moderators, we varied between fitting Subgroup Correlated Effects Plus (SCE+) and Correlated Multivariate Effects Plus (CMVE+) models.¹⁸ The SCE+ model is given by

$$T_{ij} = \sum_{c=1}^C d_{ij}^c (\mathbf{x}_{ij}\beta_c + u_{cj} + v_{cij}) + e_{ij} \quad (6)$$

¹⁸ The latter model is only described in the first version preprint of Pustejovsky & Tipton (2021). It can be found at <https://osf.io/preprints/metaarxiv/vyfcj/>.

where $\text{Var}(u_{cj}) = \tau_c^2$, $\text{Var}(v_{cij}) = \omega_c^2$, $\text{Cov}(u_{bj}, u_{cj}) = 0$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2 \sum_{c=1}^C d_{hj}^c d_{ij}^c$. Here d_{ij}^c is an indicator of whether a given outcome falls within the given subgroup, c . The SCE+ model makes the simplifying assumptions that effect size estimates from the same study falling in the same subgroup are correlated, whereas effect size estimates from the same study falling into different subgroup categories are uncorrelated.

Correlated Multivariate Effects plus (CMVE+) model

Unlike the SCE+ model, the CMVE+ model is based on the more realistic assumptions that effect size estimates from the same study falling in the same subgroup are correlated and that effect size estimates from the same study falling into different subgroup categories are correlated, as well. The CMVE+ model is given by

$$T_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \sum_{c=1}^C (d_{ij}^c u_{cj} + d_{ij}^c v_{cij}) + e_{ij} \quad (7)$$

where $\text{Var}(u_{cj}) = \tau_c^2$, $\text{Cov}(u_{bj}, u_{cj}) = \psi_{bc}\tau_b\tau_c$, $\text{Var}(v_{cij}) = \omega_c^2$, $\text{Cov}(v_{bij}, v_{cij}) = \zeta_{bc}\omega_b\omega_c$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. ψ_{bc} and ζ_{bc} are the correlations between the random effects at the study and effect size level, respectively. A further advantage of using the CMVE+ model is that by allowing for correlation among the random effects, it is possible to investigate if interventions that have a large impact on outcome A also tend to have a large impact on outcome B.

Constraints of the CMVE+ model

Although the CMVE+ model is assumed to represent the “golden” standard in terms of precision compared to the SCE+ model, it only works under rather narrow conditions when

- 1) there are few multivariate dimensions
- 2) there are a substantial number of studies and effect sizes available from each dimension.
- 3) there are a substantial number of studies having effect sizes from each possible pair of outcome dimensions.

In particular, these conditions were only met for our moderator analysis regarding the differential effects of collaborative instruction between Arts vs. STEM subjects. To investigate and decide if the categorical subgroup moderator was viable to be fitted to the CMVE+ model, we used what we call *overlapping tables*, as suggested in the Supplementary Material linked to Pustejovsky & Tipton (2021). Find these analyses in Supplementary Section S4 in Chapter II.

Further advantages of using CHE models

A significant advantage of the CHE models is that they allow reviewers to make reliable multi-contrast tests between subgroup effect sizes because they fit all subgroup dimensions into one model (Pustejovsky & Tipton, 2021; Tipton & Pustejovsky, 2015, see also Supplementary Section S3 in Chapter II). This is not possible when splitting subgroup analyses so that they conduct separate meta-analyses for each outcome, which is a common approach used in social science reviews (Polanin, 2013; Tipton et al., 2019b).

Limitations of the CHE models

Although the CHE model family has a number of advantages compared to other common methods for handling dependent effect sizes, they also come with certain limitations. The most important limitation is that the estimation of the individual random variance component (i.e., in our case, the estimation of the between- and within-study variance components) is strongly impacted by the assumed (constant) sample correlation. Thus, any substantial interpretation of the exact magnitude and the relative magnitude of the between- and within-study variance should be made with caution for the CHE model. However, it is important to note that the total variance estimate is generally insensitive to the assumed constant sample correlation (see, for instance, Supplementary Figure S10 in Chapter II). Furthermore, the sensitivity of the variance component estimates also complicates the estimation of prediction intervals for the CHE model since it is generally not viable to combine robust and non-robust estimators, as it would require estimating a prediction interval for μ of the CHE model. Some researchers might see this as a strong limitation of the CHE models because they consider prediction intervals as one of the most important sources of information in meta-analysis (Borenstein, 2019). Despite this limitation, we/I have followed the argument put forward by Pustejovsky & Tipton (2021, p. 437), “that it is better to apply a working model that captures the structure of one’s data—even if the variance component estimates are sensitive to

assumptions—than to use one that imposes stronger and less plausible assumptions (i.e., that there is no within-study heterogeneity [or no correlation among within-study effect sizes]).”

Additionally, a limitation of the CHE model family is that it is still unclear how they perform when moderators are strongly imbalanced, which would, for example, be the case if one or few studies contribute with the majority of effect sizes to the analysis. However, an advantage of the CHE model is that they apply Satterthwaite’s degrees of freedom (Satterthwaite, 1946), which can provide diagnostic information regarding the certainty in the variance estimation. Generally, Satterthwaite degrees of freedom estimates below five indicate that there might be substantial uncertainties attached to the variance estimation.

Finally, the fact that researchers can choose between a number of CHE working models makes them subject to what is called the “researcher degrees of freedom”—the meta-analytical equivalent to *p*-hacking—to which the model selection is based on whether the models yield statistically significant results or not. However, as we strive to show with our review, preregistration of a protocol, including analytical plans, can overcome this issue.

Current methodological limitations for meta-analysis of dependent effect sizes

As is apparent from the above section, meta-analytical methods to handle dependent effect sizes do still have some clear boundaries. To push some of these boundaries, the thesis introduces new power approximation formulas for meta-analysis of dependent effect sizes and introduces the *PO-MADE* R package that aims to increase the accessibility of the conduct of this type of analysis. Thereby the thesis implicitly aims to replace previous power approximation methods that were based on the assumption of independent effect sizes (Hedges & Pigott, 2001) and previous tools for conducting power analysis for meta-analytical models (Harrer et al., 2019). Yet, this only represents a minor contribution to remedying the lack of statistical methods for handling dependent effect sizes. For example, methods for conducting precision analyses (Rothman & Greenland, 2018) and power analyses of subgroup models still need more attention. Currently, these methods can only be conducted via the use of Monte Carlo Simulation (MCS) (Morris, White, & Crowther, 2019). The main concern about this approach, however, is the level of complexity which probably discourages most applied reviewers from conducting such analyses. It is, therefore, critical that the

community of meta-analysis methodologists develops new and more accessible methods for applied reviewers.

Further methods that still need more attention in the presence of dependent effect sizes are *selection models* (Hedges & Vevea, 2005), *prediction intervals* (Borenstein, 2019), and methods for *conducting reliable Wald tests* (Joshi et al., 2022; Tipton & Pustejovsky, 2015) *across multiple imputed datasets*, just to mention some of the limitations we experienced during the conduct of our review. Moreover, it is important to expand the software for conducting power-analysis for meta-analysis of dependent effect sizes so that these are available to non-R users as well.

3. Educational Theory

Although the overall focus of the thesis is on the methodological conduct of systematic reviews and meta-analysis, the substantive analysis of our review was strongly motivated by educational theories. Essentially, without profound (educational) theory, it would not have been possible to answer the second research question of the thesis since a part of conducting a state-of-the-art meta-analysis entails aligning the meta-analysis to the theory or theories of the area under review (Pigott, 2012). Therefore, in our review, educational theories regarding the effectiveness of collaborative models of instruction on student achievement significantly informed and guided the development of our search string, our coding scheme, and the conduct of statistical hypothesis testing. Specifically, we drew heavily on three seminal articles on co-teaching¹⁹ (Cook & Friend, 1995; Friend, 2008; Murawski & Swanson, 2001) and one on teacher assistants²⁰ (Muijs & Reynolds, 2003) to develop our review.

¹⁹ Defined as: two or more professionals delivering substantive instruction to a diverse, or blended, group of students in a single physical space” (Cook & Friend, 1995, p. 2). The term ‘professionals’ in this regard specifically refers to the collaboration between a general and a special education teacher, such as a speech-language clinician, reading specialist, bilingual teacher, or occupational therapist.

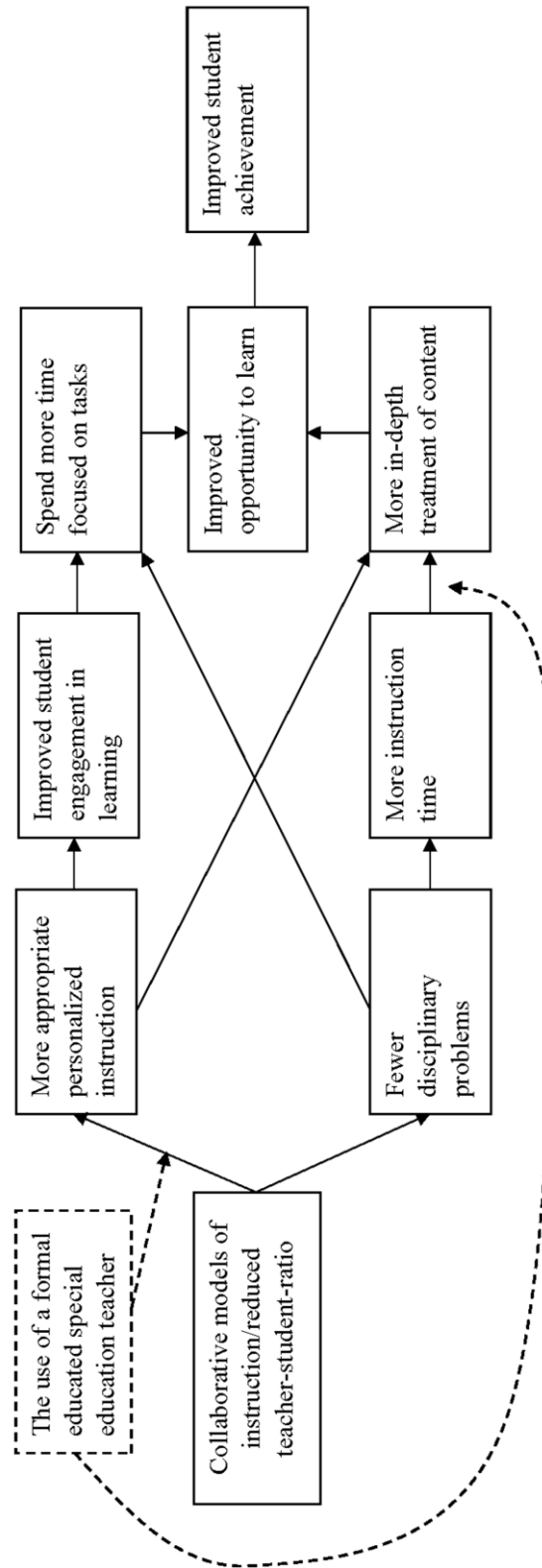
²⁰ Defined as: in-class collaboration between a general education teacher and adults/paraprofessional educators without a formal teacher education such as pedagogues, (voluntary) parents, etc. (Blatchford et al., 2011).

Testing theory through meta-analysis

Common for all collaborative models of instruction is that they draw on the same causal theory about the effect of reducing the student-teacher ratio, as depicted in Figure 1 (see Blatchford et al., 2011; Cook & Friend, 1995, pp. 3–4; Muijs & Reynolds, 2003, pp. 221–222 for overlaps in theory). Figure 1 is developed with inspiration from Filges, Sonne-Schmidt, & Jørgensen (2015) since models of collaborative instruction strongly overlap with theories on class-size reduction as well. The assumed common causal mechanism underlying the effectiveness of collaborative models of instruction is that the reduction of the student-teacher ratio has a positive effect on two focal lines of factors. Each of which, in turn, opens different paths to increased student learning. As illustrated in Figure 1, the first hypothesis is that having two in-class teachers can increase student learning by reducing the number of disciplinary problems, making time for more instruction and in-depth treatment of the content, which then improves learning conditions. On the other hand, it is assumed that reducing the student-teacher ratio increases student learning by providing the opportunity for teachers to make more appropriate personalized instruction for each student, augmenting student engagement and self-confidence, and ensuring that students spend more time on tasks, which in the end increases learning conditions.

In addition to the above-presented main assumptions, the co-teaching theory adds a number of assumptions about under which (narrow) conditions co-teaching can work or works most effectively. According to the co-teaching theory, one of the most important components underpinning the effectiveness of co-teaching is the use of a formally educated special education teacher (Cook & Friend, 1995; Friend, 2008). Hereto, it is assumed that an equal share of instruction between the general and special education teachers is vital for increasing student learning since it combines the general teacher's in-depth knowledge of the curriculum with the specialized knowledge of the special education teacher about customizing the instruction the needs of the individual student. Therefore, it is also presumed that this approach significantly enhances the appropriate personalized instruction for each student and that it is the most optimal approach to capitalize on the increased instruction time. These two assumptions are depicted by the dashed square and lines in Figure 1.

FIGURE 1. Causal diagram for the impact of collaborative models of instruction on student achievement



Note: Inspired by Filges et al. (2015). Bold lines and squares indicate the common causal mechanism underlying the theory of all collaborative models of instruction, whereas the dashed square and lines indicate where it is assumed in the co-teaching literature that the use of formally educated special education teacher augment the effects of collaborative instruction.

However, we were quite skeptical about the empirical foundation of these strong assumptions put forward by the co-teaching theory, and we wanted to use meta-analysis to test if adding a special education teacher relative to non-formal teacher-educated personnel would explain a large difference in effect sizes, as shown in Figure 1. Consequently, we carefully coded the exact two-teacher compositions used within and across all studies so that this potential difference in effect could be tested.

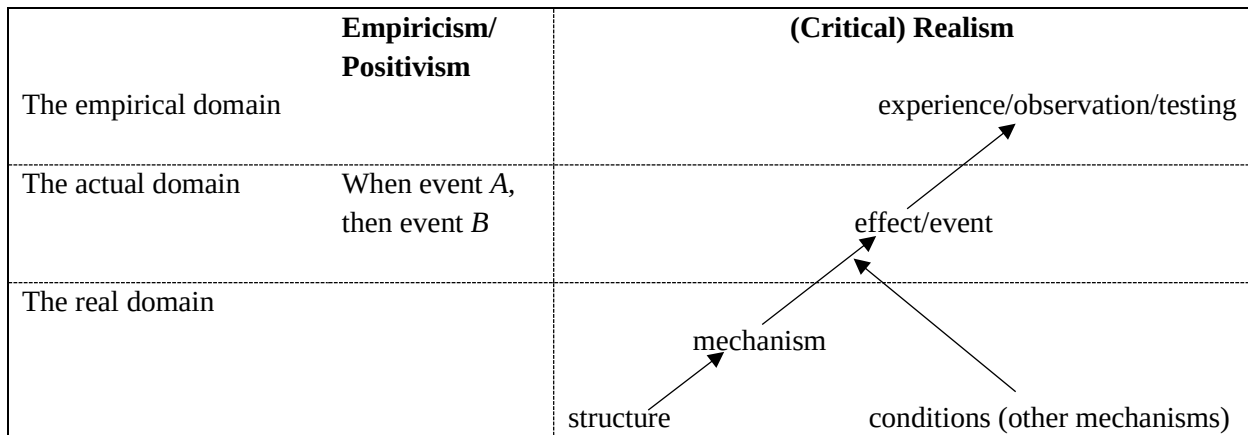
Furthermore, besides making assumptions about the effectiveness of particular two-teacher compositions, the co-teaching theory further assumes that collaborative instruction is only effective when; 1) co-teaching *training* is provided, 2) time for *co-planning* is provided, 3) using a *variety* of co-teaching models, 4) co-teaching is provided for *more than a year*, 5) provided for *two sessions per week*, 6) teacher collaboration is *voluntary*, and 7) the two teachers have a *sound working relationship* (Cook & Friend, 1995; Dafolo, 2019; Friend, 2008). To further test the co-teaching theory, we drew on these assumptions to develop our data extraction scheme questions in order to test differential effects across the factors. After pilot testing our scheme on eight studies, we dropped the investigation of assumptions 6 and 7 since these factors were not mentioned frequently in the (pilot) literature—yet, I also consider it a key part of systematic reviewing to show the boundaries of the given literature under review. All other factors were retrievable for more than 50% of the eligible studies. However, since these factors were never fully reported across all included studies in the review, we had to impute plausible values for studies containing missing information, inducing an unknown degree of error in our investigations. The main takeaway from our theoretical endeavors was, therefore, that more knowledge is needed about moderating effects of collaborative models of instruction and that the assumptions made in co-teaching literature are not well supported empirically.

The underlining aim of conducting tests like the ones presented above was to show how meta-analysis can contribute to new theoretical explanations for the effectiveness of collaborative models of instruction, as suggested by Cook et al. (1992). To gain a more in-depth theoretical understanding of the effectiveness of collaborative models of instruction, we would have liked to test interaction effects between these moderators of theoretical importance. However, the number of included studies and calculated effect sizes did not provide enough statistical power to conduct these types of tests (Cook et al., 1992).

4. Philosophy of Science

This thesis is based on a (critical) realist perspective on science, society, education, and causality (Sayer, 1999).²¹ This implies that although we find persistent and robust effects across various student populations, settings, treatment compositions, and outcomes—suggesting that collaborative models of instruction have high external validity—we do not perceive our results to represent a causal social law that works for all students in all circumstance and across all contexts. Instead, we acknowledge that the effectiveness, efficiency, and efficacy of social interventions are strongly context dependent. As depicted in Figure 2, the capacity of a social intervention to produce its effects strongly depends on the underlining condition surrounding the causal chains from the intervention to its effect, like the causal mechanisms of collaborative models of instruction illustrated in Figure 1 above. If the causal chains break, no effects will appear. Say, for example, that two teachers are neither able to increase the net instruction time nor to provide more personal instruction, then it is rather unlikely that they will bring about the expected effects of collaborative instruction.

FIGURE 2. A realist view of causation



Note: Inspiration from Sayer (1999) and Buch-Hansen & Nielsen (2005, p. 27). *The real domain* contains all humans and objects' capacities to function in certain ways (over time, these might change), whether or not these are actualized, whereas *the actual domain* contains all events that have ever happened/been potentialized, and *the empirical domain* contains the human experiences, observation, and measuring of the actual events. The *empirical fallacy* is to collapse all domains into one.

²¹ This view is based on the conviction that nature and human actions, including education, represent objective realities that are comprised by causal connections and effects that exist independently of the human perception of these entities but can in part be experienced, observed, tested and measured. I will mainly focus on my view on causality in this section since this part of the realist theory played the largest role in the project compared to the general realist theory of science.

This view on causation further explains why there is nothing mysterious about research studies not showing the exact same effect since all studies are surrounded by different contextual supporting mechanisms. To adequately illustrate the connection between contexts and social interventions, it can be helpful to use the notion of *causal principles* (CP), suggested by Cartwright & Hardie (2012, p. 26) and Kvernbekk (2016). To illustrate the CP for the effects of collaborative models of instruction on student achievement across studies, let $y_j(i)$ be the effect on student achievement and $y_j(i)_0$ be the student baseline level before the intervention for student i in study j , for all students $i = 1, \dots, N$ and studies $j = 1, \dots, J$. Then, let b_j denote *support factors* consolidating the effectiveness of the collaborative models of instruction intervention x_j (for example, these support factors could be more released instruction time and more personalized instruction, as given in Figure 1), $z_j(i)$ represent other factors that have an impact on $y_j(i)$ that does not concern x_j (for example, teachers' and students' abilities, support from parents, etc. These can also be considered as confounding factors/covariates), and $u_j(i)$ denote further random unknown factors that might contribute to increasing student achievement but which do not interact with either the intervention or its support factors, all for students $i = 1, \dots, N$ and studies $j = 1, \dots, J$. The CPs governing the effectiveness of collaborative models of instruction between students across the included studies in the meta-analysis can then be expressed as

$$\begin{aligned}
 \text{CP}_1: y_1(i) &= y_1(i)_0 + b_1 x_1(i) + z_1(i) + u_1(i) \\
 \text{CP}_2: y_2(i) &= y_2(i)_0 + b_2 x_2(i) + z_2(i) + u_2(i) \\
 &\vdots \\
 \text{CP}_J: y_J(i) &= y_J(i)_0 + b_J x_J(i) + z_J(i) + u_J(i)
 \end{aligned} \tag{8}$$

As Equation (8) here illustrates, from the realist view, all students included in our meta-analysis are, in principle, governed by different contextual conditions that determine the effectiveness of collaborative models of instruction. This also means that even in the same treatment group in the same study, some student might increase their academic achievement while others might not, simply because different conditions (i.e., y_{j_0} , b_j , z_j , and u_j) between students might support or obstruct the working of collaborative models of instruction (Cartwright, 2007). That said, it is important to notice that in our meta-analysis, we are looking for *systematic, average effects* of x_j , i.e., we aim to detect the overall proportional difference(s) between students that experienced an

increased achievement when exposed to collaborative instruction compared to single-taught environments across a substantial variation in both systematic and unsystematic preconditions and contexts. Again, “this does not mean that systematic [effects] respect constant[s]” (King et al., 1994, p. 62). More precisely, the inferences made in our review pertain to generalized causal inferences (Cook et al., 2002) that might support predictions for the effectiveness of collaborative models of instruction at the political level if/when disseminated to all or large parts of a school system. Therefore, our meta-analysis does not provide a blueprint for the effectiveness of collaborative instruction at the student level since this strongly bears on the individual educational context (i.e., y_{j0} , b_j , z_j , and u_j).

Randomization and counterfactual reasoning

Equation (8) further helps to illustrate why we favored randomized studies in our review since it can show how *randomization* provides a means to ensure balances of observable and non-observable covariates/confounders (i.e., y_{j0} , b_j , z_j , and u_j), guarding against systematic differences between the intervention and control group (Rosenbaum, 2017, pp. 10–11), which in turn strongly increases the likelihood of finding the true effect of an intervention.

On another line, a key conception not only belonging to randomized studies but to all control group designed studies is the notion of *counterfactual reasoning* (Pearl, 2009; Pearl & Mackenzie, 2018), which I/we consider to be the most reliable ground for causal knowledge. For example, without the use of reliable control groups, it would be almost impossible to decipher the effect of collaborative models of instruction from the effect of students natural maturing and development.²² Thus, subtracting the control group effect from the treatment group effect aims to isolate the treatment effect by removing all parts of the effect not coming from the treatment.

In sum, educational statistics and quantitative methods are often connected with positivism/empiricism (Chakravartty, 2011) and are criticized for disseminating a simplistic regularity view of causality that does not represent true educational processes (Biesta, 2007; Hammersley, 1997). However, I strive to show with the thesis that it is possible to make reliable generalized causal inferences without neither subscribing to a simplistic regularity view of causality nor

²² In fact, this was the main reason why we did not include single-group pre-posttest designed studies in our meta-analysis.

abandoning causality in education, as is argued by critiques of evidence-based research (Biesta, 2007; Hammersley, 1997; Korsgaard, 2020).

5. Open Science – Preregistration, Open Material, and Open Data

The principles of open science²³ have guided all articles of this thesis in various ways. The following section is devoted to describing the reasons behind and the value of open science and open data procedures in relation to the thesis.

Preregistration, open materials, and open data in systematic reviews and meta-analyses

We preregistered the analysis plan²⁴ [see Protocol in Chapter II] for our systematic review and meta-analysis via the PRISMA-P statement (Moher et al., 2009) and the Open Science Framework (OSF)²⁵ template for two broad reasons. *First*, this prevented our analyses from being subject, either consciously or unconsciously, to questionable practices such as *p*-hacking (see definition in Section 2) and HARKing (Kerr, 1998, p. 196), i.e., “hypothesizing after the results are known”. *Second*, it provides a means to make a clear distinction between *confirmatory analyses*, i.e., analyses planned prior to the data analysis, and *exploratory analysis*, i.e., analyses planned during the data analysis. This ensures that other researchers can closely investigate the difference between our initial plan and the final analysis (Moreau & Gamble, 2020; Nosek et al., 2018). It was not that we did not diverge from our initial plan, as will almost always be expected. However, we carefully documented all deviations from the protocol to ensure full transparency.

To further improve transparency, all background materials linked to our review and meta-analysis have been open sourced via OSF. We did that for various reasons. First and foremost, to ensure the trustworthiness of the review. It was especially important for me, as a single-coder of most of the effect sizes, the data extraction, and the risk of bias assessment, to ensure that these procedures were fully transparent and that other researchers can check their accuracy. For the same reason, I documented the exact pages from where all data were extracted since this has been a great

²³ i.e., the idea of making all parts of research (including data, analysis codes, other background materials, software, publication) accessible to the public.

²⁴ We have two protocols because the CHE models were developed after we made the first version. The second version incorporate our change of models.

²⁵ See <https://osf.io>.

frustration of mine when I read other systematic reviews and meta-analyses. In this regard, open-sourcing of our material also aims to avoid cognitive algebra, as discussed in Section 2. Furthermore, meta-analysis has been criticized for not being reproducible due to the lack of transparency in effect size calculations, and we wanted to overcome this issue (Maassen et al., 2020).

Yet, we also shared all of our data and analysis codes so that other researchers can either reproduce/test our results, conduct further relevant analyses, or easily update the review when more studies get published. Hereto, as I/we show in the third article of the dissertation, data sharing is a pivotal means, among other things, for making power analysis for meta-analysis of dependent effect sizes a common and reliable practice in systematic reviews. Lastly, one of the chief aims of making all codes available was to inspire future meta-analyses and provide the review community with solutions for complex coding tasks in meta-analysis, just as I have been inspired by other researchers' codes during this dissertation (in this regard, special thanks go to James E. Pustejovsky, Megha Joshi, and Wolfgang Viechtbauer).

Open data and codes for statistical simulation studies

Likewise, open data and codes are all-important for the credibility of simulation studies to avoid being subject to “selective reporting of only the most favorable (or unfavorable) configurations of data-generating mechanisms, running the simulations many times under different seeds and selecting the most favorable” (Morris et al., 2019, p. 2096). It was further important to ensure transparency of our codes since we especially favor the CHE-RVE model relative to other models for handling dependent effect sizes. Thereby, had we not shared the codes, other researchers could potentially and rightly criticize us for just reporting results that favored the performance of the CHE-RVE model.

Open codes for software/package developments

Finally, open science played a key role in the development of the *POMADE* R package, presented in the third article of the thesis. Besides ensuring the transparency of the package, the sharing of all package functions via GitHub enables researchers to report bugs and make suggestions for how we can improve the package (Wickham, 2015).

The use of R(Studio)

Our use of the statistical environments R (R Core Team, 2022) and RStudio (RStudio Team, 2015) has a vital impact on the reproducibility of the thesis since it is free of charge. Consequently, most analyses of the thesis can be reproduced by downloading R and RStudio and either pressing Ctrl + ENTER or Ctrl + Alt + R, depending on whether an R script or Rmarkdown document is used. That way, the accuracy of all analyses included in the thesis can more easily be assessed.

6. Methodology

The thesis had to overcome a number of methodological challenges, too. In this section, I give a brief overview of the five most important ones that I/we encountered during my dissertation work, and I present how I/we attempted to tackle these issues.

Challenge 1: Finding all relevant literature

One of the greatest threats to the validity of systematic reviews and meta-analyses is the failure to locate all studies and materials relevant to the research topic. This can induce serious biases of various kinds, ultimately compromising the conclusion and generalizability of the given review and meta-analysis (Reed & Baxter, 2009). Therefore, *high-quality literature searches* and *screenings of the found literature* are critical to mitigate this type of bias in systematic reviewing and meta-analysis. In the following subsections, I will describe how we strived to ensure these two parts of our review, presented in Chapter II.

Search procedures

Search string

Since we aimed at conducting a systematic review across several social science disciplines and different types of interventions, we spent much time thoroughly developing a search string that equally covered all pertinent conceptions and terms differently used across the literature. To this end, we scrutinized both seminal articles across co-teaching and teacher assistant literature and different disciplines of social science (e.g., education and political science) to pick relevant keywords for our search string. Thereafter, we consulted the local librarian to optimize the relevance of the search string. Finally, we pilot tested the search string to optimize the *precision* of the search,

i.e., to ensure a proper dimension between *the records of interest* and *all records retrieved from the search* (see Figure 5.3. in Brunton et al., 2017; Kugley et al., 2016).

Relevant databases

Selecting all relevant bibliometric databases pertinent for the literature search is critical to ensure adequate subject coverage across the social science literature and, thereby, to ensure that the literature searches yield the most relevant records (Reed & Baxter, 2009). To this end, we first searched a range of systematic review databases to locate previous reviews concerning collaborative models of instruction and thus to find out how our review could contribute to the topic area. Second, we searched bibliometric databases in education and the social sciences that are commonly used in reviews in education (Dietrichson, Bøg, Eiberg, Filges, & Jørgensen, 2016; Filges et al., 2015; Polanin, Espelage, & Grotper, 2018), and that are recommended by the Campbell Collaboration (Kugley et al., 2016). It was important for us to ensure that we also screened gray literature²⁶ in our search to avoid inducing publication biases²⁷ into our results (Rothstein et al., 2005). Therefore, the ProQuest Dissertation and Thesis database and Google Scholar (Haddaway et al., 2015) played a key role in our literature search. Although our Google Scholar search is not reproducible, we applied Google Scholar based on the philosophy that increasing the number of studies included in the review and meta-analysis was of higher priority than reproducibility.

Citation tracking

As a further technique to identify all relevant literature, we conducted *citation tracking/snowballing* from all previous literature reviews (see Figure S1 in the Supplementary Material in Chapter II) and journal articles regarding the effects of collaborative models of instruction on student learning included in the meta-analysis.

²⁶ i.e., literature that has not been published in peer reviewed scientific journals, such as dissertations and conference papers (White, 2009, p. 61).

²⁷ “Publication bias is the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies” (Rothstein et al., 2005, p. 1). The most popular example is when researchers conduct selective reporting, depending on statistical significant results.

Author solicitation

Although it was initially planned that we should conduct comprehensive author and expert solicitations, we chose not to spend time on this technique to locate relevant records. The main reason for not doing so was to save time which was needed to complete the present dissertation on time, but also, in part, because previous research showed that only 12% of primary authors replied to solicitations, of which only 0.5% provided pertinent information (Polanin et al., 2020).

Screening procedures

Yet another critical source to avoid biases in systematic reviews and meta-analysis is to assure that the abstract and full-text screenings of the literature identify all relevant studies. For this purpose, we double-screened all located abstracts and all relevant full-text literature independently. Previous research has shown that the use of a second reviewer increases the number of eligible studies located during the abstract screening by approximately 6-10% and a further 6-10% during the full-text screening (Stoll et al., 2019). Moreover, we also used this procedure to avoid being subject to cognitive algebra, as described in Section 2 of this overview article. Importantly, no studies were excluded by a single author. We tracked all of our screening in Covidence to ensure full transparency of inclusion and exclusion reasons.

Challenge 2: Extracting all relevant and reliable information

Although all abstract and full-text screenings were double-coded, we were not able to double-code all parts of our review due to resource limitations. Therefore, most of the *data extraction*, *effect size calculations*, and *risk of bias assessments* were conducted solely by me. This could potentially have induced a degree of bias in the review, for example, because I can have overlooked relevant information, I can have made computational errors or coding errors, or I can have made wrong decisions during the risk of bias assessments. Yet, I/we introduced a number of quality assurance mechanisms to alleviate these issues. I will briefly describe these procedures in the following subsection.

Quality assurance procedures

Data extraction

To mitigate potential biases of single-coder data extraction, all studies received data extraction twice. Although work-intensive, it was all-important to ensure the accuracy of the data extraction because flawed extractions might induce serious bias in our results. Therefore, it was also important for me to ensure that all quotes relevant to the data extraction were thoroughly documented in the data extraction scheme. Concretely, I documented the page numbers to warrant transparency of my extraction. As previously mentioned, all data extraction is, furthermore, open-sourced so that other researchers can scrutinize and critically investigate my extractions for potential errors.

Effect size calculation

Since effect size estimates constitute the backbone of meta-analyses, it was decisive for me to ensure reliable and transparent effect size computations. It is common to find meta-analyses in education in which all effect sizes and their standard errors are calculated as if they are coming from simple research designs only (see Equations (4.18) and (4.20) in Borenstein et al., 2009, pp. 26–27). This can lead to a number of deficits when applied to standardized mean differences estimated from more advanced research designs, such as repeated measures designs or clustered designed studies (Pustejovsky, 2016). For the former sets of designs, using simple effect size calculation formulas yield overly conservative standard errors, and for the latter, it yields too small standard errors. Therefore, it was important for us both to ensure that all effect size calculations were tailored to the specific study design and that all effect sizes represent the same unit of analysis (Hedges, 2007; Taylor et al., 2021).²⁸ As a consequence, effect size calculation becomes more complex, which, in turn, increases the likelihood of either making computational errors or coding errors. To ensure that cluster design adjustments were rigorously conducted, I built a function that could handle this correction. This means that all potential *formula errors* pertaining to clustered bias adjustments can be detected exclusively from this function. Furthermore, Bethany H. Bhat conducted quality tests for the 12 studies that required the most complex effect size calculations and for the coding that led to the final analyses (Hofner et al., 2016). This procedure is recommended by the Campbell Collaboration (2019) in cases where the full use of double-coders is not

²⁸ See Supplementary Section S1 for a detailed explanation of the effect size calculation procedure.

possible. Independently of these quality checks, the single-coder extraction and coding might have prompted that I have missed some relevant information for the effect size calculation, e.g., information relevant to estimating pre-posttest correlations. However, to reduce the severity of this bias, all studies excluded due to lacking opportunities for effect size calculation were made by two reviewers. Finally, all effect size calculations are open-sourced so that other researchers can further inspect these for potential errors.

RoB assessment: Mitigating the garbage in, garbage out/no causes in, no causes out critique

Besides finding all relevant studies and information within eligible studies, it was pivotal for us to ensure the validity and reliability of the information going into the meta-analysis to avoid that we do not induce unnecessary biases by including various types of literature and research (Egger et al., 2003). Meta-analyses are often accused of being subject to the *garbage in, garbage out* critique (Borenstein et al., 2009), to which it is argued that including many low-quality studies that are not able to reliably detect causal effects induce fundamental errors in the meta-analysis and thus obstruct reliable inference. For that reason, this critique is also described as the “no causes in, no causes out” critique (Cartwright & Hardie, 2012, p. 38). To accommodate this issue, I/we conducted comprehensive risk of bias (RoB) assessments (Higgins et al., 2019) for all treatment and control studies (i.e., 128 studies) concerning the effects of collaborative models of instruction. For non-randomized studies, we used the ROBINS-I tool (Sterne et al., 2016) to exclude those studies that were considered to be of critical risk of bias. Yet, to avoid that no study was erroneously omitted, all exclusions were double-checked by two reviewers. However, all RoB assessments for included studies were solely conducted by me. Consequently, this might potentially have induced some degree of error. However, all RoB assessments are open-sourced so that others can critically assess this issue.

To recapitulate, and as previously mentioned, no studies were excluded without agreement between at least two reviewers. Therefore, the single-coder and -rater procedures used throughout the review should mainly have had an impact on internal errors of the included studies. It can, for example, be the case that I have judged a study to be of moderate risk of bias when it, in fact, is serious or that I did not find all relevant information, wrongly producing a missing value on a given variable/characteristic, etc.

Challenge 3: Ensuring accurate statistical estimation in our meta-analysis

As also shown in Chapter III of the thesis, a key source of bias in systematic reviews and meta-analyses relates to the accuracy of the methods used to estimate effect sizes and fit meta-analytical models. As discussed and shown throughout this introductory chapter (cf. Section 2, in particular), we have strived to use state-of-the-art methods for effect size calculation and modeling of dependent effect sizes to reduce this bias. However, three issues could still occur, potentially compromising the accuracy of the results of the meta-analysis that are independent of the performance of the statistical methods used. *First*, since meta-analysis bears on many small decisions made by the reviewers (Moreau & Gamble, 2020), its results can depend on these idiosyncratic decisions. Therefore, to challenge and investigate the impact of our review decisions, we conducted a range of sensitivity analyses in which we changed the effect size calculation assumptions and applied inclusion criteria (see Figures 6-7 and Supplementary Figure S13 in Chapter II). *Second*, although we strived to guard against publication bias via our search for gray literature, selective reporting might still appear in the included set of studies (Pigott et al., 2013). If small sample studies systematically report larger effects, for example. This could potentially have induced an upward bias in our results. To investigate and test for selective reporting and/or small-study effects,²⁹ we conducted three publication bias tests that are suitable in the presence of dependent effect sizes, as suggested by Rodgers & Pustejovsky (Rodgers & Pustejovsky, 2021). All these tests were based on either modified sampling variance components or transformed effect size estimates to remove the artificial correlation between the standardized mean differences and their standard errors (Pustejovsky & Rodgers, 2019). *Third*, missing values in moderator variables in meta-analysis is a prevailing issue in almost all meta-analyses (Pigott, 2019) since studies rarely report all relevant information required by the reviewers. This potentially compromises the credibility of subgroup and meta-regression analyses. To reduce this potential bias, we applied multiple imputation techniques (Rubin, 1987; Van Buuren, 2018), which has been shown in most case to be reliable when less than 50 percent of the relevant information is missing in meta-analysis data (Diaz, 2020; Schauer et al., 2021). Yet, it cannot remove all bias, but it might be a better solution than its alternative and might still produce useful information in terms of estimating/indicating potential causal signs (Cook et al., 1992).

²⁹ i.e., "the tendency for the smaller studies in meta-analysis to show larger treatment effects" (Sterne et al., 2005, p. 75).

Challenge 4: Developing and evaluating statistical methods for meta-analysis of dependent effect sizes

During my/our endeavor to conduct a state-of-the-art meta-analysis, I/we encountered the methodological boundaries of meta-analysis of dependent effect sizes when trying to conduct power analyses for the statistical models applied in our review. My first response to this challenge was to approximate power by applying Monte Carlo Simulation³⁰ based on pilot data, as similarly suggested by Morris et al. (2019) and Pustejovsky (2019a). From this basis, I began approximating power and developing *traffic light power plots* (see Figure 5 in Chapter IV) for the correlated-effects (CE) models (Hedges et al., 2010b; Tipton, 2015; Tipton & Pustejovsky, 2015), using relevant pilot data from a previous meta-analysis about the effects of co-teaching on student achievement (Khoury, 2014, pp. 74–76).³¹ However, using simulation to generate power approximations is a resource-intensive method that might take weeks to conduct, even with access to very powerful computers, strongly restricting the general use of this method. I, therefore, consulted James E. Pustejovsky and Terri D. Pigott to investigate the viability of using simulation to approximate power for meta-analysis of dependent effect sizes. From these discussions, it was clear that a simpler method could be developed, and we began to work on the second article of the thesis. Hereto, Pustejovsky developed and provided the new approximation formulas, and we then instead used a part of my original simulation study to evaluate the performance of these new approximations for the CE-RVE and CHE-RVE models. We primarily opted to evaluate the approximation via a simulation study because these studies are considered to be “an invaluable tool for statistical research, particularly for the evaluation of new methods and for the comparison of alternative methods.” (Morris et al., 2019, p. 2074). Besides statistically evaluating the newly developed power approximation formulas, we also wanted to conduct the simulation study because it was at that time yet unknown how original power approximation, developed by Hedges & Pigott (2001), performed in terms of predicting power for models handling dependent effect sizes.

³⁰ Defined as, “computer experiments that involve creating data by pseudo-random sampling from known probability distributions” (Morris et al., 2019, p. 2074).

³¹ Since we changed our models during the conduct of our review, these power analyses were omitted from the final material.

Performance assessment

To adequately evaluate and test the performance of the new and old power approximations methods, we used the average rejecting rate³² for each set of replication as the main performance criterion (Joshi & Pustejovsky, 2020; Morris et al., 2019) to be compared to the power approximated from the formulas. To explain this procedure, let K be the number of iterations for each condition of design parameters³³ and p_k be the p value of simulation/iteration k , for $k = 1, \dots, K$. Then the definition of the rejection rate for a specific level- α is given by (Joshi, 2021)

$$\rho_\alpha = \Pr(p_k) < \alpha$$

The rejection rate and power for each set of simulated design factors are given by

$$r_\alpha = \frac{1}{K} \sum_{k=1}^K I(p_k < \alpha)$$

Furthermore, we calculated the Monte Carlo Standard Errors (MCSE) to encapsulate the uncertainty in the estimation of the rejection rate. These are given by (Joshi & Pustejovsky, 2021; Morris et al., 2019)

$$r_\alpha MCSE = \sqrt{r_\alpha(1 - r_\alpha)/K}$$

We calculated the rejection rate and power for $\alpha = .01, .05$, and $.1$ (we only concentrated on the conventional level- α , i.e., $\alpha = .05$), and for each set of conditions/design factors, $K = 4000$.

³² Defined as: “the rejection rate of a hypothesis test captures the proportion of times the p-value is below a specified α level—that is, the proportion of times we reject the null hypothesis. When the specified effect size is zero, we can examine Type 1 error rates and when the magnitude of the effect is greater than zero, we can examine power” (Joshi & Pustejovsky, 2021).

³³ We simulated 768 unique conditions (see all design factors in Table 2 in Chapter III). Furthermore, we applied a data-generating process in which the true error structure followed the correlated-hierarchical effects (CHE) working model from Equations (2).

Using graphs to evaluate the performance

Graphical displays were the main method used to investigate patterns for the Type I error rates and power for meta-analysis of dependent effect sizes (see Figures 2-4 in Chapter III), mainly because “[the] primary advantage of graphical displays of performance is that it is easier to quickly spot patterns (...) but also because “it becomes possible to present raw data estimates (...) as well as performance results summarizing them” (Morris et al., 2019, p. 2090). By combining facet grid plots from the *ggplot2* R package (Wickham, 2016) with different colors, shapes, and line types for different design factors, we were able to illustrate patterns that would otherwise have been difficult to tabulate or estimate (Pustejovsky, 2017).

Challenge 5: Making complex methods accessible to applied reviewers

A major barrier to the general applicability of the newly developed power approximation formulas is that they bear on a range of assumptions of different complexity. Hereto, no clear guidelines were developed in the second article of the thesis, potentially restricting the use of these methods even further. It was, therefore, a major challenge of the thesis to make these rather complex methods accessible to a broader audience of researchers. Therefore I developed the *POMADE* R package to make the implementation simple for other researchers. The advantage of the package solution is that it can easily be shared and provides simplified functions for the power approximations and for plotting the power estimates across different assumptions put forward by the reviewers. In this regard, I spent quite some time innovating the *traffic light power plot* developed in the third article of the thesis so that applied reviewers can conduct power analyses across a range of different assumptions about their model while simultaneously illustrating the model and conditions they most likely expect to find. As with the simulations study, the combined use of facet grid plots and different colors, shapes, and line types for different design factors play a major role in encompassing both aims of the traffic light power plot (see Figure 5 in Chapter IV).

However, it should be noticed that still more work is needed in terms of making power approximation more accessible to reviewers since not all researchers use R. Future directions could profitably focus on the development of a Shiny application (Wickham, 2021), allowing reviewers to have a point-and-click solution.

7. Summary and Discussion of Findings

In this final section, I will cover the main scientific contributions, discuss the limitations, and point to future directions related to each of the three enclosed articles of the thesis. Since the three chapters of the thesis all represent stand-alone articles that provide *main* contributions to different areas of social science, i.e., *education*, *statistics*, as well as *guideline* and *software developments*, these expositions are outlined separately for each article. However, all articles share the common aim of overcoming and improving the conduct of systematic reviewing and statistical meta-analysis in education and beyond with various means that are described below.

Article 1 (Chapter II)

The first article of the thesis presented in Chapter II represents a large-scale systematic review and meta-analysis of the effects of collaborative models of instruction on student achievement. The article was motivated by the curiosity to challenge, in part, the extensive use of narrative syntheses in previous reviews regarding co-teaching (see Table S1 in Chapter II) and, in part, the assumption that the research base regarding the connection between student achievement and co-teaching should be scarce. Hereto, the article makes several different contributions of both theoretical and methodological concerns. The first contribution of the article is to show that many more studies are available than previously anticipated. Specifically, we found 128 treatment and control group designed studies, of which 52 were excluded due to critical risk of biases according to the Cochrane risk of bias tools. In particular, we found more studies within all historical periods previously reviewed. In total, we meta-analyzed 76 studies, including 96 independent samples of students and 280 short-term effect sizes.³⁴ Unlike previous meta-analyses of co-teaching, 86% of the included effect sizes in our meta-analysis were either covariate or pretest-adjusted, significantly reducing the potential bias from confounding from included non-randomized studies (Morris, 2008). Next, the review adds contributions to the understanding of the mean effect and the differential effects of collaborative models of instruction across moderators of either theoretical or methodological concern. We found a moderate, positive, and statistically significant mean effect of $\bar{g} = 0.11$, 95% CI[0.035, 0.184] of collaborative instruction compared to single-taught controls. Next, we showed that the effects of collaborative models of instruction are generally robust across most subgroup

³⁴ i.e., effect sizes based on outcomes measured less than three months after the end of the intervention.

analyses. Of specific theoretical interest is our finding that the effect does not hinge on the two-teacher composition, starkly contradicting assumptions made in the co-teaching literature, asserting that collaborative teaching only works under narrow conditions involving formally educated special education teachers³⁵ (Cook & Friend, 1995; Friend, 2008). In this regard, we argue that the scalability of collaborative models of instruction might be easier than often assumed since non-formal teacher-educated assistants are probably easier to recruit and will, in most cases, induce lower wage costs. Furthermore, we also showed that the effect does not vary across factors that were considered in the literature as critical pre-conditions for the effectiveness of co-teaching.

With regard to methods, the article aimed to improve on prior meta-analyses of collaborative models of instruction by properly accounting for various dependencies among effect sizes coming from studies reporting multiple eligible results. To this end, we used CHE-RVE models (Pustejovsky & Tipton, 2021) with the objective of underpinning the trust of the detected effects. Moreover, a side-effect of accounting for dependent effect sizes was that it maximized the use of all relevant information retrievable from the included studies compared to methodological approaches including one effect size per study only, as done in two out of the three previous meta-analyses of collaborative models of instruction (i.e., Murawski & Swanson, 2001; Willett et al., 1983). Finally, all parts of the review have been open sourced in order to overcome the previous critique directed at meta-analysis for being intransparent (Maassen et al., 2020).

Limitations

Although we aimed to conduct a comprehensive systematic review, including a state-of-the-art meta-analysis, the article has several limitations. The most obvious limitations of the review are that it only studies the effects of student achievement and that it is unable to test a range of moderator factors anticipated in the co-literature to be all-important for the effectiveness of co-teaching. Moreover, many of the included studies might have low internal validity³⁶ because the implementation of the intervention was often poorly documented and because it was uncertain how confounding factors were controlled—however, most studies controlled for baseline differences

³⁵ Such as speech-language clinicians, reading specialists, bilingual teachers, or occupational therapists.

³⁶ i.e., “[t]he validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables *were* manipulated or measured.” (Cook et al., 2002, p. 38)

between the treatment and control groups, which to some degree reduces this issue. In addition, most studies were conducted in the U.S., with the rest representing educational systems of high-income countries only (The World Bank, 2022). This potentially restricts the generalizability of our findings to other school contexts, such as school systems in low and middle-income countries.

The article also has some methodological limitations. Some parts of the review were single-coded or -rated, potentially inducing some degree of bias in the information going into the meta-analysis. However, we strived to accommodate these issues by applying a range of quality checks (as described in the above Methodology Section), for example, by conducting quality tests of the accuracy of the most complex effect size calculations. It is also important to note that we ensured that the single-coded and -rater procedures only could have an impact on internal errors. In other words, all study exclusions were always done with an agreement between two authors to ensure that we did not exclude relevant studies. Finally, the publication bias tests that we utilized all have limitations, which means that they are either too conservative or liberal in terms of controlling the nominal Type I error rate (Pustejovsky & Rodgers, 2019; Rodgers & Pustejovsky, 2021).

Future direction for reviews and meta-analysis of collaborative models of instruction

While we find a robust moderate effect of collaborative models of instruction on student achievement, future meta-analyses should concentrate on investigating the effect of collaborative instruction on other outcomes, such as student well-being, and social and behavioral measures, since student achievement are clearly not the only educational reason for introducing collaborative models of instruction.

Next, it will be critical to keeping updating the review (Campbell Collaboration, 2019; Elliott et al., 2021). Since we finalized our last literature searches in June 2020, new eligible studies have already surfaced that need to be included in future reviews (see, for example, Hemelt et al., 2021; Jones & Winters, 2022). We expect to update the review within a five-year period, as suggested by the Campbell Collaboration (2019).

Our review further points to some vital future directions for primary research regarding the effects of collaborative instruction. Certainly, more experimental studies are needed in the co-teaching literature. Currently, it is primarily based on quasi-experimental and observational studies with a potentially lower internal validity. Hereto, it is pivotal to investigate the differential effects

and contexts that might have a moderating impact on the efficacy of collaborative models of instruction (Bryan, Tipton, & Yeager, 2021; Hedges, 2018; Tipton & Hedges, 2017). Particularly, it is critical to gain a better understanding of the differential effect between general and special education students since the evidence provided in our review was ambiguous. Finally, more focus should also be placed on investigating the long-term effects, and the cost-benefit of collaborative models of instruction since this type of research is generally absent in the literature.

Article 2 (Chapter III)

The chief contribution of the second article of the thesis presented in Chapter III is that it introduces new power approximations formulas for tests of the mean effect size from the most common models to handle dependent effect sizes, i.e., the CE model, CHE models, and MLMA models. These were developed from the need I experienced during my research on the first article of the thesis. However, this article provides several further contributions. Besides developing new methods for power approximation in meta-analysis, it also evaluated the performance of the new and more complex formulas via a comprehensive simulation study. This is a procedure that has not previously been used for assessing the performance of power approximation in meta-analysis (Hedges & Pigott, 2001, 2004). In effect, we show that the original power approximation based on the assumption of independent effect sizes performs inadequately to predict power for models using study-mean effect sizes, although these are independent. Most importantly, we show that the new-developed approximation formulas can close to exactly predict the power of the CE, CHE, and MLMA models when based on reliable pilot data and when the models are correctly specified. In this regard, we further show that making power approximations based on balanced assumptions (i.e., that the average sampling variance and the number of effect sizes per study are constant across studies), in the presence of substantial dependencies among effect sizes, overestimates power by 10-20%. Moreover, the article compared the nominal Type I error rates across the eight most common models for handling dependencies among effect sizes, showing that models involving RVE (i.e., CE-RVE, MLMA-RVE, and CHE-RVE models) perform most adequately, independently of the number of included studies. Lastly, we compared the relative power between the models using RVE, showing minor power gains for the CHE-RVE model when the true data generating model follows the correlated-hierarchical effects structure. However, these are only small gains when the true mean effect size is small or moderate, and the main takeaway from these results is that meta-

analysts should routinely guard against misspecification via RVE to ensure the accuracy of meta-analytical results, backing up previous simulation study findings (Fernández-Castilla et al., 2020).

Limitations

The main limitation of the new approximation functions is that they are complex, making a clear barrier to the applicability of the methods. This was the main reason for developing the third article of the thesis; see the exposition below. Furthermore, our simulation study does have a number of limitations. Although the approximation formulas perform adequately when based on pilot data, this can also induce bias in the approximations if the utilized pilot data is not representative of the target population of studies. Furthermore, we only investigated the performance of the approximations from one data generating mechanism, i.e., the CHE structure, and we did not investigate how the approximation functions perform when they are misspecified, such as assuming $\rho = 0.8$, when the true ρ is 0.2. Finally, the applicability of the approximation formulas is restricted by their main focus on standard mean differences (Hedges, 1981)—though these can be used for Fisher’s z-transformed correlation coefficients, as well.

Future directions for approximation formulas for meta-analysis of dependent effect sizes

The second article of the thesis yields at least five suggestions for future research for *a priori* approximation in meta-analysis. *First*, future research should focus on power approximations for other measures than standardized mean differences such as log odds ratios or relative risk ratios. *Second*, it would be beneficial to investigate how to approximate power for subgroup analyses in meta-analysis of dependent effect sizes, as Hedges & Pigott (2004) did for the original independent power approximation formulas. *Third*, in a similar vein, more knowledge is needed to understand how the original power approximation formulas empirically perform when used for predicting truly independent effect sizes, i.e., when all studies yield one effect size only. *Forth*, as indicated above, it would be interesting to understand how the newly developed formulas perform when they are mis-specified. *Fifth* and finally, since power analysis is based on arbitrary cutpoints for the power and significance levels alike, it might compel researchers to make binary interpretations of the effectiveness of given interventions. Thus, future research should revolve around developing precision approximation, i.e., an analysis that aims to approximate the number of studies needed to obtain a certain width of the confidence interval with a given probability (Rothman &

Greenland, 2018). Currently, this type of analysis is entirely absent from the field of meta-analysis and could potentially overcome the narrow focus on statistical significance (Lakens et al., 2018).

Article 3 (Chapter IV)

While the second article concentrates on the technical development and quality assurance of power approximations for meta-analysis of dependent effect sizes, it did not tackle the practical challenges that applied reviewers potentially will meet for obtaining the relevant quantities required to approximate power reliably. Therefore, the third article of the thesis aims to make three contributions in this regard. *First*, it contributes with common guidelines for how and where applied reviewers can find the relevant information needed to conduct reliable power approximations for meta-analysis of dependent effect sizes. *Second*, it introduces the *POMADE* R package, whose main contribution is to provide functions for conducting and plotting *power approximations* as well as *the minimum detectable effect size* (MDES) and *the number of studies needed to find a given effect size considered to be the smallest of practical relevance*, with preset levels of statistical power and significance as well as with and across prespecified data and model conditions. *Third*, it introduces what we coin the *traffic light power plot* that enables applied reviewers to conduct and plot power across a range of possible data and models scenarios while at the same time allowing reviewers to clarify which assumptions and design factors they expect to find (see Figure 5 in Chapter IV). A notable advantage of the *POMADE* plot functions is that they can generate information in five to ten minutes (using only one core on an average computer), which would otherwise approximately take more than a week to obtain via simulation on a computer/server with 64 cores and 376 RAM. Hopefully, the speed of the new methods may strongly increase their applicability.

Limitations

Although the article develops simplified functions to conduct power analyses for dependent effect sizes, the barriers to the use of these are that the functions remain rather complex and depend on researchers having access to relevant pilot data. Therefore, thorough documentation will be a key feature of the final package. Furthermore, the flexibility of these power approximations might come with the potential risk that researchers choose their model based on these analyses and not from substantial information related to the research topic under review. For example, it could be the case that reviewers would select the assumed sample correlation (ρ) based on the value yielding

more power and not on the value most realistic in practice. However, it is important to remember that the power approximations are based on the assumption that the model is correctly specified. Therefore, reviewers might potentially lose power if they mis-specify their model, for instance, by setting $\rho = 0$, when it, in fact, is $\rho = .8$ (Pustejovsky & Tipton, 2021).

Future directions for guidelines and software developments for power analysis for meta-analysis of dependent effect sizes

The fact that the new *POMADE* package can only be used in the statistical learning environment R restricts its use. Therefore, future package improvements should focus on making the package functions available to researchers not applying R. For example, it could be an idea to introduce a Shiny application (Wickham, 2021) with simple point-and-click solutions. Furthermore, it could also be profitable for the community if packages were developed for other statistical programs as well, such as Stata or jamovi.

Finally, the guidelines put forward in this article are most easily implemented when meta-analysts have access to reliable and relevant pilot data. This makes future demands for the meta-analysis community to embrace and implement open science and open data procedures if prospective power analyses of meta-analyses should become common practice in systematic reviews.

8. References

- Aarhus University. (2010). *Rules for the PhD programme at the Graduate School, Arts*.
https://phd.arts.au.dk/fileadmin/phd.arts.au.dk/AR/Generelle_retningslinjer_UK_1-11-2012.pdf
- Achilles, C., 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>
- Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82(4), 436–476. <https://doi.org/10.3102/0034654312458162>
- Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311, 485. <https://doi.org/10.1136/bmj.311.7003.485>
- 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>
- 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>
- Bethhäuser, B. A., Bach-Mortensen, A., & Engzell, P. (2022). *A systematic review and meta-analysis of the impact of the COVID-19 pandemic on learning*. SocArXiv.
<https://doi.org/10.31235/osf.io/g2wuy>
- Biesta, G. (2007). Why “what works” won’t work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1–22.
<https://doi.org/10.1111/j.1741-5446.2006.00241.x>
- 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. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Inc.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis* (1st ed.). John Wiley & Sons.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2), 97–111. <https://doi.org/10.1002/jrsm.12>
- Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1), 5–18. <https://doi.org/10.1002/jrsm.1230>
- Bredow, C. A., Roehling, P. V, Knorp, A. J., & Sweet, A. M. (2021). To flip or not to flip? A meta-analysis of the efficacy of flipped learning in higher education. *Review of Educational Research*, 91(6), 878–918. <https://doi.org/10.3102/00346543211019122>
- Brunton, J., Stansfield, C., Caird, J., & Thomas, J. (2017). Finding relevant studies. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 93–122). Sage.
- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 5(1), 980–989. <https://doi.org/10.1038/s41562-021-01143-3>
- Buch-Hansen, H., & Nielsen, P. (2005). *Kritisk realisme*. Roskilde Universitetsforlag.
- Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. <https://doi.org/10.3368/jhr.50.2.317>
- Campbell Collaboration. (2019). *Campbell systematic reviews: Policies and guidelines. 1.4*. <https://onlinelibrary.wiley.com/pb-assets/assets/18911803/Campbell Policies and Guidelines v4-1559660867160.pdf>
- Campbell, M., Katikireddi, S. V., Sowden, A., McKenzie, J. E., & Thomson, H. (2018). Improving Conduct and Reporting of Narrative Synthesis of Quantitative Data (ICONS-Quant): Protocol for a mixed methods study to develop a reporting guideline. *BMJ Open*, 8(2), 1–5. <https://doi.org/10.1136/bmjopen-2017-020064>
- Campbell, M., Katikireddi, S. V., Sowden, A., & Thomson, H. (2019). Lack of transparency in reporting narrative synthesis of quantitative data: A methodological assessment of systematic reviews. *Journal of Clinical Epidemiology*, 105, 1–9.

- <https://doi.org/10.1016/j.jclinepi.2018.08.019>
- Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, 2(1), 11–20.
<https://doi.org/10.1017/s1745855207005029>
- Cartwright, N. (2011). Predicting “it will work for us”: (Way) beyond statistics. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 750–768). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199574131.003.0035>
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199841608.001.0001>
- Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *J Eval Clin Pract*, 16(2), 260–266. <https://doi.org/10.1111/j.1365-2753.2010.01382.x>
- Chakravartty, A. (2011). Scientific realism. *Stanford Encyclopedia of Philosophy*.
<https://plato.stanford.edu/entries/scientific-realism/>
- 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>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.
<https://doi.org/10.4324/9780203771587>
- 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>
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Cengage Learning, Inc.
- Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. Russell Sage Foundation. <https://www.jstor.org/stable/10.7758/9781610441339>
- Cooper, H. (2015). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed., Vol. 2). Sage.
- Cooper, H., & Hedges, L. V. (2019). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 3–15). Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and*

- meta-analysis* (3rd ed.). Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>
- Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14(2), 165–176. <https://doi.org/10.1037/a0015565>
- Dafolo. (2019). *Marilyn Friend om co-teaching*. <https://www.youtube.com/watch?v=4UUdXUJQ4PU>
- DCU. (2013). *Konceptnotat juni 2013*. https://edu.au.dk/fileadmin/edu/Udgivelser/Clearinghouse/Konceptnotat_Clearinghouse_2013.pdf
- Deeks, J. J., Higgins, J. P. T., Altman, D. G., & Group, C. S. M. (2019). Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. S. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (2nd ed., pp. 241–284). Wiley Online Library. <https://doi.org/10.1002/9781119536604>
- Diaz, K. (2020). *Multiple imputation for handling missing data of covariates in meta-regression*. Columbia University.
- Dietrichson, J., Bøg, M., Eiberg, M., Filges, T., & Jørgensen, A.-M. K. (2016). Protocol for a systematic review: Targeted School-Based Interventions for Improving Reading and Mathematics for Students With or At-Risk of Academic Difficulties in Grade K to 6: A Systematic Review. *Campbell Systematic Reviews*, 12(1), 1–60. <https://doi.org/10.1002/CL2.165>
- DPU. (2022). *Pædagogisk indblik*. <https://dpu.au.dk/viden/paedagogiskindblik>
- Dyssegaard, C. B., & Larsen, M. S. (2013). *Evidence on inclusion*. Danish Clearinghouse for Educational Research. https://edu.au.dk/fileadmin/edu/Udgivelser/Clearinghouse/Evidence_on_Inclusion.pdf
- Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. (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>
- Elliott, J., Lawrence, R., Minx, J. C., Oladapo, O. T., Ravaud, P., Tendal Jeppesen, B., Thomas, J., Turner, T., Vandvik, P. O., & Grimshaw, J. M. (2021). Decision makers need constantly

- updated evidence synthesis. *Nature*, 600(7889), 383–385. <https://doi.org/10.1038/d41586-021-03690-1>
- EPPI-Centre. (2010). *EPPI-Centre methods for conducting systematic reviews*.
<https://www.betterevaluation.org/sites/default/files/Methods.pdf>
- 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>
- Filges, T., Sonne-Schmidt, C. S., & Jørgensen, A. M. K. (2015). Protocol: Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, 11(1), 1–42. <https://doi.org/10.1002/CL2.148>
- 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>
- 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>
- Friese, M., & Frankenbach, J. (2020). p-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, 25(4), 456–471. <https://doi.org/10.1037/met0000246>
- Furenes, M. I., Kucirkova, N., & Bus, A. G. (2021). A comparison of children's reading on paper versus screen: A meta-analysis. *Review of Educational Research*, 91(4), 483–517. <https://doi.org/10.3102/0034654321998074>
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3–8. <https://doi.org/10.2307/1174772>
- Glass, G. V. (2000). *Meta-analysis at 25*. <https://www.gvglass.info/papers/meta25.html>
- Gleser, L., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). Russell Sage Foundation. <https://doi.org/10.7758/9781610441384>
- Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3), 399–412.

<https://www.jstor.org/stable/2680773>

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLOS ONE*, 10(9), e0138237. <https://doi.org/10.1371/journal.pone.0138237>

Hammersley, M. (1997). Educational research and teaching: A response to David Hargreaves' TTA lecture. *British Educational Research Journal*, 23(2), 141–161. <https://doi.org/10.1080/0141192970230203>

Hargreaves, D. H. (1996). Teaching as a research-based profession: Possibilities and prospects. *The Teacher Training Agency Annual Lecture 1996*, 1–12. [https://eppi.ioe.ac.uk/cms/Portals/0/PDF reviews and summaries/TTA Hargreaves lecture.pdf](https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/TTA%20Hargreaves%20lecture.pdf)

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands-on guide*. PROTECT Lab. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Hattie, J. (2009). *Visible learning – A synthesis of over 800 meta-analysis relating to achievement*. Routledge.

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. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, 2(3), 167–171. <https://doi.org/10.1111/j.1750-8606.2008.00060.x>

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, 36(3), 346–380. <https://doi.org/10.3102/1076998610376617>

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. (2019a). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 281–298). Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>
- Hedges, L. V. (2019b). The statistics of replication. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 15(S1), 3–14. <https://doi.org/10.1027/1614-2241/a000173>
- 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. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88(2), 359–369. <https://doi.org/10.1037/0033-2909.88.2.359>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(3), 203–217. <https://doi.org/10.1037/1082-989X.6.3.203>
- Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426–445. <https://doi.org/10.1037/1082-989X.9.4.426>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010a). Erratum: 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., Tipton, E., & Johnson, M. C. (2010b). 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>
- Hemelt, S. W., Ladd, H. F., & Clifton, C. R. (2021). Do teacher assistants improve student outcomes? Evidence from school funding cutbacks in North Carolina. *Educational Evaluation and Policy Analysis*, 43(2), 280–304. <https://doi.org/10.3102/0162373721990361>
- 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>
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177. <https://doi.org/10.1111/j.1750-8606.2008.00061.x>
- 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>
- Ioannidis, J. P. (2005). Differentiating biases from genuine heterogeneity: Distinguishing artifactual from substantive effects. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 287–302). Wiley Online Library. <https://doi.org/10.1002/0470870168>
- Ioannidis, J., Patsopoulos, N. A., & Rothstein, H. R. (2008). Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*, 336(7658), 1413–1415. <https://doi.org/10.1136/bmj.a117>
- Jones, N., & Winters, M. A. (2022). Are two teachers better than one? The effect of co-teaching on students with and without disabilities. *Journal of Human Resources*. <https://doi.org/10.3368/jhr.0420-10834R3>
- Joshi, M. (2021). *Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with small number of studies* [The University of Texas at Austin]. <https://repositories.lib.utexas.edu/handle/2152/86861>
- Joshi, M., & Pustejovsky, J. E. (2020). *simhelpers: Helper functions for simulation studies* (0.1.0). <https://cran.r-project.org/web/packages/simhelpers/index.html>
- Joshi, M., & Pustejovsky, J. E. (2021). *Simulation performance criteria and MCSE*. <https://cran.r-project.org/web/packages/simhelpers/vignettes/MCSE.html>
- 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*, 1–21. <https://doi.org/10.1002/jrsm.1554>
- Karseth, B., Sivesind, K., & Gita, S.-K. (2022). *Evidence and expertise in nordic education policy*. Springer. <https://link.springer.com/content/pdf/10.1007%2F978-3-030-91959-7.pdf>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

- 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]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/1817570306?accountid=14468> NS
- Kidron, Y., & Lindsay, J. (2014). *The effects of increased learning time on student academic and nonacademic outcomes: Findings from a meta-analytic review*. Regional Educational Laboratory Appalachia. <https://ies.ed.gov/ncee/rel/Products/Publication/3603>
- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry*. Princeton University Press.
- 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>
- Korsgaard, M. T. (2020). Exemplarity and education: Retuning educational research. *British Educational Research Journal*, 46(1), 1357–1370. <https://doi.org/10.1002/berj.3636>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- KSU. (2021). *Strategi for Kunnskapssenter for Utdanning (KSU) ved Universitetet i Stavanger*. University of Stavanger. https://www.uis.no/sites/default/files/2021-03/uis_kunnskapssenter.pdf
- Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.-M. K., Hammerstrøm, K., & Sathe, N. (2016). Searching for studies: A guide to information retrieval for Campbell. *Campbell Systematic Reviews*, 13(1), 1–73. <https://doi.org/10.4073/cm.2016.1>
- Kvernbekk, T. (2016). *Evidence-based practice in education. Functions of evidence and causal presuppositions*. Routledge. <https://doi.org/10.4324/9780203774830>
- LaFever, K. M. (2012). The effect of co-teaching on student achievement in ninth grade physical science classrooms [University of Missouri – St. Louis]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1697496661?accountid=14468> NS
- 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>

- Lakens, D., Adolphi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Hasselman, F., Ziano, I., & Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3), 168–171. <https://doi.org/10.1038/s41562-018-0311-x>
- Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, 10(1), 83–98. <https://doi.org/https://doi.org/10.1002/jrsm.1316>
- Lipsey, M. W. (2007). Unjustified inferences about meta-analysis. *Journal of Experimental Criminology*, 3(3), 271–279. <https://doi.org/10.1007/s11292-007-9037-x>
- 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*.
- Littell, J. H. (2008). Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children and Youth Services Review*, 30(11), 1299–1317. <https://doi.org/10.1016/j.childyouth.2008.04.001>
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>
- 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>
- McKenzie, J. E., & Brennan, S. E. (2019). Synthesizing and presenting findings using other methods. In J. P. T. Higgins, J. Thomas, J. Chandler, M. S. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (2nd ed., pp. 321–347). Wiley Online Library.
- Melendez-Torres, G. J., O'Mara-Eves, A., Thomas, J., Brunton, G., Caird, J., & Petticrew, M. (2017). Interpretive analysis of 85 systematic reviews suggests that narrative syntheses and meta-analyses are incommensurate in argumentation. *Research Synthesis Methods*, 8(1), 109–118. <https://doi.org/10.1002/jrsm.1231>
- Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate,

- W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6), 559–572.
<https://doi.org/10.1080/13645579.2016.1252189>
- 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>
- Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*.
<https://doi.org/http://dx.doi.org/10.1037/met0000351>
- 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>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.
<https://doi.org/10.1002/sim.8086>
- 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., & 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>
- Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences - PNAS*, 115(11), 2600–2606.
<https://doi.org/10.1073/pnas.1708274114>
- OECD. (2004). *National review on educational R&D: Examiners' report on Denmark*.

- <https://www.oecd.org/education/ceri/33888206.pdf>
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- Pearl, J., & Mackenzie, D. (2018). *The book of why - The new science of cause and effect*. Basic Books.
- Petticrew, M., & Roberts, H. (2008). *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.
- Pigott, T. D. (2012). *Advances in meta-analysis*. Springer.
- 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>
- Pigott, T. D., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods*, 3(4), 257–268. <https://doi.org/10.1002/jrsm.1051>
- 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
- Polanin, J. R., Espelage, D. L., & Grotmeter, J. (2018). *The consequences of school violence: A systematic review and meta-analysis review protocol*. <https://osf.io/6hak7/>
- Polanin, J. R., Espelage, D. L., Grotmeter, J. K., Valido, A., Ingram, K. M., Torgal, C., El Sheikh, A., & Robinson, L. E. (2020). Locating unregistered and unreported data for use in a social science systematic review and meta-analysis. *Systematic Reviews*, 9(1), 116. <https://doi.org/10.1186/s13643-020-01376-9>
- Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews*. <https://www.lancaster.ac.uk/media/lancaster-university/content->

- assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf
- Pustejovsky, J. E. (2016). *Alternative formulas for the standardized mean difference*.
<https://www.jepusto.com/alternative-formulas-for-the-smd/>
- Pustejovsky, J. E. (2017). *Designing Monte Carlo simulations in R*.
<https://jepusto.github.io/Designing-Simulations-in-R/>
- Pustejovsky, J. E. (2019a). *Simulating correlated standardized mean differences for meta-analysis*. <https://www.jepusto.com/simulating-correlated-smds/>
- Pustejovsky, J. E. (2019b). *Sometimes, aggregating effect sizes is fine*.
<https://www.jepusto.com/sometimes-aggregating-effect-sizes-is-fine/>
- Pustejovsky, J. E. (2020). *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, Vienna, Austria. <https://www.r-project.org/>
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 103(1), 111–120. <https://doi.org/10.1037/0033-2909.103.1.111>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed., Vol. 1). Sage.
- Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 73–101). Russell Sage Foundation. <https://doi.org/10.7758/9781610441384>
- Reiss, J. (2018). Against external validity. *Synthese*, 196(8), 3103–3121.
<https://doi.org/10.1007/s11229-018-1796-6>
- Riley, R. D., Stewart, L. A., & Tierney, J. F. (2021). *Individual participant data meta-analysis: A handbook for healthcare research*. Wiley Online Library.
<https://doi.org/10.1002/9781119333784>

- 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>
- Rosenbaum, P. R. (2017). *Observation and experiment*. Harvard University Press.
- Rothman, K. J., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology*, 29(5), 599–603. <https://doi.org/10.1097/EDE.0000000000000876>
- 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., Boston, MA. <https://www.rstudio.com/>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in sample surveys*. John Wiley.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110–114.
- Sayer, A. (1999). *Realism and social science*. Sage.
- 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*. <https://doi.org/10.1093/alcalc/agaa144>
- Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage. <https://doi.org/10.4135/9781483398105>
- 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>
- Senn, S. J. (2009). Overstating the evidence – double counting in meta-analysis and related problems. *BMC Medical Research Methodology*, 9(1), 10. <https://doi.org/10.1186/1471-2288-9-10>
- Shadish, W. R., & Lecy, J. D. (2015). The meta-analytic big bang. *Research Synthesis Methods*, 6(3), 246–264. <https://doi.org/https://doi.org/10.1002/jrsm.1132>
- Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, 70, 747–770. <https://doi.org/10.1146/annurev->

psych-010418-102803

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, 32(4), 450–466.

<https://doi.org/10.1080/02680939.2017.1280183>

Southwick, K. E. (1998). The effects of the class within a class collaborative/co-teaching model on the achievement of general education students in grades three, four and five [University of Kansas]. In *ProQuest Dissertations and Theses*.

<https://search.proquest.com/docview/304420847?accountid=14468> NS

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4), 497–507. <https://doi.org/10.1016/j.shpsc.2011.07.003>

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). Wiley Online Library.

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>

Stoll, C. R. T., Izadi, S., Fowler, S., Green, P., Suls, J., & Colditz, G. A. (2019). The value of a second reviewer for study selection in systematic reviews. *Research Synthesis Methods*, 10(4), 539–545. <https://doi.org/10.1002/jrsm.1369>

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>

The World Bank. (2022). *World Bank country and lending groups*.

<https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>

Thomas, J., O'Mara-Eves, A. J., Harden, A., & Newman, M. (2017). Synthesis methods for combining and configuring textual or mixed methods data. In D. Gough, S. Oliver, & J.

- Thomas (Eds.), *An introduction to systematic reviews*. Sage.
- 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., & Hedges, L. V. (2017). The role of the sample in estimating and explaining treatment effect heterogeneity. *Journal of Research on Educational Effectiveness*, 10(4), 903–906.
- 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>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, 10(2), 161–179. <https://doi.org/10.1002/jrsm.1338>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194. <https://doi.org/10.1002/jrsm.1339>
- 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>
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <https://doi.org/10.3102/1076998609346961>
- Valentine, J. C., Wilson, S. J., Rindskopf, D., Lau, T. S., Tanner-Smith, E. E., Yeide, M., LaSota, R., & Foster, L. (2017). Synthesizing evidence in public policy contexts: The challenge of synthesis when there are only a few studies. *Evaluation Review*, 41(1), 3–26. <https://doi.org/10.1177/0193841X16674421>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC press. <https://stefvanbuuren.name/fimd/>
- 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>
- 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>
- Vembye, M. H., & Jensen, H. S. (2022). Et kritisk blik på kausalitet og evidens i den danske evidensbevægelse og et muligt bud på en forbedring. Et meta-review af forskning fra Dansk Clearinghouse for Uddannelsesforskning. In L. Qvortrup & J. Christensen (Eds.), *Effektfuldhed og kausalitet i pædagogisk forskning og praksis*. Aarhus Universitetsforlag.
- Vembye, M. H., Pustejovsky, J. E., & Pigott, T. D. (2022). *Power approximations for overall average effects in meta-analysis with dependent effect sizes*. MetaArXiv. <https://doi.org/10.31222/osf.io/6tp9y>
- 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. (2022). *Likelihood ratio and wald-type tests for “rma” objects*. <https://wviechtb.github.io/metafor/reference/anova.rma.html>
- Wang, X., Welch, V., Li, M., Yao, L., Littell, J., Li, H., Yang, N., Wang, J., Shamseer, L., Chen, Y., Yang, K., & Grimshaw, J. M. (2021). The methodological and reporting characteristics of Campbell reviews: A systematic review. *Campbell Systematic Reviews*, 17(1), e1134. <https://doi.org/https://doi.org/10.1002/cl2.1134>
- White, H. (2022). Getting evidence into use: The experience of the Campbell Collaboration. *Campbell Systematic Reviews*, 18(1), e1226. <https://doi.org/https://doi.org/10.1002/cl2.1226>
- White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 51–71). Russell Sage Foundation. <https://doi.org/10.7758/9781610441384>
- Wickham, H. (2015). *R packages: Organize, test, document, and share your code*. O’Reilly Media, Inc. <https://r-pkgs.org/>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. <https://cran.r-project.org/web/packages/ggplot2/index.html>
- Wickham, H. (2021). *Mastering Shiny*. O’Reilly Media, Inc. <https://mastering-shiny.org/>
- 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>

Williams, R. (2012). *Using robust standard errors to combine multiple estimates with meta-analysis* [Loyola University Chicago].

https://ecommons.luc.edu/cgi/viewcontent.cgi?article=1404&context=luc_diss

WWC. (2020). *WWC procedures and standards handbook* (4.1). Institute of Education Sciences.

<https://ies.ed.gov/ncee/wwc/Handbooks>

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

WWC. (2022). *WWC procedures and standards handbook* (5.0). Institute of Education Sciences.

<https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-HandbookVer5.0AppIES-508.pdf>

Chapter II

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

Mikkel H. Vembye, Felix Weiss, & Bethany H. Bhat

Find latest version of this chapter at <https://osf.io/preprints/metaarxiv/mq5v7/>

Abstract

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 their effects on students' academic achievement and how these effects are moderated by theoretically and practically relevant factors. Although previous research and reviews assert that the evidence base is scarce, we found 128 treatment and control group designed 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 yielding 280 short-term effect sizes, of which 82% are pretest-adjusted. We found a moderate, positive, and statistically significant mean effect of $\bar{g} = 0.11$, 95% CI[0.035, 0.184] 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 for interventions with trained teachers and assistants without teacher education. This implies a potential for the expansion of the intervention at lower costs than often expected. Moreover, factors that are highlighted in the co-teaching literature as preconditions for the effectiveness of collaborative instruction did not explain much variation in effect sizes. Finally, we did not find any clear evidence for publication bias or small study effects. Notably, a large number of the studies that we could draw upon were non-randomized studies. Therefore, future research could profitably concentrate on conducting more rigorous experimental research, especially relevant co-teaching interventions.

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

Introduction¹

Collaborative models of instruction have been applied since the 1950s (Willett et al., 1983), but their popularity has increased over the last decades in many school systems throughout high-income countries (Andersen et al., 2018; Blatchford et al., 2011; Friend, 2008; Muijs & Reynolds, 2003). Such models comprise co-teaching between general and special education teachers, but also the use of teacher assistants and paraprofessionals. This development has been fueled by different legal acts and declarations (e.g., IDEA, 2022; NCLB, 2002; UNESCO, 1994) that warrant the right for all students to receive high-quality general education, regardless of differences and difficulties. Furthermore, collaborative models of instruction are popular since they are more flexible 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 can transcend what can be done alone by a single teacher, might also have contributed to their expansion (Bacharach et al., 2010; Friend, 2008). However, still 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. Furthermore, there exists a particular lack of research and reviews investigating the differential effect between various two-teacher instruction models, e.g., between co-teaching and teacher assistant interventions. In order to overcome this knowledge gap, this systematic review includes and concentrates on various versions of collaborative instruction interventions, which allows 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 this lacking understanding of the effects of collaborative models on instruction on student achievement (Cook, McDuffie-Landrum, Oshita, & Cook, 2017; Friend, 2008; Iacono et al., 2021; Murawski & Swanson, 2001; Reinhiller, 1996). Yet, we aim to challenge this view by demonstrating that there exists a large body of literature, including numerous studies with a design that is appropriate for drawing causal conclusions. Despite applying more restricted inclusion criteria than prior reviews, we found 128 studies published between 1984 and 2020. Based on these publications, we investigated the overall mean

¹ Find the pre-registered protocols and codes for reproducing all parts and analyses of this paper at <https://osf.io/fby7w/>.

effect of collaborative models of instruction on student achievement, but also how these effects varied across focal moderators that were highlighted in the methodological and theoretical literature as important factors to explain the differential effects of collaborative models of instruction.

Definition of Terms and Mechanisms for the Effect of the Intervention

The underlying definitions of this review follow the common use in the literature as we observed it 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 of instructional and/or behavioral interventions*. This definition encompasses all of the included compositions of two-teacher instruction that are specified further in this section. Importantly, it is neither restricted to certain types of two-teacher compositions/teacher actions nor to specific groups of students. This way, we aim to include personnel without formal teacher education, i.e., paraprofessionals, pedagogues, or parent volunteers.

Through our literature search, we identified three versions of collaborative models of instruction that fall under the overall definition, but which were largely separated from each other in the literature. These were studies regarding *co-teaching*, *teacher assistants/aides*, and *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 literature that we located refers to *co-teaching* interventions. We define co-teaching in line with Cook and Friend (1995, p. 2) as “*two or more professionals delivering substantive instruction to a diverse, or blended, group of students in a single physical space.*” The term “professionals” in this regard refers to the collaboration between a formally educated general education teacher and a formally educated special education teacher, such as a speech-language clinician, reading specialist, bilingual teacher, or occupational therapist. The theoretical foundation of the co-teaching literature often highlights that co-teaching is context-dependent and only works effectively under narrow conditions. For example, Friend (2008, p. 17) states that “[c]o-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". Hence, common planning time 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 are, in turn, assumed to be vital components for improving student learning by making full and complimentary use of the professional competencies of the general and the special educator, e.g., by combining the general teacher's in-depth knowledge of the curriculum and the specialized knowledge of the special education teacher about customizing the 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) since they do not take full advantage of the competencies of both educators. Hereto, it is emphasized that co-teaching should ideally be executed by using a variety of co-teaching models to work most effectively (see Cook & Friend, 1995, pp. 5–6 for an overview of all 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 keep the co-teachers engaged in doing effective co-teaching (Cook & Friend, 1995; Friend, 2008; Scruggs et al., 2007).

As a quite concrete guideline, it is occasionally suggested that co-teaching works best when provided to students for two 60-90 minutes sessions per week (Friend in Stanek, 2017). Hereto, it is also hypothesized that co-teaching only works properly when provided for more than a year since it is a heavy developmental type of program, which takes time to learn for the co-teachers (see Friend in Dafolo, 2019).

Besides the hypothesized propositions mentioned above, the co-teaching literature also expects a positive effect from the improved student-teacher ratio (Cook & Friend, 1995, pp. 3–4)—like other interventions such as class-size reductions (see Filges et al., 2018, and Supplementary Figure S28 for the causal theory behind the models). One hypothesis for 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. An-

other hypothesis is that students are given a more appropriate and differentiated/personalized instruction, which allows for a deeper presentation of the content as well as increased student engagement.

Teacher Assistants/Aides

Another set of collaborative models of instruction found in the literature search was *teacher assistants/aides* interventions. We define the teacher assistant(s) (TA) intervention as *an in-class collaboration between a general education teacher and adults/paraprofessional educators without a formal teacher education such as pedagogues, (voluntary) parents, etc.* (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 teacher education. Thereby, the teacher roles for this model are assumed to be clearer since the support personnel always takes 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 largest difference is that the TA literature assumes that shared instructional responsibility and the formal education of the second teacher is not a prerequisite for the effectiveness of the intervention.

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

- TAs can function as role models that show students that the content is valued by adults other than the teachers.
- TAs provide the opportunity for more adult interaction, which can scaffold student learning, provide more in-depth learning, and ensure that students are more active.
- TAs can facilitate that students spend more time focused on tasks by improving behavioral issues and thus increase students’ learning time.

- TAs can facilitate more concentrated whole-class activities by improving classroom management.
- TAs can increase the amount of immediate feedback and praise given to all students, boosting the students' confidence and motivation, working habits, and willingness to finish off tasks.

Team Teaching

Lastly, we found a set of studies including compositions of two-teacher instruction not falling into the categories presented above, i.e., two regular/general and formally educated in-class teachers (e.g., two math teachers). We refer to this category as *team teaching*², which we define as *two or more general education teachers sharing instructional and/or behavioral responsibilities of students in the same physical space*. Since this model is not widespread in the literature, a specific causal theory for this model of instruction is scantily developed. However, we assume that it works similarly to the two other models, i.e., that it can reduce disciplinary problems ensuring more instructional time as well as increasing student-teacher interaction allowing students to receive more personalized instruction. An advantage of this model over the other two, however, could be that students receive a broader and more in-depth content knowledge due to the potentially complementary knowledge of the two general teachers. A further benefit might be that it is less likely that one of the teachers is ascribed the assisting role due to a lack of content knowledge, which often seems to happen for the special education teacher (Scruggs et al., 2007). This might even be more pronounced in later grades when more advanced content knowledge is required.

Previous Reviews

The systematic reviews most closely related to our review are Murawski & Swanson (2001), Khoury (2014), and Willet et al. (1983). 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.

² This definition should not be confused with Cock & Friend's (1995) "team teaching" model.

MS01 investigated the effects of co-teaching on various outcomes measures from students with special needs, including academic achievement. They found a large³ mean effect size of 0.40. However, this must be seen against the background of a rather small sample of six studies, of which five were single-case pre-posttest studies, which is a less conservative design (Cheung & Slavin, 2016), meaning that it generally yields larger effects. MS01 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.

The only prior review that applied meta-regression was authored by Khoury (2014), and the review investigated the effect of co-teaching on academic achievement outcomes among students with special needs. It found partly a relatively large mean effect size of 0.28, and partly that the effect of co-teaching did not vary across school levels or subjects, and neither about study types or the type of comparison group. Finally, the review suggested that the effect was stronger the longer students received 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 control for the nominal Type I error rate (Becker, 2000; Hedges et al., 2010; Van den Noortgate et al., 2013).

As the only prior review, Willet et al. (1983) included all types of in-class two-teacher instruction studies conducted from 1950 to 1983 and found a small, moderate mean effect size of 0.06 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.

Both Khory and Willet et al. 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 mean effect size estimation.

³ When we interpret effect sizes throughout the paper, we use Kraft's (2020) empirical guidelines and benchmarks for interpreting effect sizes related to causal research on education interventions with standardized achievement outcomes.

Narrative reviews and syntheses

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

One review, authored by Scruggs et al. (2007), is purely based on qualitative research. From interviews with and observations of co-teachers, they found that special teachers frequently report that they are given subordinated assistant roles and that they think administrative support is pivotal to facilitate the relevant training and 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 this review.

Contribution of the Review

This review goes beyond previous review studies in various ways. *First*, we aimed to fill the long synthesis gap since 1983, when the first comprehensive review of collaborative models of instruction was conducted. *Second*, we apply more clear-cut inclusion criteria 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 among treatment and the 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 aimed at a more comprehensive review combining and testing theoretically and empirically similar concepts. Therefore, we included studies with different compositions of two-teacher instruction as well as different samples of students comprising general education and special needs students to understand the differential effects of collaborative models of instruction. To overcome limitations encountered in previous reviews investigating differential effects of co-teaching, we have used state-of-the-art meta-analysis methods to handle dependent effect sizes, i.e., studies contributing multiple outcomes (Joshi et al., 2022; Pustejovsky & Tipton, 2021; Rodgers & Pustejovsky, 2021; Tipton & Pustejovsky, 2015). *Fourth*, we aimed to improve on previous meta-analyses by also accommodating common critiques against effect size calculation in meta-analysis of being obscure (Maassen et al., 2020) by ensuring unprecedented transparency. Concretely, this means that all parts of the review, including effect size calculation 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 guideline and the recommendations put forward by Pigott & Polanin (2019). Supplementary materials include completed PRISMA checklists. The review has been pre-registered at the Open Science Framework (OSF), see <https://osf.io/ur2bs> [The link is currently inactive due to an embargo; it will be activated before publication. Find blinded versions of the two linked protocols at <https://osf.io/fby7w/>].

Inclusion and Exclusion Criteria

Study designs

To draw on the most reliable research for causal inferences, we included treatment-control group designed studies only. Hence, single-case pre-posttest designed studies were excluded. Included were (cluster and/or blocked) randomized controlled trials (RCTs), quasi-experimental studies (QES), and observational studies (OBS). We characterize RCTs as studies in which the researchers both control the random assignment of students into the treatment and control groups (either individually or in clusters, e.g., classrooms or schools) and initiate the implementation of the intervention. QES represent studies in which the researcher(s) initiate the implementation of the treatment but do not randomly assign students to the intervention groups. Finally, OBS are studies where the

researchers neither have an influence on the implementation process of the intervention nor control the randomized assignment. Such studies might, however, still draw on randomization if, for example, schools randomly assigned students to classrooms before the treatment.

To avoid that we induce more bias than we prevent by including study designs with varying quality, we applied strict rules for non-randomized studies to be included (Egger, Juni, Bartlett, Holenstein, & Sterne, 2003). We only allowed posttest effect sizes from QES and OBS to be included if baseline equivalence was assured. If not assured, QES and OBS either had to provide baseline/pretest achievement or covariate-adjusted⁴ measures from which we could compute pretest- and/or covariate-adjusted effect sizes (Taylor et al., 2021). Otherwise, we considered non-equivalent groups studies that only reported posttest scores to be of critical risk of bias due to confounding. 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

All types of collaborative models of instruction 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”). 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 two weeks of treatment, i.e., at least ten school days. Studies where students had two teachers but the instruction was executed to a group of students *outside* the main classroom as well as studies where students were divided into distinct classes that received instruction in two different classrooms were excluded (e.g., see Jang, 2006a, 2006b). For an overview of teacher assistant interventions provided outside the main classroom, which were excluded, see Farrell et al. (2010, p. 440).

Eligible control groups for this review can be categorized into three groups; 1) *non-inclusive general education single-taught classrooms*, i.e., general education students only, compared

⁴ Find the list of focal covariates/confounding factors in our pre-registered protocol at <https://osf.io/fby7w/>.

to general students from co-taught classrooms, 2) *inclusive general education single-taught classrooms*, i.e., a blended student composition, either compared to general and/or special needs students in co-taught classrooms, and 3) *special education classrooms* such as resource rooms and pull-out classrooms compared to students with special needs in co-taught classrooms. We did not include two-teacher interventions conducted in special education school settings. Moreover, we did not allow collaborative teaching to be compared to single-taught classrooms using reduced class sizes, as in the Project STAR (Finn & Achilles, 1990).

Student populations

The eligible population sample for this review was students in grades one to twelve who attended primary or secondary schools, including both public and private as well as boarding schools. This also included special education schools as well since these functioned as control schools. We did not find any studies including private school settings only. Studies based on students in kindergarten, vocational or post-secondary education were excluded. Overall, we included three types of student samples, 1) students with special educational needs and/or disabilities, 2) general education students, and 3) 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 (socio-economic status)” students but these were not formally characterized as special needs students, we either amalgamated these results with the general student results, when possible, or interpreted the results as coming from 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 (The World Bank, 2022). To exemplify, we excluded the Iranian study by Aliakbari (2013). Furthermore, we only included documents and studies written in English, Danish, Swedish, Norwegian, or German. Yet, 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 achievements, such as skills in reading, writing, and mathematics. Eligible outcomes were all types of academic achievement tests, including *state- or nationwide standardized tests, norm-referenced commercial tests, grades, leaving examinations, marks for the year's work, large-scale assessment tests, teacher-developed tests, researcher-developed tests, and textbook tests*.

We solely considered Arts, Social Science, and STEM subjects to be eligible, such as language arts (LA), social studies, history, science, biology, and mathematics. In later analyses, we roughly dichotomized effect sizes into Arts & Social Science vs. 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. Notably, we allowed IQ tests to function as a proxy for student achievement if these were used as pretest or baseline measures. We divided analyses between posttest and follow-up measures. The latter was characterized as effects measured three months or later after the end of the intervention. If studies reported effects across various time points, we included all of them.

Search procedures

The search string that we developed for our electronic searches was inspired by the previous review studies as well as a number of 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 pre-registered protocol at <https://osf.io/fby7w/>. We conducted an electronic search 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 & Theses Global. The main source for grey literature was the database ProQuest Dissertations & Theses Global, which identified a large number of dissertations. Beyond this systematic search, we did 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 the replies contained the demanded information (Polanin et al., 2020; Schauer, Diaz, Lee, & Pigott, 2020), we opted to change this initial plan due to the seemingly low chances of a successful extension of 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 and reasons for exclusion of references alike were conducted and documented 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, Schmid, & Edler, 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 appeared, these were resolved 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 used data extraction scheme was closely developed in line with the co-teaching literature and theory (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, Beuchert-Pedersen, Nielsen, Thomsen, et al., 2018; Andrews-Tobo, 2009; Fontana, 2005; Muijs & Reynolds, 2003; Murawski, 2006). Hereto, we optimized the data extraction scheme by reducing the number of extraction characteristics whenever certain characteristics

were not retrievable from the pilot studies. That, for example, led us to exclude the variable “*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 are available for critical inspection and future updates at <https://osf.io/fby7w/>.

Risk of Bias (RoB) assessment

To further assure that the accuracy of the review was not compromised by including study designs of varying quality, we conducted comprehensive risk of bias (henceforth, 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 non-randomized studies, we applied the RoB 2 tool for RCTs (Sterne et al., 2019), the RoB 2 CRCT tool for cluster RCTs (Eldridge et al., 2021), and the ROBINS-I tool for non-randomized studies (Sterne et al., 2016). To ensure comparability between the three RoB assessment tools, we required that non-randomized studies should either provide raw data or a pre-registered protocol in order to receive a low risk of bias 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 Supplementary Section S6 (online only).

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 *wild-meta* (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 the Hedges's g estimator (Hedges, 1981). We coded effect sizes so that positive values indicated a treatment effect, i.e., 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; Wilson, 2016; WWC, 2020, 2021). The majority of the effect sizes was based on either pretest or covariate-adjusted computation techniques (Morris, 2008; Morris & DeShon, 2002; Pustejovsky, 2016; Taylor et al., 2021). Calculating covariate- and/or pretest-adjusted effect sizes in most cases requires information about the correlation between the covariate(s) and the outcome measures, ρ_{cor} , which are infrequently reported in primary studies but can be obtained from other measures that are usually provided (Pustejovsky, 2020a; Wilson, 2016). Whenever ρ_{cor} was impossible to derive from reported results, we imputed ρ_{cor} following the guideline put forward by the What Works Clearinghouse (2020; henceforth WWC).

All calculated effect sizes were standardized by the *total variance*, here denoted as g_T . This means that all effect sizes both encompass variance from 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 at one of these levels only were converted to ensure that they represent the same unit of analysis, i.e. g_T (Hedges, 2007). This also entailed conducting *approximate cluster bias corrections* on all effect sizes coming from multi-sited 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 & Hedberg (2007), as suggested by Hedges (2007), to conduct these 2-level conversions. To further ensure a common unit of analysis across effect sizes and to 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 of the review, see Supplementary Section S1 (online only).

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*, there are six studies that reported results from multiple non-overlapping 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 coming from non-overlapping 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*, we have four studies that contained both of the above-mentioned dependency structures, which means that they reported multiple outcomes from multiple non-overlapping samples.

A challenge for 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 remedy 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 et al., 2020; Vembye, Pustejovsky, & Pigott, 2022). RVE implies the use of *working models* that tentatively aims 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, ρ , 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 mis-specified. Furthermore, we applied the “CR2” small-sample corrector (Joshi et al., 2022; Tipton, 2015; Tipton & Pustejovsky, 2015) to 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.

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., 2022). The CHE-RVE model both takes into account the multi-level 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) while simultaneously accounting for both hierarchical and correlated effects dependence structures. Commonly, CHE models entail assuming a constant sample correlation, ρ , between effect size estimates coming from the same study. However, we obtained ρ by estimating Pearson’s correlation from studies that both reported math and language arts scores, as suggested by Kirkham et al. (2012). We estimated $\rho = .706$. That appears to be plausible since it closely resembles the sample correlations obtainable from the Project STAR-data (Achilles et al., 2008) across 1st-, 2nd-, and 3rd-grade students either assigned to the teacher’s aides or single-taught arm, which were $\rho_{grd1} = .718$, $\rho_{grd2} = .722$, and $\rho_{grd3} = .735$. Using restricted maximum likelihood techniques (Viechtbauer, 2005), we estimated two sources of heterogeneity, i.e., the standard deviations at the effect size level (also known as the within-study SD, ω) and at the study level (also known as the between-study SD, τ). Larger standard deviations indicate larger amounts of variabilities among effect sizes than would be expected from sampling error alone. See Supplementary Section S2 (online only) 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 $\rho = 0.95$. 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 ρ (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 if 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 the small sample corrections. 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. Hereto, we conducted a sensitivity analysis in which we re-estimated 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 re-estimated the mean effect size model while changing inclusion criteria by blockwise excluding the following categories of studies or effect sizes: observational studies, non-randomized studies, single-sited studies (i.e., one treatment one control class, only), large-scale studies with sample sizes above 1000 students, gray literature, serious risk of bias assessed effect sizes, non-US 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 the study by Dwyer (2018) was considered as an outlier.

Publication bias testing

We conducted three complementary publication bias and/or small study effects tests, as suggested by Hedges & Vevea (2005). This included *Trim-and-Fill* tests based both on all individual effect sizes and effect sizes aggregated to the study level, *Egger's regression* tests accounting for dependent effect sizes using the CHE-RVE model (Egger, Smith, Schneider, & Minder, 1997; Rodgers & Pustejovsky, 2021), and *step-function selection model* tests. For the latter we used both 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. 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 otherwise have created an artificial correlation among the standardized mean differences and their variability measures which consequently would have induced the risk of yielding flawed evidence for publication bias and/or small study effect. Moreover, we apply contour-enhanced funnel plots for illustrating potentially publication bias/small study effects (Peters et al., 2008). As a sensitivity analysis, we also made contour-enhanced funnel plots based on transformed measures which represent an alternative method for handling the artificial correlation between standardized mean differences and their variability measures (Pustejovsky & Rodgers, 2019). See Supplementary Section S8 (online only) for an elaboration of the conducted publication bias tests.

Moderator analyses

To investigate if focal moderators of methodological and theoretical relevance were able to 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, 1) subgroup analyses based on categorical variables without missing values (i.e., fully reported information across all studies), 2) subgroup analyses based on categorical moderators with missing values, and 3) 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 study characteristics; *type of intervention*, *subjects*, *test modes*, *grade levels*, *the type of control group*, and *the type of control group used for effect sizes calculation for samples of special needs students only*. Across these models, we varied between fitting Subgroup Correlated Effects Plus (SCE+) or the Correlated Multivariate Effects Plus (CMVE+) working models (find detailed information about the use and the embedded assumptions of these models and the

reasons for shifting across models in Supplementary Sections S2-S4 [online only]). As with the mean effect size models, these models included heterogeneity at the effect size and study level alike (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). We used both to check for consistency between these two Wald-tests, but the main interpretation was placed on the CWB values.

For the second set of models, we investigated if 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 common planning time was provided against studies reporting no provision of common planning time. *Second*, we tested differences between studies in which co-teaching training was provided against studies in which no training was provided. The SCE+ working model was the only model used for these analyses.

For the latter set of models, we investigated whether the *duration* and *intensity* of the intervention as well as if 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 a severe amount of 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 is documented in Sections S3 and S4 in the Supplementary Material (online only).

Dealing with missing data

To handle missing values on moderators variables, we used multiple imputation with 50 imputations and 50 iterations (Pigott, 2019; Van Buuren, 2018). We applied Exploratory Missingness Analysis (EMA) 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 if they had more than 50 percent missing values or if the missingness structure of the variables was correlated with the effect sizes and their variance. Find the EMA at <https://osf.io/fby7w/>.

Since no methods have yet been developed to reliably pool multi-contrast 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—for obtaining 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 & 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 without a fair chance of finding true mean difference between moderator categories.⁵

Deviations from Preregistration Protocol

The final review diverges in several ways from our initial pre-registered protocol. The initial plan was to use the EBSCO database for literature search, too. However, we experienced problems using our rather extensive search string in this database and thus excluded it. Next, 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. In addition, we did not conduct sensitivity analysis for publication bias aligned with the CHE model (Mathur & VanderWeele, 2020) and weighted average of the adequately powered studies (Stanley et al., 2017). However, as we will show later on, publication bias might not be the most pressing issue for this review due to the large number of included gray literature. Moreover, we

⁵ We sought statistical advice for this matter.

did not apply multi-level multiple imputation because the nesting structure of the missing values did not allow us to conduct this type of imputation since the covariates rarely varied within studies.

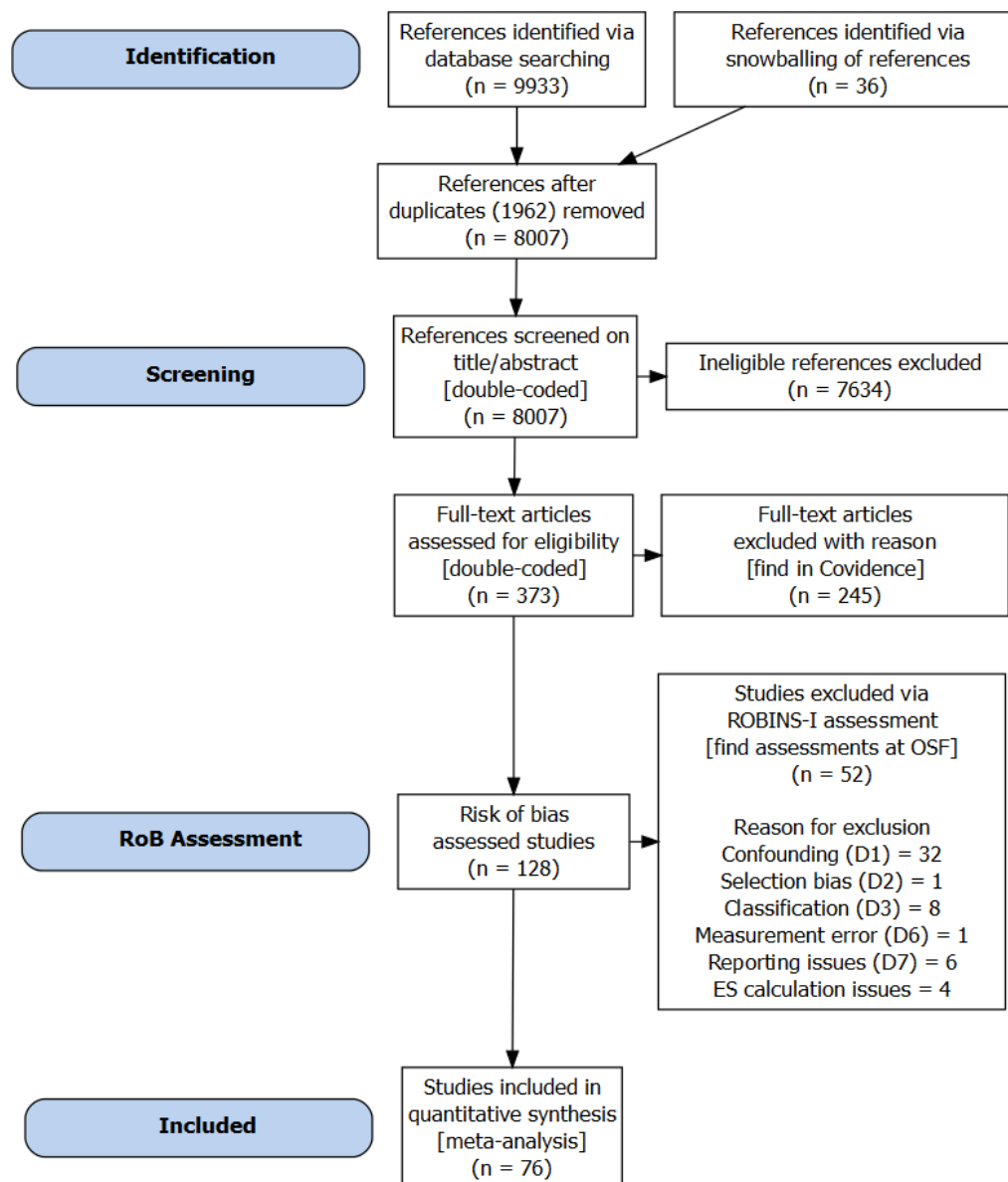
Beyond the protocol, we added several exploratory analyses to the review that were not mentioned in the protocol but that were able to strengthen the review. Concretely, we added sensitivity analyses, including an analysis using ICC values from the pretest covariate models—instead of the unconditional models—from the population representing all schools from Hedges & Hedberg (2007) for cluster bias correction of effect size, an analysis investigating the impact of large studies (i.e., sample sizes above 1000 individual students), an analysis omitting single-sited studies (based on the assumption that they likely misguide more than they inform), an analysis examining US studies only, and an analysis in which outliers were removed. We also conducted an analysis investigating the difference between special needs student effect sizes based on general and special education control groups to investigate if one of the alternative service delivery models outperformed the other. Lastly, we conducted a sensitivity analysis of the subgroup analyses based on studies that only analyze the effect of co-teaching, i.e., the collaboration 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 9969 potentially relevant references from databases- and snowballing searches. After removing 1962 duplicates, the first and second author title and abstract screened 8007 references independently. The proportionate agreement between authors was 93.74% with Cohen's $\kappa = .448$, 95% CI [.447, .449], which indicates a weak agreement according to guidelines commonly referred to (cf. McHugh, 2012; Orwin & Vevea, 2009). However, we have no profound concerns about this value since κ was mainly driven by start-up disagreements to which one of the authors applied a more inclusive screening strategy than the other. Subsequently, we full-text screened 372 studies independently. We excluded 245 studies for various reasons. The most common reason for exclusion at this stage was ineligible study designs (65, most of which were single-case pre-post designed studies), 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 eligible reasons, which artificially reduced the agreement rate between

authors. We then risk of bias (RoB) assessed 128 studies, and of these, 52 studies were excluded due to a critical RoB assessment for at least one domain in the ROBINS-I tool. The most common reason (i.e., for 25 studies) for a critical ROBINS-I judgment was posttest-designed studies not ensuring baseline equivalence among the intervention and control groups or not providing relevant covariate/pretest/baseline measures. At last, the final meta-analytic dataset included 76 studies.

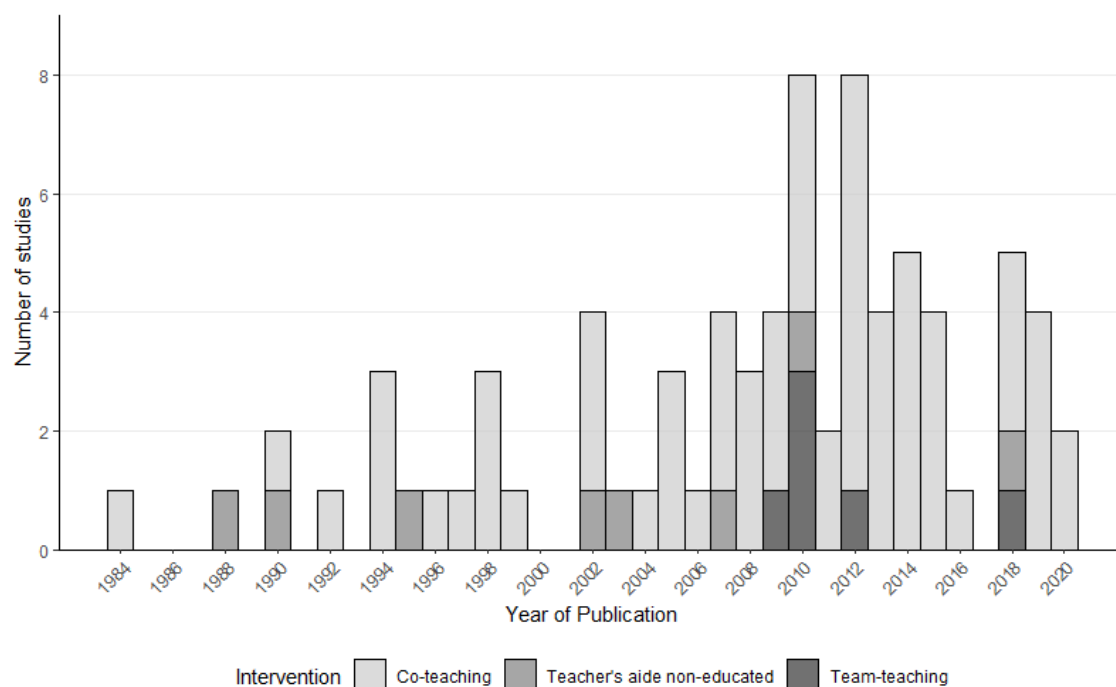
FIGURE 1. PRISMA flow chart showing the search, screening, and exclusion process



Descriptive Statistics

Figure 2 exhibits the included studies by year of publication and the type of intervention. It appears that there has been an increasing number of studies investigating the academic effects of collaborative models of instruction after 2000, i.e., 61 studies (80%) were conducted in this period. A contributing factor to this significant increase might have been the request for more data raised in the previous review by MS01. However, we were also able to detect seven additional studies in the period from 1990 to 2000 concerning the effects of co-teaching on special needs students' achievement that MS01 did not locate in the database at that time, albeit we applied more narrow inclusion criteria. This might point to significant improvement of the quality of databases since that time.

FIGURE 2. Number of studies included in the meta-analysis by year and intervention



Notice: 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.

Study and sample level characteristics

Tables 1 and 2 and Supplementary Tables S4-S8 (online only) present descriptive statistics regarding the study, sample, and effect size level, respectively. The final meta-analysis was based on 96

non-overlapping samples from 76 studies. These studies yielded results from 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, while 23 (30%) and 18 (23%) studies were on middle- and high school students, respectively. Across samples, the studied grades ranged from 1 to 11 (mean = 5.49). Consequently, 12-grade students were absent in this review (see Supplementary Figure S1 [online only]). Studies were predominately evaluating the co-teaching intervention, with co-teaching studies (65) outnumbering studies on teacher assistants (8) and team-teaching (6) by far. The size of studies was quite heterogeneous. There were many small-sample studies within the co-teaching literature ($M = 198$, $SD = 515$), while teaching assistant studies, in contrast, were often based on large samples ($M = 1915$, $SD = 3264$), primarily driven by three large-scale cluster-randomized trials (i.e., 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 Supplementary Material (online only). The mean duration of the intervention was 37.34 weeks ($SD = 23.77$), which is close to one year of schooling. The duration, however, varied substantially between studies, ranging from 3 to 160 weeks. The average number of sessions (i.e., 45-minute sessions) per week was 11.4. However, this characteristic was infrequently reported.

Approximately half of the included studies reported whether co-teaching training was provided prior to the treatment only, with 27 studies reporting the use of training and ten studies reporting no training. Aligned with theoretical recommendations for the practice of co-teaching, it was common for studies (59) to document whether common planning time was provided for the collaborative educators. Only six of these studies reported that there was no common plan time.

Regarding research designs, only nine studies were RCTs. Thus, the vast majority of studies (67) were quasi-experimental or observational studies. A distinct feature of this review is that 59 (76 %) studies were characterized as gray literature, including 55 dissertations (71%) and four conference papers (5.3%).

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 either language arts (LA, 165 effect sizes, 53 studies) or mathematics (104 effect sizes, 44 studies) achievement tests. 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 coming from general education and blended samples of students, respectively. Most frequently, effect sizes (i.e., 86.2%) were obtained from standardized achievement test measures. Only eight effect sizes represented follow-up effect size estimates (i.e., effects measured more than three months after the end of the intervention). Further, only one study (i.e., Andersen et al., 2018) reported intention-to-treat effect sizes. Hence, we concentrated on treatment-on-the-treated effects exclusively.

We applied 2-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). Of the calculated effect sizes, 82% were adjusted for pretest differences among students, and 88% were adjusted for focal covariate differences (find further univariate descriptive information in Supplementary Section S5 [online only]).

TABLE 1. Descriptive Percentages for the Included Studies.

<i>Study level characteristics</i>	<i>Studies (J)</i>	<i>Samples (I)</i>	<i>Effect sizes (K)</i>	<i>Percentage_J</i>
Study context				
US 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 (grade 1-5)	36	52	148	0.474
Middle school (grade 6-8)	23	25	84	0.303
High school (grade 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

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

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
<i>Effect size level characteristics</i>	<i>J</i>	<i>I</i>	<i>K</i>	<i>Percentage_K</i>
Outcome characteristics				
LA 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

<i>Characteristics</i>	<i>J</i>	<i>I</i>	<i>K</i>	<i>Mean_I</i>	<i>SD</i>	<i>Range</i>
Sample characteristics						
Number of students	76	96	290	391	1286	10-10781
Effective sample size ¹	76	96	290	18	20	5-113
Intervention group	76	96	290	157.96	494.648	5-4016
Control group	76	96	290	232.835	812.151	5-6765
Sample size (co-teaching)	65	80	236	198	514.973	10-4368
Sample size (teacher assistant)	8	11	36	1915	3264.341	54-10781
Sample size (team-teaching)	6	8	18	518	1186.489	46-3450
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	4.21	1-27
Mean obtainable pre-posttest ρ	24	29	90	0.611	0.168 ³	-0.036-0.92

Note. 1) Calculated via $4/\sigma_T$, where σ_T is the standard deviation both containing individual and cluster level heterogeneity. 2) This mean was aggregated to the study level. 3) This mean was calculated at the effect size level.

Risk of Bias

Figures 3 and 4 depict weighted summary plot results of the RoB assessment for the non-randomized and for randomized studies, respectively. Specifically, 67 studies were assessed via the ROBINS-I tool, while nine were assessed either by 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 Supplementary Section S6 (online only) for further details about the RoB assessment, including unweighted plots and separated plots for quasi-experimental and observational studies.

As illustrated in Figure 3, ROBINS-I assessed effect sizes most frequently received 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 judgment was made on the consideration that it might be unrealistic to expect that the pretest adjustment (or focal covariate adjustment) controls 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 encompass a low risk of confounding. In contrast, most RoB 2 assessed effect sizes received a low risk of bias due to randomization, as appears from Figure 4.

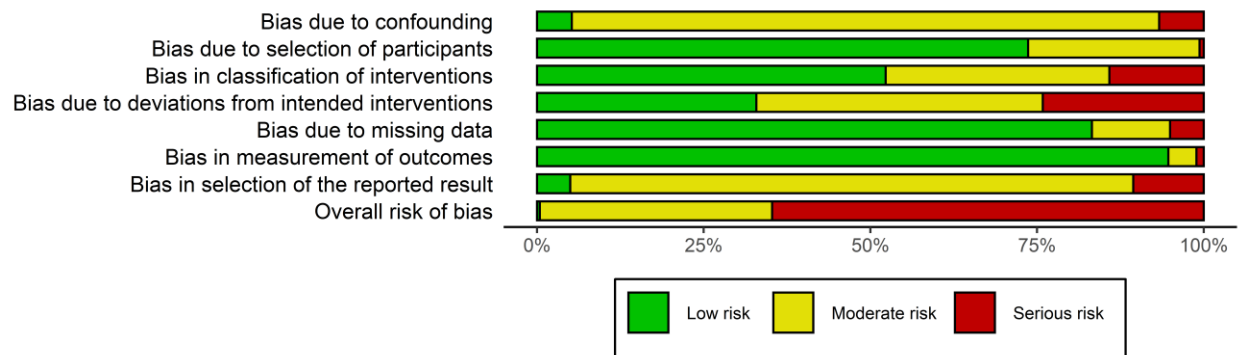
The majority of the included studies across all research designs were judged to be of moderate risk of bias due to (selective) reporting because none of the included studies were pre-registered. We only considered studies that provided the raw data behind the analyses to be of low risk of bias due to reporting. For RCTs, the main reason for not being rated as low overall risk of bias was the lacking pre-registration.

Notably, 45 out of 67 non-randomized studies contained at least one effect size assessed with an overall serious risk of bias. In total, 135 ROBINS-I assessed effect sizes received an overall serious risk of bias. The most common reason (i.e., 24.4% of the ROBINS-I assessed effect sizes, see Supplementary Table S12 [online only]) for a serious RoB assessment was due to limited descriptions of the implementation process (D4). For the same reason, ~10% of the ROBINS-I assessed effect sizes received a serious RoB assessment due to classification. We often made this

judgment because the control group was scantily 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.

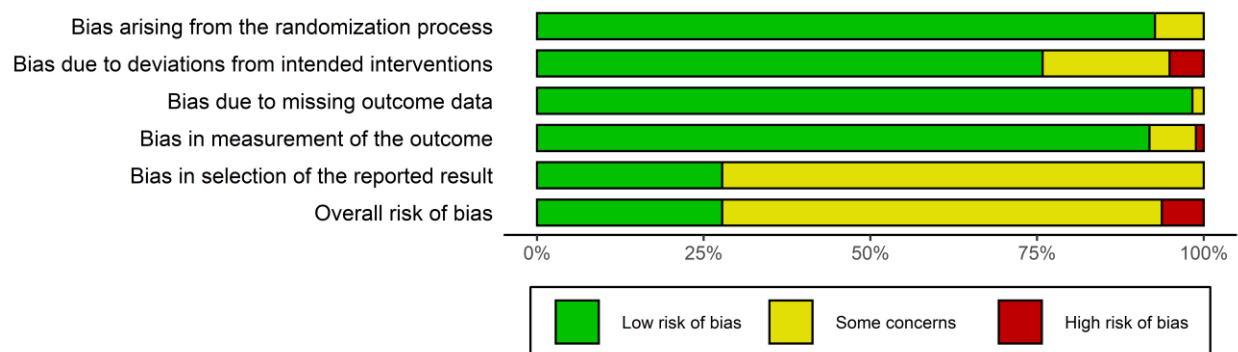
In sum, risk of bias assessment needs to be taken serious for the sample of studies going into this meta-analysis. Half of the included effect size estimates were assessed to have a serious overall risk of bias. This result was mainly driven by the fact that most of the included studies applied non-randomized research designs. In our first set of subgroup analyses, we contrast the differences between serious and non-serious risk of bias effect sizes to investigate the impact on the main results of including studies and effect sizes considered to be of serious risk of bias.

FIGURE 3: ROBINS-I Weighted Summary Plot



Note: This plot contains information related to 225 effect sizes coming from 67 non-randomized 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



Note: This plot contains information related to 55 effect sizes coming from 9 randomized studies of which 26 effect sizes come from four RCTs and 29 effect sizes come from five cluster RCTs.

Meta-Analysis

Mean effect size estimation

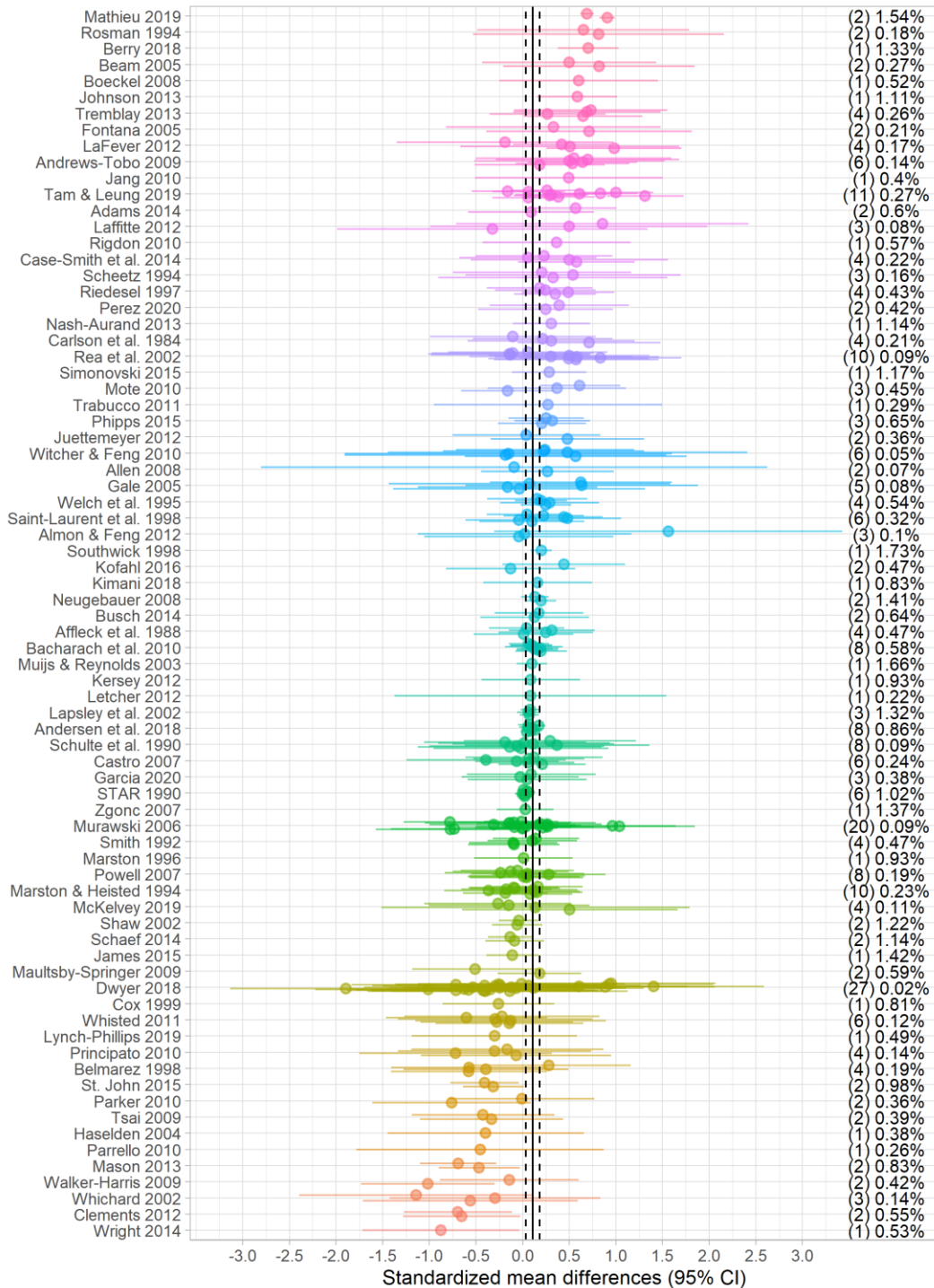
The overall mean effect size from our meta-analysis summarizes a total of 280 effect sizes across 96 samples from 76 studies. For the mean effect size analysis, we excluded 10 effect sizes, i.e., the few (eight) follow-up effect size estimates and two effect sizes from the special needs student sample 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), $t(40.3) = 2.97$, $p = 0.005$, 95% CI[0.035, 0.184]. 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 the average co-taught students had a better achievement score 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. On the student level, this translates to that a typical student from the control group would be expected to have had a percentile gain of 4.4% had the student received a collaborative model of instruction (WWC, 2020).

We found a substantial amount of heterogeneity among effect sizes, $Q(279) = 1164.2$, $p < .0001$, $I^2 = 91.73$, 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 0.274.⁶ This suggests that both the study and effect level covariates might be able to explain differences in effect sizes estimates across studies, which in turn justified all of our planned moderator analyses (cf. Pustejovsky & Tipton, 2021).

⁶ The total SD is calculated by the square root of the sum of the within- and between-study variance.

FIGURE 5. Mean Effect Size Forest Plot across Dependent Effect Sizes



Note: 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.

Sensitivity analyses

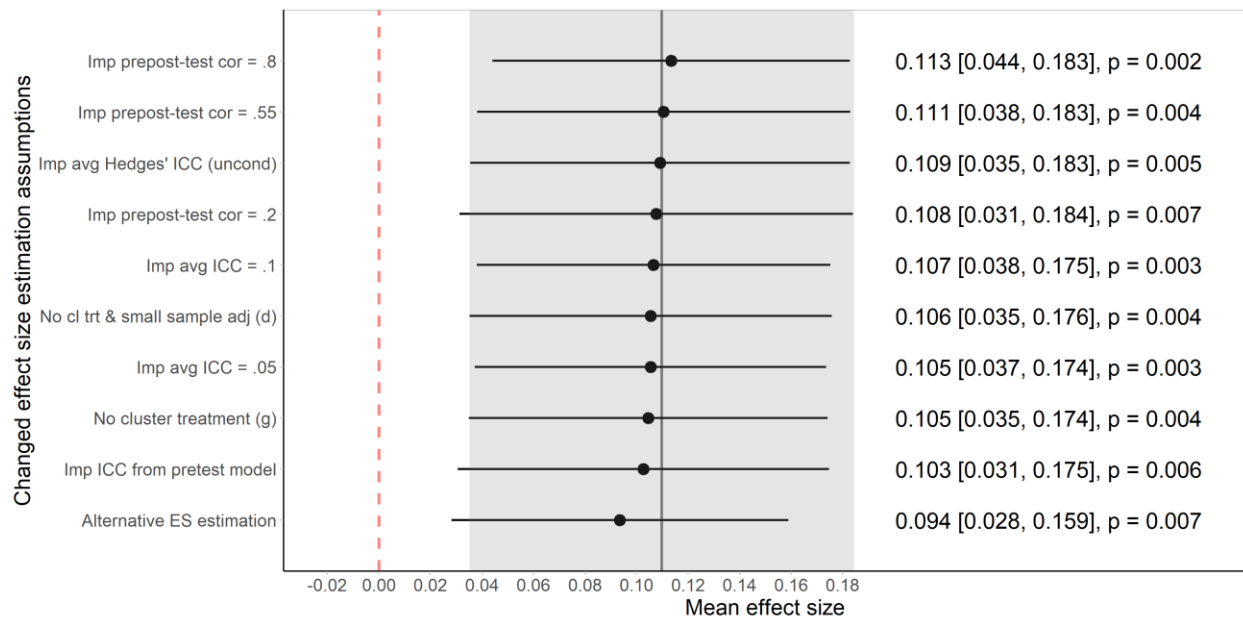
The overall mean effect size ($\bar{g}$) was insensitive to changing assumptions about the sampling correlation, ρ , among effect sizes from the same study, with estimates varying from 0.107, 95% CI[0.0332, 0.180] assuming $\rho = .0$ to maximum 0.111, 95% CI[0.034, 0.187] when $\rho = .6$. The total variance component estimate was, to a large degree, substantially insensitive to changing ρ . Yet some variation was detected, with the total SD ranging from 0.232 to 0.347. By contrast, the individual variance components were heavily sensitive to the magnitude of ρ (see Supplementary Figure S10 [online only]). Therefore, the relative magnitude of the individual variance component estimates should be interpreted with caution. The same patterns appeared for the conducted leave-one-study-out analyses in which $\bar{g}$ happened not to be substantially influenced by any single study. $\bar{g}$ ranged from 0.088, 95% CI[0.031, 0.144] when omitting Mathieu (2019) to 0.122, 95% CI[0.05, 0.194] when omitting Mason (Mason, 2013). The estimated total variance component estimate was generally insensitive to omitting any single study, ranging from 0.243 to 0.285. Yet the 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 Supplementary Figures S11 and S12 [online only]).

Figures 6 and 7 display how the mean effect size estimation was influenced by changing assumptions related to the effect size calculation and inclusion criteria of the review, respectively. Generally, $\bar{g}$ was agnostic to alterations of assumptions related to the effect size calculation procedure, with $\bar{g}$ ranging from 0.094, 95% CI[0.028, 0.159] when using the most extreme alternative effect sizes from studies reporting multiple results eligible for different effect size calculations to 0.113, 95% CI[0.044, 0.183] for calculations based on an imputed constant pre-posttest correlation of .8 for studies reporting difference-in-differences results without providing the pre-posttest correlation. The variance estimation slightly varied when changing effect size calculation assumptions, with the total SD ranging from 0.248 to 0.377 (see Supplementary Figure S13 [online only]).

Overall, $\bar{g}$ was not *substantially* influenced by changing any inclusion criteria. Under all changed conditions, $\bar{g}$ remained in the moderate effect size interval (cf. Kraft, 2020), with $\bar{g}$ ranging from 0.068, 95% CI[-0.002, 0.138] when re-estimated on RCT studies only to 0.144, 95% CI[0.066, 0.221] when observational studies were excluded. As indicated by the confidence interval in Figure 7, $\bar{g}$ only just turned statistically insignificant when the re-estimation was based on

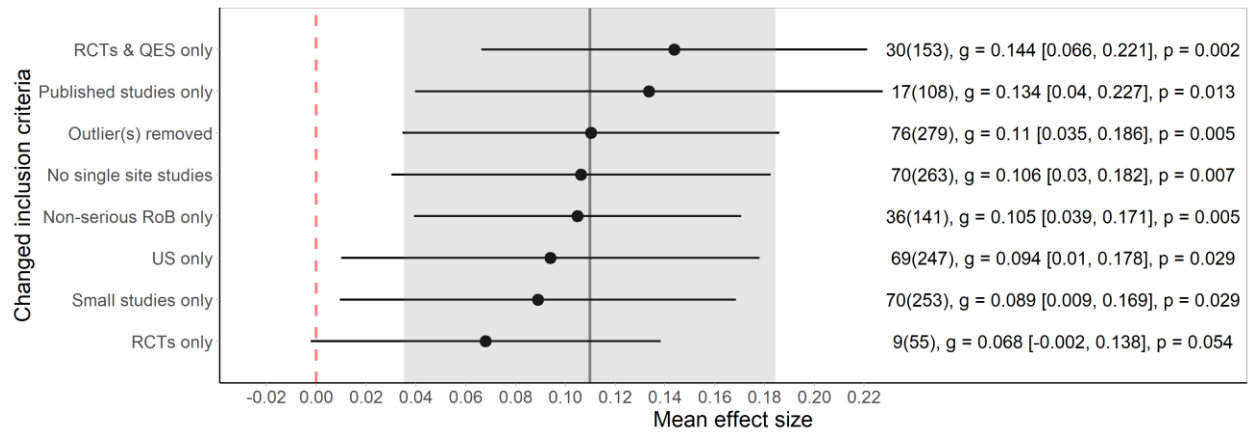
RCT studies only. However, this sensitivity analysis was based on the smallest number of studies and effect sizes (i.e., nine studies and 55 effect sizes) relative to the rest of the analysis. This substantially reduced the power of that model. In line with theoretical expectations on publication bias (Cheung & Slavin, 2016; Rothstein et al., 2005), we found that $\bar{g}$ slightly increased when omitting gray literature (i.e., studies not published in scientific peer-reviewed journals). Interestingly, we found that $\bar{g}$ is not impacted by removing all effect sizes assessed with an overall serious risk of bias, despite this reduced the sample by 40 studies and 139 effect size estimates. Contrary to our assumption, omitting all large-scaled studies with sample sizes larger than 1000 slightly reduced $\bar{g}$. We can, therefore, conclude that $\bar{g}$ was not mainly 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 more or less agnostic to the changed inclusion criteria, although the study-level SD was to a greater extent influenced by the changed inclusion criteria (see Supplementary Figure S13 [online only]).

FIGURE 6: Sensitivity analysis changing effect size estimation assumptions



Note: The right side of the figure presents the overall average effect size and its confident interval as well as the related p value for the re-estimated model. The solid line indicates the overall original main result and the gray region demarcates its confident interval.

FIGURE 7: Sensitivity analysis changing inclusion criteria

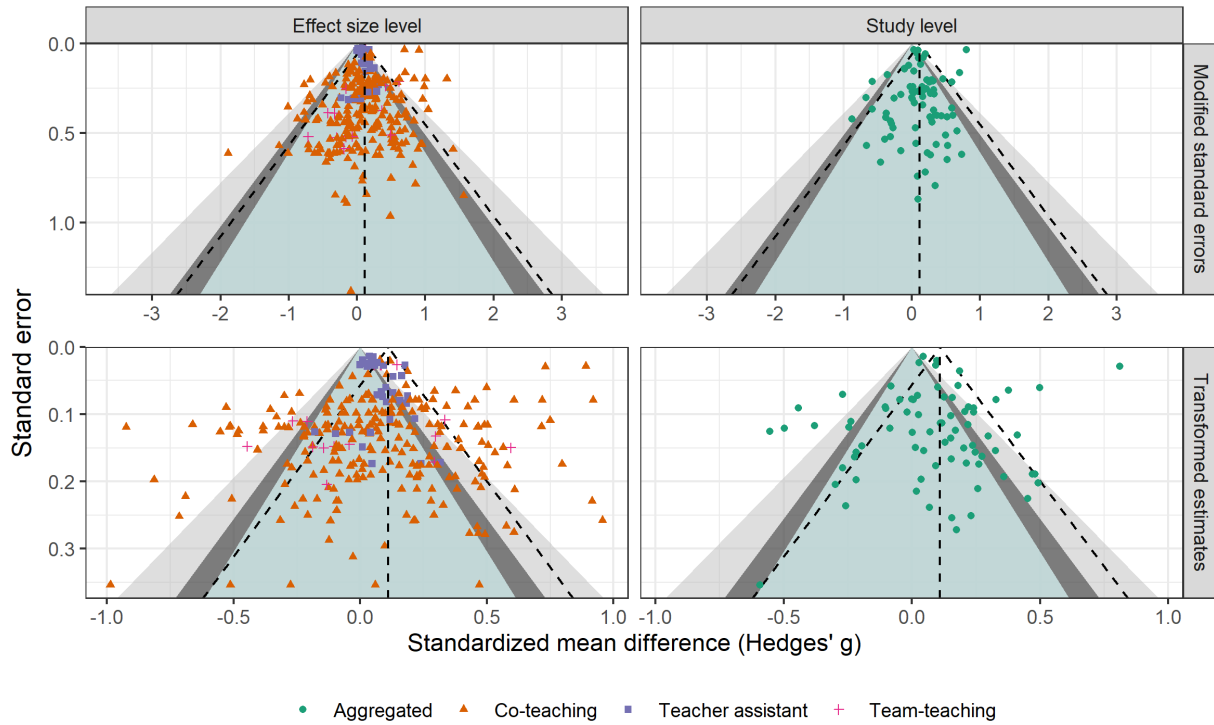


Note: The right side of the figure presents the number of studies and effect sizes in parenthesis and the overall average effect size and its confident interval as well as the related p value for the re-estimated models. The solid line indicates the overall original main result and the gray region demarcates its confident interval.

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 8 depicts contour-enhanced funnel plots conducted at the effect size and at the study level using modified standard errors and transformed effect size estimates, respectively. These plots indicate no small-study effects/publication bias. Overall, we did not find any systematic indication of publication bias or small study effects based on the applied tests, i.e., *Trim-and-Fill test*, *cluster-robust Egger's regression tests*, or *selection models test* (see Supplementary Section S11 for the concrete results [online only]). Hence, we do not expect publication bias to have had any substantial influence on the mean effect size estimation, which might not be a big surprise (Pigott, Valentine, Polanin, Williams, & Canada, 2013) since more than 77% (i.e., 57 studies) of the included studies represent gray literature.

FIGURE 8. Contour-enhanced funnel plots



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

Subgroups analyses

While the above-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 bias* characteristics. In order to study such potential heterogeneities, we conducted a comprehensive range of meta-regression analyses. Table 3 reports subgroups analyses of focal moderator variables that are categorical and did not contain any missing values (see Supplementary Section S9 for relevant subgroup forest plots [online only]). 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. Their rather small number did not allow for

reliably estimating the mean effect size for this subgroup dimension. We further excluded a number of effect sizes and studies because their research design did not allow for the exploration of moderate effects. This means that we excluded all of the 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 Schaef (2014) since it allowed us to explore differences across student samples.

As can be seen from Table 3, we generally found rather robust effects of collaborative instruction across most of the conducted subgroup analyses. Most of the individual effects are not statistically different from zero, as could 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, i.e., between 0.05 to < 0.20 . We only found two statistically significant average group differences, both for the unconditional (i.e., the models without controls) and the covariate-adjusted models. These were between covariate-adjusted and posttest-only effect sizes, with $F(1, 7.8) = 10.2, p = 0.013$ (CWB $p = 0.001$), and between OBS, QES, and RCTs, with $F(2, 12) = 4.86, p = 0.028$ (CWB $p = 0.004$). The former test showed, and in contradiction 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. These results were equivalent when controlling for differences across subject, grade level, and student sample characteristics. Notably, substantial heterogeneity remained across the majority of the conducted subgroup analyses.

The difference between the two-teacher compositions was below practical relevance and not statistically significant. The average subgroup effect size ranges from 0.067, 95 CI[-0.012, 0.147] for teacher assistant interventions to 0.12, 95% CI[0.020, 0.22] for co-teaching interventions in the unconditional model. We neither found any statistical nor substantial important difference between subjects categories, with the average group effect sizes ranging from 0.076, 95% CI[-0.021, 0.173] for STEM and 0.137, 95% CI[0.056, 0.217] for Arts & Social Science outcomes

in the unconditional model; with both results considered to be moderate in size. From the unconditional model, we did find a small effect (cf. Kraft, 2020), not statistically distinct from zero for effect sizes based on samples of general education students of 0.038, 95% CI[-0.081, 0.157], and a moderate effect statistically distinct from zero for effect sizes premised upon samples of special needs students of 0.143, 95% CI[0.016, 0.269]. However, we did not find a statistically significant difference between the two means, $F(1, 33.6) = 1.61, p = 0.213$ (CWB $p = 0.220$), which suggests that general education and special needs students might benefit equally from collective instruction. However, the effects for general students are substantially smaller, which might be of practical relevance if confirmed as statistically significant. Yet, the present models did not have enough statistical power to draw a firm statistical conclusion for this matter.

Results differed across grade levels, with $\bar{g} = 0.122$, 95% CI[0.047, 0.198], $p = 0.004$ for elementary school outcomes, $\bar{g} = 0.08$, 95% CI[-0.106, 0.267] for middle school outcomes, and $\bar{g} = 0.046$, 95% CI[-0.133, 0.226] for high school outcomes. However, the results did not reveal any statistically significant differences, $F(2, 25.6) = 0.398, p = 0.676$ (CWB $p = 0.697$), for the unconditional model. Although the high school means effect fell within the small effect interval, it can be considered as 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 found in annual gains across subjects from nationally normed tests in the U.S. (Lipsey et al., 2012).

Further, we did not find any practical relevant subgroup mean differences between risk of bias, the study outlet, and the type of test categories, with all of the subgroup means 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 with special needs students only in order to investigate if any of the service delivery models (inclusive/general education single-taught vs. special education classrooms) could be considered superior relative to the other. We did not find any statistically or practical significant difference between subgroup means for this analysis either (see Supplementary Table S16 [online only]).

TABLE 3: Subgroup Analyses for Focal Moderators Without Missingness

Subgroup-analyses				Unadjusted effects			Covariate-adjusted effects ^a		
Coefficient	Studies	ES	Est. [95% CI]	Satt. df	SD ($\tau + \omega$)	Est. [95% CI]	Satt. df	SD ($\tau + \omega$)	
Intervention characteristics									
Co-teaching ^b	63	226	0.12* [0.02, 0.22]	46.1	0.332	0.112 [-0.012, 0.236]	31.3	0.332	
Teacher assistant	8	33	0.067 [-0.012, 0.147]	2.6	0.025	0.08 [-0.051, 0.212]	8.4	0.017	
Team-teaching	6	16	0.087 [-0.022, 0.196]	1.1	0.033	0.082 [-0.053, 0.216]	7.5	0.045	
Wald test <i>p</i> values ²			0.343 (0.361)			0.612 (0.592)			
Outcome characteristics									
Arts and social science ^c	54	162	0.137** [0.056, 0.217]	34.6	0.253	0.177** [0.059, 0.296]	21.9	0.254	
STEM	48	113	0.076 [-0.021, 0.173]	34.4	0.324	0.125 [-0.002, 0.252]	28.4	0.335	
Wald test <i>p</i> values			0.179 (0.201)			0.234 (0.266)			
Posttest ES ^b	11	35	-0.264 [-0.564, 0.035]	6.5	0.236	-0.262 [-0.59, 0.067]	9.5	0.258	
Covariate adjusted ES	67	240	0.15*** [0.075, 0.224]	26.4	0.273	0.155* [0.04, 0.269]	23.9	0.274	
Wald test <i>p</i> values			0.013* (0.001**)			0.014* (0.003**)			
Non standardized test ^b	14	39	0.107 [-0.085, 0.298]	7.4	0.096	0.082 [-0.15, 0.313]	9.7	0.099	
Standardized test	68	236	0.111** [0.033, 0.189]	34	0.286	0.104 [-0.032, 0.239]	28.6	0.287	
Wald test <i>p</i> values			0.961 (0.961)			0.827 (0.812)			
Participants characteristics									
Blended sample ^b	19	61	0.066 [-0.005, 0.137]	2.8	0.024	0.043 [-0.05, 0.136]	2.4	0.021	
General education sample	26	79	0.038 [-0.081, 0.157]	18.8	0.324	0.026 [-0.104, 0.156]	20.4	0.329	
Special needs sample	42	135	0.143* [0.016, 0.269]	33.5	0.344	0.119 [-0.007, 0.244]	33.7	0.341	
Wald test <i>p</i> values ³			0.213 (0.220)			0.279 (0.269)			
Elementary school (1-5) ^b	35	142	0.122** [0.047, 0.198]	10.7	0.078	0.137* [0.015, 0.258]	12.2	0.046	
Middle school (6-8)	23	79	0.08 [-0.106, 0.267]	18.1	0.372	0.072 [-0.121, 0.264]	19.7	0.373	
High school (9-12)	18	54	0.046 [-0.133, 0.226]	12.9	0.349	0.033 [-0.154, 0.219]	14.4	0.346	
Wald test <i>p</i> values			0.676 (0.697)			0.502 (0.515)			
Special edu control grp ^b	32	96	0.173* [0.023, 0.323]	25.8	0.314	0.137 [-0.028, 0.301]	29.3	0.317	
General edu control grp	51	179	0.076* [0.012, 0.141]	16.4	0.26	0.048 [-0.032, 0.128]	13.4	0.263	
Wald test <i>p</i> values			0.225 (0.249)			0.225 (0.297)			
Study characteristics									
Observational ^b	46	129	0.064 [-0.059, 0.187]	37.4	0.304	0.026 [-0.132, 0.185]	25.4	0.31	
QES	20	95	0.245** [0.126, 0.365]	8.8	0.321	0.225** [0.103, 0.346]	17.4	0.323	
(C)RCT	8	51	0.063 [-0.012, 0.137]	2.5	0.254	0.051 [-0.079, 0.181]	13.4	0.248	
Wald test <i>p</i> values			0.028* (0.004**)			0.011* (0.002**)			
Gray literature ^b	57	168	0.081 [-0.027, 0.189]	42.3	0.308	0.082 [-0.059, 0.224]	28.1	0.313	
Published literature	17	107	0.136* [0.04, 0.231]	6.1	0.25	0.156* [0.016, 0.296]	17.9	0.249	
Wald test <i>p</i> values			0.421 (0.416)			0.292 (0.299)			
RoB low/moderate ^b	34	136	0.11** [0.042, 0.177]	9.6	0.275	0.143 [-0.026, 0.313]	22.7	0.274	
RoB serious	46	139	0.08 [-0.038, 0.197]	35.4	0.309	0.075 [-0.062, 0.213]	28	0.313	
Wald test <i>p</i> values			0.641 (0.631)			0.384 (0.376)			

* $p < .05$, ** $p < .01$, *** $p < .001$. Bold degrees of freedom (d.f.) estimates indicate low d.f. values which in turn indicate that the given variance estimation was fragile. a) Adjusted for the grade level, student sample, and subject; b) SCE+ model; c) CMVE+ model. 1) CWB represents adjusted (CR2) cluster wild bootstrapping *p* values using 1999 replications. 2) Comparison was made between co-teaching and teacher assistant interventions only 3) Comparison were made between general and special needs students only. The table is based on 275 effect sizes across 94 samples from 74 studies.

Generally, the results from the unconditional models were mostly equivalent to models controlling for subject, grade-level, and student sample differences, with no differences across all HTZ and CWB Wald tests, as well. It should also be mentioned that we did not find any inferential discrepancies between HTZ and CWB *p* values across all types of models. Moreover, we conducted tests correcting for multiplicity⁷ by using the *false discovery rate* method (Benjamini & Hochberg, 1995; Laird et al., 2005; Polanin, 2013), without this changing our inferences when based on CWB *p* values. As a sensitivity analysis, we re-estimated this set of subgroup analyses based on co-teaching effect sizes only, i.e., collaboration inventions only comprised by general and special education teachers (see Supplementary Table S15 [online only]), without finding 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 relevance based on multiple imputed values for missing values on these variables. The first set of analyses included categorical moderators, while the second set concerned continuous variables. Table 4 reports on the comparisons between studies using vs. not using common planning time and co-teaching training vs. no training, respectively. Table 5 displays the effects of duration and intensity of the collaborative instruction as well as the average percentage of males in the sample. From these analyses, we did not find any statistically significant effects, suggesting that none of these moderators explained differences in effectiveness as otherwise indicated in the co-teaching literature.

⁷ i.e., the increased probability of committing a Type I error just by conducting multiple statistical significance tests.

TABLE 4: Subgroup analyses based on categorical theoretical relevant covariates with less than 50 percent missing values.

Subgroup	Unadjusted effects			Covariate-adjusted effects ^a		
Coefficient	Est. [95% CI]	Satt. df	SD ($\tau + \omega$)	Est. [95% CI]	Satt. df	SD ($\tau + \omega$)
Plan time						
No common plan time	-0.047 [-0.316, 0.223]	3.1	0.322	-0.085 [-0.313, 0.142]	14.1	0.319
Common plan time	0.16***[0.075, 0.245]	36.2	0.256	0.132* [0.002, 0.261]	24.6	0.261
Wald test <i>p</i> value	0.085 ^b			0.098		
Training						
No co-teaching training	0.082 [-0.060, 0.224]	14.2	0.230	0.095 [-0.090, 0.281]	11.9	0.237
Co-teaching training	0.145* [0.036, 0.254]	19.2	0.308	0.148* [0.024, 0.273]	18.1	0.308
Wald test <i>p</i> value	0.869			0.945		

p* < .05. *p* < .01, ****p* < .001. a) The below results are adjusted for student sample, grade level, and subject differences. b) All significance values in this table were based on robust *F*-tests with *q* and *J* – 1 degrees of freedom.

TABLE 5. Meta-regression for continuous focal theoretical moderators with less than 50 percent missing values.

Moderator	Model 1	Model 2	Model 3	Model 4
% Males	0.012 (0.006)			0.014 (0.008)
Duration in weeks		0.001 (0.002)		0.000 (0.002)
Intensity (sessions per week)			-0.001 (0.004)	-0.001 (0.005)
General education students				0.032 (0.067)
Aggregated sample (ref. spec. education)				0.085 (0.096)
Middle school (grade 6-8)				-0.075 (0.100)
High school (grade 9-12) (ref. primary)				-0.081 (0.102)
Arts (ref. STEM)				0.036 (0.051)
Intercept	0.037 (0.045)	0.109** (0.034)	0.125* (0.050)	0.011 (0.087)
Multiple imputation	Yes	Yes	Yes	Yes
Effect sizes	275	275	275	275
Number of studies	74	74	74	74

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

Discussion

Over the last four decades, research on the effectiveness of collaborative models of instruction on students' academic achievement has increased considerably. We demonstrated that the increase was larger than often laid out in primary research and previous reviews on the topic by evaluating 128 studies with treatment-control designs. Notably, we found more studies within all historical periods that had previously been reviewed. On this body of literature, we used state-of-the-art techniques to meta-analyze results across 96 samples of students from 76 studies, which we did

not consider to be of critical risk of bias. These yielded 280 effect sizes, of which most were based on standardized achievement outcomes from LA and math tests and pretest-adjusted measures.

Across the studies included for meta-analysis, varying on intervention, location, implementation, outcome, 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 education interventions with standardized achievement outcomes (Kraft, 2020). It remained moderate in size across all conducted sensitivity analyses and publication bias tests, with the absolute majority of these tests supporting the conclusion of the mean effect size being statistically distinct from zero. Most 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, Finn, Gerber, & Zaharias, 2000), this review provides unambiguous evidence for the effectiveness of collaborative models of instruction on student achievement.

In order to assess potential differences in the effects along the lines of moderators that are considered as theoretically or methodologically important in the literature, we fitted a range of meta-regressions models. To our surprise, we found that the effects of collaborative instruction were generally robust across the assessed moderators. The absolute majority of subgroup effects fell within the interval of a moderate effect, i.e., between 0.05 to <0.20 . This applied to the unconditional as well as the covariate-adjusted meta-regression models, controlling for student, grade, and subject differences. Interestingly, we neither found any statistical nor practical difference between interventions with special education co-teachers compared to those with teacher assistants. Therefore, in contrast to the co-teaching literature—as, e.g., portrayed by 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 education of the second teacher, and/or an equal share of teaching responsibilities between co-teachers. This suggests 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 & Social Science vs. 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 previous trends found in evaluations and benchmarks of annual gains across grade levels in the U.S. (Lipsey et al., 2012). However, we did not find any statistically significant difference between the mean effect sizes from 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 substantially *small* effect for general education students, the mean effect difference between general and special education students remained statistically insignificant. This might suggest that collaborative instruction is not only a viable model for the inclusion of students with special educational needs in general education in terms of improving their achievement, but it can benefit general education students as well.

Furthermore, we tested a range of factors that are considered in the co-teaching literature as practically relevant preconditions for co-teaching to be effective. These factors included common planning time, co-teaching training, the duration and intensity of the intervention, as well as the number of males 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 recommend being cautious about the results and seeing them as preliminary. Finding moderators and conditions for a successful implementation of co-teaching is thus an area that calls for further experimental investigation.

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. It only disappeared for subgroups in which the total number of studies and effect sizes were limited. Hence, there is still a need for further investigation into differences in the effects of different collaborative models of instruction and settings.

Limitations

Although we have performed a comprehensive literature search and attempted to offer in-depth analyses, this review has several limitations. A major limitation of this review is that we concentrated on students' academic achievement only. 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 necessarily 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, social and behavioral, and teacher satisfaction measures. Additionally, we did not inspect differential effects across subtypes of subjects, such as the differences of effects between reading, writing, and spelling outcomes which might be essential to get a more fine-grained and adequate understanding of the effectiveness of collaborative instruction.

Several caveats should be mentioned with regard to the included literature as well. In many cases, we experienced difficulties in obtaining information regarding the actual number of special needs students included in the specific general classroom. Moreover, it was often uncertain if the number of special needs students was held 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 present adults 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 were held constant. Also, it was rare for studies to control out teacher differences across the treatment and control groups. In other words, only a few studies applied 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 the given interventions. Altogether, these factors might potentially have induced some degrees of error to the mean effect size estimations and reduced the generalizability of the review.

Another limitation to the generalizability of the review is the dominance of U.S. studies and its principal limitation to education systems in high-income countries according to the World Bank definition. While we demonstrate that the interventions showed to be effective in other countries as well, there is a clear need for research into the generalizability to middle- and low-income countries in particular.

Further limited by the included body of literature, we were unable to answer a range of questions of theoretical and practical importance, 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 relation and teamwork of the co-teacher teams. These might be factors of potential relevance for future research of differential effects of collaborative models of instruction.

Although we employ state-of-the-art review methods, a number of limitations remain in this regard too. For example, most parts of the risk of bias assessment and data extraction represented single-coder and -rater procedures that 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. Lastly, it is important to notice that all publication bias tests that we employed have inherent deficits. Both the Trim and Fill and cluster-robust Egger regression methods have limited power to detect small study effects, especially when the effect is small and dependent effect sizes are present. Moreover, selection models based on dependent effect sizes aggregated to the study level do not fully control the nominal Type-I error rate (Rodgers & Pustejovsky, 2021). Therefore, all publication bias tests should be seen as an indication for the absence of reporting biases, but this possibility cannot be ruled out entirely. Nevertheless, we do not consider publication bias to be an issue of paramount importance since this review included a large amount of gray literature.

Implications for Practice and Research

Our results suggest that schools and teachers can improve student academic learning for all students by using collaborative models of instruction and that the potential is independent of the specific type of within-class collaboration. Moreover, making collaborative instruction effective might be less complicated 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, Social Science, and STEM subjects as well as grade levels both in cases where no specialists with a formal teacher education are available but also when resources for common planning time and co-teaching training are restricted. It might be that formal education of educators can have an impact on the efficacy of collaborative instruction. Our results just suggest that the difference is too small to be of practical significance. Having a spare in-class adult as such seems to be the more relevant factor than the two-teacher composition and the education of the second educator. Nev-

ertheless, it goes without saying that the effectiveness of collaborative models of instruction certainly might still benefit from a careful consideration of the local context and the concrete educators involved in the implementation as well as the execution of the intervention.

Our results also have several implications for educational research. The majority of the included studies report short-term outcomes only, i.e., outcomes either measured during the intervention or immediately after its end. Thus, future research needs to concentrate more intensively on assessing the long-term effects of collaborative instruction. Since cost-benefit analyses are completely absent in the present body of literature, future research with a focus on investigating the cost-benefit of collaborative models of instruction is needed, including the relative cost-benefit of these models of instruction compared to other related interventions such as increased instruction time and class size reduction.

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 non-randomized studies. Most (cluster) randomized trials came from the teacher assistant literature and were large-scale trials. In contrast, co-teaching studies often had small sample sizes and were based on non-randomized research designs. Thus, future co-teaching research must continue to attempt large-scale randomized controlled trials or high-quality matched-groups designs to assess the true effect of co-teaching. Overall, we argue that the need for more studies as such is less urgent. Instead, larger and more rigorously conducted studies are needed, especially for co-teaching interventions. Here, primary research that investigates the variation across focal moderating factors and/or preconditions for effective co-teaching interventions is needed in particular (Hedges, 2018). More should also be learned about the differences between the effect on general vs. special needs students because the evidence for the effectiveness of collaborative instruction on the achievement of general education students was not firmly clinched in this review.

Implication 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 with scarce resources, policy-makers and local stakeholders could profitably start searching for cheaper and more efficient interventions. That said, our results point to an increased potential for the scalability

and applicability of these interventions since the effectiveness of these models did not appear to hinge on any specific sets of two-teacher compositions. That opens the possibility of relying on the comparatively inexpensive option of employing para-professional educators.

In contrast to previously discussed alternative policy options for improving the student-teacher ratio, 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. Noticeable, collaborative instruction works equally well as other structural interventions, including increased instruction time (Kidron & Lindsay, 2014, p. 5, with $\bar{g}$ ranging from -0.04 to 0.16 across literacy and math subjects) and class reduction (Filges et al., 2018, p. 10, $\bar{g} = 0.11$, 95% CI[0.05, 0.16, $p = 0.0003$]).

Albeit 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 think that collaborative instruction can, indeed, function as a contributing factor, adding further improvement to the educational achievements of special needs students together with other relevant interventions aiming at increasing student achievement for this group of students (Dietrichson et al., 2020, 2021).

Conclusion

The findings of this systematic review and meta-analysis provide evidence for the effectiveness of collaborative models of instruction on students' academic achievement. This pertains to all students and 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 of this field of literature. A range of potentially relevant moderators was not possible to analyze, e.g., since they are inadequately documented in the present research literature. Consequently, future studies should assign more weight to study such moderators, and policy-makers should bear in mind this gap in existing evidence.

References

- Achilles, C., 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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1622473702?accountid=14468> NS
- 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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304817903?accountid=14468> NS
- 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., Knoth, 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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/622077437?accountid=14468> NS
- 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>
- 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/Campbell Policies and Guidelines v4-1559660867160.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 on co-teaching*.
<https://www.youtube.com/watch?v=4UdXUJQ4PU>
- 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]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/2164271859?accountid=14468> NS
- Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. (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_8kJiIUdbongK99sx7RFI-/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 statewide 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*, 1–15. <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). Research on the effects of team teaching upon two secondary school teachers. *Educational Research*, 48(2), 177–194. <https://doi.org/10.1080/00131880600732272>
- Jang, S.-J. (2006b). 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>
- Joshi, M., & Pustejovsky, J. E. (2022). *wildmeta: Cluster wild bootstrapping for meta-analysis*. <https://github.com/meghapsimatrix/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*, 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]. In *ProQuest*

Information & Learning (US).

<https://search.proquest.com/docview/1817570306?accountid=14468> NS

- Kidron, Y., & Lindsay, J. (2014). *The effects of increased learning time on student academic and nonacademic outcomes: Findings from a meta-analytic review*. 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>
- LaFever, K. M. (2012). The effect of co-teaching on student achievement in ninth grade physical science classrooms [University of Missouri – St. Louis]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1697496661?accountid=14468> NS
- 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*. <https://search.proquest.com/docview/62200956?accountid=14468> NS
- 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]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/1614376819?accountid=14468> NS

- Mathieu, L. (2019). An examination of special education instructional programs for English learners in New York City schools [Teachers College, Columbia University]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/2279940069?accountid=14468> NS
- 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.
- 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>
- NCLB. (2002). *No Child Left Behind (NCLB) Act of 2001*. 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
- 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]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304897842?accountid=14468> NS
- 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>
- 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, Vienna, Austria. <https://www.r-project.org/>
- 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., Boston, MA. <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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1640768820?accountid=14468> NS
- 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*.
<https://doi.org/10.1093/alcalc/agaa144>
- 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>
- 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., 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, l4898. <https://doi.org/10.1136/bmj.l4898>
- 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>
- The World Bank. (2022). *World Bank country and lending groups*.
<https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>
- 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). *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, 7-10 June 1994*.
<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. 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>
- 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>
- Vemby, M. H., Pustejovsky, J. E., & Pigott, T. D. (2022). *Power approximations for overall average effects in meta-analysis with dependent effect sizes*. MetaArXiv.
<https://doi.org/10.31222/osf.io/6tp9y>
- 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>

- 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>
- WWC. (2020). *WWC procedures and standards handbook (4.1)*. Institute of Education Sciences. <https://ies.ed.gov/ncee/wwc/Handbooks>
- 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

Appendix 1: Studies Included in Meta-Analysis

- Achilles, C. M. (1993). *The teacher aide puzzle: Student achievement issues. An exploratory study* (pp. 1–33). <https://search.proquest.com/docview/62802134?accountid=14468> NS
- Adams, S. S. (2014). Coteaching in secondary special and general education classrooms and student mathematics achievement [Walden University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1622473702?accountid=14468> NS
- Affleck, J. Q., Madge, S., Adams, A., & Lowenbraun, S. (1988). Integrated classroom versus resource model: Academic viability and effectiveness. *Exceptional Children*, 54(4), 339–348. <https://doi.org/10.1177/001440298805400408>
- 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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304817903?accountid=14468> NS
- 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>
- 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]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/622077437?accountid=14468> NS
- 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>
- Beam, A. P. (2005). The analysis of inclusion versus pullout at the elementary level as determined by selected variables [The George Washington University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304999265?accountid=14468> NS

- Belmarez, B. L. (1998). The relationship between co-teaching and the mathematic achievement of groups of seventh-grade students with and without learning disabilities [Texas A&M University - Kingsville]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304482862?accountid=14468> NS
- Berry, P. J. (2018). Co-teaching vs. resource room: Which yields better growth on oral reading fluency measures? [Indiana University of Pennsylvania]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/2155990010?accountid=14468> NS
- Boeckel, A. L. (2008). Effect of the collaborative teaching classroom on reading achievement for speech-impaired elementary students [Walden University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304380512?accountid=14468> NS
- Busch, C. W. (2014). Effects of coteaching instruction between a speech pathologist and first grade teachers [Walden University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1724446029?accountid=14468> NS
- Carlson, H. L., & Others, A. (1984). *Servicing low achieving pupils and pupils with learning disabilities: A comparison of two approaches*.
<https://files.eric.ed.gov/fulltext/ED283341.pdf>
- Case-Smith, J., Weaver, L., & Holland, T. (2014). Effects of a classroom-embedded occupational therapist-teacher handwriting program for first-grade students. *American Journal of Occupational Therapy*, 68(6), 690–698.
<https://doi.org/10.5014/ajot.2014.011585>
- Castro, V. E. (2007). The effect of co-teaching on academic achievement of K–2 students with and without disabilities in inclusive and noninclusive classrooms [Fordham University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304883065?accountid=14468> NS
- Clements, T. P. (2012). How does inclusion affect african american middle school special education student performance? A comparison of disability categories [Union University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1651828510?accountid=14468> NS
- Cox, B. E. (1999). The effects of inclusion of students with learning disabilities on urban fourth-grade general education [Old Dominion University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/619444006?accountid=14468> NS

- 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]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/2164271859?accountid=14468> NS
- 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.
- Gale, P. F. (2005). Performance of students with specific learning disabilities in co-taught and pullout models of special education [The George Washington University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304997690?accountid=14468> NS
- Garcia, R. D. (2020). Effects of integrated co-teaching on 9th grade general education math students [Fairleigh Dickinson University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/2388699142?accountid=14468> NS
- Haselden, K. G. (2004). Effects of co-teaching on the biology achievement of typical and at-risk students educated in secondary inclusion settings [The University of North Carolina at Charlotte]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/305082049?accountid=14468> NS
- James, K. L. (2015). The impact of co-teaching on general education students in seventh grade math [Liberty University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1697922645?accountid=14468> NS
- Jang, S.-J. (2010). The impact on incorporating collaborative concept mapping with coteaching techniques in elementary science classes. *School Science and Mathematics*, 110(1), 86–97.
<https://doi.org/10.1111/j.1949-8594.2009.00012.x>
- Johnson, B. N. (2013). The impact of coteaching on end-of-course test scores [Tennessee State University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1773215880?accountid=14468> NS
- Juettmeyer, M. L. (2012). Coteaching and student academic performance in reading [Walden University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1095131227?accountid=14468> NS
- Kersey, D. A. (2012). Collaborative science work in the elementary classroom [Walden University]. In *ProQuest Dissertations and Theses*.

- <https://search.proquest.com/docview/1773215414?accountid=14468> NS
- Kimani, C. (2018). The impact of co-teaching for ELLs on student achievement and teacher and leader perceptions [Missouri Baptist University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/2115818755?accountid=14468> NS
- Kofahl, S. (2016). The effect of co-teaching on students with disabilities in mathematics in an inclusion classroom [Trevecca Nazarene University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/1905877564?accountid=14468> NS
- LaFever, K. M. (2012). The effect of co-teaching on student achievement in ninth grade physical science classrooms [University of Missouri – St. Louis]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1697496661?accountid=14468> NS
- Laffitte Jr., L. (2012). A comparison of pull-out and co-teaching models on the reading performance of third through fifth grade elementary students with a diagnosed specific learning disability in reading [Pepperdine University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1112071602?accountid=14468> NS
- 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*.
<https://search.proquest.com/docview/62200956?accountid=14468> NS
- Letcher, L. A. (2012). Examining the effects of coteaching on secondary language arts students [Southwest Minnesota State University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1030120144?accountid=14468> NS
- Lynch-Phillips, A. M. (2019). Special education teaching effectiveness: A comparative study of resource rooms and co-teaching [Capella University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/2406658764?accountid=14468> NS
- Marston, D. (1996). A comparison of inclusion only, pull-out only, and combined service models for students with mild disabilities. *The Journal of Special Education*, 30(2), 121–132.
<https://doi.org/10.1177/002246699603000201>
- Marston, D., & Heistad, D. (1994). Assessing collaborative inclusion as an effective model for the delivery of special education services. *Diagnostic*, 19(4), 51–67.
<https://doi.org/10.1177/073724779401900404>
- Mason, P. L. (2013). Comparing types of student placement and the effect on achievement for students with disabilities [Liberty University]. In *ProQuest Information & Learning (US)*.

<https://search.proquest.com/docview/1614376819?accountid=14468> NS

Mathieu, L. (2019). An examination of special education instructional programs for English learners in New York City schools [Teachers College, Columbia University]. In *ProQuest Information & Learning (US)*.

<https://search.proquest.com/docview/2279940069?accountid=14468> NS

Maultsby-Springer, B. M. (2009). A descriptive analysis of the impact of co-teaching on the reading/Language Arts and math achievement of selected middle school students in a Middle Tennessee school district [Tennessee State University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/613688517?accountid=14468> NS

McKelvey, E. E. (2019). The effect of co-teaching on reading achievement of students with and without disabilities [Oral Roberts University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/2371187516?accountid=14468> NS

Mote, S. Y. (2010). Does setting affect achievement of students with disabilities: comparing co-teaching to resource [Liberty University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/964170365?accountid=14468> NS

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>

Nash-Aurand, T. (2013). A comparison of general education co-teaching versus special education resource service delivery models on math achievement of students with disabilities [Liberty University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1773213717?accountid=14468> NS

Neugebauer, N. G. (2008). TAKS scores of general education students in secondary co-teach classes in a Texas school district [Texas A&M University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/89250196?accountid=14468> NS

Parker, A. K. (2010). The impacts of co-teaching on the general education student [University of Central Florida]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/889928197?accountid=14468> NS

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

- Parrello, J. (2010). The effects of co-teaching on the academic achievement of general education students [Caldwell College]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/193653348?accountid=14468> NS
- Perez, J. (2020). Impact of placement in cotaught or self-contained classrooms on student growth in reading and behavioral incidents in a Chicago suburban high school [University of St. Francis]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/2392582063?accountid=14468> NS
- Phipps, O. L. (2015). The effect of researched-based practices on reading achievement of title I students [Walden University]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/1797554570?accountid=14468> NS
- 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]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304897842?accountid=14468> NS
- Principato, K. (2010). A quantitative study of the added-value of co-teaching models implemented in the fourth grade classes of a suburban New Jersey school district [Temple University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/607931601?accountid=14468> NS
- 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>
- Riedesel, D. R. (1997). Effects of a “co-teaching inclusion model” on the achievement levels of eighth-grade regular education students [University of Houston]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304380602?accountid=14468> NS
- Rigdon, M. B. (2010). The impact of coteaching on regular education eighth grade student achievement on a basic skills algebra assessment [Walden University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/860367598?accountid=14468> NS
- Rosman, N. J. S. (1994). Effects of varying the special educator’s role within an algebra class on math attitude and achievement [University of South Dakota]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/62701265?accountid=14468> NS

- Saint-Laurent, L., Dionne, J., Giasson, J., Royer, É., Simard, C., & Piérard, B. (1998). Academic achievement effects of an in-class service model on students with and without disabilities. *Exceptional Children*, 64(2), 239–253. <https://doi.org/10.1177/001440299806400207>
- Schaefer, R. J. (2014). Exploration of co-teaching in inclusive fourth-grade classrooms as a viable option for school districts [Indiana University of Pennsylvania]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1640768820?accountid=14468> NS
- Scheetz, J. A. (1994). The effects of co-teaching on reading achievement levels of mildly handicapped elementary students [Wayne State University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304133462?accountid=14468> NS
- Schulte, A. C., Osborne, S. S., & McKinney, J. D. (1990). Academic outcomes for students with learning-disabilities in consultation and resource programs. *Exceptional Children*, 57(2), 162–172.
- Shaw, F. R. (2002). Academic achievement of students with disabilities in co-teaching, resource room, and support facilitation models [Florida Atlantic University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/276297958?accountid=14468> NS
- Simonovski, E. B. (2015). The co-teaching model and its impact on the academic gains of high school students with disabilities [The Claremont Graduate University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/1803471524?accountid=14468> NS
- Smith, C. S. (1992). Effects of elementary general education/special education team teaching on students' academic gains, social competence, and school adjustment [Michigan State University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/303975151?accountid=14468> NS
- Southwick, K. E. (1998). The effects of the class within a class collaborative/co-teaching model on the achievement of general education students in grades three, four and five [University of Kansas]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/304420847?accountid=14468> NS
- St. John, M. M. (2015). The influence of placement in a co-taught inclusive classroom on the academic achievement of general education students on the 2014 New York State ELA and mathematics assessments in grades 6-8 in a suburban New York school district [Seton Hall

- University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1733230787?accountid=14468> NS
- Tam, I. O. L., & Leung, C. (2019). Evaluation of the effectiveness of a literacy intervention programme on enhancing learning outcomes for secondary students with dyslexia in Hong Kong. *Dyslexia*, 25(1), 296–317. <https://doi.org/10.1002/dys.1626>
- Trabucco, M. (2011). The influence of co-taught inclusion in the academic achievement of third grade non-disabled students in mathematics [Seton Hall University]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/921294068?accountid=14468> NS
- Tremblay, P. (2013). Comparative outcomes of two instructional models for students with learning disabilities: inclusion with co-teaching and solo-taught special education. *Journal of Research in Special Educational Needs*, 13(1), 251–258. <https://doi.org/10.1111/j.1471-3802.2012.01270.x>
- Tsai, P.-L. (2009). Best practices of team teaching by native speaker teachers and non-native speaker teachers in Taiwanese junior high school English classes [Alliant International University, San Diego]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/305168706?accountid=14468> NS
- Walker Harris, L. (2009). Team teaching: The impact on students with disabilities in a middle school setting [Capella University]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/622091475?accountid=14468> NS
- Welch, M., Richards, G., Okada, T., Richards, J., & Prescott, S. (1995). A consultation and paraprofessional pull-in system of service delivery: A report on student outcomes and teacher satisfaction. *Remedial and Special Education*, 16(1), 16–28.
<https://doi.org/10.1177/074193259501600103>
- Whichard, S. M. (2002). Ecological variables and student outcomes in general education/special education [North Carolina State University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/305538331?accountid=14468> NS
- Whisted, M. L. (2011). Does the use of co-teaching models in algebra result in an increase in student achievement among students with disabilities and their non-disabled peers? [College of Notre Dame of Maryland]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/870954775?accountid=14468> NS

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

Witcher, M., & Feng, J. (2010). *Co-teaching vs. solo teaching: comparative effects on fifth graders' math achievement*. <https://files.eric.ed.gov/fulltext/ED533754.pdf>

Wright, R. (2014). The academic impact of co-teaching on non-disabled high school Integrated Math I students [Capella University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1615310567?accountid=14468> NS

Zgonc, K. C. (2007). The impact of co-teaching on student learning outcomes in secondary social studies classrooms implementing content enhancement routines [University of Central Florida]. In *ProQuest Information & Learning (US)*.
<https://search.proquest.com/docview/622026072?accountid=14468> NS

Appendix 2: Supplementary Material (Chapter II)

This document contains further analyses related to the paper “*The Effects of Co-Teaching and Related Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis.*” Specifically, we present more detailed information regarding previous reviews of collaborative models of instruction, the risk of bias assessment, the effect size calculation procedure, and the statistical methods used throughout the main paper. Furthermore, we present a range of tables and figures containing further information about the descriptive statistics, the distribution of the effect size estimates, risk of bias assessment features, conducted subgroup analyses, and publication bias tests.

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

<i>Study</i>	<i>Type of review</i>	<i>Intervention(s)</i>	<i>Outcomes</i>	<i>Included study designs</i>	<i>Student sample</i>	<i>Synthesis method</i>	<i>Treatment of dependency</i>	<i>Years included</i>	<i>Number of studies</i>	<i>Risk of bias assessment</i>	<i>Main conclusion</i>
Qualitative reviews											
Scruggs et al. (2007)	Systematic review	Co-teaching	Perceptions of teachers	Interview and observation studies	Mixed, i.e., general and special education students	Narrative	N/R	1995-2006	32	“considered ‘quality indicators’ as represented by Brantlinger et al. (Figure 3, p. 202)” (p. 298)	“it can be concluded that teachers and administrators were satisfied overall, (..) with co-teaching” (p. 411)
Mixed method reviews											
Alborz et al. (2009) AND Farrell et al. (2010)	Systematic review	Teaching assistants (most often including interventions where students are withdrawn from class)	Academic achievement outcomes	Treatment and control group designed studies, only AND Mixed	Mixed, i.e., general and special education students	Narrative	N/R	Unlimited -2008	35	EPPI’s “weight of evidence”	“The findings in relation to TA impacts on participation of pupils with SEN present a mixed picture.” (Alborz et al. 2009, p. 40)
Cook et al. (2017)	Narrative review	Co-teaching	Mixed	Mixed	Special education students	Narrative	N/R	1991-2015	~15	CEC (2014) Quality Indicators	“Our review of the empirical literature indicated that experimental research on co-teaching continues to be sparse and inconclusive.” (p. 246)
Dyssegaard & Larsen (2013)	Systematic review	Co-teaching and teacher assistants	Mixed	Mixed	Mixed	Narrative	N/R	1989-2012	6 (mostly reviews)	“Clearinghouse always makes quality assessments in cooperation with leading researchers in the given field” (p. 46)	Mixed effects of both service delivery models

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

Iacono et al. (2021)	Systematic review	The six co-teaching models	Academic achievement outcomes (assumed)	Mixed	Special education students	Narrative	N/R	2008-2019	21	“Hansworth’s appraisal tool (..) PEDro scale for appraising the quality of RCT (Maher et al. 2003) and McMaster Guidelines for qualitative studies (Letts et al. 2007).” (p. 5)	“Overall, we found the research base for co-teaching to be limited, both in terms of a cohesive body of studies and their quality.” (p. 11)
Lönnqvist & Sundqvist (2016)	Systematic review (of published studies only)	Co-teaching	Mixed	Mixed	Mixed, i.e., general and special education students	Narrative	N/R	2002-2015	13	Non	More beneficial than disadvantageous effects of using co-teaching on student achievement.
Van Garderen et al. (2012)	Systematic review	Co-teaching, consultation, collaborative team, cooperative teaming, or a combination of models (p. 485)	Mostly academic achievement outcomes	Mixed	Mixed, i.e., general and special education students	Narrative	N/R	Unlimited -2012	19	Non	“the lack of student outcome data in strong support of Collaboration” (p. 495)
Welch et al. (1999)	Systematic review (of published studies only)	Team-teaching (simultaneous presence of two educators in a classroom setting [p. 38])	Mixed (see Table 1, p. 40)	Mixed of qualitative and quantitative studies	Special education students	Narrative	N/R	1980-1997	40	Non	“Service delivery to students with special needs such as team teaching and problem-solving teams has not kept pace with their implementation” (p. 46)

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

Zigmond & Magiera (2002)	Narrative review	Co-teaching	Academic achievement outcomes	Treatment and control group designed studies, only	Special education students	Narrative	N/R	1990-1997	4	Non	“Despite the current and growing popularity of co-teaching, research on student outcomes in this service delivery model is very limited”
Quantitative reviews and syntheses											
Khoury (2014)	Systematic review	Co-teaching	Academic achievement outcomes	Treatment and control group designed studies, only	Special needs students	Meta-analysis + meta-regression	No treatment, i.e., assume independence among effect sizes	1992-2013	20	Non	Moderate statistical significant mean effect size, i.e., $g = 0.281$
Murawski & Swanson (2001)	Systematic review	Co-teaching	Mixed between social, behavioral, and academic outcomes	Mixed between single-case and comparison group studies	Special needs students	Meta-analysis	No treatment or aggregated means	1989-1999	6	Non	Large mean effect size, i.e., 0.40
Willet et al. (1983)	Systematic review	All kinds of collaborative models of instruction (see p. 409)	Science outcomes	Treatment and control group designed studies, only	Mixed, (assumed)	Meta-analysis	No treatment or aggregated means	1950-1983	41	Non	Moderate statistical insignificant mean effect size, i.e., $d = 0.06$

Note: N/R = Not relevant

TABLE S1. Overview of previously conducted reviews of collaborative models of instruction

S1. Effect Size Calculation

In this review, we applied a broad range of effect size calculation approaches (Borenstein, 2009; Hedges, 2007; Higgins et al., 2019; Morris, 2008; Morris & DeShon, 2002; Pustejovsky, 2016; Wilson, 2016; WWC, 2020, 2021) because the literature of collaborative models of instruction represents a diverse set of study designs and estimation techniques to deduce treatment effects (mean differences) and sampling variances. We generally computed Hedges's g via

$$g_T = J \times \left(\frac{b}{S} \right) \quad (1)$$

g_T is the effect size standardized on the *total variance* (indicated by the subscript T), i.e., containing both within- and between-cluster (e.g., classroom or school) variance. b is the mean difference, S represents the standard deviation, i.e., the standardizer, and J is a small sample bias correction equal to $1 - 3/(4 \times df - 1)$. We estimate df in different ways. For single-sited studies (i.e., studies with only one treatment and control class from the same school, i.e., a site in this review refers to classrooms) and simple, randomized trials, we calculated $df = N_t + N_c - 2$. Where N_t and N_c are the sample size of the treatment and control group, respectively. For cluster studies/all multi-sited studies (i.e., studies having two or more treatment and control classes), we estimated df via Equation F.1.2 from the WWC Procedures Handbook 4.1 (2020).

We obtained the mean difference, b , from Equation (1) in several ways. Most frequently, b was obtained as a covariate-adjusted mean difference (most frequently pretest-adjusted mean differences) from ANCOVA models, regression models, or pre-post results (i.e., difference-in-differences, henceforth DiD). For studies only presenting results in either ANCOVA, repeated measure ANOVA, or related ANOVA tables, we calculated $b = \sqrt{MS \times \left(\frac{1}{N_t} + \frac{1}{N_c} \right)}$ where MS is the mean square of the treatment. If b was obtainable in multiple ways within the same study, we obtained all possible estimates of b to check for discrepancies.

If studies reported pre-posttest scores on *different scales*, we calculated $g_T = g_{post} - r(g_{pre})$, where g_{post} and g_{pre} are the standardized mean difference of the posttest and pretest

scores, respectively. r is the pre-posttest correlation. If studies did not report the pre-posttest correlation (as they rarely do for different-scaled outcomes), we imputed $r = 1$, as suggested by WWC (2020, p. E-6). If studies provided raw data, we estimated effect sizes and variance components by fitting standardized linear regression models using all relevant covariates available. If cluster information was given (as in Achilles et al., 2008), we fitted multi-level models guarding against any misspecification via cluster robust variance estimation using the `clubSandwich` package in R (Pustejovsky, 2020b).

We most frequently obtained S in Equation (1) via the post-test standard deviations reported separately for the treatment and control groups, respectively. This is

$$S = \sqrt{\frac{(N_t - 1) \times SD_t^2 + (N_c - 1) \times SD_c^2}{N_t + N_c - 2}}$$

SD_t and SD_c represent the standard deviation for the treatment and control group, respectively. When studies reported the total posttest standard deviation across the treatment and control groups, we used this quantity.

For three studies (Affleck et al., 1988; Nash-Aurand, 2013; Rosman, 1994), we calculated $S = \sqrt{\left[\frac{MS_{error}}{1-R^2} \right] + \left[\frac{df_{error}-1}{df_{error}-2} \right]}$. Where MS_{error} is the ANCOVA mean square error, df_{error} is the model error degrees of freedom, and R is the correlation between the covariate(s) and the dependent variable. For the two latter studies, we imputed $R = .5$, and assessed the overall risk of bias to be serious, since r heavily impacts the size of the effect. For two studies (Mathieu, 2019; Muijs & Reynolds, 2003), we obtained S from level-specific variance components, (i.e., student- and school-level variance components) so that $S = \sqrt{S_{stud}^2 + S_{school}^2}$. For one study (St. John, 2015), we were unable to obtain the standard deviation of the posttest scores. Therefore, we calculated effect sizes using the pretest standard deviation as the standardizer.

A general formula expressing how we obtained the sample variance of g_T can be written as (Pustejovsky, 2016; Pustejovsky & Rodgers, 2019)

$$V_{g_T} = J^2 \times (W + P) \quad (2)$$

where the first term W in the parenthesis is the scaled/standardized sampling variance of b from Equation (1), i.e., $\left(\frac{se_b}{S}\right)^2$, which expresses the contribution of the variability of b , whereas P “captures the contribution of the variance of $[S]$ —that is, how precisely estimated is the *scale* [*standard deviation*] of the outcome” (Pustejovsky & Rodgers, 2019, p. 59). We applied several approaches to estimating W tailored to the specific study design and estimation technique (Pustejovsky, 2016). For studies reporting DiD and/or ANCOVA estimates, we estimated

$$W_{DiD} = 2 \times (1 - r) \times \left(\frac{1}{N_t} + \frac{1}{N_c}\right) \text{ and } W_{ANCOVA} = (1 - r^2) \times \left(\frac{1}{N_t} + \frac{1}{N_c}\right) \quad (3)$$

where r is either the pre-posttest or covariate-outcomes correlation. Only three studies reported r , while we were able to estimate r from 19 studies either by using Equation 31 from Wilson (2016) or the r -equation from Pustejovsky (2020a). Whenever F -values were reported from ANCOVA and ANOVA (i.e., effect sizes from 14 studies), we calculated $W = \frac{g_T^2}{F}$. This made it possible to *reliably* calculate W in cases where it was impossible to obtain an estimate of r . For one study (Wright, 2014), we computed W using the reported t -value, because $t^2 = F$. For studies in which it was neither possible to obtain r or F (i.e., effect sizes from 27 studies), we imputed r following the recommendation from WWC, meaning that we imputed $r = .5$ for DiD and $r = 1$ for ANCOVA models into Equation (3) so that W reduces to

$$\frac{1}{N_t} + \frac{1}{N_c} = \frac{N_t + N_c}{N_t \times N_c} \quad (4)$$

Equation (4) equals W for simple, independent groups designs. This yields a conservative estimate of W for pretest- and/or covariate-adjusted effect size variance estimates, but, in this case, it aims to control the nominal Type-I error rate (WWC, 2020). Equation (4) is incorporated in all variance estimations of g_T based on posttest scores only. For studies only reporting the total sample size across the treatment and control group, we calculated Equation (4) via $\left(\frac{4}{N}\right)$, where $N = N_t + N_c$,

assuming equal sample size across groups. From studies reporting non-standardized regression estimates, we estimated $W = \left(\frac{se\beta}{s}\right)^2$, and for standardized regression models, $W = se_{\beta_{std}}^2$.

We most commonly estimated the scale precision, P , from Equation (2) via $\frac{g_T^2}{2df-2}$. However, for simple, independent group design studies that did not account for clustering of students and reported post-test scores only, we calculated P from WWC's more complex Equation 5.2 and for DiD effect sizes based on different scaled pretest and post-test scores, we used WWC's Equation 3.3 (2021).

If studies reported results across subgroups that were considered to be superficial to the analysis of the review, we aggregated the results (see, e.g., Jang, 2010) to avoid inducing an artificial and unnecessary amount of within-study variability to the review. To further reduce within-study variance, we only retrieved overall test results when possible, which means that we did not calculate effect sizes from all sub-tests if these were reported along with the overall test results. In a similar line, we aggregated test results reported across any subdomains/subtests superficial to the analysis of this review by averaging mean differences and the sampling variance estimates across subscale tests. For example, for Rea et al. (2002), we average test results across reading and writing scores to obtain an overall estimate of the student ability in English language arts. Ideally, this procedure should be conducted by using the between subtest correlations, but these were not obtainable. Notice, therefore, that the reliability of averaging within-study results is based on the assumption of high correlation among subdomain tests. However, we considered the advantage to compensate the deficit of this procedure. The main reason for averaging subtest results is that it ensures increased comparability among studies since most studies report results at the aggregated test level, and secondly, it helps to reduce artificial within-study variability. We aggregated results across subtests for Carlson (1984), Rea (2002), and Schulte (1990).

Unit of Analysis

To ensure that all effect sizes represent the same unit of analysis, i.e., the standardized mean differences (SMD), representing the mean difference standardized/scaled by the *total variance* (i.e., containing both within- and between-cluster variance), g_T (Hedges, 2007; Taylor et al., 2021), we conducted various 2-level conversions of the raw calculated effect sizes.

Conversion of Single-Sited Study Estimates

Effect sizes from single-sited studies (i.e., one treatment and control group from the same school) were converted to represent effect sizes standardized by the total variance by computing $g_T = g_W * \sqrt{1 - \rho_{ICC}}$ and $V_{g_T} = (1 + (n - 1)\rho_{ICC}) \times V_{g_W} \times (1 - \rho_{ICC})$. g_W and V_{g_W} are the small-sample corrected effect size and its sampling variance estimated from the individual student scores only. ρ_{ICC} is the intraclass correlation (ICC) for the given outcome, and n is the average cluster/class size. We imputed ICC values from Hedges & Hedberg's (2007) unconditional models using the corresponding subject (i.e., mathematics or reading) and grade, following the guideline of Hedges (2007) and WWC (2020). We used the reading ICCs for all Arts subjects and the math ICCs for all STEM subjects. If effect sizes were calculated on samples aggregated across grades, we calculated the mean ICC value across the corresponding grades. The average class size, n , for single-sited studies was estimated from the average class size of the two included classes.

Conversion of Results Reported at the Cluster Level

For studies reporting estimates at the cluster/classroom level only (i.e., LaFever, 2012; Southwick, 1998), we first calculated effect sizes standardized by the between-cluster standard deviation, d_B , either from Equations 11 and 12 or 21 and 22 from Hedges (2007), depending on whether the exact class sizes were reported for all included classrooms. Then, we estimated $g_T = J \times d_B \times \sqrt{\rho_{ICC}}$ and $V_{g_T} = J^2 \times V_{d_B} \times \rho_{ICC}$, imputing ρ_{ICC} from Hedges & Hedberg's (2007) unconditional models based on the general student population as well.

Cluster Bias Adjustment

We conducted *approximate cluster design adjustment* for all multi-sited studies (i.e., studies with more than one treatment and control class), including simple RCTs, because collaborative models of instruction are provided at the class-level, which naturally creates dependencies among students sharing the same classroom, teacher, and student composition, etc., independently of the procedure of assignment (Higgins et al., 2019, p. 576). For studies not accounting for nesting of students (e.g., in classrooms or schools), we multiplied the design effect $\eta = (1 + (n - 1)\rho_{ICC})$ to W in Equation (2). ρ_{ICC} was rarely reported in primary studies (only in Achilles et al., 2008; Mathieu, 2019; Muijs & Reynolds, 2003). Consequently, we imputed ICC values from Hedges & Hedberg's (2007) unconditional models using the corresponding subject and grade. If no value of the average

cluster/class size was obtainable, we imputed the average cluster/class size to be 5 for special needs student samples, 18 for samples with general students only, and 23 for blended samples of students, respectively. For studies properly accounting for clustering, we multiplied as upwards bias corrector $\gamma = 1 - \frac{2(n-1)\rho_{ICC}}{N_t + N_c - 2}$ to W in Equation (2). Independently of the cluster treatment, we multiplied the upward-bias corrector $\sqrt{\gamma}$ to g_T in all multisited studies, as suggested by Equation 5.1 in Appendix E of the WWC Procedures Handbook (WWC, 2021). To illustrate the cluster bias correction procedure for the sample variance estimates, Equation (2) can generally be described by

$$V_{g_T} = J^2 \times (W \times \xi + P) \quad (5)$$

where ξ either represents η or γ , as given above. It is important to emphasize that the cluster bias correction is substantially based on approximation, “[h]owever, making no correction for the effects of clustering at all corresponds to assuming that $[\rho_{ICC}] = 0$ is often very far from the case, and thus it may introduce more serious biases in the computation of variances than using values of $[\rho_{ICC}]$ that are slightly in error” (Hedges, 2007, p. 260). A further justification for using approximate cluster bias adjustment is that it guards against lower-quality studies (assuming that lower-quality studies rarely account for clustering of students) getting disproportionately more weight relative to more rigorously conducted studies when estimating average effect sizes. Cluster bias correction was further important to do in order to; 1) most reliably estimate between-study variance, 2) determine the weights used to estimate the overall average effect size, μ , 3) assess the uncertainty of the estimation of μ , and 4) assess the extent of uncertainty in the between-study variance estimate. Whereas robust variance estimation (RVE) can handle scenario 3), the three other scenarios (1, 2, and 4) hinge on the assumption that the sample variance estimation is reasonably accurate.

S2. Mean Effect Size Estimation

Because we expected at the planning stage of the review (see our protocols) both to find correlated and hierarchical effects dependence structures among effect sizes but also because we expected to find true random variation among effect sizes both at the between- and within-study levels across

all models, our models draw on the correlated-hierarchical effects (CHE) working models. In particular, we applied three different random-effects models from the newly developed CHE family (Pustejovsky & Tipton, 2021) to estimate treatment effects and the corresponding variance components. To explicate the models used throughout the paper, assume that we have a collection of J studies, each reporting $k_j \geq 1$ effect size estimates. Then let T_{ij} be the effect size estimate i from study j with a corresponding sample error σ_{ij} , for $i = 1, \dots, k_j$ and $j = 1, \dots, J$. We assumed that T_{ij} is an unbiased estimate of the effect size parameter θ_{ij} and that σ_{ij} are fixed and known. This can be expressed as

$$T_{ij} = \theta_{ij} + e_{ij} \quad (6)$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = s_{ij}^2$. We assumed across all models that effect sizes coming from different studies were uncorrelated, so $\text{cor}(e_{hj}, e_{il}) = 0$ when $j \neq l$.

To explain potential sources of heterogeneity via meta-regression, we assumed the effect size estimates represent a sample from some underlying population of effects and that the average effect sizes can be explained by a set of covariates or predictors (as Pustejovsky & Tipton, 2021). Therefore, let $\mathbf{x}_{ij}$ denote a row vector of p covariates and β denote a vector of p regression coefficients, so that the meta-regression model can be express as

$$T_{ij} = \mathbf{x}_{ij}\beta + u_{ij} + e_{ij} \quad (7)$$

where u_{ij} represent the variation not accounted for by the covariates.

CHE model

To estimate the overall average effect size, $\bar{g}$, we applied the regular CHE model with effect sizes nested in studies. This model usually assumes that there is a constant sample correlation, ρ , between effect size i and m for $i, m = 1, \dots, k_j$ and $j = 1, \dots, J$. The CHE model is given by

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + v_{ij} + e_{ij} \quad (8)$$

where $\text{Var}(u_{ij}) = \tau^2$, $\text{Var}(v_{ij}) = \omega^2$, $\text{Var}(e_{ij}) = s_j^2$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. τ and ω represent the between-study and within-study SDs, respectively, and $s_j^2 = \frac{1}{k_j} \sum_{i=1}^{k_j} s_{ij}^2$. For the intercept-only model, $\mathbf{x}_{ij}$ reduces to a vector of 1's. As mentioned above, the CHE models imply assuming a constant sampling correlation, ρ . However, we obtained ρ by estimating Pearson correlation from all those studies that provided both mathematics and language arts scores, as suggested by Kirkham et al. (2012).

Albeit our data contains ten studies reporting multiple non-overlapping samples, we did not model variability at the sample level of our model(s) since we conducted a *likelihood ratio test* of the variance components, showing that the fit of the four-level model was not significantly better than the three-level model (Viechtbauer, 2022).

S3. Subgroup Analyses and Meta-Regression

For subgroup analyses investigating if the effects of collaborative models of instruction vary as a function of a range of pre-specified categorical predictors/moderators, we either applied the Subgroup Correlated Effects Plus (SCE+) model or the Correlated Multivariate Effects Plus (CMVE+) model. Although the latter model potentially yields more precise estimates and most adequately captures the true dependency among effect sizes within and across subgroup dimensions, simulation results suggest that this model only works under certain narrow conditions (Pustejovsky & Tipton, 2021). We decided whether the CMVE+ model was feasible by using what we call *overlapping tables*, suggested in the supplementary material to Pustejovsky & Tipton (2021). Find these analyses in Tables S2 and S3 and the condition for when the CMVE+ model works reliably in the next section (S4) below. The SCE+ is given by

$$T_{ij} = \sum_{c=1}^c d_{ij}^c (\mathbf{x}_{ij}\beta_c + u_{cj} + v_{cij}) + e_{ij} \quad (9)$$

where $\text{Var}(u_{cj}) = \tau_c^2$, $\text{Var}(v_{cij}) = \omega_c^2$, $\text{Cov}(u_{bj}, u_{cj}) = 0$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2 \sum_{c=1}^C d_{hj}^c d_{ij}^c$. Here d_{ij}^c is an indicator of whether a given effect size falls within the given subgroup. The SCE+ model is based on the assumption that outcomes from the same study falling into the same subgroup category are correlated but outcomes from the same study falling into different subgroup categories are assumed to be independent.

Next, the CMVE+ model is given by

$$T_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \sum_{c=1}^C (d_{ij}^c u_{cj} + d_{ij}^c v_{cij}) + e_{ij} \quad (10)$$

where $\text{Var}(u_{cj}) = \tau_c^2$, $\text{Cov}(u_{bj}, u_{cj}) = \psi_{bc}\tau_b\tau_c$, $\text{Var}(v_{cij}) = \omega_c^2$, $\text{Cov}(v_{bij}, v_{cij}) = \zeta_{bc}\omega_b\omega_c$, and $\text{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. ψ_{bc} and ζ_{bc} are the correlations between the random effects at the study and effect size level, respectively. Unlike the SCE+ model, the CMVE+ model both assumes that effect sizes from the same study falling into different subgroups and effect sizes from the same study falling into the same subgroup are correlated. Contrary to original subgroup analyses that treat each subgroup model as independent (and thereby exclude the opportunity for statistical comparison between subgroup means), these above-presented models allow us to model all subgroups in one model “by interacting the covariates ($\mathbf{x}_{ij}$) with indicators for each [subgroup] category (d_{ij}^c) and similarly interacting the random effect-effects terms (u_{cj}) with indicators for each category” (Pustejovsky & Tipton, 2021). By using these models, we were, therefore, able to conduct reliable (multiple-contrast) Wald tests, including HTZ Wald tests (Tipton & Pustejovsky, 2015), as well as Wald tests based on cluster wild bootstrapping (CWB) with 1999 replications (Joshi et al., 2022). We added the latter tests since recent simulations studies have shown that CWB less conservatively controls Type-I errors and yields more power relative to HTZ (Joshi et al., 2022).

For continuous covariates, we fitted the CHE model presented in Equation (8). All moderators used for investigation were deduced from the theoretical and methodological literature regarding relevant moderating factors such as important theoretical constructs and study features. Find the list of all relevant moderators we expected to investigate in the pre-registered protocol attached to this review.

S4. Model Selection

The main difference between the SCE+ and CMVE+ working models is that the latter allows effect sizes from the same study that falls into different subgroup categories to be correlated. Although superior to the SCE+ working model in terms of statistical accuracy, the CMVE+ model only works adequately under restricted conditions. Specifically, the condition under which the CMVE+ model works is when

- 1) there are few multivariate dimensions
- 2) there are a substantial number of studies and effect sizes available from each dimension.
- 3) there are a substantial number of studies having effect sizes from each possible pair of outcome dimensions.

To decide if the two latter conditions were in place for the CMVE+ model, we applied overlapping tables, as presented below.

TABLE S2. Overlapping table for the subject variable (accepted for CMVE+)

Subject	Arts and Social Science	STEM
Arts and Social Science	54 (168)	28 (85)
STEM	28 (70)	48 (113)

Note: Number of studies and number of effect sizes (in parenthesis) by co-occurrence of dependent variable types.

TABLE S3. Overlapping table for the type of student sample (not accepted for CMVE+)

Sample	Blended	General students	Special needs students
Blended	19 (61)	-	1 (2)
General students	-	26 (79)	12 (32)
Special needs students	1 (2)	12 (32)	42 (141)

Note: Number of studies and number of effect sizes (in parenthesis) by co-occurrence of student samples.

We only fitted the subject covariate to the CMVE+ model since it was the only moderator variable accommodating all of the necessary conditions, as shown in Table S2.

S5. Descriptive Statistics

FIGURE S1. Number of studies included in the meta-analysis by mean grade of study

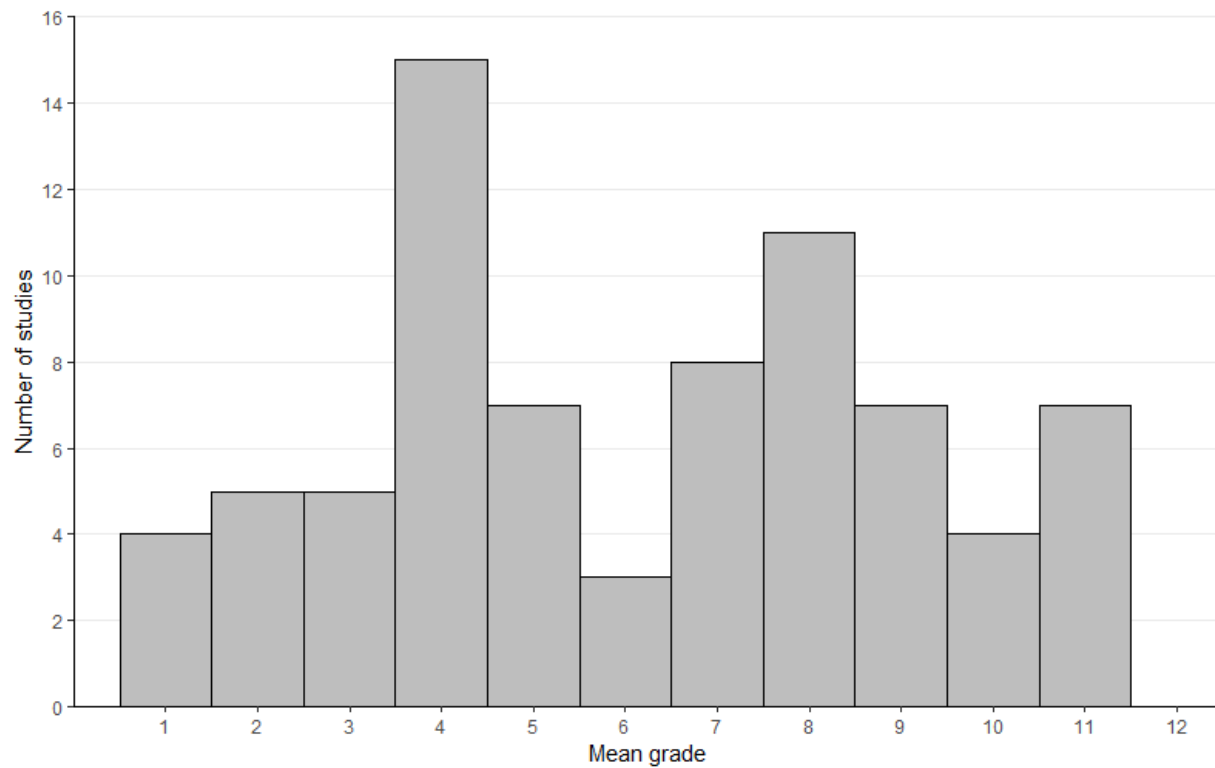


Figure S1 displays the 76 included studies by the mean grade of the study. It shows that most grades are well represented. However, 12th-grade students were absent from the pool of included studies.

FIGURE S2. Primary study sample sizes

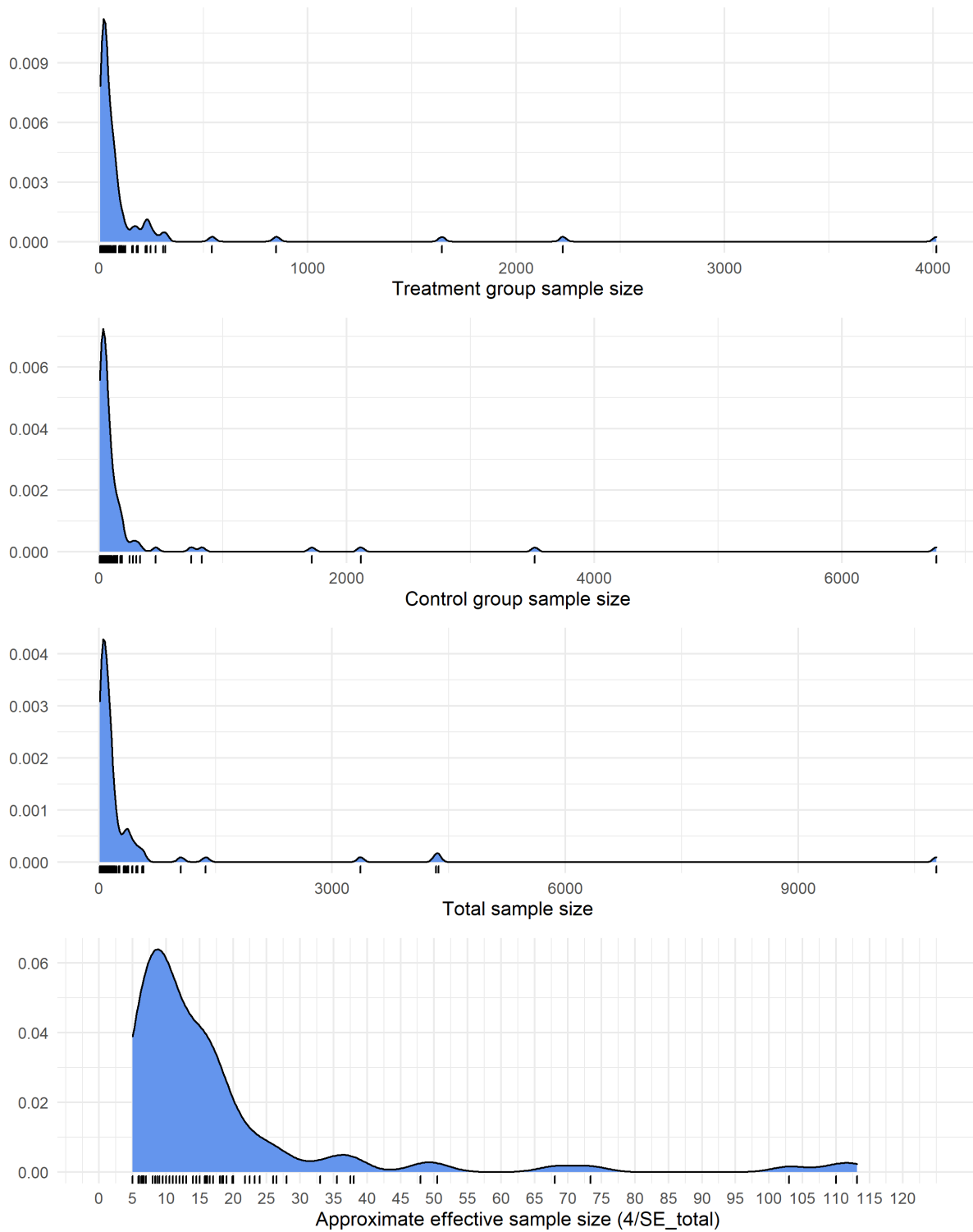


TABLE S4. Distribution of treatment group sample sizes from primary studies

Mean	SD	P0	P25	P50	P75	P100
158	494.60	5	19	37	79	4016

TABLE S5. Distribution of control group sample sizes from primary studies

Mean	SD	P0	P25	P50	P75	P100
232	812.20	5	20	55	120	6765

TABLE S6. Distribution of total sample sizes from primary studies

Mean	SD	P0	P25	P50	P75	P100
391	1286	10	44	102	201	10781

TABLE S7. Distribution of effective sample sizes from primary studies

Mean	SD	P0	P25	P50	P75	P100
18	20.27	5	8	12	18	113

Estimation techniques used in primary studies

Table S8 below describes from which kind of estimation techniques effect sizes were obtained. Most often, pre-test adjustment was conducted via difference-in-differences techniques¹, i.e., 149 (50%) of the effect sizes are calculated from some kind of pre-posttest score means. A substantial amount of the effect sizes were further calculated from some kind of adjusted means, i.e., 69 (23%) effect sizes from ANCOVA and 41 (13%) via regression, respectively. Only 10 (3%) effect sizes were obtained from ANOVA models and 29 (9%) effect sizes from post-test means, respectively.

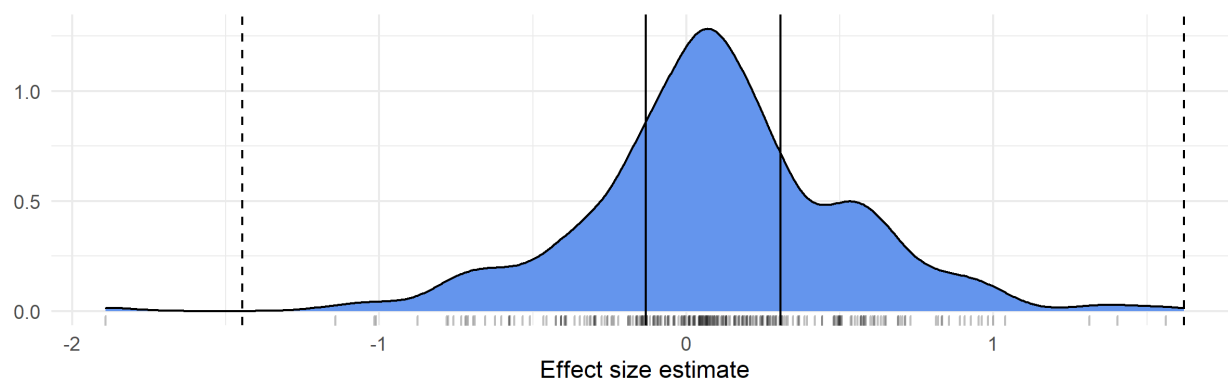
TABLE S8. Further description of effect size characteristics

Effect size characteristics	Studies (<i>J</i>)	Effect sizes (<i>K</i>)	Percentage _{<i>K</i>}
% ES from ANCOVA	17	69	0.238
% ES from ANOVA	4	12	0.041
% ES from Diff-in-Diffs	41	139	0.479
% ES from raw posttest means	10	29	0.1
% ES from regression	11	41	0.141

¹ This also includes studies reporting of gain, growth, and development scores.

Outcomes distributions and outlier tests

FIGURE S3. Empirical distribution of effect size estimates

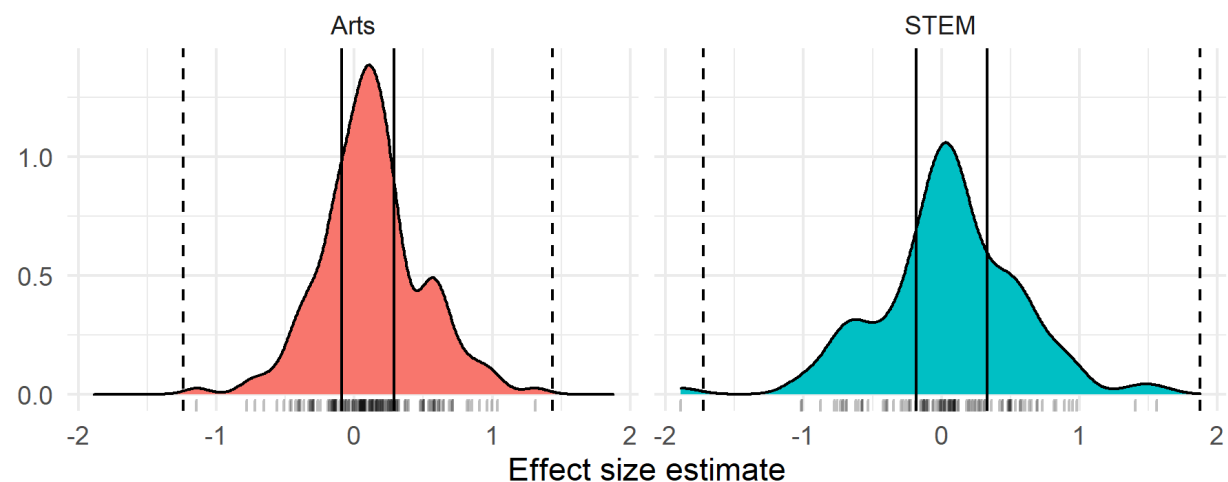


Note: Distribution of the estimated effect sizes. Solid lines indicate the lower and upper quartiles, respectively. “Dashed lines indicate the 1st quartile minus three times the inter-quartile range and the 3rd quartile plus three times the interquartile range. Effect sizes outside of the range of dashed lines would be considered outliers according to Tukey’s (1977) definition” (Winters et al., 2022, supplementary material).

TABLE 9. Distribution of effect size estimates

Mean	SD	P0	P25	P50	P75	P100
0.09	0.42	-1.89	-0.13	0.08	0.31	1.56

FIGURE S4. Empirical distribution of effect size estimates across arts and STEM outcomes

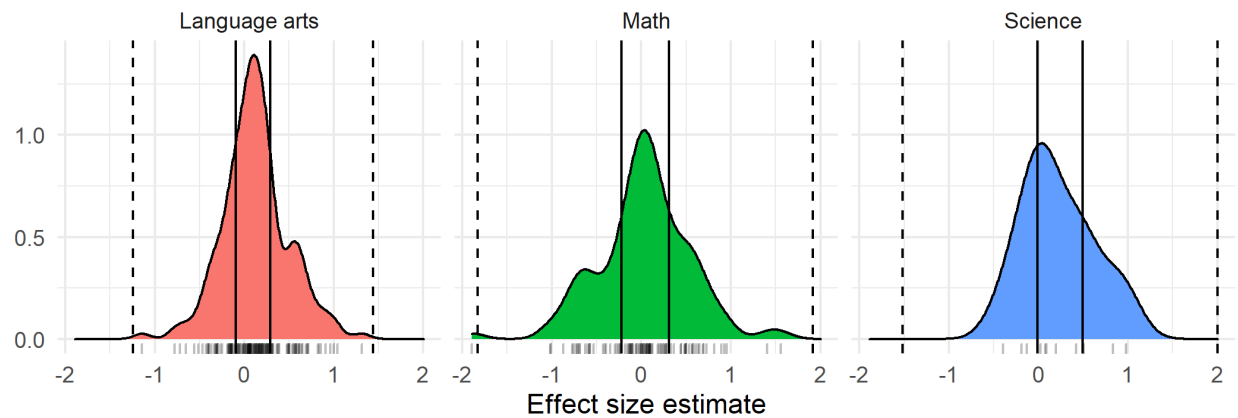


Note: Solid lines indicate the lower and upper quartiles, respectively. “Dashed lines indicate the 1st quartile minus three times the inter-quartile range and the 3rd quartile plus three times the interquartile range. Effect sizes outside of the range of dashed lines would be considered outliers according to Tukey’s (1977) definition” (Winters et al., 2022, supplementary material).

TABLE 10. Distribution of effect size estimates across arts and STEM outcomes

Subject	Mean	SD	P0	P25	P50	P75	P100
Arts	0.12	0.37	-1.14	-0.09	0.10	0.29	1.31
STEM	0.05	0.51	-1.89	-0.18	0.05	.33	1.56

FIGURE S5. Empirical distribution of effect size estimates across various subject outcomes



Note: Solid lines indicate the lower and upper quantiles, respectively. “Dashed lines indicate the 1st quartile minus three times the inter-quartile range and the 3rd quartile plus three times the interquartile range. Effect sizes outside of the range of dashed lines would be considered outliers according to Tukey’s (1977) definition” (Winters et al., 2022, supplementary material).

TABLE 11. Distribution of effect size estimates across various subject outcomes

Subject	Mean	SD	P0	P25	P50	P75	P100
LA	0.13	0.37	-1.14	-0.09	0.11	0.29	1.31
Math	0.03	0.52	-1.89	-0.22	0.05	0.31	1.56
Science	0.22	0.4	-0.4	-0.01	0.09	0.5	0.98

S6. Risk of Bias (RoB) Assessment

Extended description of RoB assessment

We slightly modified the used assessment schemes so that they included further questions in the reporting domain (D7) regarding whether any evidence suggested error-prone reported results (e.g., studies reporting extremely small standard deviations, etc.) or if any measures of variability were retrievable from the study. Studies receiving moderate or serious judgments for four domains or more were assessed to have a serious risk of bias overall. Domains with a substantial amount of

missing information were rated as ‘high’ risk of bias, but studies were not excluded. In addition, we considered primary study data, where 20% of the treatment variable observations were missing in the realized sample, to be of serious risk of bias due to missingness.

For a few studies, we excluded a part of the total number of effects. For example, we excluded effect sizes for “At-risk”-students for Haselden (2004) and Saint-Laurent (1998). In the former case because the treatment for the “At-risk”-students were provided outside the general classroom, and for the latter, because the “At-risk”-student control group was comprised of both students from special education settings and students from single-taught general education settings which made a substantial interpretation of these effects infeasible. However, we had no reason to question the accuracy of the general student comparisons in the respective studies. For Maultsby-Springer (2009), we excluded two out of four effect sizes due to reporting errors that surfaced when we calculated the pre-posttest correlation, ρ , from the reported paired t -tests.² We had firm no reason to suspect the accuracy of the remaining effect sizes, and, therefore, they were included. On the other hand, we excluded one study (Christie, 2020) in which only one result out of twelve effect sizes seemed to be trustworthy based on the paired t -test values.³ The serious ROB assessment was often reached because results substantially diverged, for example, between effect sizes calculated from raw pre-posttest scores and adjusted means and ANCOVA statistics. Only effect sizes from two studies belonging to the RCT family received a high overall RoB assessment. One study (i.e., Garcia, 2020) received high risk of bias due to deviation of the intended intervention, and one study (i.e., Parrello, 2010) because of high risk of bias due to measurement error. Parrello used grade scores from treatment classrooms in which she both taught and graded the students.

Extra RoB figures and tables

Figures S6 and S7 illustrate unweighted summary RoB plots for the ROBINS-I and RoB 2 assesses studies and effect sizes, respectively. Figures S8 and S9 illustrate ROBINS-I RoB plots disaggregated between quasi-experimental and observational studies, respectively. The concrete assessments behind the plots are presented in Tables S12 and S13.

² See the effect size calculation at <https://osf.io/fby7w/>.

³ Find the calculation in effect size calculation document at <https://osf.io/fby7w/>.

FIGURE S6. Unweighted ROBINS-I assessment summary plot

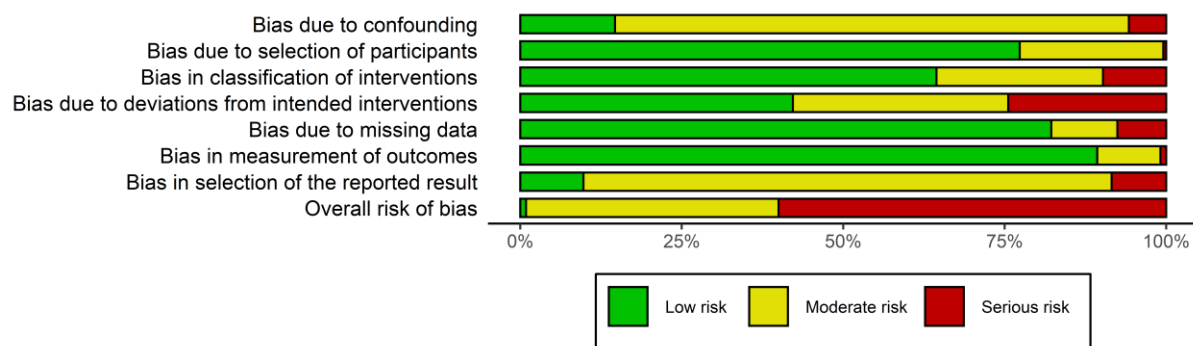


FIGURE S7. Unweighted ROB 2 assessments summary plot

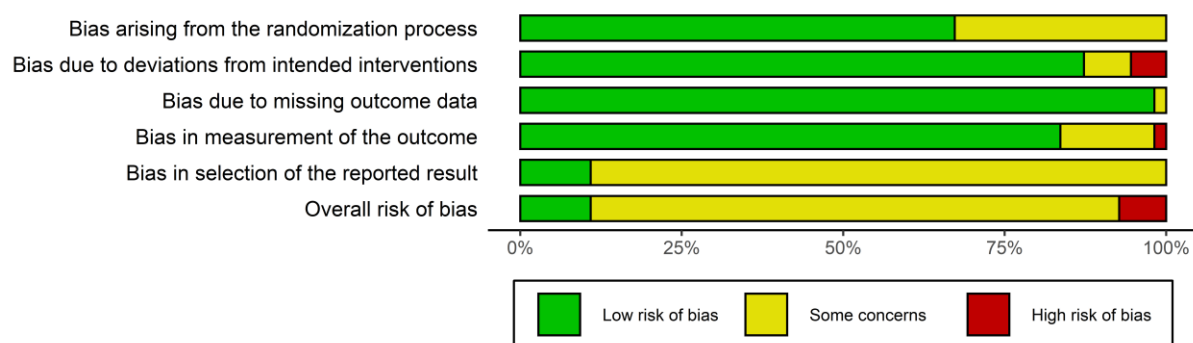


FIGURE S8. Weighted ROBINS-I assessment summary plot for quasi-experimental studies, only

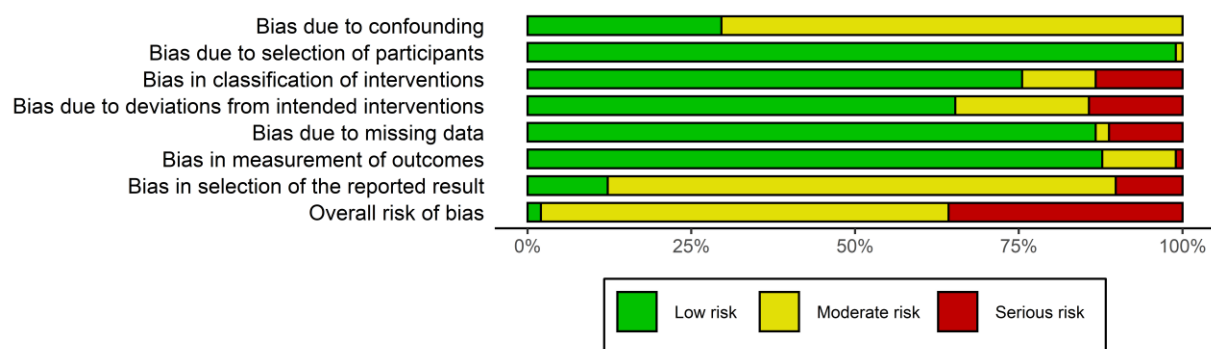


FIGURE S9. Weighted ROBINS-I assessment summary plot for observational studies, only

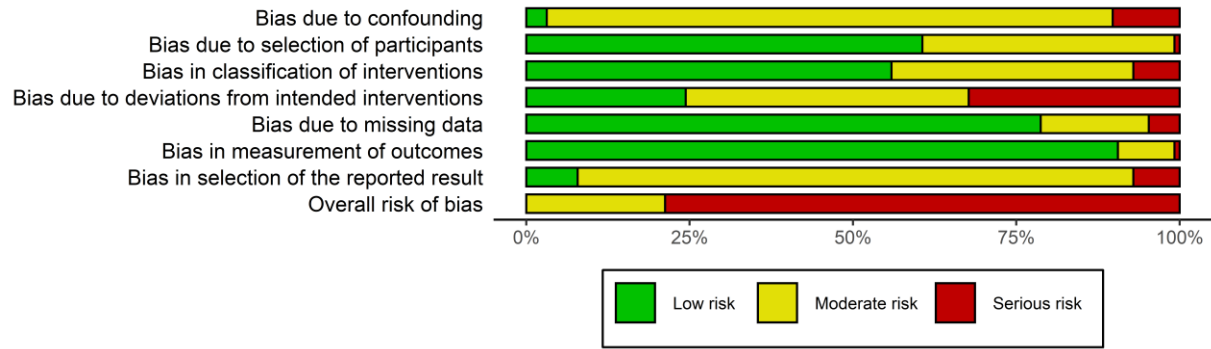


TABLE S12. ROBINS-I Table – percent of effect sizes across domains and judgments

RoB as- sessment	D1 J, K (%)	D2 J, K (%)	D3 J, K (%)	D4 J, K (%)	D5 J, K (%)	D6 J, K (%)	D7 J, K (%)	Overall J, K (%)
Low	5, 33 (14.7%)	50, 174 (77.3%)	44, 145 (64.4%)	28, 95 (42.2%)	54, 185 (82.2%)	62, 201 (89.3%)	6, 22 (9.8%)	1, 2 (0.9%)
Moderate	60, 179 (79.6%)	17, 50 (22.2%)	21, 58 (25.8%)	25, 75 (33.3%)	7, 23 (10.2%)	8, 22 (9.8%)	55, 184 (81.8%)	29, 88 (39.1%)
Serious	6, 13 (5.8%)	1, 1 (0.4%)	8, 22 (9.8%)	14, 55 (24.4%)	6, 17 (7.6%)	2, 2 (0.9%)	8, 19 (8.4%)	45, 135 (60%)

Note. J = number of studies. K = number of effect sizes. NA, i.e., missing information judgments were given a serious risk of bias judgment.

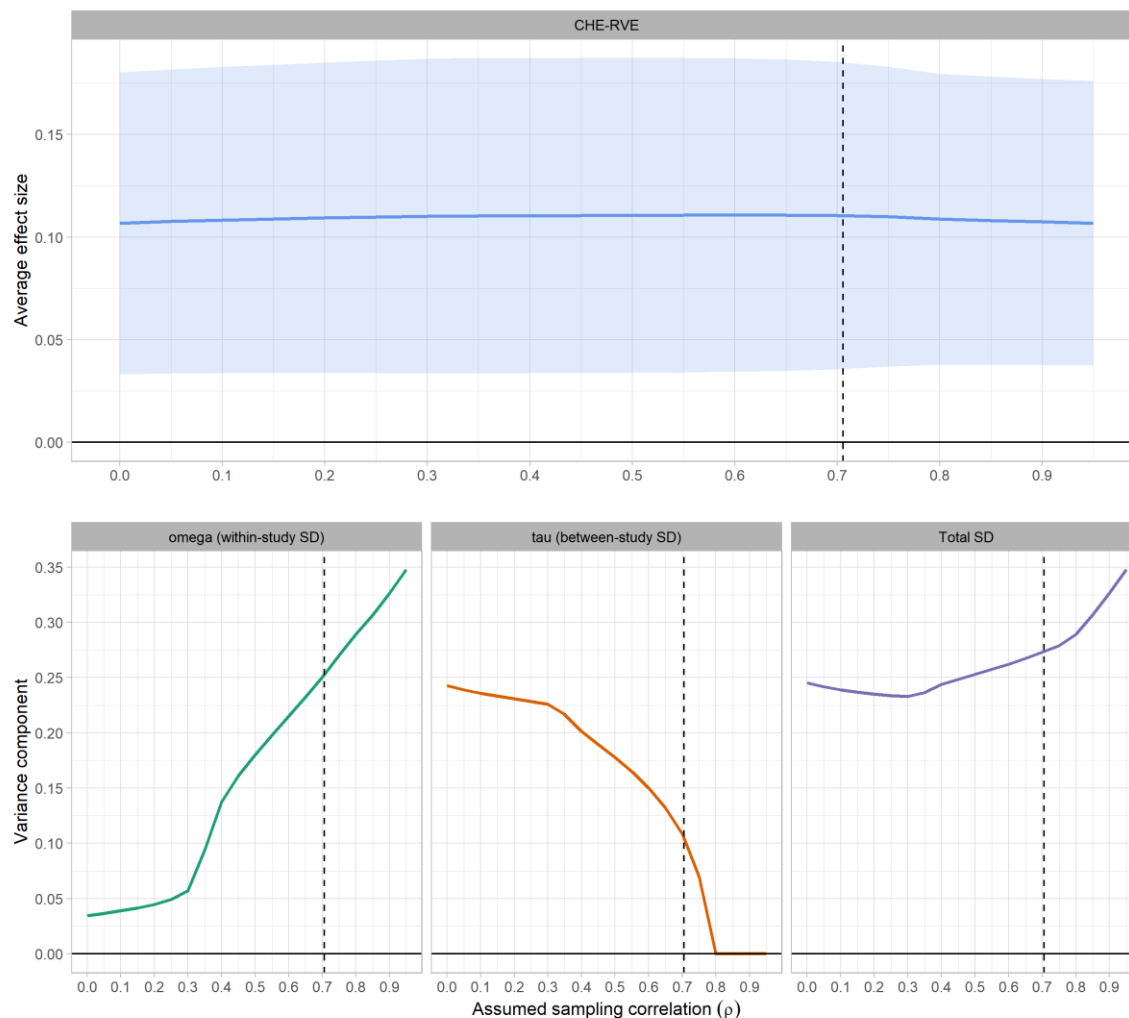
TABLE S13. RoB 2 Table – percent of effect sizes across domains and judgments

RoB assessment	D1 J, K (%)	D2 J, K (%)	D3 J, K (%)	D4 J, K (%)	D5 J, K (%)	Overall J, K (%)
Low	8, 37 (67.3%)	6, 48 (87.3%)	9, 54 (98.2%)	7, 46 (83.6%)	1, 6 (10.9%)	1, 6 (10.9%)
Some concerns	2, 18 (32.7%)	2, 4 (7.3%)	1, 1 (1.8%)	3, 8 (14.5%)	8, 49 (89.1%)	6, 45 (81.8%)
High	NA	1, 3 (5.5%)	NA	1, 1 (1.8%)	NA	2, 4 (7.3%)

Note. J = number of studies. K = number of effect sizes. NA, i.e., missing information judgments were given a serious risk of bias judgment.

S7. Sensitivity Analysis of Mean Effect Size

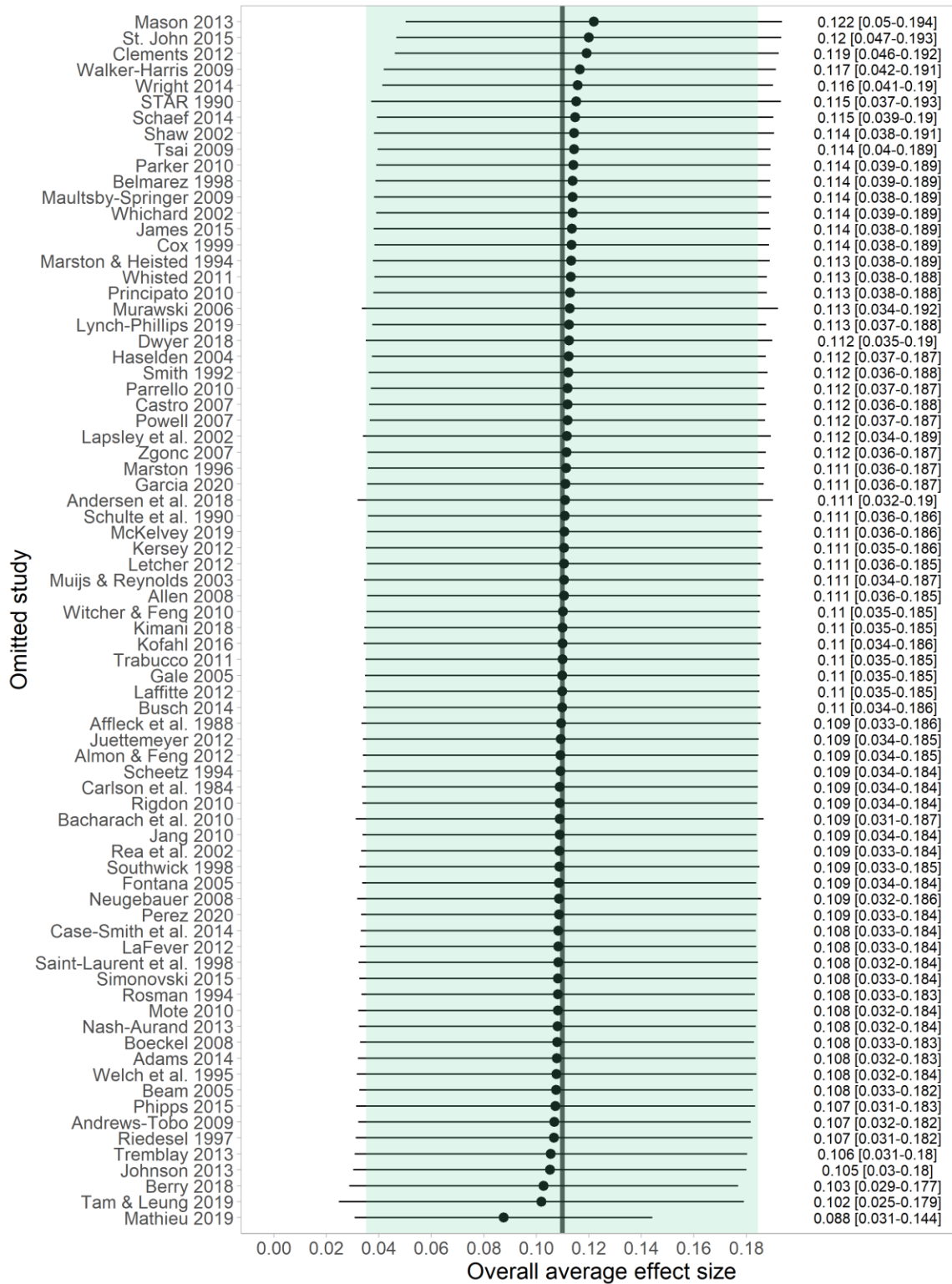
FIGURE S10. Sensitivity of meta-analysis parameter estimates to the assumed value of the sampling correlation (ρ) between effect size estimates.



Model Check – Leave-One-Study-Out Analyses

Figure S11 below displays the impact on the overall mean effect size by leave-one-study-out at a time. Leaving one study out generally does not have any significant impact on the estimation of the weighted overall mean effect size, $\bar{g}$. For all studies, the substantial interpretation of the magnitude of $\bar{g}$ does not change, and all models yield statistically significant results. However, leaving out Mathieu (2019) had quite an impact on the point estimate and its precision. As can be seen from Figure S11, Mathieu did have a substantial significant impact on the between-study SD, τ . In fact, τ dropped to zero when Mathieu was omitted from the analysis. This can suggest that the between-study variance estimation is fragile, but it can also indicate that Mathieu estimated a true different effect relative to the rest of the included studies. Furthermore, leaving out Mathieu did have an impact on the within-study variance, but to a lesser extent than for τ , and it did not change the overall conclusion that a substantial amount of true variation is present in the data. Yet the total SD seemed to be constant, even when leaving Mathieu out, confirming the presence of true random variation within this body of literature. This is further supported by the fact that the I^2 and Q -statistics seem to be more or less insensible to leaving out any study from the weighted mean effect size analysis.

FIGURE S11. Impact of leaving one study out on the average effect size



Note: Dashed lines and shade indicate the estimated values and the confidence interval from the overall average effect size of the CHE model presented in the paper.

FIGURE S12. The impact of leaving one study out on heterogeneity quantities

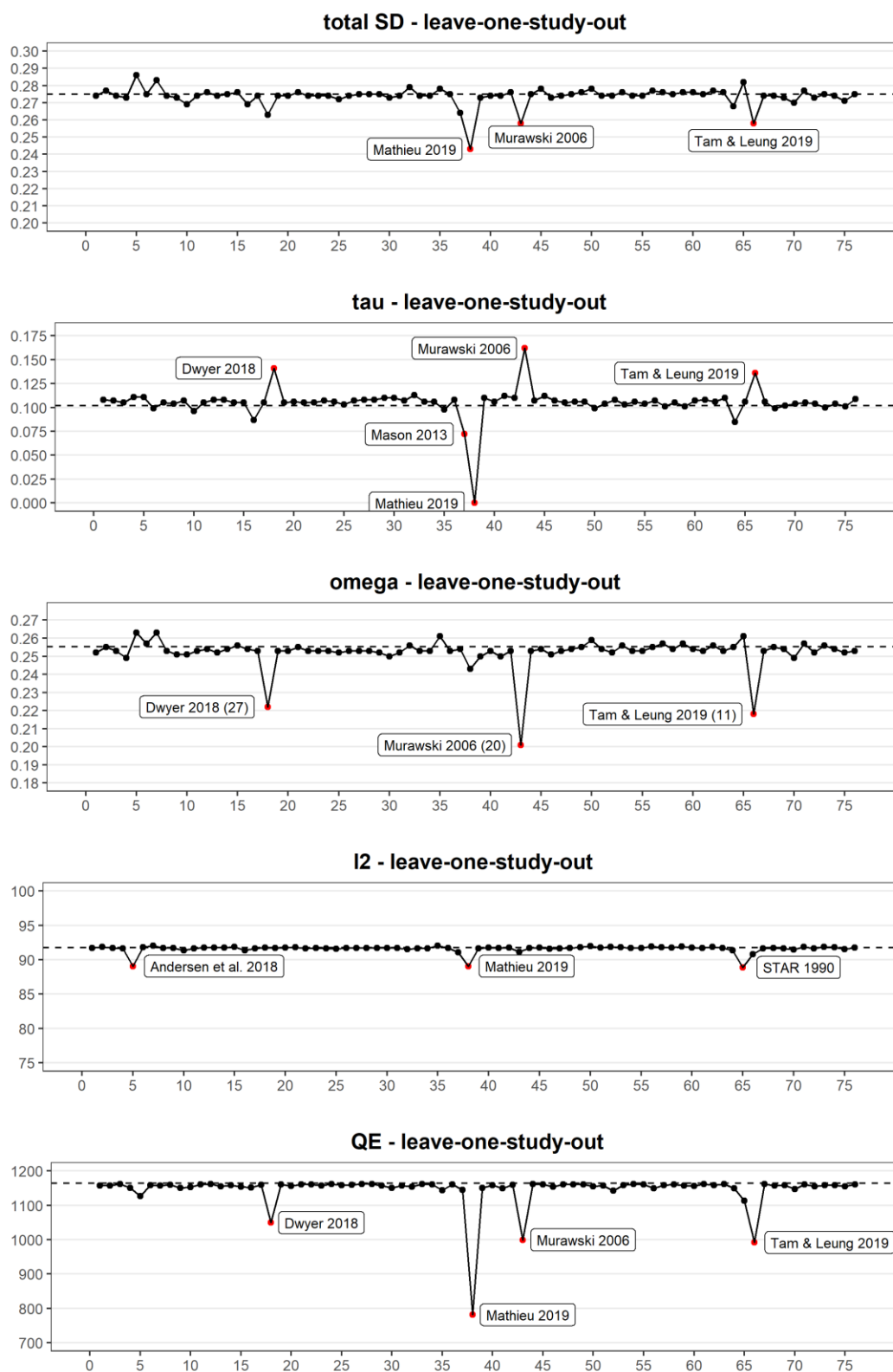
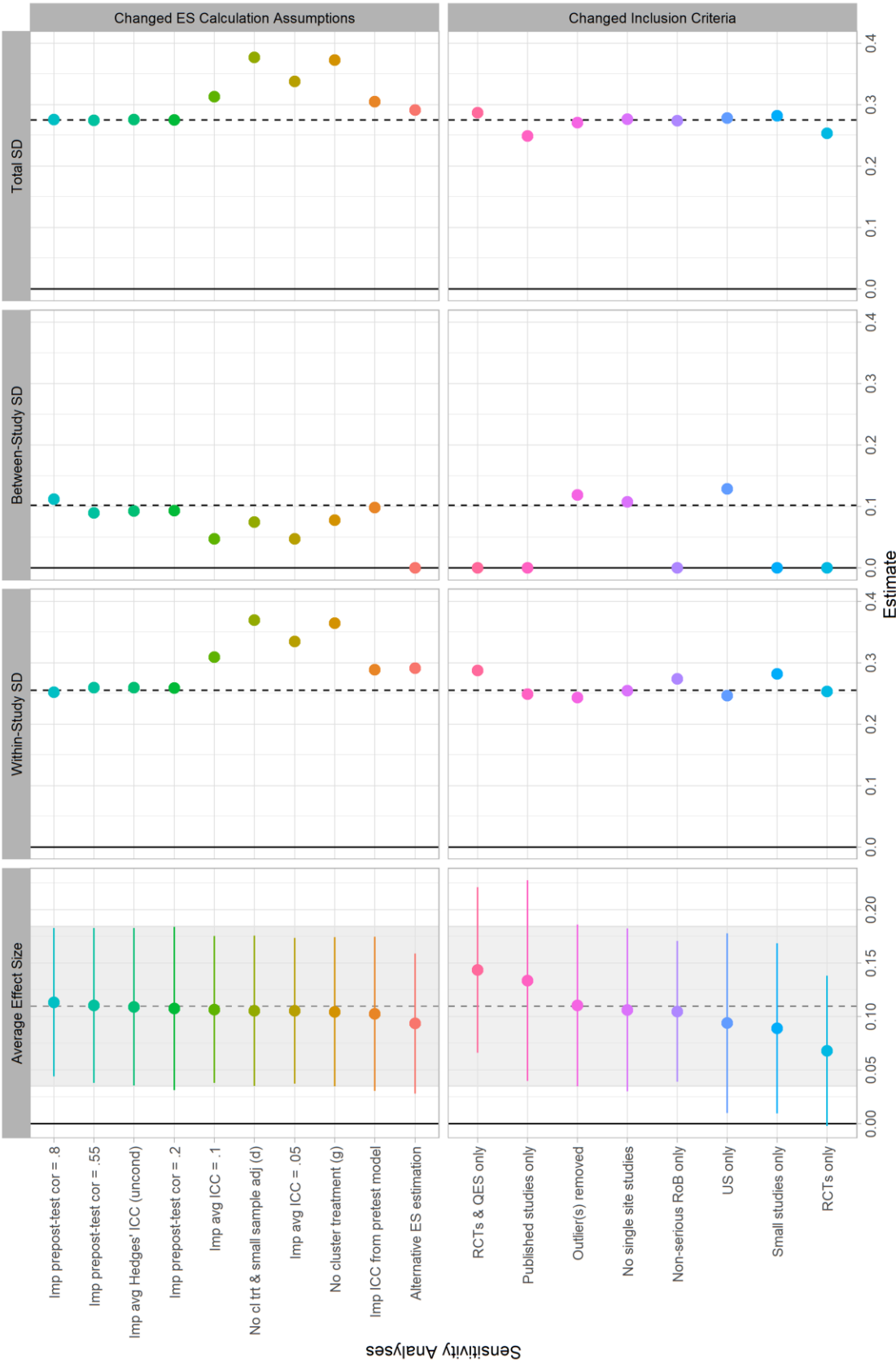


FIGURE S13. Sensitivity analyses changing effect sizes calculation assumptions and inclusion criteria



Note: Dashed lines and shades indicate the estimated values and the confidence interval from the overall average effect size of the CHE model presented in the paper.

S8. Correlation Matrix

TABLE S14. Correlation matrix for covariates

Moderators	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
(1) Aides	1																							
(2) Cotaught	-.79	1																						
(3) Team teach	-.09	-.53	1																					
(4) Special stud	-.16	.21	-.12	1																				
(5) Agg sample	.4	-.51	.28	-.52	1																			
(6) General stud	-.18	.23	-.12	-.62	-.34	1																		
(7) STEM	.06	-.02	-.05	-.23	.03	.22	1																	
(8) Arts & Social Sci.	-.06	.02	.05	.23	-.03	-.22	-1	1																
(9) Primary	.18	-.09	-.1	-.21	.24	.02	-.08	.08	1															
(10) High sch.	-.15	.16	-.04	-.05	-.18	.21	.15	-.15	-.51	1														
(11) Secondary	-.06	-.04	.15	.28	-.11	-.21	-.04	.04	-.66	-.31	1													
(12) Observational	-.17	.11	.05	.3	-.19	-.16	-.06	.06	-.23	-.02	.27	1												
(13) QES	.01	.04	-.08	-.24	.09	.18	.14	-.14	.38	-.22	-.22	-.68	1											
(14) RCT	.2	-.19	.04	-.09	.13	-.01	-.09	.09	-.17	.31	-.08	-.45	-.35	1										
(15) Non standardized	-.15	.08	.08	.14	.03	-.19	-.09	.09	.1	-.07	-.05	.06	-.1	.05	1									
(16) Standardized	.15	-.08	-.08	-.14	-.03	.19	.09	-.09	-.1	.07	.05	-.06	.1	-.05	-1	1								
(17) Special edu crt	-.06	.1	-.08	.75	-.39	-.46	-.19	.19	.02	-.15	.11	.31	-.28	-.06	.18	-.18	1							
(18) General edu crt	.06	-.1	.08	-.75	.39	.46	.19	-.19	-.02	.15	-.11	-.31	.28	.06	-.18	.18	-1	1						
(19) Gray literature	-.3	.23	.04	-.17	-.08	.26	.21	-.21	.02	.02	-.04	.38	-.02	-.46	-.15	.15	-.01	.01	1					
(20) Journal article	.3	-.23	-.04	.17	.08	-.26	-.21	.21	-.02	-.02	.04	-.38	.02	.46	.15	-.15	.01	-.01	-1	1				
(21) Posttest es	.06	-.05	.0	-.03	.01	.02	.04	-.04	.0	-.02	.02	-.01	-.14	.18	.09	-.09	.04	-.04	.08	-.08	1			
(22) Covar-adj es	-.06	.05	.0	.03	-.01	-.02	-.04	.04	.0	.02	-.02	.01	.14	-.18	-.09	.09	-.04	.04	-.08	.08	-1	1		
(23) Low/mod RoB	.19	-.19	.03	-.36	.21	.21	.09	-.09	.14	.12	-.26	-.54	.23	.41	-.26	.26	-.25	.25	-.18	.18	.04	-.04	1	
(24) Serious RoB	-.19	.19	-.03	.36	-.21	-.21	-.09	.09	-.14	-.12	.26	.54	-.23	-.41	.26	-.26	.25	-.25	.18	-.18	-.04	.04	-1	1

Note: Bold numbers indicate correlation above 0.5.

S9. Moderator Forest Plots

FIGURE S14. Forest plot by type of intervention

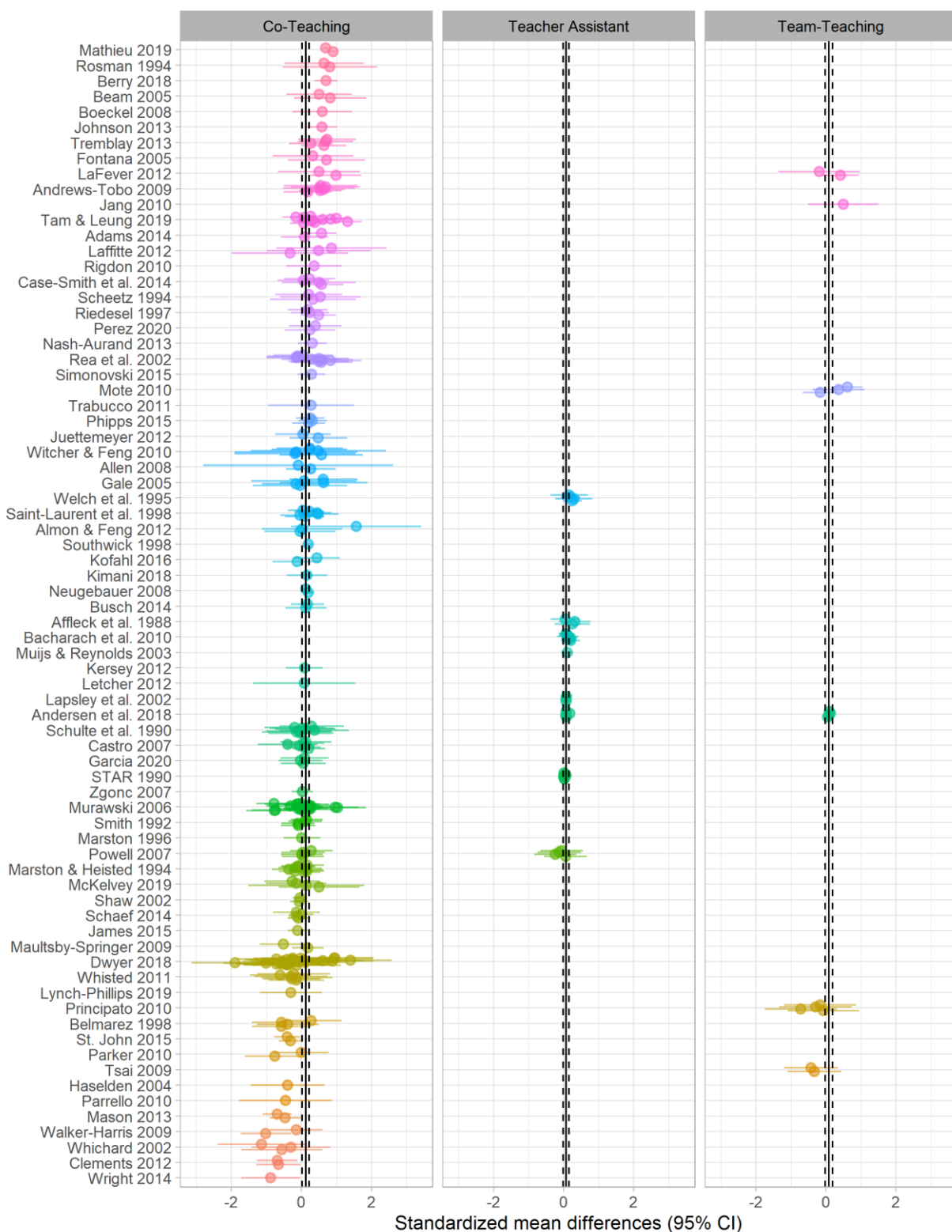


FIGURE S16. Forest plot by subject

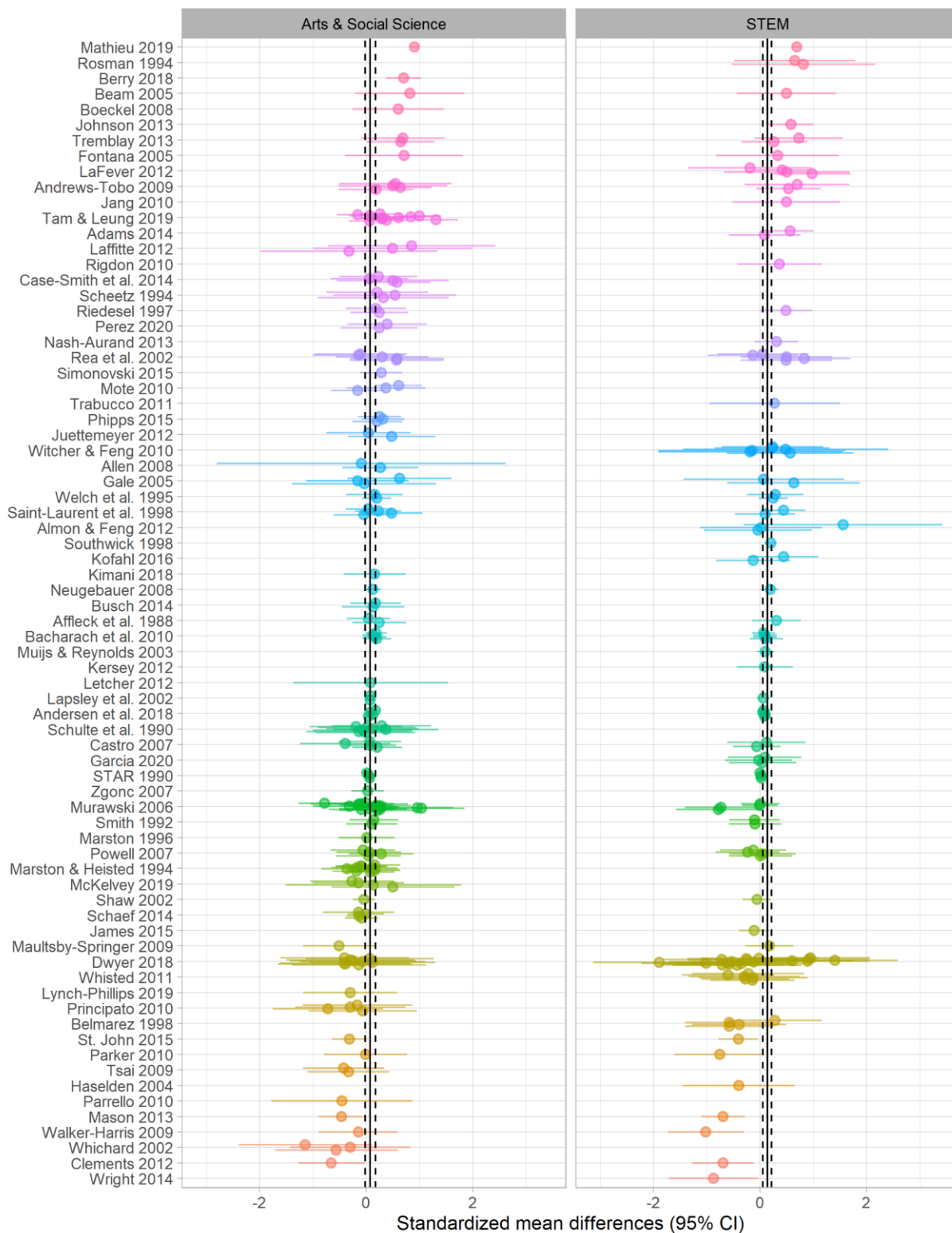


FIGURE S17. Forest plot by type of effect size

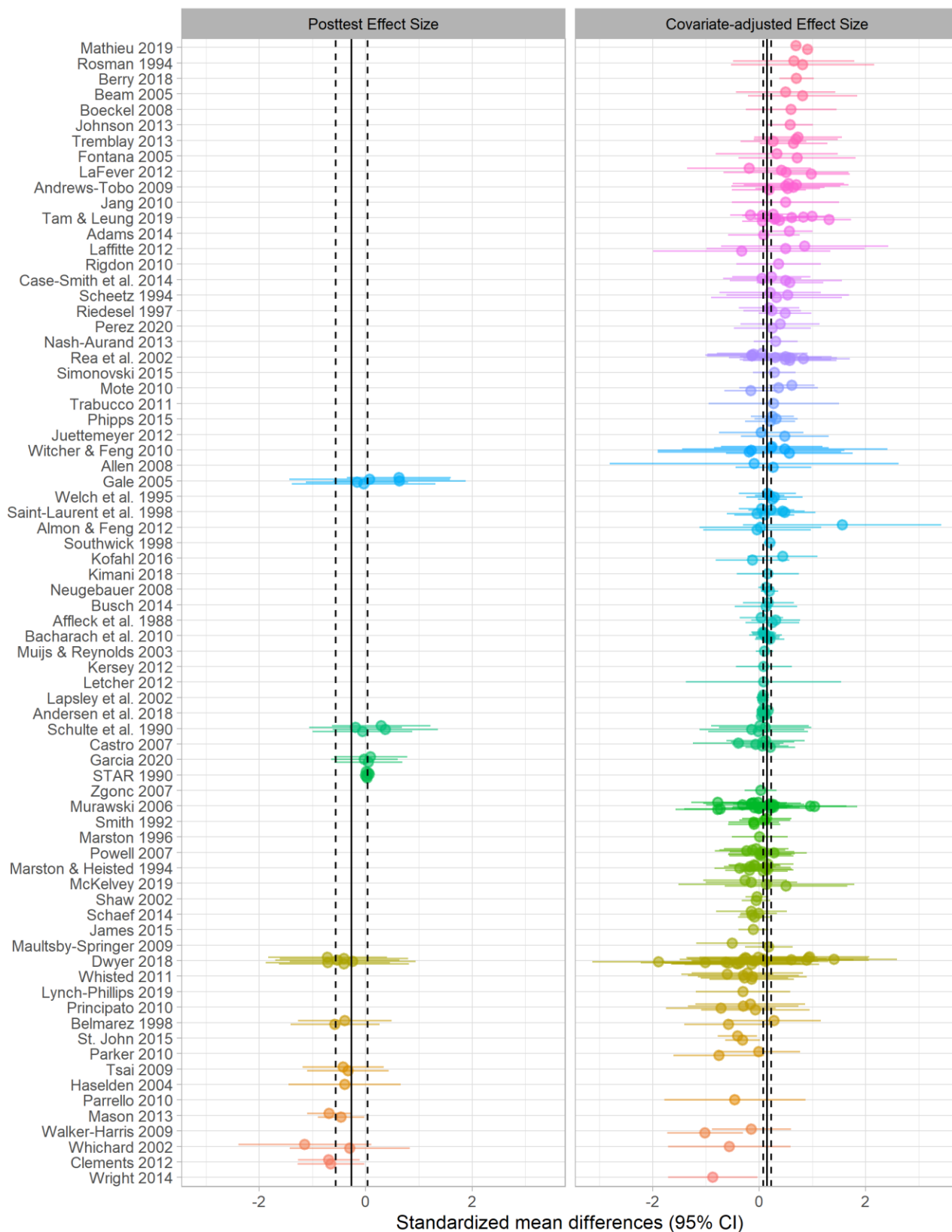


FIGURE S18. Forest plot by type of achievement test



FIGURE S19. Forest plot by student sample



FIGURE S20. Forest plot by grade level



FIGURE S21. Forest plot by type of research design

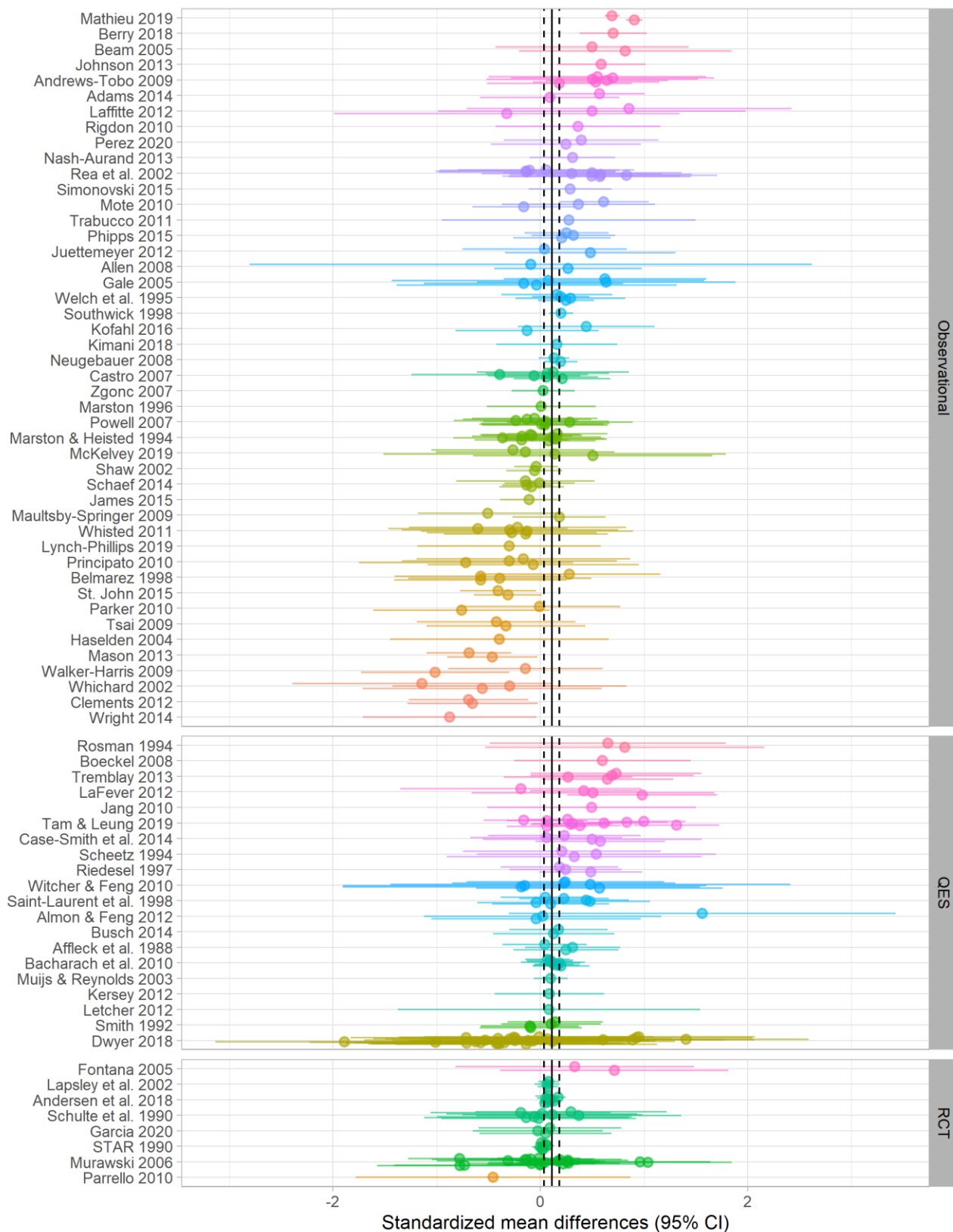


FIGURE S22. Forest plot by study outlet

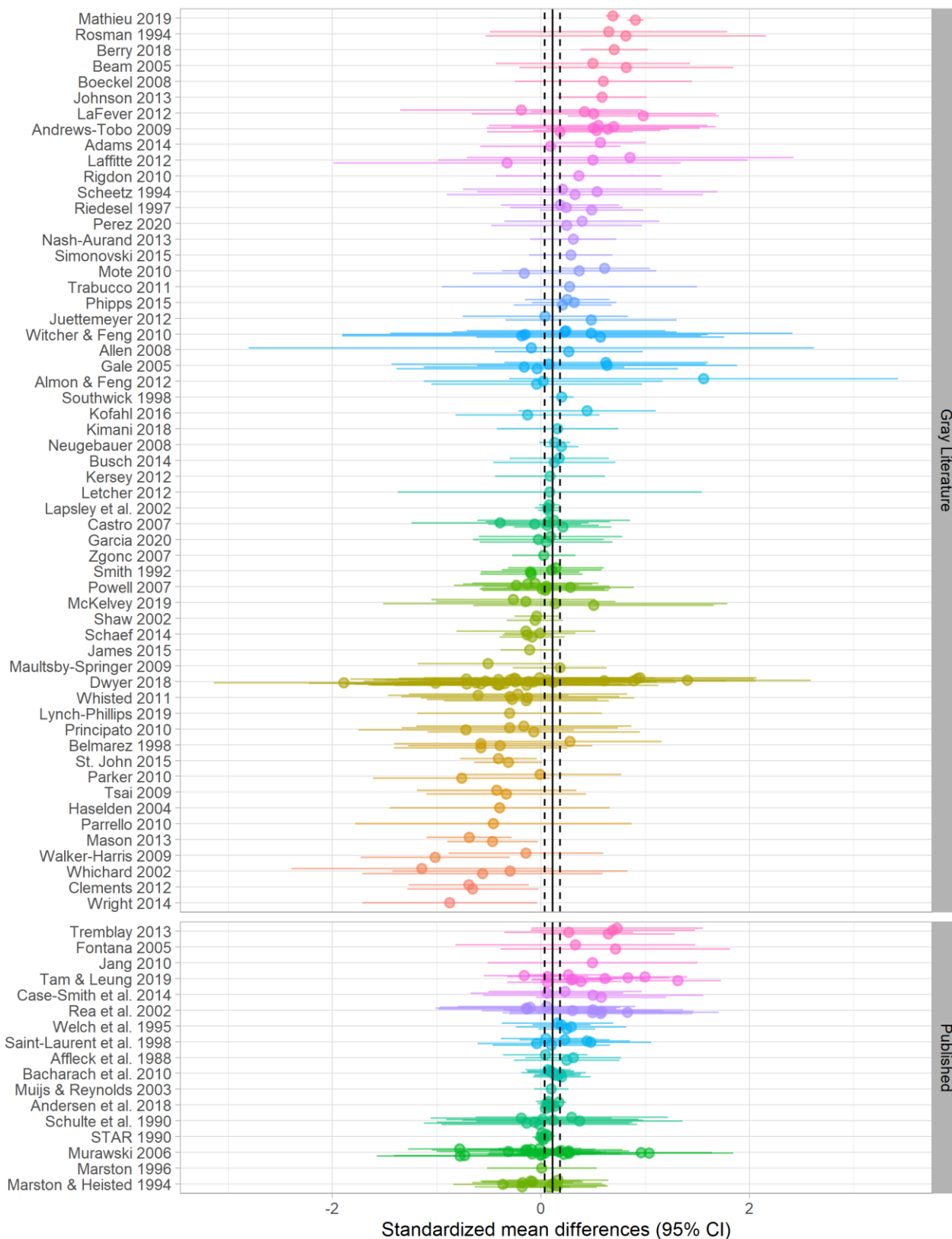
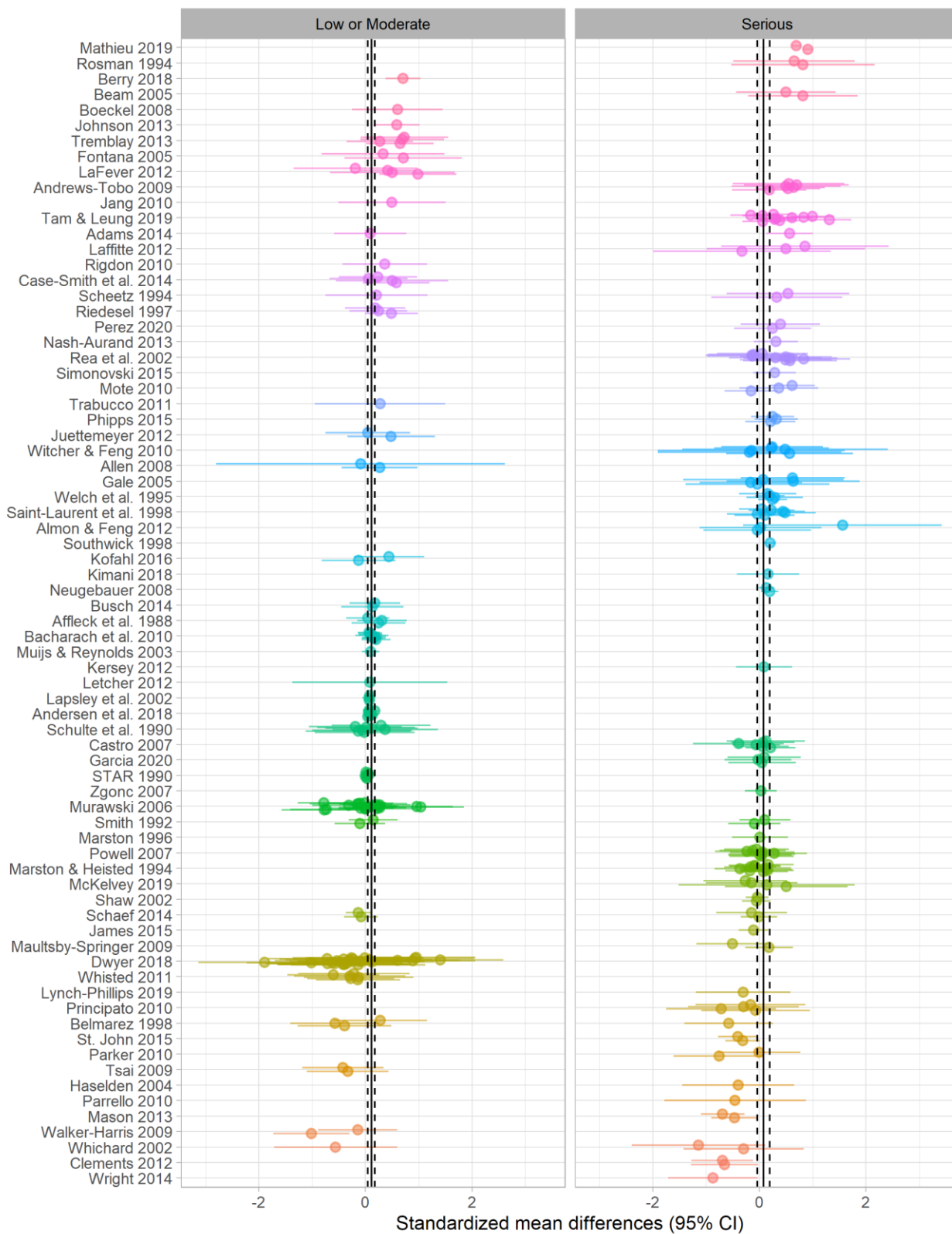


FIGURE S23. Forest plot by received overall risk of bias assessment



S10. Sensitivity Analysis of Moderator Analyses

TABLE S15: Subgroup Analyses for Focal Moderator without Missingness (Co-Teaching Studies Only)

Subgroup-analyses		Unadjusted effects			Covariate-adjusted effects ^a		
Coefficient	Studies	ES	Est. [95% CI]	Satt. df	SD ($\tau + \omega$)	Est. [95% CI]	SD ($\tau + \omega$)
Outcome char.							
Arts and social science ^c	45	134	0.151** [0.044, 0.257]	36.6	0.295	0.175* [0.037, 0.312]	15.9
STEM	40	92	0.076 [-0.057, 0.209]	34.7	0.411	0.114 [-0.04, 0.268]	22.7
HTZ Wald test <i>p</i> value			0.222			0.315	
Posttest ES ^b	9	27	-0.387* [-0.709, -0.065]	5.1	0.108	-0.429* [-0.766, -0.091]	10.9
Covariate adjusted ES	58	199	0.164** [0.065, 0.263]	38.7	0.32	0.1 [-0.045, 0.245]	21.4
HTZ Wald test <i>p</i> value			0.005**			0.007**	
Non standardized test ^b	12	35	0.117 [-0.091, 0.325]	6.2	0.097	0.063 [-0.191, 0.317]	8.8
Standardized test	58	191	0.118* [0.011, 0.225]	42.3	0.351	0.074 [-0.086, 0.233]	22.2
HTZ Wald test <i>p</i> value			0.988			0.919	
Participants char.							
Aggregated sample ^b	11	28	0.119 [-0.082, 0.321]	4.1	0	0.074 [-0.154, 0.302]	10.9
General education sample	25	76	0.025 [-0.101, 0.15]	17.8	0.337	0.015 [-0.145, 0.175]	19.3
Special education sample	39	122	0.139 [-0.001, 0.278]	31.2	0.364	0.107 [-0.068, 0.283]	25.6
HTZ Wald test <i>p</i> value ¹			0.209			0.324	
Elementary school (1-5) ^b	27	112	0.144* [0.031, 0.256]	16.2	0.223	0.102 [-0.04, 0.245]	18.1
Middle school (6-8)	20	63	0.09 [-0.131, 0.311]	16.5	0.414	0.066 [-0.171, 0.303]	20.3
High school (9-12)	18	51	0.058 [-0.131, 0.247]	12.6	0.35	0.042 [-0.181, 0.264]	16.5
HTZ Wald test <i>p</i> value			0.690			0.829	
Design char.							
Observational ^b	42	112	0.067 [-0.064, 0.199]	34.2	0.314	0.027 [-0.155, 0.209]	17.4
QES	16	80	0.309** [0.161, 0.456]	8.4	0.374	0.27** [0.106, 0.435]	13.9
(C)RCT	5	34	0.048 [-0.149, 0.246]	2.5	0.365	0.017 [-0.267, 0.3]	5.1
HTZ Wald test <i>p</i> value			0.055			0.053	
Gray literature ^b	53	150	0.092 [-0.025, 0.21]	40.5	0.325	0.073 [-0.087, 0.233]	22.3
Published literature	10	76	0.23* [0.014, 0.446]	6.2	0.324	0.182 [-0.065, 0.43]	11.1
HTZ Wald test <i>p</i> value			0.221			0.348	
Special edu control grp ^b	29	84	0.166 [-0.004, 0.336]	23.6	0.339	0.101 [-0.095, 0.297]	22.8
General edu control grp	42	142	0.085 [-0.03, 0.199]	27.5	0.336	0.03 [-0.105, 0.165]	21.4
HTZ Wald test <i>p</i> value			0.414			0.473	
RoB low/moderate ^b	26	102	0.155* [0.016, 0.293]	18.9	0.377	0.144 [-0.063, 0.352]	23.4
RoB Serious	43	124	0.076 [-0.049, 0.201]	33.2	0.32	0.068 [-0.078, 0.214]	19.5
HTZ Wald test <i>p</i> value			0.371			0.391	

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. a) Adjusted for grade level, student sample, and subject differences; b) SCE+ model; c) CMVE+ model. 1) Comparison was made between general and special education students only. The table is based on 226 effect sizes from 63 co-teaching studies, i.e., studies in which a general and a special education teacher collaborated.

S11. Publication Bias

We conducted three publication bias or small study effects tests. This includes *Trim-and-Fill* tests based both on all the individual effect sizes and effect sizes aggregated to the study level, *CHE Egg Sandwich* tests accounting for dependent effect sizes using the correlated-hierarchical effects models (Rodgers & Pustejovsky, 2021), and *step-function selection model* tests using three cutpoints (i.e., $p = 0.05, 0.10$, and 0.50) and two cutpoints (i.e. $p = 0.025$ and 1) based on effect sizes aggregated to the study level. For all tests, we either used a corrected estimate of the standard error, i.e. $SE_{adj} = \sqrt{J^2 \times (W \times \xi)}$ or variance $V_{adj} = J^2 \times (W \times \xi)$ to avoid the artificial correlation between the effect size and the sampling variance induced by the scale precision estimate, P , presented in Equation (5) (Hedges & Olkin, 1985; Pustejovsky & Rodgers, 2019). For sensitivity analysis purposes, we also transformed effect sizes to avoid artificial correlation among g and SE_g . For this purpose, we used Equation 3 from (Pustejovsky & Rodgers, 2019).

Results

As can be seen in Figures S24 and S25, we did not find any sound evidence for publication bias or small study effects from the *Trim-and-Fill* analyses. Further, we conducted two cluster-robust Egger's regression tests fitting the modified effect size standard errors and variance estimates to the CHE-RVE model, respectively. From these analyses, we found $p = .261$ for the former model and $p = .465$ for the latter, further indicating an absence of small study effects and/or publication bias. From the employed selection model with cutpoints at $p = .05, .10$, and $.50$, using average study effect sizes, $\bar{g}$ turned out to be statistically insignificant but remained moderate in size, i.e., 0.077 , 95% CI $[-0.150, 0.303]$ with a total SD of 0.282 . However, in the sensitivity analysis with the selection model using cutpoints at $p = 0.025$ and $p = 1$, the overall average effect size $\bar{g}$ retained to be statistically significant, and $\bar{g}$ increased to 0.184 , 95% CI $[0.0659, 0.303]$ with a total SD of 0.238 , clearly underpinning the impact of the chosen selection model.

FIGURE S24. Trim and Fill funnel plot with modified standard errors across individual and study mean effect size estimates

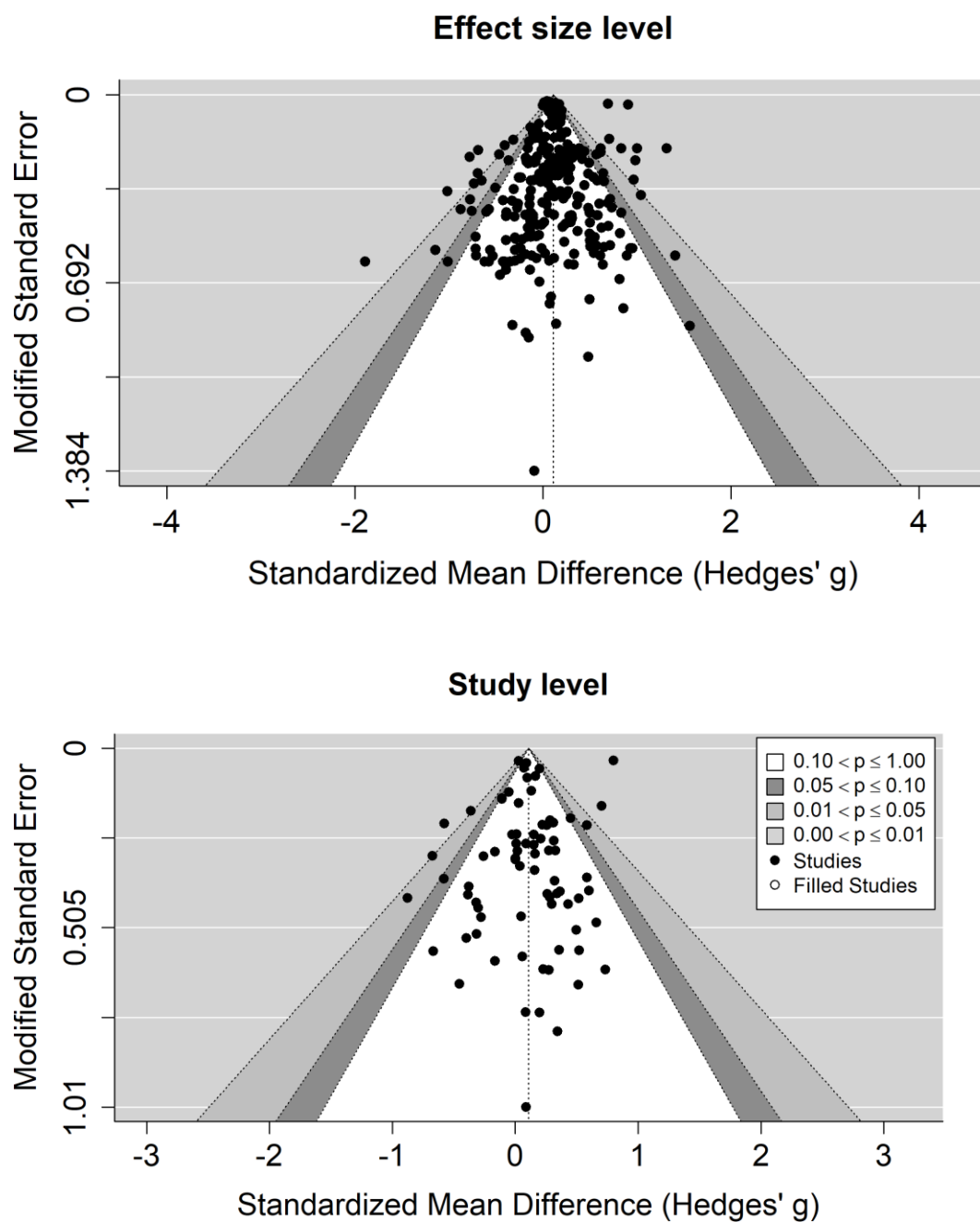


FIGURE S25. Trim and Fill funnel plot with transformed individual and transformed mean effect size estimates

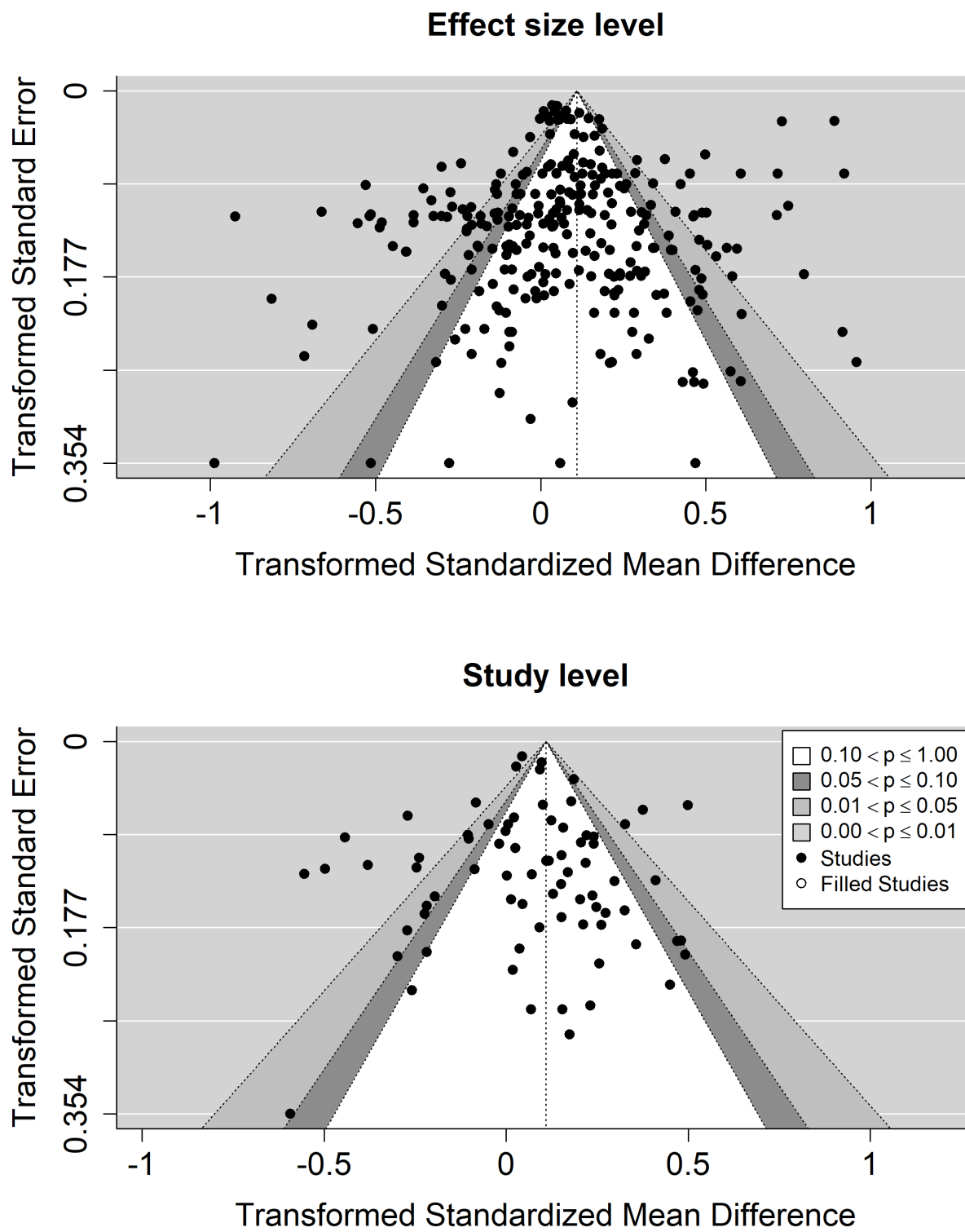


FIGURE S26. Selection model with three cutpoints ($p = 0.025, 0.10, 0.50$, and 1)

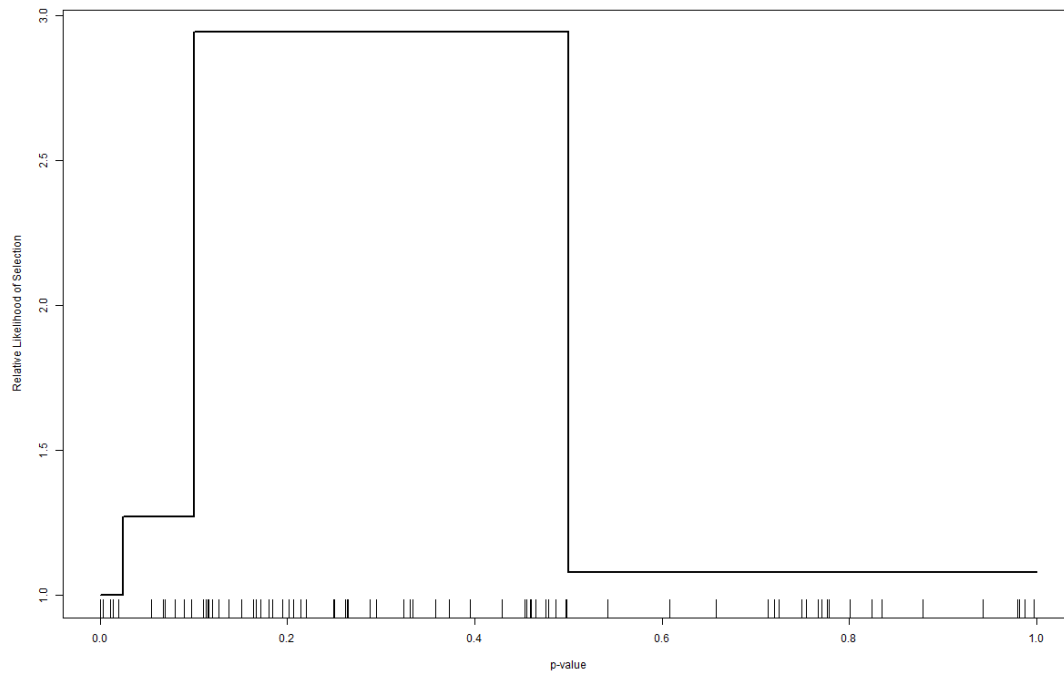
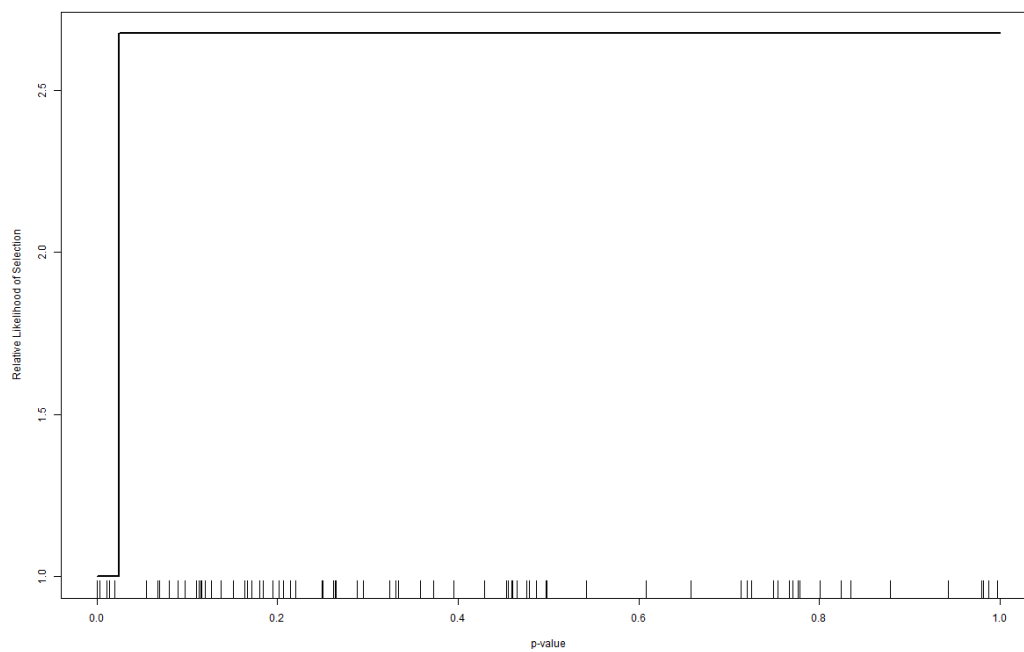


FIGURE S27. Selection models with cutpoints ($p = 0.025$ and 1)



S12. Exploratory Analyses

Table S16 presents subgroup mean differences between effect sizes based on different control groups used to calculate effect sizes from samples of special needs students to investigate if one of the alternative service delivery models was better than the other. We did not find any statistical differences between the two groups, suggesting that special education and single-taught general education classrooms have equal effects on the academic achievement of students with special needs or disabilities.

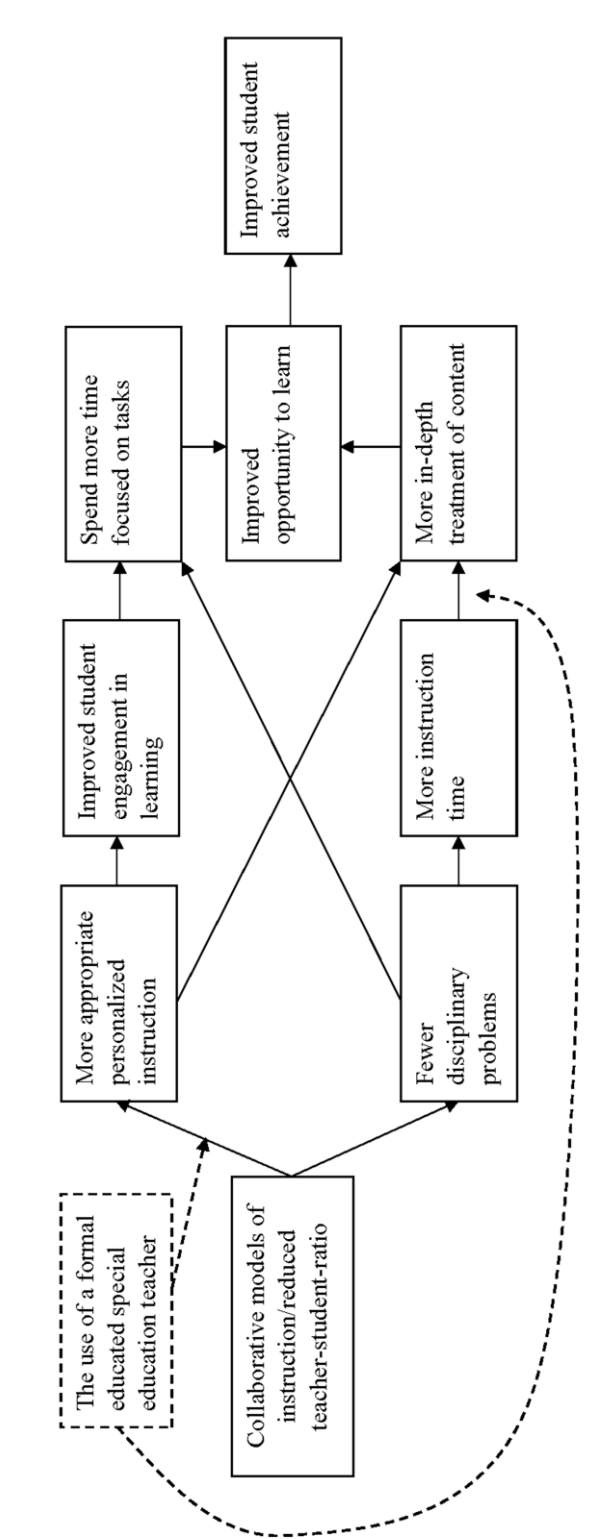
TABLE S16: Subgroup analysis between special and general educations control groups

Subgroup		Unadjusted effects				Covariate-adjusted effects ^a		
Coefficient	Studies (J)	ES (K)	Est. [95 % CI]	Satt. df	SD ($\tau + \omega$)	Est. [95 % CI]	Satt. df	SD ($\tau + \omega$)
Control group								
General education	11	39	0.086 [-0.222, 0.395]	7.6	0.421	-0.034 [-0.338, 0.271]	12.6	0.402
Special education	32	96	0.178* [0.025, 331]	26	0.311	0.070* [-0.114, 0.256]	17	0.317
HTZ Wald test			0.556			0.491		
<i>p</i> values (CWB)			(0.579)			(0.494)		

* $p < .05$. ** $p < .01$, *** $p < .001$. Note. a) The below results are adjusted for school level and subject differences.

S13. Causal Theory of Collaborative Models of Instruction

FIGURE S28. Causal diagram for the impact of collaborative models of instruction on student achievement



Note: Inspired by Filges et al. (2015). Bold lines and squares indicate the common causal mechanism underlying the theory of all collaborative models of instruction, whereas the dashed square and lines indicate where it is assumed in the co-teaching literature that the used of formally educated special education teacher augment the effect of collaborative instruction.

References

- Achilles, C., 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>
- Affleck, J. Q., Madge, S., Adams, A., & Lowenbraun, S. (1988). Integrated classroom versus resource model: Academic viability and effectiveness. *Exceptional Children*, 54(4), 339–348. <https://doi.org/10.1177/001440298805400408>
- Alborz, A., Pearson, D., Farrell, P. T., & Howes, A. J. (2009). *The impact of adult support staff on pupils and mainstream schools*. EPPI-Centre. [http://eppi.ioe.ac.uk/cms/Portals/0/PDF reviews and summaries/Support staff Rpt.pdf?ver=2009-05-05-165528-197](http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/Support%20staff%20Rpt.pdf?ver=2009-05-05-165528-197)
- 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>
- Carlson, H. L., & Others, A. (1984). *Servicing low achieving pupils and pupils with learning disabilities: A comparison of two approaches*. <https://files.eric.ed.gov/fulltext/ED283341.pdf>
- Christie, H. M. (2020). The impact of co-teaching on high school English achievement for both general education and special education students in a large suburban high school [University of St. Francis]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/2395333551?accountid=14468> NS
- 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>
- Dyssegaard, C. B., & Larsen, M. S. (2013). *Evidence on inclusion*. Danish Clearinghouse for Educational Research. https://edu.au.dk/fileadmin/edu/Udgivelser/Clearinghouse/Evidence_on_Inclusion.pdf
- 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>
- Garcia, R. D. (2020). Effects of integrated co-teaching on 9th grade general education math students [Fairleigh Dickinson University]. In *ProQuest Dissertations and Theses*.

- <https://search.proquest.com/docview/2388699142?accountid=14468> NS
- Haselden, K. G. (2004). Effects of co-teaching on the biology achievement of typical and at-risk students educated in secondary inclusion settings [The University of North Carolina at Charlotte]. In *ProQuest Dissertations and Theses*.
- <https://search.proquest.com/docview/305082049?accountid=14468> NS
- 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., & 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.
- 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>
- Iacono, T., Landry, O., Garcia-Melgar, A., Spong, J., Hyett, N., Bagley, K., & McKinstry, C. (2021). A systematized review of co-teaching efficacy in enhancing inclusive education for students with disability. *International Journal of Inclusive Education*, 1–15. <https://doi.org/10.1080/13603116.2021.1900423>
- Jang, S.-J. (2010). The impact on incorporating collaborative concept mapping with coteaching techniques in elementary science classes. *School Science and Mathematics*, 110(1), 86–97. <https://doi.org/10.1111/j.1949-8594.2009.00012.x>
- 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*, 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]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/1817570306?accountid=14468> NS
- 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>

- LaFever, K. M. (2012). The effect of co-teaching on student achievement in ninth grade physical science classrooms [University of Missouri – St. Louis]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1697496661?accountid=14468> NS
- Lönnqvist, E., & Sundqvist, C. (2016). Samundervisning som inkluderande arbetssätt i skolan - Fördelar och nackdelar för elever. *Nordic Studies in Education*, 36(1–2016), 38–56. <https://doi.org/10.18261/issn.1891-5949-2016-01-04>
- Mathieu, L. (2019). An examination of special education instructional programs for English learners in New York City schools [Teachers College, Columbia University]. In *ProQuest Information & Learning (US)*. <https://search.proquest.com/docview/2279940069?accountid=14468> NS
- Maultsby-Springer, B. M. (2009). A descriptive analysis of the impact of co-teaching on the reading/Language Arts and math achievement of selected middle school students in a Middle Tennessee school district [Tennessee State University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/613688517?accountid=14468> NS
- 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., & 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>
- Nash-Aurand, T. (2013). A comparison of general education co-teaching versus special education resource service delivery models on math achievement of students with disabilities [Liberty University]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/1773213717?accountid=14468> NS
- Parrello, J. (2010). The effects of co-teaching on the academic achievement of general education

- students [Caldwell College]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/193653348?accountid=14468> NS
- 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>
- 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>
- 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>
- 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>
- Rosman, N. J. S. (1994). Effects of varying the special educator's role within an algebra class on math attitude and achievement [University of South Dakota]. In *ProQuest Dissertations and Theses*. <https://search.proquest.com/docview/62701265?accountid=14468> NS
- Saint-Laurent, L., Dionne, J., Giasson, J., Royer, É., Simard, C., & Piérard, B. (1998). Academic achievement effects of an in-class service model on students with and without disabilities. *Exceptional Children*, 64(2), 239–253. <https://doi.org/10.1177/001440299806400207>
- Schulte, A. C., Osborne, S. S., & McKinney, J. D. (1990). Academic outcomes for students with learning-disabilities in consultation and resource programs. *Exceptional Children*, 57(2), 162–172.
- 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>
- Southwick, K. E. (1998). The effects of the class within a class collaborative/co-teaching model on the achievement of general education students in grades three, four and five [University of Kansas]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/304420847?accountid=14468> NS
- St. John, M. M. (2015). The influence of placement in a co-taught inclusive classroom on the academic achievement of general education students on the 2014 New York State ELA and mathematics assessments in grades 6-8 in a suburban New York school district [Seton Hall University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1733230787?accountid=14468> NS
- 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., & 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.
- van Garderen, D., Stormont, M., & Goel, N. (2012). Collaboration between general and special educators and student outcomes: A need for more research. *Psychology in the Schools*, 49(5), 483–497. <https://doi.org/10.1002/pits.21610>
- Viechtbauer, W. (2022). *Likelihood ratio and wald-type tests for “rma” objects*.
<https://wviechtb.github.io/metafor/reference/anova.rma.html>
- 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>
- 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>

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

- 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>
- Wright, R. (2014). The academic impact of co-teaching on non-disabled high school Integrated Math I students [Capella University]. In *ProQuest Dissertations and Theses*.
<https://search.proquest.com/docview/1615310567?accountid=14468> NS
- WWC. (2020). *WWC procedures and standards handbook (4.1)*. Institute of Education Sciences.
<https://ies.ed.gov/ncee/wwc/Handbooks>
- 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
- Zigmond, N., & Magiera, K. (2002). Co-teaching. *Current Practice Alerts*, 6(6), 1–4.
http://ppsacademicsupport.weebly.com/uploads/2/9/0/4/29048495/co-teachinginfo_ld.pdf

Appendix 3: OSF Preregistered Protocol (Second Version)

First version registered June 8, 2020. Find at <https://bit.ly/3nhVX3H>.

Second version registered November 1, 2021 (this version).

Study Information

1. Title:

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

1.1. *Identification:* This report is the second pre-registered protocol linked to “*The Effects of Co-teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis*” study. This report mainly draws on the PRISMA-P advice and checklist complemented by the setup of the OSF (Open Science Framework) pre-register [template](#). This updated protocol was primarily made due to major amendments to the analytical strategy because a new method has been developed after the outset of this study (Pustejovsky & Tipton, 2021). Furthermore, we provide more details on our data extraction procedure, effect size calculation (including cluster design adjustment of effect sizes), the risk of bias assessment, and the analytical strategy and rationale of the study.

1.2. *Update:* Minor parts of this meta-analysis function as an update/follow-up study of previously conducted meta-analyses authored by Christopher Khoury’s (Khoury, 2014) and Wendy W. Murawski & H. Lee Swanson’s (2001), respectively. Yet this project has a broader focus on comparing the general effects of several kinds of two-teacher approaches on both general students and special needs students in the pre-defined context of primary and lower secondary schools (i.e. grades 1-12). This contradicts the previous meta-analyses since they only concentrate upon the effects of co-teaching on students with special needs/disabled students. The present meta-

analysis has a wider aim of looking at the general effects of co-teaching/team-teaching¹ on student achievement for all students. However, one of the planned subgroup analyses of this study—in which we compare the special needs student vs. general student effect sizes—will be close-to similar to the previously conducted meta-analyses and can be seen as a kind of replication² of the former conclusions in the field of the effects of co-teaching for students with special needs (see section 20). Still, our approach differs since we partly use subgroup data from the primary studies.

One further key difference between this meta-analysis and the previously conducted meta-analyses is that we apply more rigorous inclusion criteria with regard to eligible study designs. We only allow studies that draw on counterfactual designs, i.e. studies in which some kind of control group is applied. For studies that do not apply randomization, we require for them to be eligible to somehow ensure baseline equivalence either by conducting baseline analyses or providing pre-test/baseline measures/results. This stands in stark contrast to the prior meta-analyses (see Cook, McDuffie-Landrum, Oshita, & Cook (2016) for an overview). Furthermore, our effect size calculation will be quite different from the previous meta-analyses, since we go beyond the textbook examples of the computation of effect size which has been used in the prior analysis, and which has been shown to be biased when applied to various common counterfactual designs, especially pre-posttest and covariate adjusted design (Pustejovsky, 2016). Further, this study is distinguished from previous reviews since we conduct a (2-level) cluster design

¹ Co-teaching in this regard refers to any kind of two-teacher teaching, and does not solely refer to co-teaching as a concept related to special training activities such as inclusion of students with special needs in the mainstream classroom. Notions like tema-teaching and co-teaching will appear in the text interchangeably.

² Replication denotes the definition from Hedges (2019, pp. 3–4) which is as follows: “An important distinction is that between reproducibility and replicability. Reproducibility concerns whether another investigator can obtain the same results when given the first investigator’s research report and their data (and possibly the computer code they used to analyze the data). Replicability concerns whether another investigator can obtain the same results when they obtain their own (new) data by attempting to repeat the study that was carried out by the first investigator. A key difference between reproducibility and replicability is that the former involves whether two investigators can obtain the same answers when given the same data, but replicability involves whether two investigators can obtain the same answers from two different datasets. Furthermore, it is important to distinguish between “*direct replication*, which involves the replication of an experimental procedure” and “*conceptual replication*, which involves the repetition of earlier research work with different methods” (Hedges, 2019, p. 4). Hedges’ definition aligns with the definition of IES & NSF (2018). The above definition, though, contradicts the definitions coined by Freese & Peterson (2017) and Cartwright (1991), in which the denotation of the two concepts are reversed. However, the more substantial definitions of the concept are on a general level identical.

adjustment for all studies/effect sizes that do not account for clustering at the class level (Hedges, 2007). We apply multivariate/multilevel methods with robust variance estimation (RVE) that allow multiple effect sizes from the same study to be included in the analysis without making false assumptions of independence among effect sizes (Pustejovsky & Tipton, 2021), which has been the case in previous meta-analyses. It should, furthermore, be mentioned that this study also functions as a replication of great parts of a systematic review regarding inclusion in education conducted by the Danish Clearinghouse for Educational Research (Dyssegaard & Larsen, 2013; Dyssegaard, Larsen, & Tiftikçi, 2013). The underlying intention of our study is to depict the clear differences between the conduct of narrative synthesis vs. meta-analysis.

2. *Registration (adopted from PRIMA-P checklist):*

The study is registered at OSF (Open Science Framework). For more details, see <https://cos.io/prereg/>. Questions regarding the pre-registration contact prereg@cos.io

3. *Authors:* Mikkel Holding Vembye (PI), Felix Weiss (CI)

3.1. *Contact (PRISMA-P):*

Mikkel Holding Vembye

Aarhus University

E-mail: mihv@edu.au.dk or mikkel.vembye@gmail.com

ORCID-ID: <https://orcid.org/0000-0001-9071-0724>

Scan:



3.2. *Contributions (PRISMA-P)*: This meta-analysis represents independent research from the respective authors. As principal investigator, Vembye will have a greater (working) share in the project/article than Weiss. This project aims to be a key part of Vembye's Ph.D. project. Consequently, we deem it natural that Vembye takes a larger involvement in the project. Following the advice of Pigott & Polanin (2019), Weiss' main role is to make quality assessments of all screening and analysis parts of the project, i.e. literature retrieval, abstract screening, full-text screening, code-book assessment. We will seek statistical advice for the final statistical analyses. Consequently, more authors might be involved in the final project. It is pivotal to notice that we are not able to double code all parts of the full-text extraction and risk of bias (RoB) assessment due to time/resource constraints of the project. Vembye will lead and do all data extraction.

4. *Amendments (adopted from PRIMA-P checklist)*:

This pre-registered protocol is the second one linked to this study. It reflects close-to-final ideas behind the conduct of the study prior to the final data analysis. Any divergence between this protocol and the final analysis will be documented in the final paper.

5. *Support (adopted from PRIMA-P checklist)*:

5.1. *Sources*: No external financial sources are connected to this study.

5.2. *Sponsor*: The study has no organizational or institutional funders or sponsors. For clarity, see section 5.1. The study is a part of an Open Call scholarship received from Aarhus University (Application no. 22606592).

5.3. *Role of sponsor or funder*: None

6. *Introduction*

6.1 *Description/rationale:*

Co- and team-teaching—which we broadly define as two or more teachers/adults sharing the responsibility of the within-class instruction/teaching and/or support of the students—is widely used in various countries around the world including the US and many parts of Europe, especially in the Scandinavian countries (Andersen, Beuchert-Pedersen, Nielsen, Thomsen, et al., 2018; Cook et al., 2016; Friend, 2017). The co-teaching literature often contends that the co-teaching model (i.e. collaboration between a general educator and a special educator in the same physical space/classroom) is a thoroughly tested instruction model and thereby an evidence-based teaching practice that has a positive and substantial impact on student (academic) achievement, most pronounced on students with special needs (Friend, 2017). It is argued that co-teaching outperforms alternative modes of instruction, such as special education classrooms and inclusion of students with special needs in mainstream classrooms with only one general teacher (Friend, 2017). However, the empirical evidence underpinning this narrative seems to be meager (Cook et al., 2016; Murawski & Lee Swanson, 2001). From a policy perspective, the co-teaching literature—primarily defined as a delivery model directed toward students with special needs—further begs the question of what the effects are on general student achievement? The aim of this study is, therefore, partly to understand the overall effect on student achievement of collaborative models of instruction on general and special needs students in grades 1-12, and partly to examine how the effects vary across the general and special needs students. Our goal is to examine whether and with what effect co- and team-teaching approaches can function for all students as a flexible and subject-specific alternative to reducing the teacher-student/adult-student ratio without reducing class sizes (Filges, Sonne-Schmidt, & Nielsen, 2018; Glass & Smith, 1979; Hedges & Stock, 1983).

Another focal aim of this study is to investigate if and how effects vary as a function of the intervention model, i.e. whether the composition of the two or more teachers/adults has varying effects on student achievement. Such knowledge might be important from a management and political perspective since great costs can be attached to co-teaching between two formal-educated teachers. In order to examine this question, we aim to combine

the area of co-teaching (Friend, 2017; Murawski & Lee Swanson, 2001) with the fields of team-teaching, teacher's aides/teacher assistants (i.e. two teacher instruction strategies not rigorously defined to contain collaboration between a general and special educator, but widely defined, i.e. all kinds of collaboration between two of legal-age educational personnel simultaneously delivering instruction and support to the students in the same physical classroom) (Andersen, Beuchert-Pedersen, Nielsen, Thomsen, et al., 2018; Blatchford, Russell, & Webster, 2012; Willett, Yamashita, & Anderson, 1983).³ The study will draw on the random-effects model because we believe that this amalgamation of different fields of literature creates “unidentifiable sources of variability (i.e., unmeasured covariates)” (Valentine, Pigott, & Rothstein, 2010, p. 217). We anticipate finding substantial between-study and within-study random effects since we allow a diverse set of study designs to enter from different content areas to be meta-analyzed. More random effects might be added to our models when we obtain more accurate knowledge about the dependence structure of the final dataset.

Yet another key part of this study is to investigate factors that moderate effects of two-teacher instruction on different levels, i.e. *study context* (e.g. urban vs. rural), *study design characteristics* (e.g. type of treatment and comparison group, research design, etc.), *outcome assessment* (i.e. type of test instrument, effect calculation mode, etc.), *participant characteristics* (e.g. grade, type of student—general vs. special students, etc.), *intervention characteristics* (e.g. duration, subject taught, etc.), and *Risk of Bias indicators* (Higgins et al., 2019). In Table 1, we map the full list of (theoretical and methodological important) moderator variables that we try to locate during our data extraction, although we do not expect to find enough information for all variables. This way we pursue to examine the informational boundaries of this field of literature. Table 1, further, presents the assumptions we have regarding the relationships we expect to find between the moderator variable and the student achievement. We aim to test moderator effects through meta-regression (T. Pigott, 2012; Tipton, Pustejovsky, & Ahmadi, 2019) in order to avoid confounding between

³ See Hattie for a similar fusion of the literature http://www.visiblelearningmetax.com/influences/view/co~team_teaching. Although major mistakes are made in the mean effect size calculation. Murawski & Swanson (2001) reports a mean effect size of 0.4 not 0.31, and Willett et al. (1983) effect size estimate predicated upon 41 study not 130.

given covariates. Moreover, all effect sizes will be interpreted in relation to relevant benchmarks since effects sizes vary substantially across year groups of student, subjects, type of intervention, and type of student (e.g. special vs. mainstream students), etc. (Bloom, Hill, Black, & Lipsey, 2008; Hill, Bloom, Black, & Lipsey, 2008; Kraft, 2020; Lipsey et al., 2012). Ideally, we would like to employ all available moderators in our meta-regression models.

Last, we will test for publication bias in the co- and team-teaching literature. Following the recommendation made by Hedges & Vevea (2005, p. 161), we will apply a range of different tests for publication bias such as *Egger's regression* (Egger Sandwich i.e. Egger's regression test using robust variance estimation, i.e. the correlated hierarchical effects model (CHE), and with variance-stabilized effect sizes (Pustejovsky & Rodgers, 2019; Pustejovsky & Tipton, 2020; Rodgers & Pustejovsky, 2019) *Funnel plot assessment* (Fernández-Castilla et al., 2020), *Trim and Fill test with multiple outcomes*, *Selection Models*, and if we have time *sensitivity analysis for publication bias in meta-analysis*⁴ (Mathur & VanderWeele, 2020), and if possible *weighted average of the adequately powered (WAAP) studies* (Stanley, Doucouliagos, & Ioannidis, 2017). The last test functions as a sensitivity analysis. We will only conduct the WAAP test if we have a sufficient number of adequately powered effect sizes.

7. Objective (adopted from PRIMA-P checklist):

7.1. Main research questions: ***Do collaborative models of instruction have a positive, substantial significant impact on students' academic achievement?***

7.1.1. Research sub-questions:

Does the magnitude effect of collaborative models of instruction vary as a function of theoretical and methodological focal moderator variables⁵?

⁴ See <https://cran.r-project.org/web/packages/PublicationBias/index.html>

⁵ These are defined from theory discussed within collaborative models of instruction literature and from empirical findings from the meta-analytical literature.

8. *Hypotheses:*

8.1. **Main hypothesis:**

1) We assume to find a positive relationship between co-/team-teaching and student achievement. Premised upon the previously conducted meta-analyses, we expect that collaborative models of instruction have an overall average positive effect of approximately around 0.2 to 0.4 standard deviations across all populations of students.⁶ This includes non-disabled/general students as well as students with special needs.

Table 1 provides an overview of all possible covariates that would ideally be studied as sub-hypotheses in a complete/ideal meta-analysis. We aim at testing as many of them as possible, obviously restricted by the number of available studies and the information about them. Thus the list serves as an ideal model, and we do expect some variables to be excluded from the final analysis due to a severe lack of required information in the co- and team-teaching literature. Premised partly upon the meta-analytical literature (Rothstein, Sutton, & Borenstein, 2005) and partly on the content-specific literature of co- and team-teaching (Cook et al., 2016), the variables presented in Table 1. represents what we consider to be the most important factors to examine in the fields of meta-analysis and collaborative models of instruction, respectively. Most of our hypotheses are deduced from empirical findings in the fields of meta-analysis and co-/team-teaching, respectively. We do not elaborate on any exact magnitude of the direction on the presented variables from Table 1. However, we recognize that minor statistical significant effects do not necessarily have substantial practical importance. Table 1 is inspired by Dietrichson et al. (2017), Pigott (2012), and Lipsey (2009). Notice, red text in Table 1 indicates that the given variable has been removed from the original protocol and thereby also from the final analysis.

⁶ This does not mean that the efficacy of the intervention cannot vary as a function of the population composition (Borenstein et al., 2009). That is also why we conducted several test for heterogeneity to better understand the variability of the efficacy of the two-teacher intervention.

Table 1. Moderator Variables and Related Hypotheses of the Directions of the Variables

Variable	Moderator Type	Data Type	Description	Direction of Assumed Relationship (- 0 +)
Subject	Outcome assessment	Categorical	1) STEM subjects 2) Arts subject (including ELA and social sciences)	+ (We expect to find substantial differences between the science, math, with language arts (ELA) yielding larger effects)
Type of test	Outcome assessment	Categorical	1) Standardized 2) non-standardized test	- (We assume that standardized test yields smaller effect sizes than non-standardized test because standardized tests usually represent broader subject-related content)
Time from baseline	Outcome assessment	Continuous	Number of months from baseline	
Follow-up (more than three months from the end of the intervention)	Outcome assessment	Categorical	1) Yes 2) No	- (we expect the intervention effect to fade out over time)
Covariate adjusted effect size	Effect size calculation	Categorical	1) Yes 2) No	- We expect covariate adjusted effect size to yield smaller effect size relative to effect sizes calculated from posttest scores.
Pre-test adjusted effect size	Effect size calculation	Categorical	1) Yes 2) No	- We expect pretest adjusted effect size to yield smaller effect size relative to non-pretest adjusted effect sizes
Date of publication	Study level	Continuous	Year of publication	0 (We expect to find no impact of the year of publication on student achievement)
Publication type	Study level	Categorical	1) Published (scientific journal/peer review) 2) Unpublished ⁷	+ (Larger effect sizes for published literature are expected, see Cheung & Slavin (2016))
Time to publication	Study level	Continuous	Time from initiations of intervention to publication	- (the effect size will decrease as a function of

⁷ Unpublished research in this sense refers to “not independently edited or unrefereed” (White, 2009, p. 61).

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

				the time gap from the start of the intervention to publication i.e. large time gaps are expected to yield smaller effect sizes)
Design	Methodological	Categorical	2) RCT (including cluster and/or block randomized trials) 1) Quasi-experimental study design (QES) 0) Observational study	- (C)RCTs are expected to yield smaller effect sizes since we expect that confounding factors inflate the effect sizes of QES and observational studies)
Randomization (is contained in the design variable)	Methodological	Categorical	1) Random cluster 2) Random individual 3) Matched 4) Convenience 5 Other	- (In a similar vein, we expect randomized designs to yield lower effect sizes)
Risk of Bias	Methodological	Categorical	Overall serious risk of bias vs. not serious risk of bias	- (We expect less serious risk of biased studies to yield smaller effects relative to studies that have been assessed to be of serious risk/high risk of bias)
Treatment	Intervention characteristic	Categorical	1) Co-taught 2) Team-taught (general-general teachers) 3) Teacher's aide	0 (We hypothesize that the type of program does not have any substantial effect, because we suppose that having one more educational personnel in the classroom by itself will have almost equal effects independently of the formal education of the personnel.
Duration	Intervention characteristic	Continuous	Duration in weeks. One school year = 10 months of teaching	+ (from the co-teaching literature we expect duration to have a positive impact on the efficacy of co- and team-teaching).
Intensity of intervention	Intervention characteristic	Continuous	Sessions per week	+ (hours/sessions per week receiving two-teaching approaches are expected to have a positive impact on student achievement)

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

Implementation (covered in the des- ing variable)	Intervention characteristic	Categorical	1) Monitored 2) Not-monitored	+ (Well-monitored stud- ies are expected to yield greater effect sizes)
Training	Intervention characteristic	Categorical	Yes or no	+ (trained staff is ex- pected to have a positive impact on the imple- mentation of co-teach- ing and thereby on stu- dent achievement)
Quality of collabora- tion (if applicable)	Intervention characteristic	Categorical	1) Good 2) Bad 3) Other/not gauged	+ (The co-teaching liter- ature suggest that the collaboration between the co-teachers is a piv- otal component for suc- cess, therefore we ex- pect to find a positive relationship)
Planning time	Intervention characteristic	Categorical	Planning time vs. no plan- ning time	+ (The co-teaching liter- ature suggest that the planning time prior to co-teaching is a pivotal component for success, therefore we expect to find a positive relation- ship between the amount of planning time and student achieve- ment)
Same teacher(s) across arms	Intervention characteristic	Categorical	Same teacher(s) vs. not the same teacher	- (we expected that re- search designs in which the the same teachers are utilized across arms will yield lower effect size because the design controls the impact of random teacher factors.)
Teacher experience	Intervention characteristic	Continuous	Average years of experi- ence	0 (we don't expect ex- perience to have a sig- nificant effect on the ef- ficacy of the interven- tion. This is evidenced by Andersen, Beuchert- Pedersen, Nielsen, & Thomsen (2018))
Comparison group	Intervention Characteristic	Categorical	1) Single-taught general classroom 2) Special education class- room/pull-out arrangement	+ (We expect that effect sizes will be greater when the comparison group is based on one- teacher class-rooms)

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

Country	Study context	Categorical	Country of which the study is conducted	0
Type of school	Study context	Categorical	1) Public school(s) 2) Private school(s) 3) A mixture of public and private schools	0
Location	Study context	Categorical	1) Urban or suburban 2) Rural 3) Mixed	0
Grade	Participants and sample characteristics	Categorical	1) 1-5 2) 6-8 3) 9-12	- (We expect effect sizes to decrease as a function grade)
Student sample	Participants and sample characteristics	Categorical	1 General students (GS) 2 Special needs students (SNS) 3 Aggregate across SNS and GS	+ (We expect special needs students to gain most from collaborative models of instruction)
Gender	Participants and sample characteristics	Continuous	Percent of males in the sample	+ (We suppose that male thrives more from co-teaching than girls because distracting behavior will more frequently be reduced for boys than girls)
Race and ethnicity	Participants and sample characteristics	Continuous	Percent of migrant or ethnic/racial minority students in the sample	- (We expect that greater amounts of migrants students represented in the sample decrease the effect sizes)
Special needs students	Participants and sample characteristics	Continuous	Percentage of students classified as special needs student in the sample (only relevant)	- (We expect co- and team-teaching to be less efficient when large amounts of special needs students are represented in the sample. This hypothesis is suggested by the co-teaching literature)
SES composition	Participants and sample characteristics	Categorical	1) Low SES 2) Low-middle SES 3) Middle SES 4) Middle-upper SES 5) Upper SES 6) Labeled as "mixed" 999) Can't tell	+ (we anticipate to find, that populations containing a greater share of students with the advantaged socioeconomic ground will gain more from co-/team-teaching strategies)

Due to time constraints and the expectation that we won't be able to reach much information about these factors, we have removed the *time to publication*, *same teacher across arms*, *time from baseline*, *quality of collaboration*, *ethnicity*, and *SES* variables from the original protocol. Table 1 does not resemble the final coding schemes in all its details. Find the final coding schemes on OSF (<https://osf.io/vtjqs/>). Table 1 represents a coarse-grained description of the covariates of the study, only. Below, we present the variables, we assume to include in our subgroup analyses and final meta-regression model. If a factor from Table 1 is not represented in the below list of focal variables, it either indicates that we do not anticipate that the will not vary enough the be relevant for the model or that we do not expect to find enough relevant information about this regard.

Covariates expected to be included in the final meta-regression model

Study design characteristics

- 1) Design
- 2) Publication status
- 3) Risk of bias

Effect size characteristics

- 4) Covariate adjusted effect size

Outcome assessment

- 5) Subject (and subject x grade)
- 6) Type of test
- 7) Follow-up effect size

Intervention characteristics

- 8) Treatment
- 9) Control group
- 10) Duration
- 11) Intensity
- 12) Plan time
- 13) Training

Sample characteristics

14) Grade

Study context

15) Location

Design Plan

9. *Study type:*

10.

10.1. A systematic review and a meta-analysis.

11. *Blinding:*

11.1. Blinding is always a complicated matter in educational field experiments since it will often be obvious for the participant/students whether they receive the intervention or not. Consequently, we will not exclude any studies due to the lack of blinding.

11.1.1. If some sort of blinding has been used, we will record this.

12. *Study design:*

12.1. *Aggregated meta-analysis* (Cooper & Patall, 2009; T. Pigott, 2012; T. Pigott, Williams, & Polanin, 2012; Riley et al., 2008).⁸

⁸ Meta-analysis should not be confused with other related concept like meta-synthesis and meta-narrative reviews. Meta-analysis in this regard “refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the finding” (Glass, 1976, p. 3). Meta-analysis ” refer specifically to statistical analysis in research synthesis and not to the entire enterprise of research synthesis [/systematic reviewing]” (Cooper et al., 2009, pp. 6–7).

Sampling Plan

13. *Existing data:*

We have finalized the data collection, and we will initiate the final statistical analyses just after we have re-registered this second protocol.

14. *Sample size:*

After our literature searches, snowballing across references we have found approximately 135 studies on which effect sizes can be calculated. However, we only include 75 due to our risk of bias assessment. We will conduct a sensitivity analysis in which we exclude all studies that entail a high/serious risk of bias to examine how the high/serious risk of bias studies affect and potentially inflate the average effect size.

15. *Eligibility criteria for the meta-analysis (added from the PRISMA-P checklist):*

15.1. *Population, Intervention, Comparison, Outcome (PICO) + Time:*

Table 2 – PICO+T statement of the study

The PICO + Time statement of the study

Population	All students enrolled in private or public primary or lower secondary schools, including special schools (as the control group). Geographically limited to “high-income countries” according to the World Bank Classification in 2020. ⁹ We apply this broad definition of the relevant population since we would like to test whether the effects of co-teaching vary between different age groups/year groups/grades. Percent of disabled students in the study population will be recorded as well, if possible. We will only test/include the country covariate in our model if we encounter that effect sizes vary substantially across countries. We do not expect this to be the case. We anticipate that most studies have been conducted in the US.
Intervention variables	Having at least two of legal age teachers (age ≤ 18) in-class during significant parts of a session, i.e. approximately 50 percent of a session. This criterion is set quite arbitrarily, and we will modify this if better arguments

⁹ <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>

come to the fore. Notice, if a class has two teachers but one is pulling out a specific group of students (e.g. students with special needs), we consider such strategies as special education. Nevertheless, we have left this category widely open in order to ensure an inclusive approach towards the body of literature concerning various kinds of two-teacher instruction strategies (Cooper, 2015). Consequently, “having two of legal age teachers during the instruction time” refers to any kind of combination of adults/teachers. It includes e.g. a combination of a general and special education teacher as well as two general education teachers or a general education teacher working together with a teaching assistant/aide without a formal teaching degree. However, it is pivotal that the intervention is provided in-class during regular school time. Two-teacher for after-school or holiday programs will be excluded. For the final analysis, we might end up with a more concise definition of the intervention, but we take this inclusive/open/abstract approach to “allow unexpected operationalizations to get caught in [our] ... search net.” (Cooper, Hedges, & Valentine, 2009, p. 23). See Cooper (2015, p. 37) for similar arguments and the definition of *multiple operationism*.

Comparison groups

Student groups who exclusively have received/receive instruction in general classrooms from *a single* teacher or another type of educational personnel, for example, substitute teachers, and students who have received less than two weeks of co-teaching throughout their entire schooling (see section Time below for further argument regarding this demarcation). Further, we allow special education classrooms to be comparison group since these are often assumed in the literature to be less efficient compared to the inclusive co-teaching classroom for special needs students. We focus on these two comparisons since we consider these being the most relevant and natural alternatives to two-teacher instruction, especially for special needs students. Furthermore, these comparisons most adequately resemble the “treatment as usual” whereas we consider interventions such as reducing class sizes as an alternative intervention to reducing the student-adult ratio. Therefore, we do not employ class size reduction as an eligible comparison since this comparison responds to a question different from the one we aim to answer. In our final model, we will test if effect sizes vary as a function of the control group used to calculate the effect sizes. See section 20 for elaboration.

If/when multiple interventions are compared to the same control group, we account for the dependency between the comparisons via robust variation estimation (RVE). RVE allows maximum use of information retrieved on each study. (Pigott & Polanin, 2019, p. 12; Pustejovsky & Tipton, 2020). To make the most fine-grained analysis of studies with multiple interventions compared to the same control group, we strive to calculate the covariance of effect size measure via equation 19.19 from Gleser

& Olkin (2009) and Wei & Higgins (2013), which will be used for constructing the final covariance matrix for the CHE model.¹⁰ It can be the case that, we will drop constructing individual covariance matrix simply because it can be complex to construct and will take too much time.

**Outcome variables/
Dependent measures**

All kinds of *academic achievement tests* such as *grades, leaving examination, marks for the year's work, national test, large-scale assessment test, teacher-developed test, researcher-developed test, textbook test*, etc. If present, we allow IQ-tests to function as a proxy for student achievement as well if it is measured prior to the posttest scores.

Key outcomes for this study are reading/language arts/foreign language achievement, mathematics achievement, and science achievement. One of the key goals of the study is to test whether effect sizes vary across the taught subjects. Among other things, we record the type of test/measurement (i.e. standardized vs. non-standardized) and we will control for this factor in our final (hierarchical) meta-regression model (Lipsey, 2009; Pigott, 2012, p. 22).

Notice: Due to time constraints, we do not focus upon other potentially important factors such as; *social outcomes, attitudinal outcomes, illegal absence, referrals*, and *mental health-related outcomes*. The effects of two-teacher instruction on the above-mentioned outcomes begs further research.

Time

2-weeks period and above “not including pre- and post-testing” (Murawski & Lee Swanson, 2001, p. 259). We copied this delimitation from the Murawski & Swanson meta-analysis. To substantiate the argument, we do assume and argue that the “isolated” effect of co-teaching is small or non-existent if the students receive less than two weeks of co-teaching during their school attendance.

Further study characteristic:

Eligible study designs for the meta-analysis: Studies to be included will be those using quantitative data for identifying the effects of co-teaching and allowing us to calculate counterfactual effect sizes (i.e. those containing control groups) for co- and team-teaching interventions. This includes studies/designs as:

- 1) (Cluster or/and block) Randomized controlled trials: Where districts, school,

¹⁰ CHE model here refers to the entire family of models mentioned in Pustejovsky & Tipton (2020).

classrooms, or students are assigned randomly to either the treatment or control group

2) Quasi-experimental designs: Where alternative procedures for allocation are used, e.g. date of birth, convenience, etc., but the intervention has been manipulated by the researcher.

3) Observational studies: Where participants or groups are assigned to conditions non-randomly and the researcher does not have any impact on the delivery of the intervention. These include *Regression Discontinuity Designs*, *Propensity Score Matching*, *Exact Matching*, *Matching in general*, *Instrument variable*, *Natural Experiments*, *Difference-in-differences techniques*, etc. (Dietrichson, Bøg, Eiberg, Filges, & Jørgensen, 2016).

We will examine the differences between these designs since it is well known in the educational research that these designs tend to yield substantially different effect sizes with QES and observational studies yielding the largest effect sizes (Cheung & Slavin, 2016). Furthermore, for quasi-experimental and observational studies to be included in this review, the study must somehow examine baseline equivalence or provide some pretest achievement scores. This demarcation of the review is inspired by the below quotations provided by Morris (2008, p. 365):

“The PPC [Pretest-Posttest-Control] design has a number of advantages over other common designs in evaluation research. The posttest only with control design (POWC) has participants assigned to treatment and control conditions, but participants are measured only after administration of the treatment. In quasiexperimental designs, preexisting differences between groups could artificially inflate or obscure differences at posttest, casting suspicion on results from the POWC design. In contrast, the PPC design allows researchers to control for preexisting differences, allowing estimates of treatment effectiveness even when treatment and control groups are nonequivalent (Cook & Campbell, 1979; S. B. Morris & DeShon, 2002). Even in experimental designs, where preexisting group differences are controlled through random assignment, there are advantages to the PPC design. The use of repeated measurements in the PPC

design allows each individual to be used as his or her own control, which typically increases the power and precision of statistical tests (Hunter & Schmidt, 2004)."

- 15.2. *Setting of studies:* We include countries in the "High Income" lending group according to the World Bank¹¹ since we deem these school systems to be more similar compared to school systems in "Low- and Middle Income Countries" – in particular in terms of funding.
- Following Murawski and Swanson (2001, p. 259), we require the interventions to occur in the same physical space, i.e. in the general classroom. As previously mentioned, tutoring of a second teacher outside the classroom or regular school time is not included to count as a two-teacher intervention.
- 15.3. *Timespan of studies:* We are exclusively interested in studies conducted in the period from 1984-2020. We have selected 1984 to be our starting point because an extensive focus on two-teacher approaches begins to surface at this point, and the Project STAR is initiated around the same time. Furthermore, we aim to cover the period following the team-teaching meta-analysis authored by Willet et. al. (1983).
- 15.4. *Language of studies:* We allow studies in English, German, Danish, Swedish, and Norwegian to be included in the present systematic review.
- 15.5. *Publication status of studies:* We allow unpublished¹² and grey literature such as dissertations, conference abstracts/papers, and working papers, etc since it is well known that effect sizes from published and non-published literature vary substantially (Cheung & Slavin, 2016). We will use this information to test for publication bias and related issues (Rothstein et al., 2005).

¹¹ <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>

¹² Unpublished research in this sense refers to "not independently edited or refereed" (White, 2009, p. 61).

16. *Information sources:*

Describe all intended information sources (such as electronic databases, contact with study authors, trial registers, or other grey literature sources) with planned dates of coverage.

Notice: Due to time constraints, we have not been able to screen all databases hosted by EBSCO. Consequently, we have only screened 8006 references instead of approximately 14000 references¹³. However, we have done an extensive amount of snowballing across references, and hence we hope that we will uncover most of the potential missing studies from the EBSCO databases. Nevertheless, this assumption certainly requires future investigation.

16.1. *Databases and data hosts including trial registers and sources for grey literature:*

Due to resource and time constraints, we have not searched all databases and database engines stated in the original preregistered database protocol. The databases and the database engines that we have searched are listed below:

- Scopus
- ProQuest
- APA PsycArticles®
- APA PsycInfo®
- Australian Education Index
- Ebook Central
- EconLit
- Education Database
- ERIC
- Periodicals Archive Online
- ProQuest Dissertation & Thesis Global (for grey literature)
- Web of Science
- Social Science Citation Index (SSCI)
- Science Citation Index Expanded
- Book Citation Index

¹³ Duplicates might be present in this pool of references

- Emerging Sources Index
- British Educational Research Index on EBSCO

Notice: If time allows for it we will fill in the PRISMA-S(earch strategy)

16.2. *Other retrieval – contact with study authors:*

We intend to contact key authors in the field after we have compiled all studies from the literature and the snowballing of references.

17. *Search strategy:*

Scopus search string

TITLE-ABS-KEY("teacher aid*" OR "teacher's aid*" OR "teacher assistant" OR "educational assistant" OR "co*teach*" OR "co*taught" OR "cooperative taught" OR "collaborat* teach*" OR "team teach*" OR "second teacher" OR "team taught*" OR "team-based teach*" OR "classroom teacher collaborat*" OR "cooperative teach*" OR "pull-in instruction" OR "parallel teaching" OR "joint* instruction" OR "team instruction" OR "collaborative instruct*" OR "pull-in teaching" OR "co*instruction" OR "two teacher organi*ation" OR "paraprofessional teaching assistants" OR "extra teacher" OR "spare teacher" OR "station teaching" OR "joint* teach*" OR "transdisciplinary team approach" OR "alternative teaching" OR "alternative teaching" OR "consultation teaching" OR "consultation instruction" OR "co-planned teaching" OR "co-planned instruction*" OR "interdisciplinary team teach*" OR "team teaching instruction" OR "clustering of teachers" OR "co-taught class*" OR "collaborat* teaching" OR "cooperative teaching school*" OR "cooperative teaching class*" OR "complementary instruct*" OR "team taught class*" OR "team-taught*" OR "two-teacher approach" OR "two-teacher strategy" OR "collaborative team teaching" OR "team-teaching school*" OR "team-teaching class*") AND (TITLE-ABS-KEY("RCT" OR "randomized control*" OR "randomised control*" OR "randomised experiments" OR "randomized experiments" OR "experiment" OR "quasi-experimental" OR "fixed effect*" OR "random effect*" OR "large-scale assessment" OR "meta-analysis" OR "systematic review" OR "synthesis" OR "cohort stud*")

OR "pre-test" OR "post-test" OR "case-control" OR "case series" OR "efficacy" OR "treatment" OR "intervention" OR "effect*" OR "outcome*" OR "correlat*" OR "academic achievement" OR "achievement" OR "high school drop-out" OR "upper secondary school drop-out" OR "marks for the year's work" OR "year-end grades" OR "end of year marks" OR "leaving examination" OR "final exam*" OR "achievement test" OR "social outcome*" OR "attitude*" OR "mental health" OR "grade point average" OR "average mark*" OR "transcript" OR "effect size*" OR "grad*" OR "mark*" OR "predict*" OR "association" OR "case stud*" OR "observation*" OR "cluster random*" OR "survey" OR "matching" OR "matched" OR "impact*" OR "performance" OR "consequence*" OR "test*" OR "grade transcript" OR "transcript of record" OR "absence" OR "influen*") AND NOT(KEY("higher educ*" OR "kindergarten" OR "college" OR "undergraduate" OR "post*secondary" OR "pre*school" OR "vocational education")) AND (LIMIT-TO (AFFILCOUNTRY, "United States") OR LIMIT-TO (AFFILCOUNTRY, "United Kingdom") OR LIMIT-TO (AFFILCOUNTRY, "Australia") OR LIMIT-TO (AFFILCOUNTRY, "Canada") OR LIMIT-TO (AFFILCOUNTRY, "Germany") OR LIMIT-TO (AFFILCOUNTRY, "Italy") OR LIMIT-TO (AFFILCOUNTRY, "Netherlands") OR LIMIT-TO (AFFILCOUNTRY, "Spain") OR LIMIT-TO (AFFILCOUNTRY, "France") OR LIMIT-TO (AFFILCOUNTRY, "Japan") OR LIMIT-TO (AFFILCOUNTRY, "Taiwan") OR LIMIT-TO (AFFILCOUNTRY, "South Korea") OR LIMIT-TO (AFFILCOUNTRY, "Sweden") OR LIMIT-TO (AFFILCOUNTRY, "Switzerland") OR LIMIT-TO (AFFILCOUNTRY, "Israel") OR LIMIT-TO (AFFILCOUNTRY, "Belgium") OR LIMIT-TO (AFFILCOUNTRY, "Singapore") OR LIMIT-TO (AFFILCOUNTRY, "Finland") OR LIMIT-TO (AFFILCOUNTRY, "Norway") OR LIMIT-TO (AFFILCOUNTRY, "Ireland") OR LIMIT-TO (AFFILCOUNTRY, "Hong Kong") OR LIMIT-TO (AFFILCOUNTRY, "Greece") OR LIMIT-TO (AFFILCOUNTRY, "Saudi Arabia") OR LIMIT-TO (AFFILCOUNTRY, "New Zealand") OR LIMIT-TO (AFFILCOUNTRY, "Portugal") OR LIMIT-TO (AFFILCOUNTRY, "Austria") OR LIMIT-TO (AFFILCOUNTRY, "Denmark") OR LIMIT-TO (AFFILCOUNTRY, "Poland") OR LIMIT-TO (AFFILCOUNTRY, "United Arab Emirates") OR LIMIT-TO (AFFILCOUNTRY, "Czech Republic") OR LIMIT-TO (AFFILCOUNTRY, "Croatia") OR LIMIT-TO (AFFILCOUNTRY, "Chile") OR LIMIT-TO (AFFILCOUNTRY, "Slovenia")

OR LIMIT-TO (AFFILCOUNTRY,"Cyprus") OR LIMIT-TO (AFFILCOUNTRY,"Hungary") OR LIMIT-TO (AFFILCOUNTRY,"Qatar") OR LIMIT-TO (AFFILCOUNTRY,"Kuwait") OR LIMIT-TO (AFFILCOUNTRY,"Slovakia") OR LIMIT-TO (AFFILCOUNTRY,"Oman") OR LIMIT-TO (AFFILCOUNTRY,"Estonia") OR LIMIT-TO (AFFILCOUNTRY,"Malta") OR LIMIT-TO (AFFILCOUNTRY,"Latvia") OR LIMIT-TO (AFFILCOUNTRY,"Lithuania") OR LIMIT-TO (AFFILCOUNTRY,"Iceland") OR LIMIT-TO (AFFILCOUNTRY,"Luxembourg") OR LIMIT-TO (AFFILCOUNTRY,"Puerto Rico") OR LIMIT-TO (AFFILCOUNTRY,"Bahrain") OR LIMIT-TO (AFFILCOUNTRY,"Trinidad and Tobago") OR LIMIT-TO (AFFILCOUNTRY,"Barbados") OR LIMIT-TO (AFFILCOUNTRY,"Uruguay") OR LIMIT-TO (AFFILCOUNTRY,"Liechtenstein") OR LIMIT-TO (AFFILCOUNTRY,"Bermuda") OR LIMIT-TO (AFFILCOUNTRY,"British Indian Ocean Territory") OR LIMIT-TO (AFFILCOUNTRY,"Monaco") OR LIMIT-TO (AFFILCOUNTRY,"Undefined")) AND (LIMIT-TO (SUBJAREA,"SOC") OR LIMIT-TO (SUBJAREA,"PSYC") OR LIMIT-TO (SUBJAREA,"PHYS") OR LIMIT-TO (SUBJAREA,"MATH") OR LIMIT-TO (SUBJAREA,"BUSI") OR LIMIT-TO (SUBJAREA,"ARTS") OR LIMIT-TO (SUBJAREA,"NEUR") OR LIMIT-TO (SUBJAREA,"DECI") OR LIMIT-TO (SUBJAREA,"ECON") OR LIMIT-TO (SUBJAREA,"CHEM") OR LIMIT-TO (SUBJAREA,"MULT") OR LIMIT-TO (SUBJAREA,"Undefined")) AND (LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015) OR LIMIT-TO (PUBYEAR,2014) OR LIMIT-TO (PUBYEAR,2013) OR LIMIT-TO (PUBYEAR,2012) OR LIMIT-TO (PUBYEAR,2011) OR LIMIT-TO (PUBYEAR,2010) OR LIMIT-TO (PUBYEAR,2009) OR LIMIT-TO (PUBYEAR,2008) OR LIMIT-TO (PUBYEAR,2007) OR LIMIT-TO (PUBYEAR,2006) OR LIMIT-TO (PUBYEAR,2005) OR LIMIT-TO (PUBYEAR,2004) OR LIMIT-TO (PUBYEAR,2001) OR LIMIT-TO (PUBYEAR,2000) OR LIMIT-TO (PUBYEAR,1999) OR LIMIT-TO (PUBYEAR,1998) OR LIMIT-TO (PUBYEAR,1997) OR LIMIT-TO (PUBYEAR,1996) OR LIMIT-TO (PUBYEAR,1995) OR LIMIT-TO (PUBYEAR,1994) OR LIMIT-TO (PUBYEAR,1993) OR LIMIT-TO (PUBYEAR,1992) OR LIMIT-TO (PUBYEAR,1991) OR LIMIT-TO (PUBYEAR,1990) OR LIMIT-TO (PUBYEAR,1989) OR LIMIT-TO (

PUBYEAR,1988) OR LIMIT-TO (PUBYEAR,1987) OR LIMIT-TO (PUBYEAR,1986)
OR LIMIT-TO (PUBYEAR,1985) OR LIMIT-TO (PUBYEAR,1984) OR LIMIT-TO (PUBYEAR,1983)) AND (LIMIT-TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"German") OR LIMIT-TO (LANGUAGE,"Norwegian") OR LIMIT-TO (LANGUAGE,"Swedish") OR LIMIT-TO (LANGUAGE,"Danish"))

PROQUEST full search: 7907 hits

Name:

twoteacher_proquestsearch

Searched for:

noft("teacher aid*" OR "teacher's aid*" OR "teacher assistant" OR "educational assistant" OR "co*teach*" OR "co*taught" OR "cooperative taught" OR "collaborat* teach*" OR "team teach*" OR "second teacher" OR " team taught*" OR "team-based teach*" OR "classroom teacher collaborat*" OR "cooperative teach*" OR "pull-in instruction" OR "parallel teaching" OR "joint* instruction" OR "team instruction" OR "collaborative instruct*" OR "pull-in teaching" OR "co*instruction" OR "two teacher organi*ation" OR "paraprofessional teaching assistants" OR "extra teacher" OR "spare teacher" OR "station teaching" OR "joint* teach*" OR "transdisciplinary team approach" OR "alternative teaching" OR "alternative teaching" OR "consultation teaching" OR "consultation instruction" OR "co-planned teaching" OR "co-planned instruction*" OR "interdisciplinary team teach*" OR "team teaching instruction" OR "clustering of teachers" OR "co-taught class*" OR "collaborat* teaching" OR "cooperative teaching school*" OR "cooperative teaching class*" OR "complementary instruct*" OR "team taught class*" OR "team-taught*" OR "two-teacher approach" OR "two-teacher strategy" OR "collaborative team teaching" OR "team-teaching school*" OR "team-teaching class*") AND noft("RCT" OR "randomized control*" OR "randomised control*" OR "randomised experiments" OR "randomized experiments" OR "experiment" OR "quasi-experimental" OR "fixed effect*" OR "random effect*" OR "large-scale assessment" OR "meta-analysis" OR "systematic review" OR "synthesis" OR "cohort stud*" OR "pre-test" OR "post-test" OR "case-control" OR "case series" OR "efficacy" OR "treatment" OR "intervention" OR "effect*" OR "outcome*" OR "correlat*" OR "academic achievement" OR "achievement" OR "high school drop-out" OR "upper secondary school drop-out" OR "marks for the year's work" OR "year-end grades" OR "end of year marks" OR "leaving examination" OR "final exam*")

Chapter II: Meta-Analysis of the Effects of Collaborative Models of Instruction

OR "achievement test" OR "attitude*" OR "mental health" OR "grade point average" OR "average mark*" OR "transcript" OR "effect size*" OR "grad*" OR "mark*" OR "predict*" OR "association" OR "case stud*" OR "observation*" OR "cluster random*" OR "survey" OR "matching" OR "matched" OR "impact*" OR "performance" OR "consequence*" OR "test*" OR "grade transcript" OR "transcript of record" OR "absence" OR "influenc*") NOT TI,IF,AB("higher educ*" OR "kindergarten" OR "college" OR "undergraduate" OR "post*secondary" OR "pre*school" OR "vocational education") AND pd(1983-2020)

Limited by:

Date: From 1983 to 2020

Databases:

9 databases searched

- APA PsycArticles®
- APA PsycInfo®
- Australian Education Index
- Ebook Central
- EconLit
- Education Database
- ERIC
- Periodicals Archive Online
- ProQuest Dissertations & Theses Global

These databases are searched for part of the query.

Notes:

Saved: June 15 2020

Web of science:

- Exclusions by keywords cannot be made here, so we cannot exclude e.g. kindergarten: have to live with some more hits.

((TS=("teacher aid*" OR "teacher's aid*" OR "teacher assistant" OR "educational assistant" OR "co*teach*" OR "co*taught" OR "cooperative taught" OR "collaborat* teach*" OR "team teach*"

OR "second teacher" OR " team taught*" OR "team-based teach*" OR "classroom teacher collaborat*" OR "cooperative teach*" OR "pull-in instruction" OR "parallel teaching" OR "joint* instruction" OR "team instruction" OR "collaborative instruct*" OR "pull-in teaching" OR "co*instruction" OR "two teacher organi*ation" OR "paraprofessional teaching assistants" OR "extra teacher" OR "spare teacher" OR "station teaching" OR "joint* teach*" OR "transdisciplinary team approach" OR "alternative teaching" OR "alternative teaching" OR "consultation teaching" OR "consultation instruction" OR "co-planned teaching" OR "co-planned instruction*" OR "interdisciplinary team teach*" OR "team teaching instruction" OR "clustering of teachers" OR "co-taught class*" OR "collaborat* teaching" OR "cooperative teaching school*" OR "cooperative teaching class*" OR "complementary instruct*" OR "team taught class*" OR "team-taught*" OR "two-teacher approach" OR "two-teacher strategy" OR "collaborative team teaching" OR "team-teaching school*" OR "team-teaching class*")) AND (TS=("RCT" OR "randomized control*" OR "randomised control*" OR "randomised experiments" OR "randomized experiments" OR "experiment" OR "quasi-experimental" OR "fixed effect*" OR "random effect*" OR "large-scale assessment" OR "meta-analysis" OR "systematic review" OR "synthesis" OR "cohort stud*" OR "pre-test" OR "post-test" OR "case-control" OR "case series" OR "efficacy" OR "treatment" OR "intervention" OR "effect*" OR "outcome*" OR "correlat*" OR "academic achievement" OR "achievement" OR "high school drop-out" OR "upper secondary school drop-out" OR "marks for the year's work" OR "year-end grades" OR "end of year marks" OR "leaving examination" OR "final exam*" OR "achievement test" OR "social outcome*" OR "attitude*" OR "mental health" OR "grade point average" OR "average mark*" OR "transcript" OR "effect size*" OR "grad*" OR "mark*" OR "predict*" OR "association" OR "case stud*" OR "observation*" OR "cluster random*" OR "survey" OR "matching" OR "matched" OR "impact*" OR "performance" OR "consequence*" OR "test*" OR "grade transcript" OR "transcript of record" OR "absence" OR "influenc*")) NOT (TS=("higher educ*" OR "kindergarten" OR "college" OR "undergraduate" OR "post*secondary" OR "pre*school" OR "vocational education")))) AND **LANGUAGE:** (English OR Danish OR German OR Norwegian OR Swedish)

Indexes=SCI-EXPANDED, SSCI, A&HCI, ESCI Timespan=1983-2020

British Education research Index on EBSCO

DE(("teacher aid*" OR "teacher's aid*" OR "teacher assistant" OR "educational assistant" OR "co*teach*" OR "co*taught" OR "cooperative taught" OR "collaborat* teach*" OR "team teach*" OR "second teacher" OR "team taught*" OR "team-based teach*" OR "classroom teacher collaborat*" OR "cooperative teach*" OR "pull-in instruction" OR "parallel teaching" OR "joint* instruction" OR "team instruction" OR "collaborative instruct*" OR "pull-in teaching" OR "co*instruction" OR "two teacher organi*ation" OR "paraprofessional teaching assistants" OR "extra teacher" OR "spare teacher" OR "station teaching" OR "joint* teach*" OR "transdisciplinary team approach" OR "alternative teaching" OR "alternative teaching" OR "consultation teaching" OR "consultation instruction" OR "co-planned teaching" OR "co-planned instruction*" OR "interdisciplinary team teach*" OR "team teaching instruction" OR "clustering of teachers" OR "co-taught class*" OR "collaborat* teaching" OR "cooperative teaching school*" OR "cooperative teaching class*" OR "complementary instruct*" OR "team taught class*" OR "team-taught*" OR "two-teacher approach" OR "two-teacher strategy" OR "collaborative team teaching" OR "team-teaching school*" OR "team-teaching class*")) AND DE (("RCT" OR "randomized control*" OR "randomised control*" OR "randomised experiments" OR "randomized experiments" OR "experiment" OR "quasi-experimental" OR "fixed effect*" OR "random effect*" OR "large-scale assessment" OR "meta-analysis" OR "systematic review" OR "synthesis" OR "cohort stud*" OR "pre-test" OR "post-test" OR "case-control" OR "case series" OR "efficacy" OR "treatment" OR "intervention" OR "effect*" OR "outcome*" OR "correlat*" OR "academic achievement" OR "achievement" OR "high school drop-out" OR "upper secondary school drop-out" OR "marks for the year's work" OR "year-end grades" OR "end of year marks" OR "leaving examination" OR "final exam*" OR "achievement test" OR "social outcome*" OR "attitude*" OR "mental health" OR "grade point average" OR "average mark*" OR "transcript" OR "effect size*" OR "grad*" OR "mark*" OR "predict*" OR "association" OR "case stud*" OR "observation*" OR "cluster random*" OR "survey" OR "matching" OR "matched" OR "impact*" OR "performance" OR "consequence*" OR "test*" OR "grade transcript" OR "transcript of record" OR "absence" OR "influenc*"))) NOT DE (("higher educ*" OR "kindergarten" OR "college" OR "undergraduate" OR "post*secondary" OR "pre*school" OR "vocational education"))

➔ Restricted by year of publication

18. *Study records:*

18.1. *Data management:* Covidence is the main management tool we use for literature retrieval, literature management (such as duplicate assessment, abstract screening, and full-text screening). We use excel for the data extraction and the Risk of Bias (RoB) assessment. Will we document all effect size calculations via Rmarkdown so that readers of this review can assess and follow all parts of the computation. Find the all effect size calculations at <https://osf.io/awb5s/>.

18.2. *Selection process:*

Stage characteristic:

- 1) Defining the problem: done by the authors exclusively.
- 2) Literature search: To control the process of the search, Weiss has conducted the literature search. However, we have been supervised by our institutional affiliated librarian Lars Jakob Jensen from Aarhus University.
- 3) Screening, Coding, and judging of eligibility of the literature: has been done by the author independently of each other. See section 17.3 below.
- 4) Analyses and interpretations of the literature: will solely be conducted by all the authors in collaboration.

18.3. *Data collection procedures: (added from the PRISMA-P checklist):*

As far as possible, we followed advice from Cooper et al. (2009) and Higgins & Thomas (2019) regarding data collecting. Due to resource constraints, however, fully independent double coding will only be carried out for the abstract and full-text screening. We have piloted tested all of our coding tools.

Variables/ Data items

19. *Measured variables:*

See Tables 1 and 2

19.1. *Intervention/independent variable:*

See Table 2 for an overview.

19.2. *Control variables/participant/sample population:*

See Table 2 for an overview.

19.3. *Study-test variables:*

See Tables 1 and 2 for an overview.

Analysis Plan

This section aims to describe the statistical analytical strategy of the study.

20. *Effect size calculation strategies (added ourselves)*

20.1. *Obtaining effect size data from primary studies:*

If studies report multiple estimates eligible for effect size calculation, we always obtain the estimates from the model controlling most covariates. Furthermore, if a study reports estimates from all or some sub-tests and the aggregate measures of a given test battery from the same subject, we will only make use of the aggregated estimates, since we have no assumptions about how the effects of collaborative models of instruction might vary as a function of the specific content areas, and which goes beyond the major aim of this review.

Even if fully randomized, we will always prioritize pre-posttest/covariates adjusted measures above posttest measures only. If a study both reports pre-posttest

measures and mean gain scores, we will calculate effect sizes from the raw pre-posttest measure, and use the mean gain scores to investigate if any divergence appears among the two types of effect sizes. In a similar vein, we commit ourselves to retrieving all information relevant for different effect size calculation approaches so that we via a sensitivity analysis can test the impact of the given method of effect size calculation on our estimation.

We will strive to extract all relevant information that studies provide regarding the correlations between repeated measures (e.g. pre-posttest correlation), between-outcome correlation, and/or the covariance of multiple outcomes, respectively. This information is partly key to our difference-in-differences effect size calculation and partly key to the construction of our covariance-variance matrices behind our multivariate CHE models (Pustejovsky & Tipton, 2021). See section 20 for elaboration. Furthermore, we use the pre-posttest correlations to impute into similar studies from which we cannot obtain the pre-posttest correlation but in which pre-post tests are reported.

If studies provide subgroup analyses of the type of student (general vs. special needs students) we will retrieve this data.

20.2. *How to calculate effect size*

We apply several computational methods for the effect size calculation since general textbook examples of effect size calculation of standardized mean difference can be severely biased when applied on various common control group designs and reported estimates encountered in education and the social sciences, especially pre-posttest designs and covariate-adjusted measure. We draw on Borenstein (2009), Higgins & Thomas (2019), Pustejovsky (2016), Hedges (2007), the WWC Procedure Handbook (2020, 2021), and Wilson (2016) for our effect size computations. We will use the pooled post-test standard deviation for all effect sizes. If only the pretest standard deviation is available, we will use that quantity for the effect size calculation. If we encounter that a large number of effect sizes only can be obtained

via alternative standard deviations or effect sizes such as Glass' delta, we will control for this factor in our models. We don't expect this to be the case. All effect size calculations will be available on the associated OSF webpage.

If a study yields a pre-posttest measure but does not provide enough information from which we can calculate the pre-posttest correlation, we will impute the pre-posttest correlation as suggested by WWC (See 2020 Appendix E).

When pre-posttest correlations can be calculated from the study, we pool the correlations by transforming the correlations to the Fisher's z scale and taking the weighted mean of these estimates. Afterward, we convert it back to a mean correlation measure again (Borenstein, Hedges, Higgins, & Rothstein, 2009, pp. 99–100).

We will calculate all effect sizes which is possible given the data from the study to examine how effect sizes vary across different methods of effect size computation. As already mentioned, we will employ alternative computations to perform a sensitivity analysis concerning the impact of the computational methods on our results. However, if studies provide more than one alternative effect size measure, we will extract the most divergent one, only. The alternative effect size/outcome vector (alternative dependent variable) will consist of *the* alternative effect size if only one alternative exists or the most extreme effect size measure if more than one alternative appears. For studies in which we can only calculate effect sizes via one computational method, the effect size(s) will be copied from the “correct” effect size/outcome vector to the alternative effect size/outcome vector.

Effect size adjustment

Effect sizes will be based on Hedges' g , which is a part of the d -family but in which Cohen's d is corrected for its tendency to overestimate the effect size in small sample sizes (Borenstein et al., 2009, p. 27). We consider this to be an important correction since we expect a great number of smaller studies to be included in the final dataset.

Cluster design adjustment

All studies that do not account for clustering on the class level will be cluster design corrected on the missing level (Hedges, 2007). For the cluster design adjustment, we expect that we most often have to impute a value for the intraclass correlation (ICC) (if not reported by the study authors). We retrieve the imputed ICC from the unconditional models from Tables 2 and 3 in Hedges & Hedberg (2007) suggested by Hedges (2007). Notice that we use ICCs from mathematic achievement when we cluster-correct science effect sizes, and ICCs from reading achievement when we cluster-correct general language arts effect sizes, respectively.

We follow the WWC (2021) and Cochrane (2019) guidelines for estimating cluster design adjusted effect sizes. We cluster design correct studies although students have been individually randomized to the treatment or control group because collaborative models of instruction are delivered at the class level, which likely will produce dependency among students with clusters (Higgins et al., 2019, p. 576). Find all formulas in WWC's Supplement material to appendix E¹⁴.

21. *Statistical models (required)*

21.1. *Weighted mean effect size model (intercept-only meta-regression model)*

In the final analysis, we will construct step-wise models meaning that we first fit random-effects models for the mean effect size model via the correlated-hierarchical effects (CHE) model which is simply a meta-regression intercept-only model. Second, we anticipate to fit sub-group models (meta-regression models with one focal covariate and possibly with some focal control factor) by either using the multivariate SCE model (Subgroup Correlated Effects) or the CMVE model (Correlated Multivariate Effects) depending on whether the covariate contains a substantial amount of within-study variance which is necessary for the latter model to be applicable. Lastly, we fit a meta-regression model with all focal covariates employing the CHE model. As all the above models entails, we use RVE to guard against model misspecifications (Pustejovsky & Tipton, 2020).

¹⁴ https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-41-Supplement-508_09212020.pdf

It might happen that we will build models that contain more random effect(s) and corresponding variance component(s) than just the true between-study and true within-study effects and variances.

The main advantage of using the combination of multivariate/multilevel modeling and RVE is that we simultaneously can estimate the true heterogeneity of effect size within-study and between-study via restricted maximum likelihood (REML) and guard against misspecifications of the model. Further, we can use fully-inverse weights under the working model (i.e. using all true variance of the working model to create the weights).¹⁵ These are less biased and most efficient, (i.e. most precise) compared to the correlated effects model (the original RVE approach we intended to use) that only applies “approximately efficient” weights which are not fully inverse-variance weights with respect to their working model (Pustejovsky & Tipton, 2020) and which ignores the true within-study variance. Being able to estimate true variances components provide helpful diagnostic information on which level more moderators are needed for understanding the true variation in the collaborative models of instruction literature. Notice that since the REML variance component estimates are quite sensitive to the choice of between-outcome correlation (ρ) for studies where this information is unknown, we might refrain from commenting on the relative magnitude of the variance components (Pustejovsky & Tipton, 2020, p. 30). However, we will use *profile likelihood plots* of the variance estimations to examine whether it is reliable to assume that the variance components are identifiable. Moreover, we will use the variances components to estimate ICC correlations across levels to understand how strongly effect sizes are correlated within clusters. Hereto, we will use the variance components to estimate “the total amount of heterogeneity in true effects” (Viechtbauer, 2020)¹⁶ in the model.

We assume a *common mean between outcomes within-study correlation* which we estimate “by calculating the Pearson correlation between the pairs of

¹⁵ To gain further insight of the weighting scheme behind these models, see <https://www.jepusto.com/weighting-in-multivariate-meta-analysis/> and http://www.metafor-project.org/doku.php/tips:weights_in_rma.mv_models

¹⁶ See <https://www.metafor-project.org/doku.php/analyses:konstantopoulos2011>

available treatment effect estimates in those studies that provide data on both outcomes” (Kirkham, Riley, & Williamson, 2012, p. 2182), i.e. in our cases math/science and reading/language outcomes. This means that we will assume a “constant sample correlation” within studies. We will conduct sensitivity analyses by varying the between outcome within-study correlation to investigate how this assumption impacts the final results. For all analyses, we apply the `metafor` package in R (Viechtbauer, 2010) to estimate the multivariate/multi-level models, and to guard against misspecifications and to impute the covariance-variance matrices we use the `clubSandwich` package in R (Pustejovsky, 2020).

Important inference statistics for our interpretation will be the *p-values and confidence intervals for the mean effect size*, *Q-statistics*, $\sigma_1^2, \dots, \sigma_n^2$, and I^2 .¹⁷ The *p-value* and confidence intervals of the mean effect size are partly used to examine whether the mean effect size is different from null, partly to assess the magnitude of the average mean effect size, and partly to explore the precision of the estimated average mean effect size (the latter two points are overlapping). We apply the small-sample adjusted *t*-test with Satterthwaite approximated degrees of freedom (Tipton, 2015). We use *Q* (and the *p-value* of *Q*), $\sigma_1^2, \dots, \sigma_n^2$, and I^2 to explore whether it is reasonable to assume that moderator mechanisms exist and to assess what percentage of the variance that can be attributed to *true variation* of effect sizes. As the prior literature in the field of co-teaching suggests (Khoury, 2014), we expect to find a significant amount of heterogeneity when assessing the between-study variance of the mean effect size mainly due to our broad inclusion criteria of eligible study designs. Since prediction intervals have not yet been developed for meta-analysis with multiple outcomes, i.e. in our case the CHE-family models, we cannot follow the advice from Borenstein (2019) as initially intended regarding that the prediction interval should be the main index for assessing the specific amount and range of heterogeneity. If any new techniques surface, we intend to apply and present these.

¹⁷ We will compute I^2 for our multi-level models like [https://www.metafor-project.org/doku.php/tips:i2_multi-level_multivariate?s\[\]=i2](https://www.metafor-project.org/doku.php/tips:i2_multi-level_multivariate?s[]=i2)

For model checking of all our models, we are conducting *leave-one-out sensitivity analysis*, i.e. that we leave out one full study at a time. Furthermore, we will check if the model substantially captures the true variation in the dataset compared to models with fewer levels (Harrer, Cuijpers, Furukawa, & Ebert, 2019).

Subgroup analysis (meta-regression models from which we interpret one covariate only)

For some individual factors of critical importance, we will estimate classical-like sub-group analyses individually. All covariates used for the sub-group analyses will subsequently also be fitted in the full meta-regression model to avoid unexpected confounding among sub-group factors and to examine the impact on the results of the sub-group analysis. We will estimate the sub-group analyses either via SCE models or CMVE model depending on whether the covariate substantially varies within studies. Otherwise, it will not be possible to reliably fit CMVE models reliably (Pustejovsky & Tipton, 2020). All covariates that vary substantially within studies will be centered to avoid contrary results within and across studies effects (E. E. Tanner-Smith & Tipton, 2014). We use these models because they both can be interpreted as classical subgroup analyses and the estimates can be statistically compared. We will estimate the model via the `metafor` package and guard against misspecification of the model by use of the `clubSandwich` package. Statistically, this means that we use small-sample adjusted *F*-tests for our multi-contrast tests (Tipton & Pustejovsky, 2015), which yields approximate Hotelling T-squared test with Zhang-type degrees of freedom (AHT(Z)). However, new research suggests that the HTZ methods can be conservative and that cluster wild bootstrapping (CWB) better “controls for Type I error rate and has more power than the HTZ test” (Joshi, 2021). Pending the availability of software or code for implementing CWB, we will use it for all inferential tests. If software/code is not yet available at the time of analysis, we will use the Approximate Hotelling T-squared test with Zhang-type degrees of freedom (AHT(Z)). If CWB software/code becomes available prior to submission of our study, we will update the results using

CWB for the primary analysis and report AHT(Z) as a sensitivity analysis. This also counts for the intercept-only model.

We will conduct subgroup analysis in which we examine the impact of several focal moderators that we deem to be of special interest to the fields of co- and team-teaching such as *subject taught*, *intervention mode*, *type of test/content*, *publication type*, *type of control group*. We will also conduct a design-specific subgroup analysis regarding the use of randomization (see Dietrichson et al. (2017) and Cooper et al. (Cooper, Hedges, & Valentine, 2019) for a similar approach). Since we expect to conduct several subgroup analyses, we run the risk of increasing the probability of committing Type I error rate merely because of the number of analyses. This fact is referred to as the “family-wise error rate” or multiplicity. Therefore, we will correct for multiplicity by using the *false discovery rate method*, suggested by Polanin (2013). This means, that we will lower the acceptable *p*-value of all meta-regression models (fitted with one or more covariates).

We use *mixed-effects models* for our subgroup analyses in which the studies within the specific subgroup are treated as random, but where the subgroups are treated as fixed (Borenstein, 2019). Generally, we will use the same statistical inference components as for the univariate analysis of the weighted mean effect size. However, we apply the pooled τ^2 estimate based on calculations of distinct τ^2 estimations from within each subgroup to examine if the sub-group indexation aims to explain a substantial part/the reduction of the between-study variance (Borenstein, 2019, p. 199). See Table 1 for our hypotheses regarding the direction of the results of the subgroup analyses.

The above analysis strategy is based on the ideal case, in which we can obtain all relevant or at least a substantial amount of information for many studies. Therefore, the exact number of covariates used for all of our meta-regression models might vary in the final study.

Meta-regression (substantial interpretation of more than one covariate)

We expect to fit the full meta-regression model by using the CHE model. In this model, we examine the relationship between several factors from different levels

i.e. *study design characteristics/level, outcome characteristics, participant characteristics, intervention characteristics, general study characteristics, and effect size calculation features* based on theory and prior empirical workings in the fields of collaborative models of instruction and meta-analysis, respectively. (see Table 1, p. 7, for elaboration) Student achievement (i.e. reading/language arts, math, and science achievement) is the main dependent variable of the study/model. Table 1 represents the ideal meta-regression model but from a realistic perspective, we don't expect that the literature of collaborative models of instruction is rich enough on information to configure such a comprehensive model. We correct our model for multiplicity.

Will present the covariance-variance matrix of the meta-regression model to examine if multicollinearity is present in the model.

*Risk of bias (RoB)*¹⁸

Since we allow for non-randomized (QES) and observational studies to be included in the meta-analysis, we both apply the risk-of-bias (RoB 2) tool for all randomized controlled trials, the RoB 2 CRCT tool for all cluster RCTs, and the ROBINS-I tool¹⁹ (Risk Of Bias In Non-randomized Studies of Interventions) for all non-randomized or observational intervention studies. Since we allow various kinds of publication to be included in the review, we find it critical to conduct RoB assessments since “trials which are difficult to locate are often of lower quality raises the worrying possibility that rather than preventing bias through extensive literature searches, bias could be introduced by including trials of low methodological quality” (Egger, Juni, Bartlett, Holenstein, & Sterne, 2003, p. iv). We exclude QES and observational studies if they retrieve a critical risk of bias judgement, and due to time resources will stop the assessment just after the first critical risk of bias judgement is reached. This means that a study might have multiple errors but this is outside our concern for this review.

¹⁸ This part is in many parts heavily inspired by Filges et al. (2015)

¹⁹ See <https://www.riskofbias.info/welcome/home>.

When we assess the risk-of-bias for non-randomized studies, we will direct a careful focus towards confounding factors of which we deem to be of special importance for the co- and team-teaching literature. These include across-arms balance or control of *age, grade level, performance at baseline, gender, socioeconomic background, number of students with special needs, and demography of school* (i.e. urban/suburban vs. rural). These factors are considered to be of critical importance for several distinct reasons. We anticipate that the cognitive function of students is heavily related to the age of the students, and therefore, we deem it to all-important that this factor is equally balanced or controlled out in non-randomized study if the given results are to be trusted. We will exclude studies instantly if they apply two different year/age groups for the treatment and control group. In a similar vein, and based on Lipsey et al. (2012), it is evident that the grade level is a significant predictor of student achievement. Consequently, this factor must be balanced between the treatment and control group in non-randomized studies if they are to be interpreted as trustworthy. Deduced from the same literature, it appears that males overall and on average have a lower academic performance relative to females, especially in language arts subjects (Bloom et al., 2008). For this reason, we consider this factor to be an important balance/control factor in non-randomized studies. Inspired by Filges et al. (2015), we assume that the teacher-ratio may be negatively correlated with students' socioeconomic background similar to what is valid in the field of class size reduction. In general, we anticipate students from more advantaged socio-economic backgrounds to have a greater gain of two-teacher instruction partly due to for example an increased amount of upward perception bias (Jæger, 2011). Socioeconomic factors which we deem to be important to be balanced in non-randomized studies are parents' level of education, family income (outside the Scandinavian countries), minority background, etc. Moreover, if some of these factors aren't balanced the given study risks contain a serious bias because students with socioeconomic disadvantaged backgrounds seem to perform unsatisfactorily in achievement tests (Filges et al., 2015, p. 19). However, we consider it to be acceptable if just one of the above-mentioned factors is balanced/controlled out.

The co-teaching literature suggests that the inclusion of too many (approximately more than five disabled students) can have a negative impact on the effect of co-teaching (Cook et al., 2016). Therefore, it seems to be relevant to take this into account when assessing important balancing/controlling factors in non-randomized studies.

The last balance/control factor, we deem to be of substantial importance is the demography of the school since much indicates that urban schools are better equipped in terms of material resources such as better libraries, etc. relative to rural schools and that this has a significant impact on student learning. Thus it might seem to be important that non-randomized studies take this factor into account to be trustworthy.

We consider pre-posttest design studies to be of moderate risk of bias due to confounding even if they don't focus on the above balance factors.

We will upload all RoB schemes to OSF.

Sensitivity Analyses

We will perform several sensitivity analyses. First, we will check the impact of the alternative-calculated effect size estimates on the final results. Second, we will test if results change if had chosen not to include high/serious risk of bias studies. Third, we will conduct sensitivity analyses for publication biases from Mathur & VanderWeele (2020) and Stanley, Doucouliagos, & Ioannidis (2017). If only a few effect size estimates from our meta-analysis database are well powered, we will not apply the *weighted average of the adequately powered* (WAAP) method. We will also conduct a *leave-one-out* analysis. Fourth, we will test if results are robust to the imputed pre-posttest correlation used for effect size calculation for pre-posttest designed studies from which it was not possible to retrieve any pre-posttest correlation to calculate the variance estimation. Finally, we will test how and if results change had we either imputed ICC values for the cluster bias adjustment differently or had ignored cluster issues.

If software or codes become available from which we can apply cluster wild

bootstrapping (CWB), we will compare the results between CWB and HTZ methods. More sensitivity analyses might be added in the final analysis if we uncover a pressing necessity.

22. Transformations (optional)

We will use *variance-stabilizing* for assessing selective reporting/publication bias/small study bias (Hedges & Olkin, 1985; Pustejovsky & Rodgers, 2019, p. 60; Rothstein et al., 2005). This is necessary for removing the artificial correlation between the variance and the effect size component which naturally occur due to the fact that the estimated effect size obtained from the given primary study is included in the formula underpinning the calculation of the variance of the standardized mean difference (Pustejovsky & Rodgers, 2019; Rodgers & Pustejovsky, 2019). The variance-stabilized estimates will be used for Egger's Sandwich test based on CHE models and not CE models as suggested in Pustejovsky & Rodgers (2019), Funnel Plot Asymmetry test with multiple outcomes, Trim and Fill test with multiple outcomes, Cumulative Analysis. We do also conduct three-parameter selection models (3PSM) either via the `weightr` package (Coburn & Vevea, 2019) or the `metafor` package in R (Viechtbauer, 2010). For this model, we apply a modified version of the precision $W_i = \text{Var}(\hat{\Delta})/\sigma_j^2$. This has shown to be the best way to apply the 3PSM model since this approach adequately controls Type I error rates and have higher convergence rates (meaning that it for most calculations correctly include the true parameter in the confidence interval) compared to variance-stabilized estimation. However, the variance-stabilizing transformation has a power advantage over all the other tests, and that is why we use variance-stabilized for all other small study tests (Pustejovsky & Rodgers, 2019). It is not yet possible to use the RVE framework for 3PSM models. Consequently, we will both apply the mean effect size of each study and repeated random sampling of one effect size from each multi-effects study, when we conduct 3PSM models as suggested by Rodgers & Pustejovsky (2019, p. 45). Moreover, we will fit the 3PSM model if find more than 40 studies.

23. Inference criteria (optional)

23.1. Is mentioned in section 19

24. Data exclusion (optional)

24.1. We exclude QES and observational studies that are judged to contain a critical risk of bias for at least one ROBINS-I domain. Furthermore, we exclude studies and effect sizes if we are entirely certain that the results are error-prone.

25. Missing data (optional)

25.1. We apply multiple imputation (MI) to handle missing data. Before our conduct of MI, we will conduct explorative missing data analyses to examine the potential missing data patterns (Schauer, Diaz, Lee, & Pigott, 2020). This will help us to construct the most accurate imputation model. To reduce biased MI results, we will handle all substantial missing data of independent variables via *multi-level multiple imputation* (Buuren & Groothuis-Oudshoorn, 2010; Van Buuren, 2018) so that we also take into account the nesting structure of our data in our handling of missing data. We expect to generate at least 20 datasets premised upon multiple imputations, and we will use Bernard & Rubin's (1999) small sample correction to pool results which are currently most adequate to use for pooling statistics premised upon cluster-robust variance estimators (Pustejovsky, 2017). If good predictors for variables with many missings are available, we will allow covariates with 60 percent missings to be included in the regression model.

When it comes to missing data for the dependent variable i.e. for the effect sizes, we conducted tests for selective reporting via several methods which are recommended in the meta-analytical literature (Rodgers & Pustejovsky, 2019; Rothstein et al., 2005). These tests include Eggers Sandwich based on CHE-models (Rodgers

& Pustejovsky, 2019), Funnel Plot Asymmetry with multiple outcomes, and Trim and Fill test with multiple outcomes²⁰. Finally, we will apply the 3PSM.

26. Exploratory analysis (optional)

26.1. See section 22.1

Other

27. Other (Optional)

27.1. This study heavily draws on the workings of (Cooper et al., 2009; Dietrichson et al., 2017; Fisher & Tipton, 2015; Hedges & Pigott, 2001, 2004; Hedges, Tipton, & Johnson, 2010; T. Pigott, 2012; T. D. Pigott & Polanin, 2019; J. R. . Polanin, Espelage, & Grotmeter, 2018; Pustejovsky & Rodgers, 2019; Pustejovsky & Tipton, 2020; Rodgers & Pustejovsky, 2019; E. Tanner-Smith, Tipton, & Polanin, 2016; Tipton, 2015; Valentine et al., 2010; Viechtbauer, 2010).

All relevant documents linked to this study including codes behind all statistical procedures will be uploaded to <https://osf.io/fby7w/>.

²⁰ In this regard, we are heavily inspired by the working of Rodgers & Pustejovsky.

References

- Allerup, P. (2006). Det kræver kørekort at bruge P-værdier: en replik til en analyse af skolestørrelser. *Pædagogisk Psykologisk Tidsskrift*, 43(6), 599–610.
- Andersen, S. C., Beuchert-Pedersen, L. V., Nielsen, H. S., & Thomsen, M. K. (2018). The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial. *Journal of the European Economic Association*.
- 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*. <https://doi.org/10.2139/ssrn.2626677>
- Barnard, J., & Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948–955. <https://doi.org/10.1093/biomet/86.4.948>
- Blatchford, P., Russell, A., & Webster, R. (2012). *Reassessing the impact of teaching assistants: How research challenges practice and policy*. Routledge.
- Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289–328.
- 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). Russel Sage foundation.
- Borenstein, M. (2019). *Common Mistakes in Meta-Analysis: And How to Avoid Them*. Retrieved from <https://books.google.dk/books?id=N4vXyAEACAAJ>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.
- Buuren, S. van, & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.
- Cartwright, N. (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy*, 23(1), 143–155.
- 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>
- Coburn, K. M., & Vevea, J. L. (2019). Package 'weightr.' Retrved from [Https://Cran.Rproject](https://Cran.Rproject).

Org/Web/Packages/Weightr/Weightr. Pdf.

- Cook, B. G., McDuffie-Landrum, K. A., Oshita, L., & Cook, S. C. (2016). Co-teaching for Students With Disabilities: A Critical and Updated Analysis of the Empirical Literature. *Handbook of Special Education*, 233–248.
- Cooper, H. (2015). *Research Synthesis and Meta-Analysis: A step-by-step approach* (Vol. 2). Sage publications.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *Handbook of Research Synthesis and Meta-Analysis* (2nd ed.). New York: Russell Sage Foundation.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis* (3rd ed.). New York: Russell Sage Foundation.
- Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14(2), 165.
- Dietrichson, J., Bøg, M., Eiberg, M., Filges, T., & Jørgensen, A.-M. K. (2016). Protocol for a systematic review: Targeted School-Based Interventions for Improving Reading and Mathematics for Students With or At-Risk of Academic Difficulties in Grade K to 6: A Systematic Review. *Campbell Systematic Reviews*, 12(1), 1–60.
- 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.
- Dyssegaard, C. B., & Larsen, M. S. (2013). *Evidence on inclusion*.
- Dyssegaard, C. B., Larsen, M. S., & Tiftikçi, N. (2013). Effekt og pædagogisk indsats ved inklusion af børn med særlige behov i grundskolen. *Systematisk Review*. København: IUP, Aarhus Universitet.
- Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment (Winchester, England)*, 7(1), 1–82. <https://doi.org/10.3310/hta7010>
- 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.

- Filges, T., Sonne-Schmidt, C. S., & Jørgensen, A. M. K. (2015). Protocol: Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *The Campbell Collaboration*.
- 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.
- Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. *ArXiv Preprint ArXiv:1503.02220*.
- Freese, J., & Peterson, D. (2017). Replication in Social Science. *Ssrn*.
<https://doi.org/10.1146/annurev-soc-060116-053450>
- Friend, M. (2017). *Co-teaching i praksis : samarbejde om inkluderende læringsfællesskaber*. (1. udgave.). Frederikshavn: Dafolo.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3–8.
- Glass, G. V., & Smith, M. L. (1979). Meta-Analysis of Research on Class Size and Achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2–16.
<https://doi.org/10.3102/01623737001001002>
- Gleser, L., & Olkin, I. (2009). Stochastically dependent effect sizes. *The Handbook of Research Synthesis and Meta-Analysis*, 357–376.
- Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hands-on Guide*. Retrieved from
https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/
- 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). <https://doi.org/10.1002/0470870168.ch9>
- 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. (2019). The Statistics of Replication. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 15(S1), 3–14.
<https://doi.org/10.1027/1614-2241/a000173>
- 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* (L. V Hedges & I. Olkin, Eds.). London: Academic Press.
- Hedges, L. V, & Pigott, T. D. (2001). The Power of Statistical Tests in Meta-Analysis. *Psychological Methods*, 6(3), 203.
- Hedges, L. V, & Pigott, T. D. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods*, 9(4), 426.
- Hedges, L. V, & Stock, W. (1983). The Effects of Class Size: An Examination of Rival Hypotheses. *American Educational Research Journal*, 20(1), 63. <https://doi.org/10.3102/00028312020001063>
- Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Erratum: Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(2), 39–65. <https://doi.org/10.1002/jrsm.17>
- 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*.
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177.
- IES, & NFS. (2018). *Companion Guidelines on Replication & Reproducibility in Education Research A Supplement to the Common Guidelines for*. The National Science Foundation & The Institute of Education Sciences, U.S. Department of Education.
- Jæger, M. M. (2011). Does cultural capital really affect academic achievement? New evidence from combined sibling and panel data. *Sociology of Education*, 84(4), 281–298.
- Joshi, M. (2021). wildmeta. Retrieved January 9, 2021, from <https://github.com/meghapsimatrix/wildmeta>
- 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). Retrieved from <https://search.proquest.com/docview/1817570306?accountid=14468> NS -
- 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/https://doi.org/10.1002/sim.5356>

- Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 0013189X20912798. <https://doi.org/10.3102/0013189X20912798>
- Lipsey, M. W. (2009). Identifying Interesting Variables and Analysis Opportunities. *The Handbook of Research Synthesis and Meta-Analysis*, 2, 147–158.
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... 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*.
- 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>
- 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>
- Murawski, W. W., & Lee Swanson, H. (2001). A Meta-Analysis of Co-Teaching Research: Where Are the Data? *Remedial and Special Education*, 22(5), 258–267.
- Pigott, T. (2012). *Advances in Meta-Analysis*. Springer Science & Business Media.
- Pigott, T. D., & Polanin, J. R. (2019). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, 0034654319877153.
- Pigott, T., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods*, 3(4), 257–268.
- Polanin, J. R. (2013). *Addressing the issue of meta-analysis multiplicity in education and psychology*. Loyola University Chicago.
- Polanin, J. R. ., Espelage, D. L. ., & Grotper, J. (2018). *The Consequences of School Violence: A Systematic Review and Meta-Analysis Review Protocol*. Retrieved from <https://osf.io/6hak7/>
- Pustejovsky, J. E. (2016). Alternative formulas for the standardized mean difference. Retrieved from <https://www.jepusto.com/alternative-formulas-for-the-smd/>
- Pustejovsky, J. E. (2017). Pooling clubSandwich results across multiple imputations.
- Pustejovsky, J. E. (2020). *clubSandwich: Cluster-robust (sandwich) variance estimators with*

- small-sample corrections. R package version 0.5.0.* Retrieved from <https://cran.r-project.org/web/packages/clubSandwich/index.html>
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71.
- Pustejovsky, J. E., & Tipton, E. (2020). *Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models*. Retrieved from <https://osf.io/x8yre/>
- Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with Robust Variance Estimation: Expanding the range of working models. *Prevention Science*, 1–14.
- Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L., & Bouillon-Bu, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, 27(11), 1870–1893.
- Rodgers, M. A., & Pustejovsky, J. E. (2019). *Evaluating Meta-Analytic Methods to Detect Selective Reporting in the Presence of Dependent Effect Sizes*.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication Bias in Meta-Analysis. In *Publication bias in meta-analysis: Prevention, Assessment and Adjustments*. Wiley Online Library.
- Schauer, J., Diaz, K. G., Lee, J., & Pigott, T. (2020). *Exploratory Analyses for Missing Data in Meta-Analyses*.
- 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>
- Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and spss. *Research Synthesis Methods*, 5(1), 13–30. <https://doi.org/10.1002/jrsm.1091>
- Tanner-Smith, E., Tipton, E., & Polanin, J. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: a Tutorial in R. *Journal of Developmental and Life-Course Criminology*, 2(1), 85–112. <https://doi.org/10.1007/s40865-016-0026-5>
- 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>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194. <https://doi.org/10.1002/jrsm.1339>
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <https://doi.org/10.3102/1076998609346961>
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
- 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>
- Viechtbauer, W. (2020). Konstantopoulos (2011). Retrieved January 8, 2021, from <https://www.metafor-project.org/doku.php/analyses:konstantopoulos2011>
- Wei, Y., & Higgins, J. P. T. (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(7), 1191–1205. <https://doi.org/10.1002/sim.5679>
- White, H. D. (2009). Scientific Communication and Literature Retrieval. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 51–71). New York: Russel Sage foundation.
- 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.
- Wilson, D. B. (2016). *Formulas Used by the “Practical Meta-Analysis Effect Size Calculator.”*
- WWC. (2020). *WWC Procedures and Standards Handbook (version 4.1)*. Retrieved from <https://ies.ed.gov/ncee/wwc/Handbooks>
- WWC. (2021). *Supplement Document for Appendix E og the What Works Clearinghouse Procedures Handbook, Version 4.1*. Institute of Education Sciences.

Chapter III

Power Approximations for Meta-Analysis of Dependent Effect Sizes

Mikkel H. Vembye, James E. Pustejovsky, & Terri D. Pigott

Accepted in *Journal of Educational and Behavioral Statistics*¹

¹ Find a revised version of this article at <https://osf.io/preprints/metaarxiv/6tp9y/>

Abstract

Meta-analytic models for dependent effect sizes have grown increasingly sophisticated over the last few decades, which has created challenges for a priori power calculations. We introduce power approximations for tests of average effect sizes based upon the most common models for handling dependent effect sizes. In a Monte Carlo simulation, we show that the new power formulas can accurately approximate the true power of common meta-analytic models for dependent effect sizes. Lastly, we investigate the Type I error rate and power for several common models, finding that tests using robust variance estimation provide better Type I error calibration than tests with model-based variance estimation. We consider implications for practice with respect to selecting a working model and an inferential approach.

KEYWORDS: *power, meta-analysis, dependent effect sizes, CHE model, robust variance estimation*

Introduction

Meta-analyses in the social and behavioral sciences typically include studies that report on multiple outcomes measured on the same sample. Recent research in meta-analysis (Pustejovsky & Tipton, 2021; van den Noortgate et al., 2013) provides models that better reflect the complex error structure of such effect size data, recognizing the dependence among effect sizes within studies and accounting for the multilevel nature of the data. As these models come into wider use, it is important to understand their performance given the complex structure of many meta-analysis data sets. One critical aspect of performance is the statistical power of the model to detect a non-null average effect size.

Power analysis in meta-analysis can provide insight about the potential utility of a planned systematic review. Conducting an *a priori* power analysis helps researchers determine whether the existing evidence base is large enough to detect an effect size of substantive importance, so that both researchers and potential funders can judge if the literature is mature enough for a systematic review. An *a priori* power analysis can also guide decisions about potential meta-analytic models. Meta-analysts are employing more complex models that reflect the multilevel and correlated nature of effect size data, and these models have greater data requirements than traditional, independent effect size models. As illustrated later in this paper, statistical power to detect a non-null average effect size may differ depending on both the nature of effect size data and the model used to approximate the distribution of effect sizes.

Available methods for calculating *a priori* power of the statistical tests used in meta-analysis are limited to models for independent effect sizes, that is, where each study contributes one independent effect size estimate to the meta-analysis (Hedges & Pigott, 2001, 2004; Jackson & Turner, 2017; Valentine et al., 2010). However, the assumption of independent effect sizes tends to hold only for narrowly-focused and smaller-scale meta-analyses (Ahn et al., 2012; Tipton et al., 2019; Tipton & Pustejovsky, 2015). As researchers adopt meta-analysis models that reflect the multivariate and multilevel nature of effect size data, information is needed about the power of these newer models, given the distinct assumptions and data structures on which they are based. In this paper, we develop new power approximations and examine the power of the test of *the mean effect size* under different strategies for modeling dependent effect sizes nested within

studies. Below we review current models for dependent effect sizes nested within studies and then discuss the aims of this research.

Models for Dependent Effect Sizes

Research syntheses in the social and behavioral sciences often include multiple effect sizes from a single primary study, leading to dependent effect sizes. Dependency can occur for a variety of reasons, for example, by studies measuring multiple relevant outcomes (e.g. math and science scores, respectively) on the same sample of individuals, or by studies reporting effect sizes across multiple independent samples (e.g. results for primary and secondary school students, respectively). In the past, researchers often handled effect size dependency through ad hoc modifications of the data. For instance, researchers might calculate a synthetic effect size for each study, averaging across different outcomes and/or time points (Tipton et al., 2019), or chose a single effect size from each study for analysis. These strategies then allowed the use of univariate meta-analysis methods.

Multivariate effect size models that reflect effect size dependencies were first introduced by Hedges & Olkin (1985) and further developed by Raudenbush, Becker, and Kalaian (1988). These methods did not see widespread use in meta-analysis because they required the correlation matrix among effect sizes, information not usually available from primary studies. A key advance in the modeling of effect size dependencies occurred when Hedges et al. (2010) introduced the use of robust variance estimation (RVE), a technique that allows for the estimation of meta-analysis models even when the exact correlation matrix among effect sizes is unknown. More recent research (Tipton, 2015; Tipton & Pustejovsky, 2015) extended this approach, providing small-sample corrections for standard errors and hypothesis tests.

A key difference between RVE and previous approaches is that inferences under earlier multivariate models were *model-based*, meaning that they required the distributional assumptions of the model to be correctly specified for hypothesis tests and confidence intervals to work properly. In contrast, RVE makes use of a *working model* for dependence among the effect sizes, which is an approximation to the dependence structure that need not be entirely correct. Initially, Hedges and colleagues (2010) introduced two working models, called the correlated effects (CE) model and the hierarchical effects (HE) model, to approximate different aspects of dependence.

They showed that even when the working model is mis-specified, it can still provide reasonably precise estimates of the mean effect size or meta-regression coefficients. Furthermore, and in contrast to model-based inference methods, RVE methods produce properly calibrated hypothesis tests and confidence intervals, even if the working model is mis-specified.

An alternative strategy, suggested by Van den Noortgate and colleagues (2013), is to use a multi-level meta-analysis (MLMA) model along with conventional, model-based inference methods. Van den Noortgate and colleagues (2013, 2014) demonstrated that model-based inferences from the MLMA work well in the presence of dependent effect sizes, even though aspects of the model may be mis-specified. They argued that the MLMA is therefore robust and, similar to the RVE approach, can be applied without knowledge of the dependence structure of the data. More recently, Moeyaert and colleagues (2017) conducted head-to-head comparisons of RVE (with a correlated effects working model) and MLMA. Their findings indicate that both methods perform similarly when the data include a large number of studies, but that RVE provided more accurate uncertainty assessments when the number of studies was limited. Further, Fernandez-Castilla et al. (2020) suggested that MLMA could be treated as a working model and combined with RVE inferential methods to provide additional robustness to model-mis-specification.

Another new strategy—coined by Pustejovsky & Tipton (2021) as the correlated-hierarchical effects (CHE) working model—recognizes both the correlated nature of effect size estimates and the multilevel structure of effect sizes nested within studies. Compared to the previously proposed CE and HE working models, the CHE working model provides researchers with the option of more closely approximating the actual structure of meta-analytic data while also guarding against mis-specification using robust variance estimation techniques.

Aims

In this article, we investigate the power of current models for handling dependent effect sizes in meta-analysis. We pursue three aims: 1) to develop approximations for the power of models that reflect the multivariate and multilevel nature of effect size data, 2) to validate these approximations using simulations, and 3) to provide guidance to researchers applying these models in terms of Type I error and power. To illustrate the approximations and to provide context for the simulation conditions, we use a recent meta-analysis conducted by Dietrichson, Bøg, Filges, and Jørgensen

(2017, henceforth DBFJ17) that investigates interventions for increasing the academic achievement (i.e. mathematics and reading) of students with low socioeconomic status (SES).

We develop new approximations for the power of several different hypothesis tests in meta-analysis of dependent effect sizes. For developing prospective power calculations, it is necessary to posit a true data-generating process. We take as a starting point the CHE model because it nests many other simpler specifications of interest. Under the CHE, we provide power approximations for 1) a model-based test based on a correctly specified working model (CHE-model); 2) a robust test based on a correctly specified model (CHE-RVE); 3) a robust test based on a simpler correlated effects (CE) working model, which is not correctly specified; 4) a model-based test based on an incorrectly specified MLMA model (MLMA-model); and 5) a robust test that uses the MLMA as a working model (MLMA-RVE). We then provide guidance for applying these approximations to a meta-analysis. Next, we test and validate the performance of the new power approximations via Monte Carlo simulation by comparing the true simulated and approximated power across various model conditions. Before describing the new power approximations, we review extant methods for statistical power in univariate meta-analysis.

Power Approximation for Univariate Meta-Analysis

Current methods for *a priori* power calculations are limited to models that include a single, independent effect size estimate from each study. Consider such a meta-analysis, based on data from J studies, where the primary aim is to test the null hypothesis that the overall average effect size μ is equal to a specific value d . Let σ_j denote the standard error of the effect size estimate from study j , for $j = 1, \dots, J$. Under a univariate random-effects model (RE), the null hypothesis $H_0: \mu = d$ would typically be tested using the Wald statistic

$$t^U = \frac{\hat{\mu} - d}{\sqrt{\hat{V}}} \quad (1)$$

where $\hat{\mu}$ is the random effects estimate of the overall average effect size and $\hat{V}$ is its estimated sampling variance (Hedges & Pigott, 2001). When the null hypothesis holds, the test statistic t^U approximately follows a central Student-t distribution with $J - 1$ degrees of freedom; when the null does not hold, its distribution is approximately a non-central Student-t distribution with non-

centrality parameter $\lambda = (\mu - d)/\sqrt{V}$ and $J - 1$ degrees of freedom, where V is the expected sampling variance (Hartung & Knapp, 2001). Power is therefore given by

$$F_t(-c_{\alpha/2, J-1} | J - 1, \lambda) + 1 - F_t(c_{\alpha/2, J-1} | J - 1, \lambda) \quad (2)$$

where $F_t(x|v, \lambda)$ is the cumulative distribution function of a non-central Student-t distribution, and $c_{\alpha, \zeta}$ is the upper α -level critical value for the central Student-t distribution with ζ degrees of freedom, so $F_t(c_{\alpha/2, \zeta} | \zeta, 0) = 1 - \alpha/2$.

The usual way of approximating *a priori* power under a univariate model is first to determine the minimum effect size of practical significance, and second to estimate the variance of the weighted overall mean effect size by estimating a) the average sampling variance of an effect size estimate in a “typical” study; b) the true between-study variance, τ^2 ; and c) the expected number of studies in the meta-analysis, J . The variance of the weighted mean effect size is approximately $V = 1/W^{RE}$, where $W^{RE} = \sum_{j=1}^J w_j^*$ and $w_j^* = (\tau^2 + \sigma_j^2)^{-1}$ are the study-specific inverse variance weights under the random effects model. If a complete balance of sample sizes is assumed, so that $\sigma_1 = \sigma_2 = \dots = \sigma_J = \sigma$, then V simplifies to $(\tau^2 + \sigma^2)/J$.

In meta-analyses of standardized mean difference effect sizes comparing two groups, the effect size estimate’s sampling variance is closely related to the overall sample size (Valentine et al., 2010). Assuming the groups are of equal size,

$$\sigma^2 \approx \left(\frac{4}{N} + \frac{\mu^2}{2(N-2)} \right), \quad (3)$$

where N is the assumed average effective sample size. Thus, if we know the average effective sample size of studies in a given area, we can approximate the average sampling variance. To arrive at a value for the between-study variance τ^2 , Pigott (2012) suggested that $\tau^2 = (1/3)\sigma^2$ could be considered a low degree of heterogeneity, $\tau^2 = \sigma^2$ could be considered a moderate degree of heterogeneity, and $\tau^2 = 3\sigma^2$ could be considered a large degree of heterogeneity.

Suppose we aim to estimate the power of the test for $H_0: \mu = 0$, with the usual level of $\alpha = .05$, in a meta-analysis of standardized mean difference effect sizes. With a low degree of heterogeneity, Pigott's guidelines would suggest a sampling variance of approximately $V = 4\sigma^2/(3J)$. Suppose that we expect to identify at least 12 studies and that the average effective sample size is $N = 100$. Therefore, $\sigma^2 \approx 4/100$ and the expected sampling variance is at most $V = 16/3600$. Using this value in Equation (2), we find power of 0.278 for an average effect of $\mu = 0.1$, power of 0.780 for $\mu = 0.2$, and power of 0.983 for $\mu = 0.3$.

Existing power approximations do not apply directly in meta-analyses involving dependent effect sizes. However, one could try applying them by calculating power assuming that there is just one effect size estimate per study—as would be the case if the meta-analyst calculated a single, synthetic effect size per study. Following this approach, we would anticipate power of 0.278 to detect an average effect of $\mu = 0.1$ in a meta-analysis of 12 studies with an average sample size of $N = 100$, regardless of whether each study included a single or multiple effect size estimates. The performance of this approximation in terms of predicting the true power of models with synthetic effect sizes is not known. Furthermore, because this approximation under-determines important quantities needed for calculating power under more complex models, we now turn to the development of new power formulas for models of dependent effect sizes.

Power Approximations for Meta-Analysis of Dependent Effect Sizes

We now describe approximations for the power of tests for an overall average effect in a meta-analysis of dependent effect sizes. We assume that the data-generating process conforms to the correlated-and-hierarchical effects (CHE) model as described by Pustejovsky and Tipton (2021). Under this data-generating process, we consider several different testing procedures, including both model-based tests and robust tests based on several distinct working models. Unlike the univariate approximations described in the previous section, we allow for sampling variances and other features to differ from study to study, so that we can examine the implications of assuming that study features are homogeneous.

Consider a collection of J studies to be included in a meta-analysis, where each study contributes k_j effect size estimates, for $j = 1, \dots, J$. Let T_{ij} denote effect size estimate i from study j , with corresponding standard error σ_{ij} , for $i = 1, \dots, k_j$ and $j = 1, \dots, J$. As usual in meta-analysis,

we shall assume that each T_{ij} is an unbiased estimator of an effect size parameter θ_{ij} and that σ_{ij} is fixed and known. These assumptions can be expressed by the model

$$T_{ij} = \theta_{ij} + e_{ij}, \quad (4)$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = \sigma_{ij}^2$. We assume that the effect size estimates from different studies are uncorrelated, so $\text{cor}(e_{hj}, e_{il}) = 0$ when $j \neq l$, but that effect size estimates from the same study may be correlated. For simplicity, we also assume that the sampling variances are constant within each study, so $\sigma_{1j}^2 = \sigma_{2j}^2 = \dots = \sigma_{kj}^2 = \sigma_j^2$, and that the correlations between sampling errors within a given study are all equal to a known constant, $\text{cor}(e_{hj}, e_{ij}) = \rho$.

Following the CHE model, we assume that the effect size parameters represent a sample from an underlying population of effects that has a hierarchical structure, according to

$$\theta_{ij} = \mu + u_j + v_{ij}, \quad (5)$$

where the study-level error term u_j has mean zero and variance τ^2 and the effect size-level error term v_{ij} has mean zero and variance ω^2 . The main parameters of the data-generating model are then the overall average effect size μ ; the between-study heterogeneity τ^2 ; the within-study heterogeneity ω^2 ; and the sampling correlation ρ . Under this model, we consider tests of the null hypothesis $H_0: \mu = d$ versus a two-sided alternative, with specified Type-I error level α .

Estimation of CHE

If one treats the model as correctly specified, then estimation of the overall average effect size μ entails first estimating the variance components and then using the estimated variance components to take an inverse-variance weighted average of the effect size estimates. Let $\hat{\tau}^2$ and $\hat{\omega}^2$ denote full or restricted maximum likelihood estimators of the variance components, which are calculated given the true sampling correlation ρ . Given values of these estimators, the overall average effect size estimate is a weighted average of the study-specific average effect size estimates, with weights given by

$$w_j = \frac{k_j}{k_j \hat{\tau}^2 + k_j \rho \sigma_j^2 + \hat{\omega}^2 + (1 - \rho) \sigma_j^2}. \quad (6)$$

The overall weighted average is then

$$\hat{\mu} = \frac{1}{W} \sum_{j=1}^J w_j \bar{T}_j, \quad (7)$$

where $\bar{T}_j = \frac{1}{k_j} \sum_{i=1}^{k_j} T_{ij}$ and $W = \sum_{j=1}^J w_j$. If the CHE model is correctly specified, then

$$\text{Var}(\hat{\mu}) \approx S = \frac{1}{W}. \quad (8)$$

The approximation here arises because, in practice, W is calculated using estimated variance components rather than known parameter values.

Model-Based Hypothesis Test

One way to test the null hypothesis $H_0: \mu = d$ is via a conventional Wald test. The model-based Wald test statistic is

$$t^M = \frac{\hat{\mu} - d}{\sqrt{1/W}}. \quad (9)$$

Consider the scenario in which the CHE model is correctly specified and the number of independent studies is large. If the null hypothesis holds, then t^M follows a standard normal distribution. If the null hypothesis does not hold, then t^M approximately follows a normal distribution with mean

$$\lambda = \sqrt{W}(\mu - d). \quad (10)$$

and unit variance. However, such large-sample approximations do not necessarily provide an adequate guide for sample sizes encountered in practice because of the uncertainty in the variance component estimates used to calculate W . It is thus desirable to develop an approximation that works even with a smaller number of studies.

In practice, researchers might use a Student-t distribution with $J - 1$ degrees of freedom as a reference distribution in the model-based tests. This is a fairly rough approximation to the sampling distribution of the model-based test. Alternatives would be to use a Satterthwaite approximation (Giesbrecht & Burns, 1985) for the degrees of freedom or Kenward and Roger (2009) approximation for the sampling variance estimator and degrees of freedom. We consider the former because it is simpler and more tractable.

We propose to approximate the power of the model-based Wald test by assuming that t^M follows a non-central Student-t distribution with non-centrality parameter λ and ζ degrees of freedom, where the degrees of freedom are determined using Satterthwaite approximation. As previously, let $F_t(x|\zeta, \lambda)$ be the cumulative distribution function of the Student-t and let $c_{\alpha, \zeta}$ be the upper α -level critical value from a central Student-t distribution. The power of the model-based Wald test against a two-sided alternative can then be approximated by

$$F_t(-c_{\alpha/2, \zeta}|\zeta, \lambda) + 1 - F_t(c_{\alpha/2, \zeta}|\zeta, \lambda). \quad (11)$$

Under the CHE model, the Satterthwaite degrees of freedom are given by

$$\zeta = \frac{st - u^2}{sy^2 + tx^2 - 2uxy}, \quad (12)$$

where

$$x = \frac{1}{W} \sum_{j=1}^J w_j^2, \quad y = \frac{1}{W} \sum_{j=1}^J \frac{w_j^2}{k_j}, \quad s = x^2 + Wx - \frac{2}{W} \sum_{j=1}^J w_j^3,$$

$$t = y^2 + \sum_{j=1}^J \frac{w_j^2}{k_j^2} + \sum_{j=1}^J \frac{k_j - 1}{(\hat{\omega}^2 + (1 - \rho)\sigma_j^2)^2} - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{k_j^2}, \text{ and } u = xy + Wy - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{k_j}.$$

If all studies include the same number of effect sizes ($k_1 = k_2 = \dots = k_J = k$) and have equal standard errors ($\sigma_1 = \sigma_2 = \dots = \sigma_J = \sigma$), we describe the meta-analytic sample as “completely balanced.” With a completely balanced sample, the weights will be equal for any values of the variance components τ^2 and ω^2 and the degrees of freedom will simplify to $\zeta = J - 1$. In a sample that is not completely balanced, ζ will be less than $J - 1$.

Note that the proposed approximation uses a Student-t critical value with the Satterthwaite degrees of freedom ζ . We acknowledge that the Satterthwaite approximation is not commonly applied in practice, nor is it available in commonly used software. In principle, one could use the power approximation with other degrees of freedom, such as by substituting the critical value $-c_{\alpha/2, J-1}$. However, such a test would have distorted Type-I error rate to the extent that the Satterthwaite degrees of freedom deviate from $J - 1$. We explore the extent of such size distortions in the simulation study.

In order to implement this power approximation prospectively, one will need to calculate weights for each of J included studies. We propose to make such calculations using assumed values for the variance component estimates $\hat{\tau}^2$ and $\hat{\omega}^2$, as well as assumptions about the sampling correlation ρ and the distribution of primary study sample sizes and effect sizes per study. We demonstrate such prospective power calculations and discuss these assumptions further at the end of this section.

Robust Hypothesis Test

Even when using Satterthwaite degrees of freedom, the model-based test will have close-to-correct Type I error only when the assumptions of the CHE working model hold. In light of the lack of information about the sampling correlations between effect size estimates, meta-analysts may prefer to use tests based on robust variance estimation methods, which maintain close-to-correct size even if the CHE model is mis-specified. With the CHE working model, a robust estimator for the variance of $\hat{\mu}$ is given by

$$V^R = \frac{1}{W^2} \sum_{i=1}^J \frac{w_j^2 (\bar{T}_j - \hat{\mu})^2}{\left(1 - \frac{w_j}{W}\right)}. \quad (13)$$

The denominator of the summand is equivalent to the CR2 small-sample correction described by Tipton (2015). When the working model is correctly specified, then V^R is an exactly unbiased estimator of $Var(\hat{\mu})$. However, even if the assumptions of the working model do not hold, V^R remains close to unbiased. A robust Wald test statistic based on V^R is

$$t^R = \frac{\hat{\mu} - d}{\sqrt{V^R}}. \quad (14)$$

Again consider the scenario in which the CHE model is correctly specified and the number of independent studies is large. If the null hypothesis holds, then t^R follows a standard normal distribution. If the null hypothesis does not hold, then t^R approximately follows a normal distribution with mean λ (as given in Equation 10) and unit variance. Thus, with a sufficiently large number of studies, the robust test has power equivalent to that of the model-based test. However, large-sample approximations do not necessarily provide an adequate guide for sample sizes encountered in practice.

Tipton (2015) proposed approximating the distribution of t^R under the null hypothesis by a Student-t distribution with ξ degrees of freedom, where ξ is derived based on a Satterthwaite approximation under the assumption that the working model is correct. Here, we propose to use the same approximation when the null does not hold, so that t^R approximately follows a non-central Student-t distribution with ξ degrees of freedom and non-centrality parameter λ . The power of the robust Wald test against a two-sided alternative can then be approximated by

$$F_t(-c_{\alpha/2, \xi} | \xi, \lambda) + 1 - F_t(c_{\alpha/2, \xi} | \xi, \lambda). \quad (15)$$

If the working model is correctly specified (and treating the variance components as known), then the degrees of freedom for the robust test are given by

$$\xi = \left[\sum_{j=1}^J \frac{w_j^2}{(W - w_j)^2} - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{(W - w_j)^2} + \frac{1}{W^2} \left(\sum_{j=1}^J \frac{w_j^2}{W - w_j} \right)^2 \right]^{-1}. \quad (16)$$

In a completely balanced sample, the degrees of freedom simplify to $\xi = J - 1$. When the sample is not completely balanced, the degrees of freedom will be less than $J - 1$ to an extent that depends on the degree of imbalance. One implication is that, for a completely balanced meta-analytic sample, the robust test has power approximately equivalent to that of the model-based test. The tests might diverge in power, however, when the primary study features are imbalanced.

RVE with CE Working Model

The original implementation of RVE introduced working models that were simplifications of the CHE model, as well as using weights that were not exactly inverse-variance under those simplified working models. The default working model, called the correlated effects (CE) model, has only a single, between-study variance component, estimated using a method-of-moments formula. Let $\tilde{\tau}^2$ denote this method-of-moments estimator. If the true data-generating process follows the CHE model, then this estimator has expectation

$$E(\tilde{\tau}^2) = \tau^2 + \omega^2 \left(\frac{1 - \sum_{j=1}^J \frac{1}{k_j \sigma_j^4}}{1 - \sum_{j=1}^J \frac{1}{\sigma_j^4}} \right). \quad (17)$$

For purposes of power calculations, we will approximate the estimator $\tilde{\tau}^2$ by its expectation. The weights used with the CE model are given by

$$\ddot{w}_j = \frac{1}{(\tilde{\tau}^2 + \sigma_j^2)}, \quad (18)$$

with overall average effect size estimator $\ddot{\mu} = \sum_{j=1}^J \ddot{w}_j \bar{T}_j / \ddot{W}$, where $\ddot{W} = \sum_{j=1}^J \ddot{w}_j$. If the CE model is applied when the true data-generating process follows the CHE model, then the variance of the overall average effect size estimator will be

$$\text{Var}(\ddot{\mu}) = \ddot{S} = \frac{1}{\ddot{W}^2} \sum_{j=1}^J \ddot{w}_j^2 \left(\tau^2 + \rho \sigma_j^2 + \frac{1}{k_j} [\omega^2 + (1 - \rho) \sigma_j^2] \right), \quad (19)$$

which will generally be larger than $1/W$.

This approximation for the power of the robust test with the CE working model entails two simplifications. First, the robust variance estimator itself is not exactly unbiased because the working model is not correctly specified (although the estimator is still asymptotically consistent as the number of studies increases). Second, the Satterthwaite degrees of freedom approximation is derived under the assumption that the working model is correctly specified (Tipton & Pustejovsky, 2015), which is not the case here. As a result, the approximation might not provide the correct Type-I error rate. Ignoring both of these complications for the time being, we propose to approximate the power of the robust test based on the CE model using the same Student-t approximation as above, but with non-centrality parameter

$$\ddot{\lambda} = \frac{\mu - d}{\sqrt{\ddot{S}}} \quad (20)$$

and degrees of freedom $\ddot{\xi}$, calculated just as in Equation (16), but with $\ddot{w}_j$ in place of w_j . In the completely balanced case, $\ddot{S} = S$, $\ddot{\lambda} = \lambda$, and $\ddot{\xi} = \xi = J - 1$, and so the test will have power equal to the other tests. If the data are not completely balanced, then the power of the CE test might diverge from that of the robust test based on the CHE working model.

Multi-level Meta-Analysis

Van den Noortgate et al. (2013, 2014) proposed handling dependent effect sizes via a multi-level meta-analysis model (MLMA), which includes both between-study and within-study random effects but ignores the possible correlation of effect size estimates drawn from the same sample. This model is a special case of the CHE, under the assumption that the correlation between sampling errors is $\rho = 0$. When the true sampling correlation is non-zero, the model is mis-

specified. However, Van den Noortgate et al. (2013, 2014) provided simulation evidence that model-based standard errors can still be accurate despite the model mis-specification.

A challenge in analyzing the power of the MLMA model is that the variance component estimates may be systematically biased when the true sampling correlation is non-zero. For purposes of power calculations, we approximate the variance component estimators using the values that minimize the Kullback-Liebler divergence between the MLMA and the true data-generating model (White, 1982). Let $\tilde{\tau}^2$ and $\tilde{\omega}^2$ denote the minimizing values of the between-study and within-study variance components, respectively. The supplementary materials provide further details about how these quantities are calculated.

The weights used with the MLMA model are then given by

$$\tilde{w}_j = \frac{k_j}{(k_j \tilde{\tau}^2 + \tilde{\omega}^2 + \sigma_j^2)}, \quad (21)$$

with overall average effect size estimator $\tilde{\mu} = \sum_{j=1}^J \tilde{w}_j \bar{T}_j / \tilde{W}$, where $\tilde{W} = \sum_{j=1}^J \tilde{w}_j$. The variance of the overall average effect size estimator is

$$\text{Var}(\tilde{\mu}) = \tilde{S} = \frac{1}{\tilde{W}^2} \sum_{j=1}^J \tilde{w}_j^2 \left(\tau^2 + \rho \sigma_j^2 + \frac{1}{k_j} [\omega^2 + (1 - \rho) \sigma_j^2] \right). \quad (22)$$

For the MLMA, the model-based variance estimator is $1/\tilde{W}$, which may be a biased estimator for $\text{Var}(\tilde{\mu})$ due to mis-specification.

In practice, the MLMA model is commonly used with model-based variance estimation and $J - 1$ degrees of freedom. However, for consistency with the other models that we have examined, we consider approximating the power of the test using Satterthwaite degrees of freedom for the model-based variance estimator. We calculate the Satterthwaite degrees of freedom using Equation (12), but substituting $\tilde{\omega}^2$ for $\hat{\omega}^2$ and $\tilde{w}_j$ for w_j . Let $\tilde{\zeta}$ denote the MLMA model-based degrees of freedom and let $\tilde{\lambda} = (\mu - d)/\tilde{S}$. We approximate the power of the model-based Wald test with Satterthwaite degrees of freedom as

$$F_t(-g \times c_{\alpha/2, \tilde{\xi}} | \tilde{\xi}, \tilde{\lambda}) + 1 - F_t(g \times c_{\alpha/2, \tilde{\xi}} | \tilde{\xi}, \tilde{\lambda}), \quad (23)$$

where $g = 1/\sqrt{\tilde{W}\tilde{S}}$. For the test with $J - 1$ degrees of freedom, we replace $c_{\alpha/2, \tilde{\xi}}$ with $c_{\alpha/2, J-1}$.

Fernandez-Castilla et al. (2020) suggested combining MLMA with robust variance estimation. We approximate the power of the robust test based on the MLMA model by following the same approach as with the CE model. We denote the Satterthwaite degrees of freedom based on the MLMA working model as $\tilde{\xi}$, calculated by using $\tilde{w}_j$ in place of w_j in Equation (16). We then approximate the power of the robust test using Equation (15), with $\tilde{\lambda}$ in place of λ and $\tilde{\xi}$ in place of ξ .

Using the Power Approximations: A Computational Example

To put each of these power approximations into practice, we need to determine the non-centrality parameters and the degrees of freedom of each of the tests. These quantities are a function of a) the number of included studies, J ; b) the parameters of the data-generating model, τ , ω , and ρ ; and c) the sample characteristics, including the primary study sample sizes and the number of effect size estimates in each primary study. We now demonstrate the mechanics of the power calculations using a hypothetical example.

Consider an on-going review in which the investigators have identified $J = 12$ studies and determined the (average) sampling variances and number of eligible outcomes available in each study. Table 1 lists these quantities. Recall that in our prior univariate power example, we assumed an average sampling variance of $\sigma^2 = 4/100$ and a low degree of heterogeneity, with $\tau = \sqrt{\sigma^2/3} = 1/\sqrt{75} = 0.115$. Let us also assume $\omega = .10$ and $\rho = .5$ and determine power under the CHE, CE, and MLMA models to detect an average effect size of $\mu = 0.1$ against the null hypothesis $H_0: \mu = 0$.

Given the assumed values of the variance components, we can calculate weights under the CHE, CE, and MLMA models, as well as under the univariate random effects model (i.e., ignoring multiplicity of effects). These weights are reported in the last four columns of Table 1. Given the CHE weights, we calculate $W = 230.65$ and $\lambda = 1.519$. For the model-based test, Equation (12) gives $x = 23.798$, $y = 9.5753$, $s = 4740.7$, $t = 23952$, $u = 1944.4$, and degrees of freedom

$\zeta = 8.37$. With these degrees of freedom, the model-based test has power of 0.271. From Equation (16), the robust test has degrees of freedom $\xi = 8.71$, leading to power of 0.273 (Equation 15).

Based on the CE weights and assumed model parameters, we calculate $\hat{\tau}^2 = 0.0176$ (Equation 17), $\hat{S} = 0.004412$ (Equation 19), $\hat{\lambda} = 1.506$ (Equation 20), and $\hat{\xi} = 8.71$. The robust test based on the CE working model therefore has power of 0.269.

Based on the MLMA weights, we calculate $\hat{\tau}^2 = 0.0305$, $\hat{\omega}^2 = 0$, $\hat{S} = 0.004488$, $\hat{\lambda} = 1.493$, $g = 1.0037$, $\hat{\zeta} = 9.54$ and $\hat{\xi} = 9.71$. Using the MLMA model, the model-based test has power of 0.267 and the robust test has power of 0.271. In this particular example, the model-based CHE test, the robust CHE test, the robust CE test, the model-based MLMA test, and the robust MLMA test all have quite similar power. Using the effective sample sizes listed in Table 1, the univariate approximation described in the previous section gives power of 0.259, slightly lower than the power of the more complex approximations.

TABLE 1. *Hypothetical Studies in a Meta-Analysis*

Study	N_j	σ_j^2	k_j	CHE weight (w_j)	CE weight ($\tilde{w}_j$)	MLMA weight ($\tilde{w}_j$)	RE weight (w_j^*)
A	28	1 / 7	1	6.02	6.23	5.77	6.40
B	32	1 / 8	3	10.00	7.01	13.87	7.23
C	40	1 / 10	2	10.71	8.50	12.43	8.82
D	48	1 / 12	3	13.85	9.91	17.17	10.34
E	56	1 / 14	4	16.54	11.23	20.70	11.80
F	64	1 / 16	2	15.34	12.48	16.21	13.19
G	80	1 / 20	2	17.91	14.79	18.03	15.79
H	96	1 / 24	1	15.38	16.87	13.87	18.18
I	128	1 / 32	2	23.94	20.46	21.70	22.43
J	180	1 / 45	3	31.76	25.10	26.41	28.12
K	192	1 / 48	5	35.93	26.00	28.88	29.27
L	256	1 / 64	2	33.28	30.08	26.13	34.53

Using the Power Approximations in Practice

Often, researchers will want to make prospective power calculations before completing the search and screening process of a systematic review. In this situation, the number of included studies and properties of those studies will not yet be known, and so the researcher will need to make

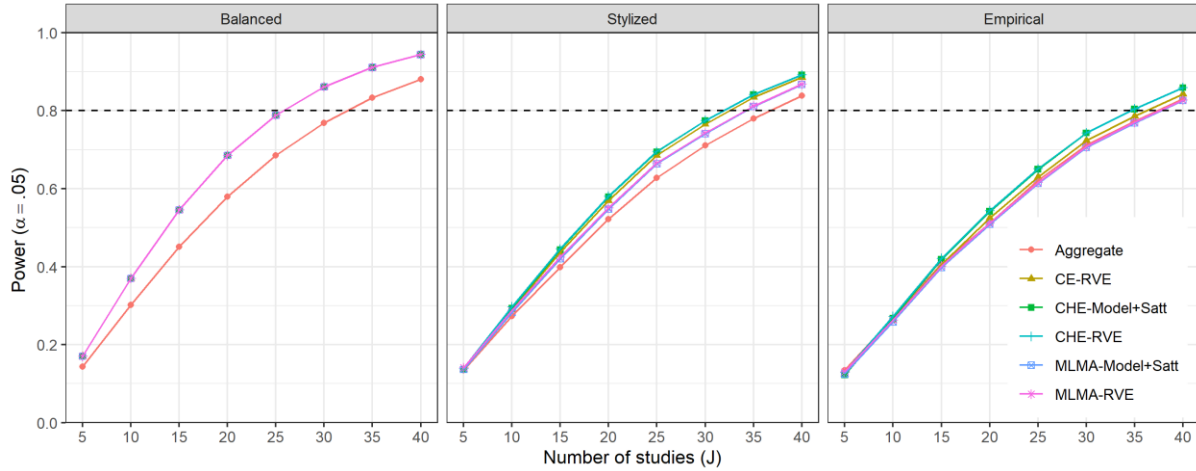
assumptions about the distribution of sampling variances and number of effect sizes per study. Assuming complete balance will generally yield optimistic power calculations (i.e., higher power than what would be expected in practice). Alternative approaches would be to simulate σ_j^2 and k_j from stylized distributions with specified parameters or to sample σ_j^2 and k_j from an empirical distribution of study characteristics—perhaps based on pilot data or previous syntheses on similar research topics. With approaches that simulate or sample study characteristics, the power approximations given in Equations (11), (15), and (23) become random quantities, with distributions governed by the distribution of σ_j^2 and k_j . For prospective power calculations, we can calculate power as the expectation over this distribution, such as by drawing many repeated samples of size J , calculating power, and then averaging over the samples.

We now demonstrate the power calculations as they might be used in practice, by developing power estimates based on the characteristics of primary studies included in the DBFJ17 meta-analysis. For purposes of illustration, we used the subsample of 77 studies (comprising 317 unique effect sizes) with effective sample sizes of no more than 500 and no more than 20 effect sizes per study. Many of the included studies were cluster-randomized trials, for which sampling variances were computed using cluster adjustment formulas from Hedges (2007). In the analytic sample of 77 studies, effective sample sizes ranged from 19 to 485, with a median of 87, a mean of 140, and a standard deviation of 125. The average sampling variance was $\sigma^2 = 0.068$. Included studies reported between 1 and 18 effect sizes, with a median of 3, a mean of 4.1, and a standard deviation of 3.5.

We calculate power to detect an average effect of $\mu = 0.1$, again assuming $\tau = 0.115$, $\omega = 0.10$, and $\rho = .5$ for sample sizes ranging from $J = 5$ to $J = 40$. Figure 1 displays the power of each model for which we have developed approximations. Each panel corresponds to a different method of determining the distribution of study characteristics. In the left panel, we assume completely balanced samples with $\sigma_j^2 = .068$ and $k_j = 4.1$, the average values of the studies in DBFJ17. Because the sample characteristics are perfectly balanced, the power of all three working models for dependent effect sizes coincide and can be calculated directly from the formulas, without re-sampling. In the middle panel of Figure 1, we determined the sample characteristics by drawing $4/\sigma_j^2$ from a gamma distribution with shape $\alpha = 1.33$ and $rate = .0095$ (which we obtained from fitting to the effective sample sizes from DBFJ17 by maximum likelihood using the

`fitdistr` function from the MASS R package, see Ripley et al., 2013) and by sampling $k_j \sim 1 + \text{Poisson}(3.1)$. In the right panel of Figure 1, we determine the sample characteristics by repeatedly sampling directly from the empirical distribution of sampling variances and number of effect sizes found in DBFJ17.

FIGURE 1. Power for finding $\mu = 0.1$ with $\tau = 0.115$, $\omega = 0.1$ and $\rho = .5$ across three different methods for obtaining n_j and σ_j^2 . For the stylized and pilot sample methods, the average power is estimated across 100 iterations. Dashed lines indicate power of 80 percent.



Across all three panels, the power of the aggregate-level approximation is notably lower because it does not account for the availability of multiple effect sizes per study. The power of the model-based and robust tests under the correctly specified CHE working model are very similar across all three panels. Because the CE working model uses weights that are not entirely efficient when the study characteristics are not balanced, the CE-RVE test has slightly lower power than the tests based on the CHE, but the difference is only noticeable when k_j and σ_j^2 are sampled from the pilot data. Similarly, the MLMA tests have lower power than the CHE tests because the MLMA tests use weights that are not entirely efficient.

Comparing across panels, the power levels of each test are substantially higher when based on balanced study characteristics than when based on the stylized distributions or empirical distributions. For instance, with $J = 25$ studies, the CHE-RVE test has power of 0.79 when assuming balanced study characteristics, but power of only 0.70 when using the stylized distribution or 0.65 when using the empirical distribution. A very similar pattern holds for the other model-based and robust tests (see Supplementary Figure S1 for further details).

Simulation Study

We used Monte Carlo simulation to validate the new power approximations and investigate the performance of different working models and inferential approaches for testing overall average effects. We designed the simulations to address three specific aims. First, we examine the accuracy of the proposed power approximations by comparing predicted power levels to simulation-based estimates of power, which fully capture the uncertainties of estimating the working models from limited data. In these analyses, we are interested both in the overall accuracy of the approximations as well as the extent to which the assumed distribution of σ_j^2 and k_j matters for obtaining accurate power estimates. Second, we evaluate the empirical Type I error rates of tests based on the different working models and inferential approaches for which we have provided power approximations. Third, we examine the relative power of tests that adequately control Type I error rates. Across all three aims, we seek to provide a basis for clearer recommendations about how to select a working model and an inferential approach in meta-analyses of dependent effect sizes.

Data Generation Process

The simulations focused on a data-generating process in which the true error structure followed the correlated-hierarchical effect (CHE) working model from Equations (4) and (5) because this model nests the simpler correlated effect model and MLMA model. The data generating procedures followed the same process as the simulations reported by Pustejovsky and Tipton (2021), except that we used the DBFJ17 data to inform the distribution of study characteristics. We imposed the same restrictions as in the example described in the previous section, after which the analytic sample was comprised of 77 studies, with an average effective sample size of 140 and an average of 4.1 effect sizes per study.

We simulated standardized mean difference (SMD) effect size estimates because this is one of the most common metrics encountered in meta-analyses in education (Ahn et al., 2012; Tipton et al., 2019). Similar to Pustejovsky and Tipton (2021), we generated effect size estimates by first simulating study-specific characteristics and effect sizes. We simulated effective sample sizes N_j and the number of effect sizes k_j by sampling from the study characteristics of DBFJ17. We then simulated true effect sizes based on Equation (5), given values of the overall average effect size μ , between-study SD τ , and within-study SD ω . We assumed that the effect size

estimates from a given study were equi-correlated with a common correlation ρ . We focus on this case in order to compare the approximations against the true simulated power when the CHE working model is correctly specified.

Given the study-specific parameters N_j , k_j , and $\boldsymbol{\delta}_j = (\delta_{1j}, \dots, \delta_{k_j j})'$, we simulated unstandardized mean difference effect size estimates for study j from a normal distribution with mean $\boldsymbol{\delta}_j$ and covariance matrix $4\boldsymbol{\Sigma}_j/N_j$, where $\boldsymbol{\Sigma}_j$ is a $k_j \times k_j$ compound symmetric matrix with unit diagonal entries and off-diagonal entries of ρ . We simulated a pooled covariance matrix for study j by drawing from a Wishart distribution with $N_j - 2$ degrees of freedom and scale matrix $\boldsymbol{\Sigma}_j$, then dividing the result by $N_j - 2$. We then calculated study-specific standardized mean differences by dividing the unstandardized mean differences by the square root of the diagonal entries in the pooled covariance matrix, then applied the Hedges' g correction. We calculated sampling variances for each effect size estimate g_{ij} as

$$V_{ij} = \left(1 - \frac{3}{4(N_j - 2) - 1}\right)^2 \left(\frac{4}{N_j} + \frac{g_{ij}^2}{2(N_j - 2)}\right)$$

This approach to simulating summary statistics is equivalent to simulating raw data from a multivariate normal distribution within each group, then calculating the effect size estimate and its variance from the raw data (Pustejovsky & Tipton, 2021).

Estimators

For each simulated dataset, we applied eight different tests that varied in terms of the working model, the variance estimator, and the method for calculating degrees of freedom (d.f.). Specifically, we calculated all five tests for which we have developed power approximations, including: the CHE working model with model-based variance and with robust variance estimation, the CE working model with robust variance estimation, the MLMA model with model-based variance and with robust variance estimation. For each of these tests, we used the corresponding Satterthwaite d.f. Because the Satterthwaite d.f. for the model-based variance estimator is novel and not typically applied in practice, we also examined tests based on the CHE working model and the MLMA working model with model-based variance and the more

conventional choice of $J - 1$ d.f. Finally, we also included a test based on the common approach of aggregating effect sizes to the study level. For the aggregated effect sizes, we used a univariate random effects model, with Knapp-Hartung adjusted standard error (Hartung & Knapp, 2001) and $J - 1$ d.f. We estimated all the above models using the `metafor` (Viechtbauer, 2010), `robumeta` (Fisher & Tipton, 2015), and `clubSandwich` (Pustejovsky, 2020) packages in R.

Experimental Design

We examined the performance of the tests using a full factorial design with 768 unique conditions. As shown in Table 2, we varied the number of independent studies from $J = 10$ to 60. These represent a small to moderate number of studies compared to sample sizes encountered in meta-analyses in education (Tipton et al., 2019). We used a maximum of 60 studies because power tended to reach ceiling levels beyond this range. We set the true average effect size to values of $\mu = 0$ (to investigate the Type I error rate) or 0.05, 0.1, or 0.2 (to examine power). The latter values represent a small, moderate, and large effect sizes for educational interventions, as suggested by Kraft (2020). We chose $\tau = 0.05, 0.2$, or 0.4 to represent a small, medium, or large amount of between-study heterogeneity, respectively. We used $\omega = 0.0, 0.05, 0.1$, or 0.2 to represent a no, small, medium, or large amounts of within-study heterogeneity. Lastly, we let values of $\rho = 0, .2, .5, .8$ represent no, small, moderate, and large levels of correlation between effect size estimates from the same study. In conditions where $\rho = 0$, the MLMA model is correctly specified (as is the CHE), whereas in conditions where $\rho > 0$, the MLMA model is increasingly mis-specified.

Table 2
Design factors for the simulation study

Factor	Parameter values
Number of studies (J)	10, 20, 40, 60
Average effect size (μ)	0.00, 0.05, 0.10, 0.20
Between-study heterogeneity (τ)	0.05, 0.20, 0.40
Within-study heterogeneity (ω)	0.00, 0.05, 0.10, 0.20
Sampling correlation (ρ)	.0, .2, .5, .8

Performance Assessment

The main performance criterion of interest was the rejection rate of each test, which we estimated by calculating the proportion of replications in which a test returned a p -value less than a specific α -level. For conditions where $\mu = 0$, the rejection rate corresponds to Type I error. For conditions with $\mu > 0$, the rejection rate is the power of the test. We calculated rejection rates of each test for $\alpha = .01, .05, .10$, although we mainly concentrate on the conventional level of $\alpha = .05$. For each simulation condition, we generated 4000 replications. For true rejection rates of .05, Monte Carlo standard errors were less than .0035; for rejection rates of .5, Monte Carlo standard errors were less than .0080.

Replication Materials

R code for replicating the simulations and numerical results from all simulation conditions are available on the Open Science Framework at <https://osf.io/auj2e/>.

Results

We describe results of the simulation study pertaining to each of the three aims.

Finding 1a: Power approximations are accurate when based on empirical study characteristics

Our first aim was to validate the proposed power approximations for meta-analysis models of dependent effect sizes. Figure 2 plots the approximated power versus the true (simulated) power for the tests based on the CHE or CE working models, where the approximation formulas and the simulation conditions are all premised upon the same parameter values. Different shapes and colors correspond to different methods of sampling k_j and N_j . Points above the 45-degree line represent conditions where the approximation over-states the true power level. Supplementary Figure S2 depicts the same comparisons for the tests based on the MLMA working model.

The approximation formulas for the CHE, CE, and MLMA working models are quite accurate when the approximations are based on sampling from the pilot data. The approximations nearly perfectly reproduce the simulated power levels for the robust tests (CHE-RVE, CE-RVE, and MLMA-RVE) when sampling k_j and N_j from the pilot data. For the CHE and MLMA model-based tests with Satterthwaite d.f., the power approximations were sometimes too optimistic

(exceeding the simulated power level), whereas using the model-based tests with $J - 1$ degrees of freedom sometimes led to overly cautious power levels. This indicates that the approximations for the RVE-based tests are more accurate than those for the model-based tests when the analyst has pilot data available.

Finding 1b: Power approximations generally over-state true power when assuming a stylized distribution or completely balanced samples.

Figure 2 and Supplementary Figure S2 also indicate that the power approximations generally over-state the true power of all models when the approximations are based either on the stylized distributions for k_j and N_j or on the assumption of complete balance. This pattern is most pronounced for the CE-RVE and the CHE and MLMA models with Satterthwaite d.f. Consequently, we cannot suggest using the approximations for these two models when sampling k_j and N_j is based on either stylized distributions or complete-balance assumptions. In contrast, the approximation formulas for the CHE-RVE and MLMA-RVE tests over-state the true power to a lesser extent, i.e. the approximations never exceed 10 percentage points more than true the power. The approximations based on stylized sample distributions for the CHE and MLMA models with $J - 1$ degrees of freedom seem also to behave adequately, although for some conditions these approximations tend to underestimate the true power. Our results suggest that power approximations premised upon the assumption of complete balanced sample characteristics generally perform poorly across all models. Therefore, when no pilot data is available to the researcher, we recommend approximating power with some kind of stylized distribution of k_j and N_j (or σ_j^2) and to anticipate that the true power may be at least 5 to 10 percentage points lower than the approximation.

Finding 1c: Simple power approximations do not accurately predict true power levels

Researchers might also wonder about how the original, simpler power approximations for univariate meta-analysis (Hedges & Pigott, 2001) perform for anticipating power in meta-analyses involving of dependent effect sizes. Figures S5-S7 in the supplementary material illustrate the performance of the univariate approximation formula to predict the true power both for the RE model estimate using synthetic effect sizes and the more complex models using RVE. From these supplementary investigations, the original power approximation performs inadequately as a means

for estimating the true power of all models handling dependency, including the RE model. Across conditions, the univariate approximations often over- or under-estimate the true simulated power by 20 percentage points or more. Thus, we do not recommend using the original univariate power approximations for estimating power of the overall average mean effect size in the presence of dependent effect sizes.

Finding 2: Robust variance estimation guards against Type I error with all working models

Figure 3 and Supplementary Figure S3 display the distribution of simulated Type I error rates for the eight different tests under consideration. Tests using $J - 1$ d.f. yielded Type I error rates that were substantially above nominal levels. This pattern is especially evident when the number of studies is small (10) to moderate (40). Even with $J = 60$ studies, the aggregated model fails to control the nominal Type I error when $\rho = 0$ or $\rho = .2$.

Tests based on model-based variance estimation and Satterthwaite d.f. (CHE-Model+Satt and MLMA-Model+Satt) appear conservative, sometimes yielding Type I error rates substantially below nominal when the number of studies is $J = 20$ or fewer. Under these scenarios, they also cover the widest range of rejection rates across the different simulation conditions (based on the width of the interquartile range of the boxplots). Concretely, this indicates that the Type I error rate of this set of models fluctuates substantially when the number of independent studies is small. Although conservative, these models should be prioritized relative to the models with Type I error rates exceeding the nominal level.

FIGURE 2. Simulated power levels versus approximated power by the C(H)E working model, for different methods of sampling k_j and N_j . Solid 45 degree lines indicate exact correspondence between approximated power and simulated power. The solid gray lines indicate 10-30 percent over- and under-estimation of the approximation. Dashed lines indicate power of 80 percent.

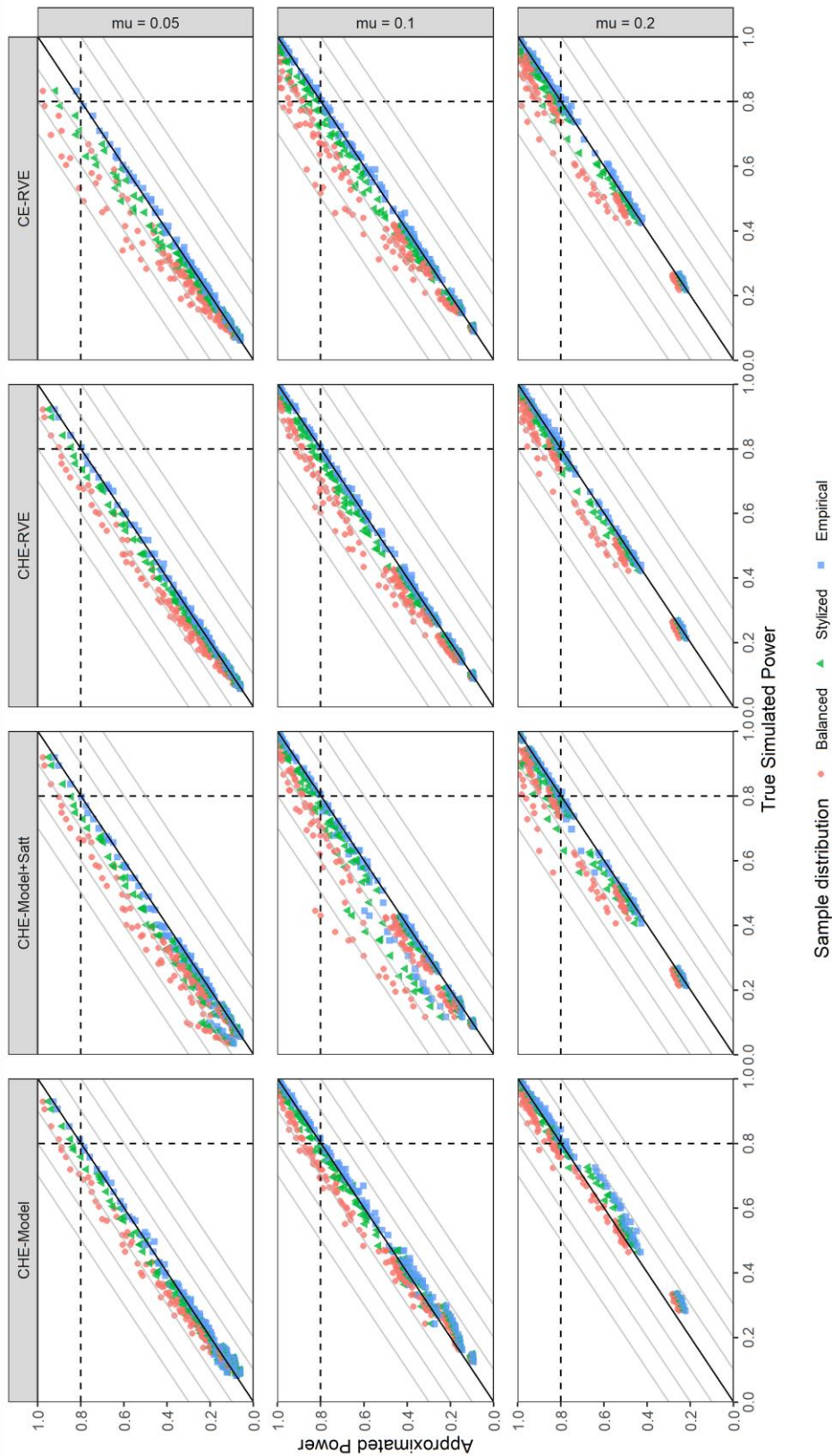
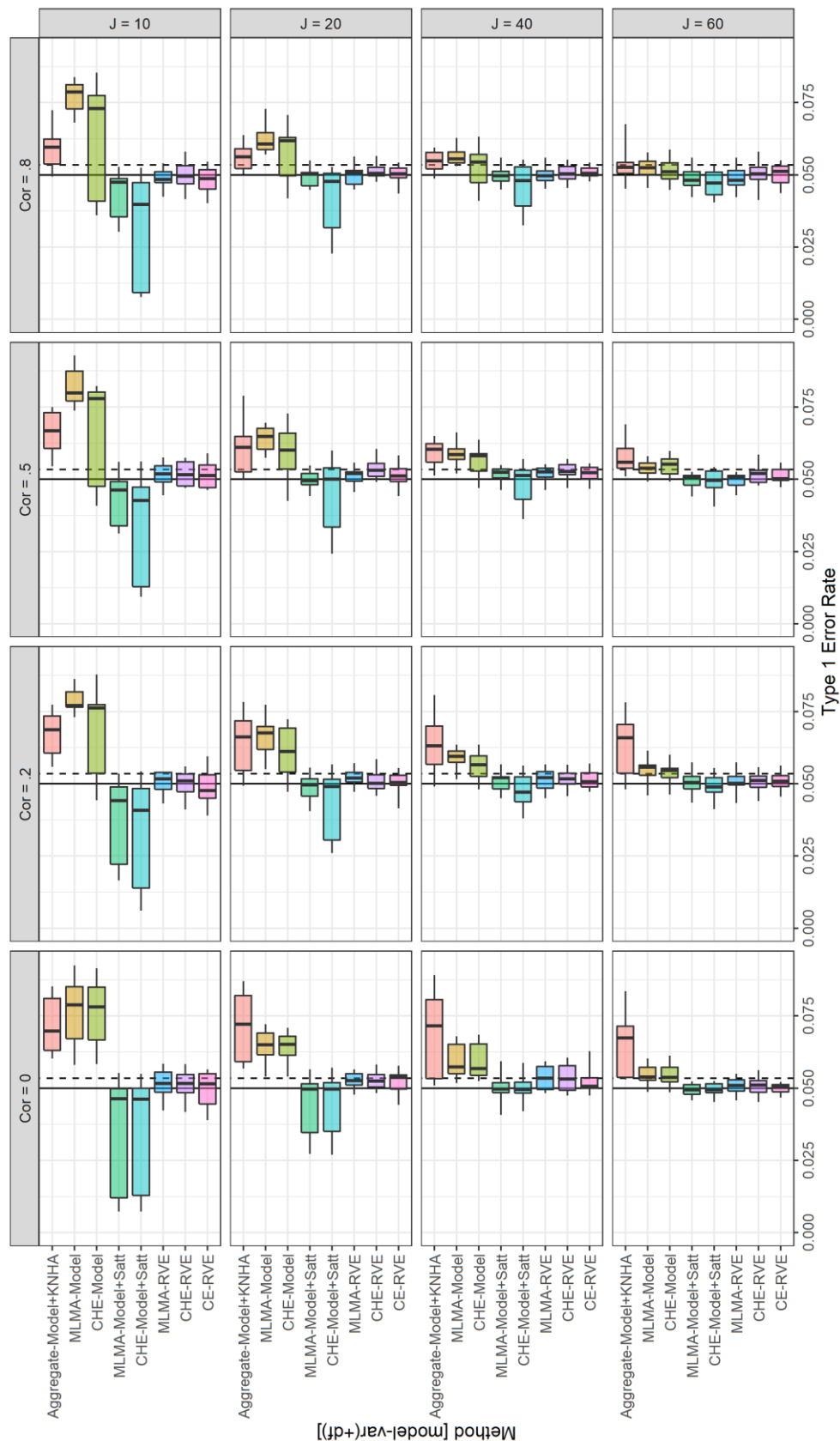


FIGURE 3. Type I error rate for $\alpha = .05$ of all estimated models by number of studies, J , and between outcomes within-study correlation, ρ . Solid lines indicate the .05 α -level and dashed lines indicate bounds for simulation error.

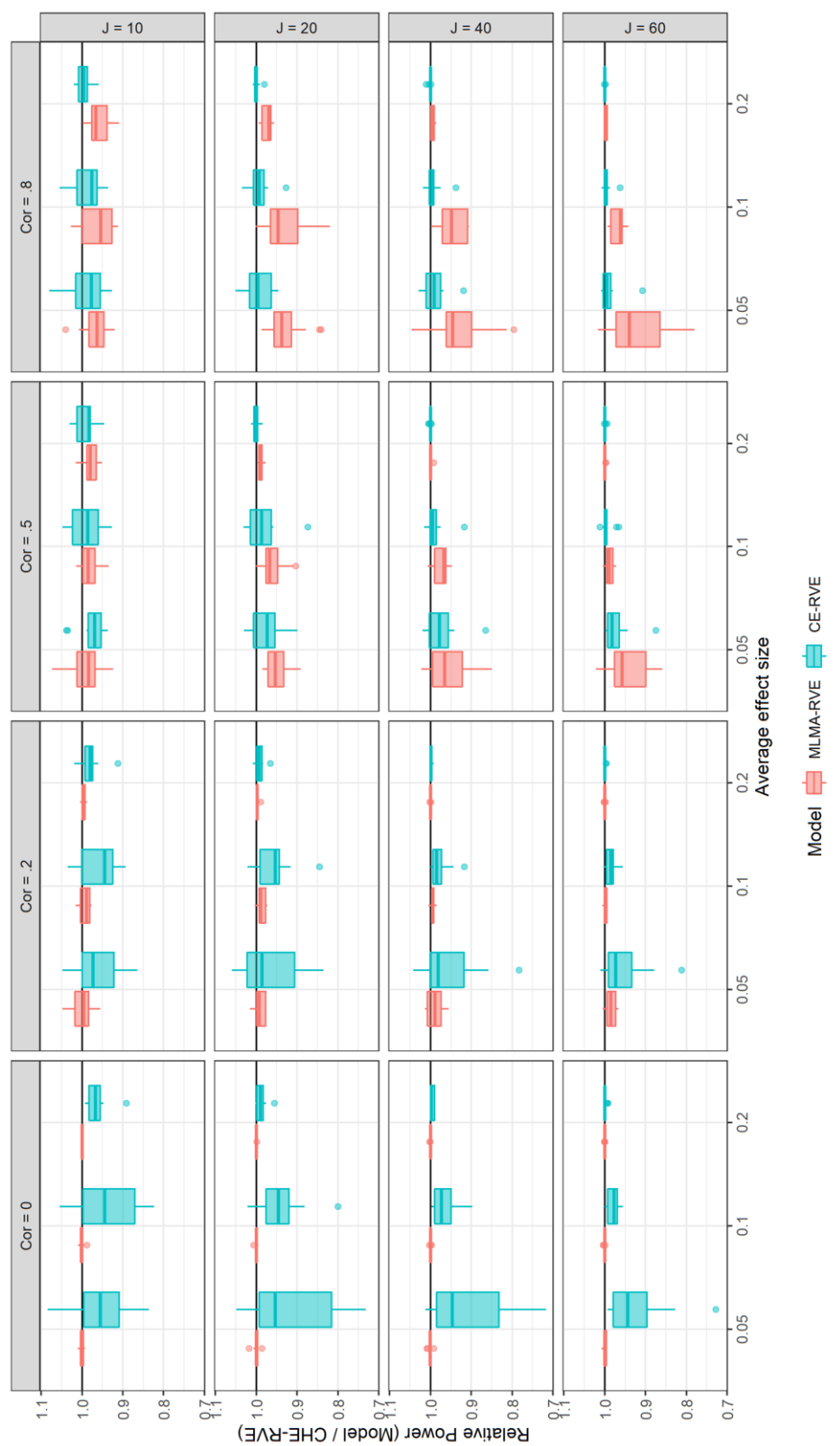


Ideally, a hypothesis testing procedure should not only control the Type I error rate so that it does not exceed the nominal level but should also come as close to the nominal level as possible. In this regard, it can be seen that all tests based on RVE with small-sample adjusted standard errors and Satterthwaite d.f. are close to or equal to the nominal rejection rate. Using small sample adjustments is particularly relevant for multilevel meta-analysis models because these methods usually use model-based tests with large-sample approximations, which can be inaccurate when the total number of studies is small. Indeed, results in Figure 3 demonstrate that the conventional MLMA test with $J - 1$ degrees of freedom requires a large number of studies ($J = 60$) to attain near-nominal Type I error—even when $\rho = 0$ so that the MLMA is correctly specified. Similar to findings from Fernández-Castilla et al. (2020), we find that combining the MLMA model with RVE to guard against misspecification adequately controls Type I error.

Finding 3: Only small power differences between RVE models

Figure 4 and Supplementary Figure S4 display the power of the CE-RVE and MLMA-RVE models, respectively, relative to the power of CHE-RVE, across varying number of studies, sizes of the within-study between outcomes correlations for small to large effect sizes and various amounts of within-study heterogeneity. Points below 1 indicate a loss of power relative to the CHE-RVE model. Under the conditions examined, one would expect that tests based on the CHE will achieve the highest possible power because they use a working model that is consistent with the true data-generating model. In contrast, the CE-RVE tests are based on a mis-specified working model and also use weights that are not fully efficient. In light of this, it is interesting that the CE-RVE tests do not lose substantial power relative to CHE-RVE. Under most conditions, the relative power of CE-RVE tests was 80% or higher, and often closer to 95%. Similarly, the MLMA model is only correctly specified when $\rho = 0$. When correctly specified, it is equivalent to the CHE working model and thus the MLMA-RVE test has power identical to that of CHE-RVE. For $\rho > 0$, the MLMA working model is mis-specified. Interestingly, though, the MLMA-RVE test still retains most of the power of the CHE-RVE test, with relative power of 90% or more. These results suggest that all models for handling dependent effects are reliable with regard to estimating the overall average effect size (even when the total number of studies is small) as long as these are guarded for any misspecifications via RVE.

FIGURE 4. Relative power between the simulated power for the CHE-RVE model and the CE-RVE and MLMA-RVE models across the different values of between outcomes within-study correlation ρ , total number of studies, J , and average effect sizes, μ , respectively. Values less than 1 indicate loss of power relative to CHE-RVE.



Discussion and Conclusion

Methods for handling dependent effect sizes have grown increasingly complex, which has created challenges for how to conduct prospective power analysis for meta-analysis. In this study, we developed new approximation formulas for several Wald-type tests based on the CHE, CE, and MLMA models, and we evaluated the performance of the approximations via Monte Carlo simulations assuming a correlated-and-hierarchical effects data-generating process. The new approximation formulas can closely match the true model power when the relevant primary study characteristics, including sample variance, σ_j^2 , or average sample size per study N_j , and the number of effect sizes per study, k_j , are sampled from pilot data with similar characteristics to the data used for the eventual meta-analysis.

We acknowledge that it will not always be possible for systematic reviewers to have access to reliable or relevant pilot data that can inform their power analysis. Therefore, we also tested the performance of power approximations when these are either based on completely balanced study characteristics (i.e. all studies have equal sampling variance and the same number of effect sizes) or on a stylized distribution of k_j and σ_j^2 . We found that most of the power approximations overestimate the true power to some extent. Approximations based on the assumption of complete balance perform worse (yielding overly optimistic power estimates) than approximations that consider imbalance across studies. We thus do not recommend researchers assume complete balance in practice. When no pilot data are available, we recommend reviewers use the approximations for working models using RVE based on stylized distributions of k_j and σ_j^2 (or N_j) because these approximations rarely overestimated the true power by more than 10 percentage points. We tentatively suggest that reviewers should anticipate a systematic power loss of 5-10 percent when conducting power analysis when using stylized distributions.

From our simulation study, we also investigated Type I error rates and, for the models that adequately controlled the nominal Type I error rate, relative power. The simulation results provide further evidence that meta-analysts should routinely guard against model misspecification by using robust variance estimation. For tests of overall average effect sizes, using robust variance estimation has little cost in terms of power. If using model-based inference, meta-analysts should use the more conservative test based on Satterthwaite degrees of freedom, particularly when the

total number of studies is small or moderate (i.e. 10-40). Our results support the previous recommendations from Tipton (2015) to routinely use both small-sample adjustments and Satterthwaite d.f. Compared to model-based variance approaches with $J - 1$ degrees of freedom, tests based on robust variance estimation more adequately control the nominal rejection rate and yield more adequate power estimates. In addition, the power differences between the CE, CHE, and MLMA models are minor when applying RVE. That said, and in line with Pustejovsky and Tipton (2021), we recommend using working models, such as the CHE, that capture the main features of the data structures that meta-analysts are likely to encounter in practice.

Predicated upon our results, we generally recommend using the new power approximation for the models using RVE, because this approximation seems to perform most reliably across all techniques for obtaining k_j and σ_j^2 (or N_j). We also recommend conducting power analysis for one (ideally pre-specified) model, only, to reduce “researcher’s degrees of freedom” in the eventual meta-analysis. We further propose to calculate power by drawing many repeated samples of k_j and σ_j^2 of size J , and then averaging the power over the samples, when approximations are not based on the complete balance assumption. Lastly, we find that the original univariate power approximation (Hedges & Pigott, 2001) performs insufficiently for purposes of estimating power of both the univariate model using synthetic effect sizes or the more complex family of models using RVE. We, therefore, recommend no longer using the univariate and more simple formulas to approximate power for models handling dependent effect sizes. Future research is needed to investigate how these univariate power formulas perform when the true data-generation process follows an independent effects structure.

The work in this article does have some clear limitations. Although we find that the approximations perform well when based on pilot data, it may be that available pilot data are not representative of the target population of studies (for instance, by imposing too much or too little imbalance in the data), which could distort the accuracy of the proposed approximations. Furthermore, our simulation results are limited by the selected data-generating model and parameters. The most clear limitation of this study is that we have only concentrated on the situation in which the CHE working model is consistent with the true data-generation process, a best-case scenario that implies that the CHE working model will have higher power than the CE or MLMA models. In future work, it might be useful to elaborate upon the power approximations

by allowing for a specific degree of model mis-specification of the working model, such as by assuming a correlation of $\rho = .6$ but allowing the true data-generating process to have a correlation of $.4 < \rho < .8$.

This study is limited in scope in that the simulations focused on the common case of standardized mean differences effect sizes. The power formulas can readily be applied to some other effect size metrics such as Fisher's z-transformed correlation coefficient, but application to metrics such as log odds ratios or risk ratios requires making further assumptions. Future research needs to develop guidance about how to implement the power calculations under a range of scenarios encountered by working meta-analysts.

In this article, we have only focused on power of tests for the overall average effect size, which clearly limits the application of the proposed methods. For testing the overall average effect size, we found that the choice of working model (CHE or CE or MLMA) leads to only minor differences in power. However, this finding may not generalize to more complex models involving moderator variables. Rather, Pustejovsky and Tipton (2021) found that using CHE can lead to substantially more precise estimates than using CE for meta-regression models with predictor variables that vary within study. Thus, the choice of working model may be more consequential for models that involve potential moderator variables.

Developing power calculations for more intricate models, such as meta-regressions with one or multiple predictors, requires making strong assumptions about the distribution of covariates across studies and effect sizes, which may be difficult to specify *a priori*. However, if reviewers have access to detailed and relevant pilot data, power analysis for meta-regression can be conducted via Monte Carlo simulation. Although not trivial, future research could focus on making power simulation for meta-regression models more accessible to the applied meta-analyst.

References

- Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82(4), 436–476. <https://doi.org/10.3102/0034654312458162>
- 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>
- 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(2), 702–717. <https://doi.org/10.3758/s13428-020-01459-4>
- Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. <https://cran.r-project.org/web/packages/robumeta/vignettes/robumetaVignette.pdf>
- Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics*, 41(2), 477–486. <https://doi.org/10.2307/2530872>
- Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12), 1771–1782. <https://doi.org/10.1002/sim.791>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(3), 203–217. <https://doi.org/10.1037/1082-989X.6.3.203>
- Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426–445. <https://doi.org/10.1037/1082-989X.9.4.426>
- 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>
- Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods*, 8(3), 290–302. <https://doi.org/10.1002/jrsm.1240>
- Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed

- effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, 53(7), 2583–2595. <https://doi.org/10.1016/j.csda.2008.12.013>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6), 559–572. <https://doi.org/10.1080/13645579.2016.1252189>
- Pigott, T. D. (2012). *Advances in meta-analysis*. Springer.
- Pustejovsky, J. E. (2020). *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., & 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>
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 103(1), 111–120. <https://doi.org/10.1037/0033-2909.103.1.111>
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). *MASS: Support functions and datasets for venables and Ripley's MASS*. <https://cran.r-project.org/web/packages/MASS/index.html>
- 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>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194. <https://doi.org/10.1002/jrsm.1339>

- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <https://doi.org/10.3102/1076998609346961>
- 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>
- 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>
- 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>
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25. <https://doi.org/10.2307/1912526>

Appendix 4: Supplementary Material (Chapter III)

Kullback-Liebler Divergence for the Multi-Level Meta-Analysis Model

In developing power approximations for the correlated-and-hierarchical effects (CHE) model, we make the simplifying assumption that the variance component estimates are equal to the corresponding parameter values. This is reasonable because we assume that the model is correctly specified, and so restricted maximum likelihood estimates will be close to unbiased. However, this simplification does not work for purposes of developing power approximations for the multi-level meta-analysis (MLMA) model because the model is mis-specified. The challenge is thus to determine the behavior of the variance component estimates when the true data-generating process follows the correlated-and-hierarchical effects model.

We propose to approximate the variance component estimates from the MLMA model by using their asymptotic limits. White (1982) demonstrated that, under suitable regularity conditions, the maximum likelihood estimator of a mis-specified model converges to the value that minimizes the Kullback-Liebler divergence (KLD) between the mis-specified model and the true data-generating process. The KLD is the expectation (under the true data-generating process) of the difference between the log likelihood of the true data-generating process and the log likelihood of the mis-specified model.

Consider the CHE data-generating process (Equation 7 in the main text) for a collection of J studies, each of which includes k_j effect size estimates. Let $\mathbf{1}_j$ denote a $k_j \times 1$ vector of 1's, let $\mathbf{I}_j$ be a $k_j \times k_j$ identity matrix, and let $\mathbf{T}_j$ be the $k_j \times 1$ vector of effect size estimates from study j , for $j = 1, \dots, J$. The CHE data-generating process can be written succinctly as

$$\mathbf{T}_j \sim N(\mu \mathbf{1}_j, \Phi_j),$$

where

$$\Phi_j = (\tau^2 + \rho \sigma_j^2) \mathbf{1}_j \mathbf{1}_j' + (\omega^2 + (1 - \rho) \sigma_j^2) \mathbf{I}_j.$$

Twice the restricted log likelihood of the CHE model is therefore

$$2 \times l_R(\tau^2, \omega^2, \rho) = c - \sum_{j=1}^J \log|\Phi_j| - \log \left(\sum_{j=1}^J \mathbf{1}_j' \Phi_j^{-1} \mathbf{1}_j \right) - \mathbf{T}' \mathbf{Q} \mathbf{T},$$

where $\mathbf{Q} = \Phi^{-1} - \Phi^{-1} \mathbf{1}(\mathbf{1}' \Phi^{-1} \mathbf{1})^{-1} \mathbf{1}' \Phi^{-1}$, and where Φ_j is a function of the variance component parameters τ^2 , ω^2 , and ρ . Note that the MLMA is a special case of the CHE, and so its restricted log likelihood is the same as above, but fixing $\rho = 0$. Let $\tilde{\tau}^2$ and $\tilde{\omega}^2$ denote the variance component parameters under the MLMA. Let Ω_j denote the variance-covariance of $\mathbf{T}_j$ under the MLMA, given by

$$\Omega_j = \tilde{\tau}^2 \mathbf{1}_j \mathbf{1}_j' + (\tilde{\omega}^2 + \sigma_j^2) \mathbf{I}_j.$$

Let $\tilde{\mathbf{Q}} = \Omega^{-1} - \Omega^{-1} \mathbf{1}(\mathbf{1}' \Omega^{-1} \mathbf{1})^{-1} \mathbf{1}' \Omega^{-1}$. The KLD between the MLMA and the CHE can then be written as

$$\begin{aligned} KLD(\tilde{\tau}^2, \tilde{\omega}^2, \tau^2, \omega^2, \rho) &= \mathbb{E} \left[\log \left(\frac{l_R(\tau^2, \omega^2, \rho)}{l_R(\tilde{\tau}^2, \tilde{\omega}^2, 0)} \right) \right] \\ &= c + \sum_{j=1}^J \log|\Omega_j| + \log \left(\sum_{j=1}^J \mathbf{1}_j' \Omega_j^{-1} \mathbf{1}_j \right) + \mathbb{E}[\mathbf{T}' \tilde{\mathbf{Q}} \mathbf{T}] \\ &= c + \sum_{j=1}^J \log|\Omega_j| + \log \left(\sum_{j=1}^J \mathbf{1}_j' \Omega_j^{-1} \mathbf{1}_j \right) + \text{tr}(\tilde{\mathbf{Q}} \Phi), \end{aligned}$$

where c is a constant that does not depend on $\tilde{\tau}^2$ or $\tilde{\omega}^2$. Denote the inverse-variance weights under the MLMA working model as

$$\tilde{w}_j = \frac{k_j}{k_j \tilde{\tau}^2 + \tilde{\omega}^2 + \sigma_j^2}, \quad \text{with} \quad \tilde{W} = \sum_{j=1}^J \tilde{w}_j.$$

and the inverse-variance weights under the CHE as

$$w_j = \frac{k_j}{k_j \tau^2 + k_j \rho \sigma_j^2 + \omega^2 + (1 - \rho) \sigma_j^2}.$$

Then we can write

$$\begin{aligned} \sum_{j=1}^J \log |\mathbf{\Omega}_j| &= \sum_{j=1}^J \left[(k_j - 1) \log(\tilde{\omega}^2 + \sigma_j^2) - \log\left(\frac{\tilde{w}_j}{k_j}\right) \right], \\ \log\left(\sum_{j=1}^J \mathbf{1}_j' \mathbf{\Omega}_j^{-1} \mathbf{1}_j\right) &= \log \tilde{W}, \end{aligned}$$

And

$$\text{tr}(\tilde{\mathbf{Q}}\mathbf{\Phi}) = \sum_{j=1}^J \left[(k_j - 1) \left(\frac{\omega^2 + (1 - \rho) \sigma_j^2}{\tilde{\omega}^2 + \sigma_j^2} \right) + \frac{\tilde{w}_j}{w_j} \left(1 - \frac{\tilde{w}_j}{\tilde{W}} \right) \right].$$

Therefore,

$$\begin{aligned} KLD(\tilde{\tau}^2, \tilde{\omega}^2, \tau^2, \omega^2, \rho) \\ &= c + \sum_{j=1}^J (k_j - 1) \log(\tilde{\omega}^2 + \sigma_j^2) - \sum_{j=1}^J \log\left(\frac{\tilde{w}_j}{k_j}\right) + \log \tilde{W} \\ &\quad + \sum_{j=1}^J (k_j - 1) \left(\frac{\omega^2 + (1 - \rho) \sigma_j^2}{\tilde{\omega}^2 + \sigma_j^2} \right) + \sum_{j=1}^J \frac{\tilde{w}_j}{w_j} \left(1 - \frac{\tilde{w}_j}{\tilde{W}} \right). \end{aligned}$$

The asymptotic limits of the variance component estimators under the MLMA are the values of $\tilde{\tau}^2$ and $\tilde{\omega}^2$ that minimize $KLD(\tilde{\tau}^2, \tilde{\omega}^2, \tau^2, \omega^2, \rho)$ for fixed τ^2 , ω^2 , and ρ . Although KLD is a complicated, non-linear objective function, it can be minimized numerically using standard algorithms. For purposes of power calculations, we find the minima using the R function `optim()`, with a limited-memory, Box-constrained quasi-Newton method, setting `method = "L-BFGS-B"`.

The asymptotic limits have simpler solutions when the study characteristics are balanced, such that $k_1 = k_2 = \dots = k_J = k$ and $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_J^2 = \sigma^2$. If the study characteristics are balanced and if $\omega^2 > \rho\sigma^2$, then the limits are given by

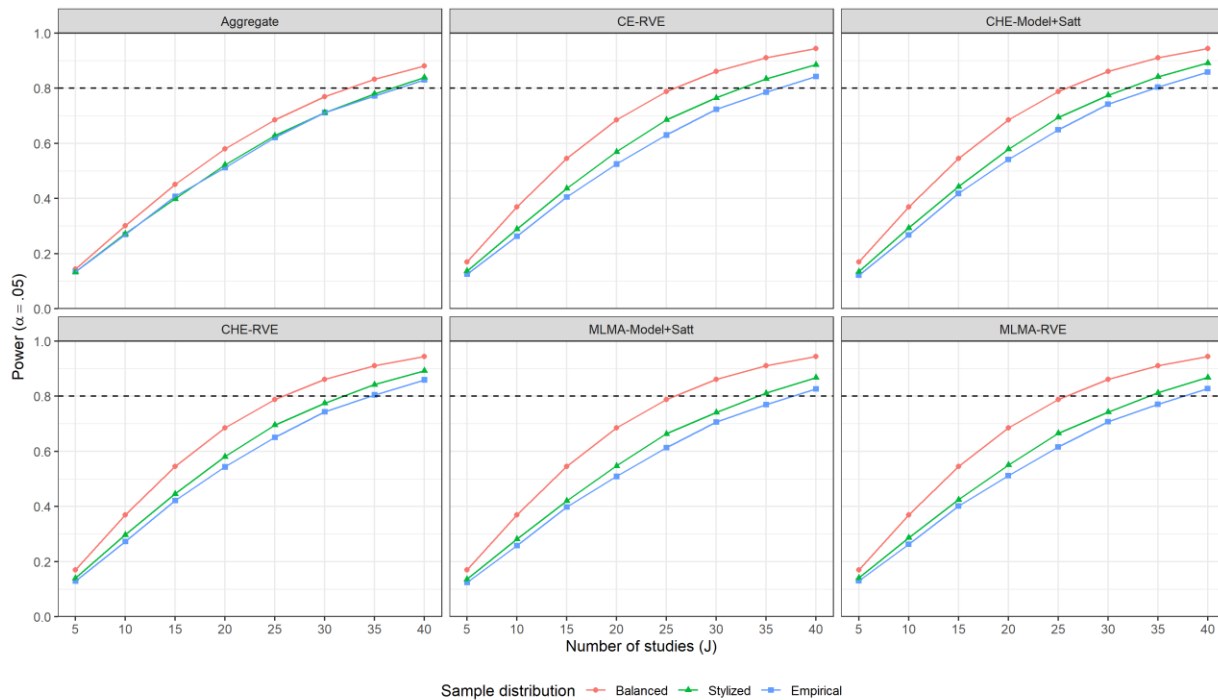
$$\tilde{\tau}^2 = \tau^2 + \rho\sigma^2 \quad \text{and} \quad \tilde{\omega}^2 = \omega^2 - \rho\sigma^2.$$

If the study characteristics are only moderately imbalanced, then we expect that the exact asymptotic limits will still be quite close to these values. More generally, we expect that the estimator for τ^2 will be positively biased and the estimator for ω^2 will be negatively biased under the misspecified MLMA model, and the size of the bias will depend on the magnitude of the sampling variances ($\sigma_1^2, \dots, \sigma_J^2$) and true sampling correlation ρ .

References

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25. <https://doi.org/10.2307/1912526>

FIGURE S1. Power approximation for $\mu = 0.1$ with $\tau = 0.115$, $\omega = 0.1$ and $\rho = .5$ across various models for handling dependent effect sizes



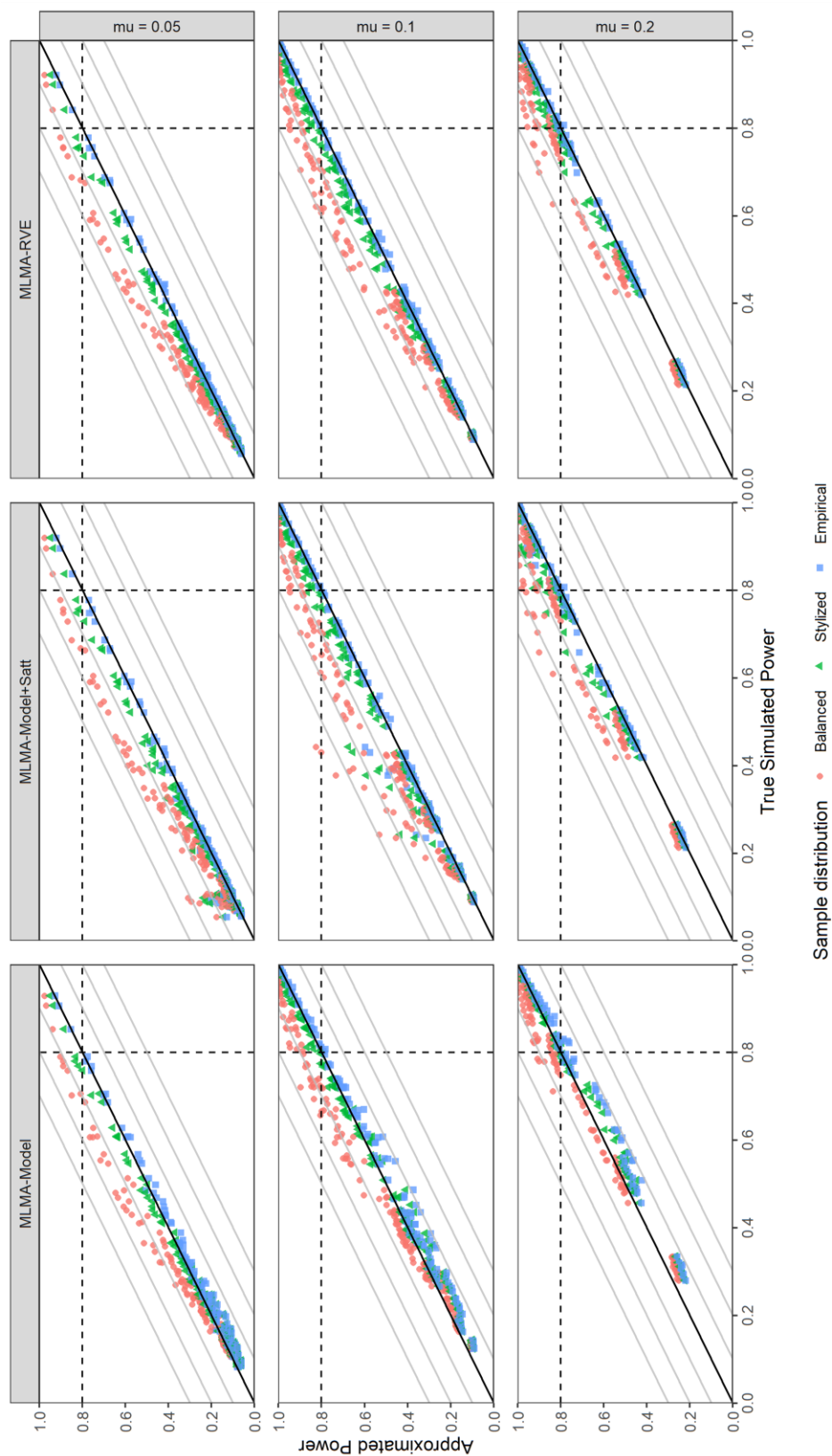


FIGURE S2. Simulated power levels versus approximated power by the MLMA working model, for different methods of sampling k_j and N_j . Solid 45 degree lines indicate exact correspondence between approximated power and simulated power. The solid gray lines indicate 10-30 percent over- and underestimation, respectively. Dashed lines indicate power of 80 percent.

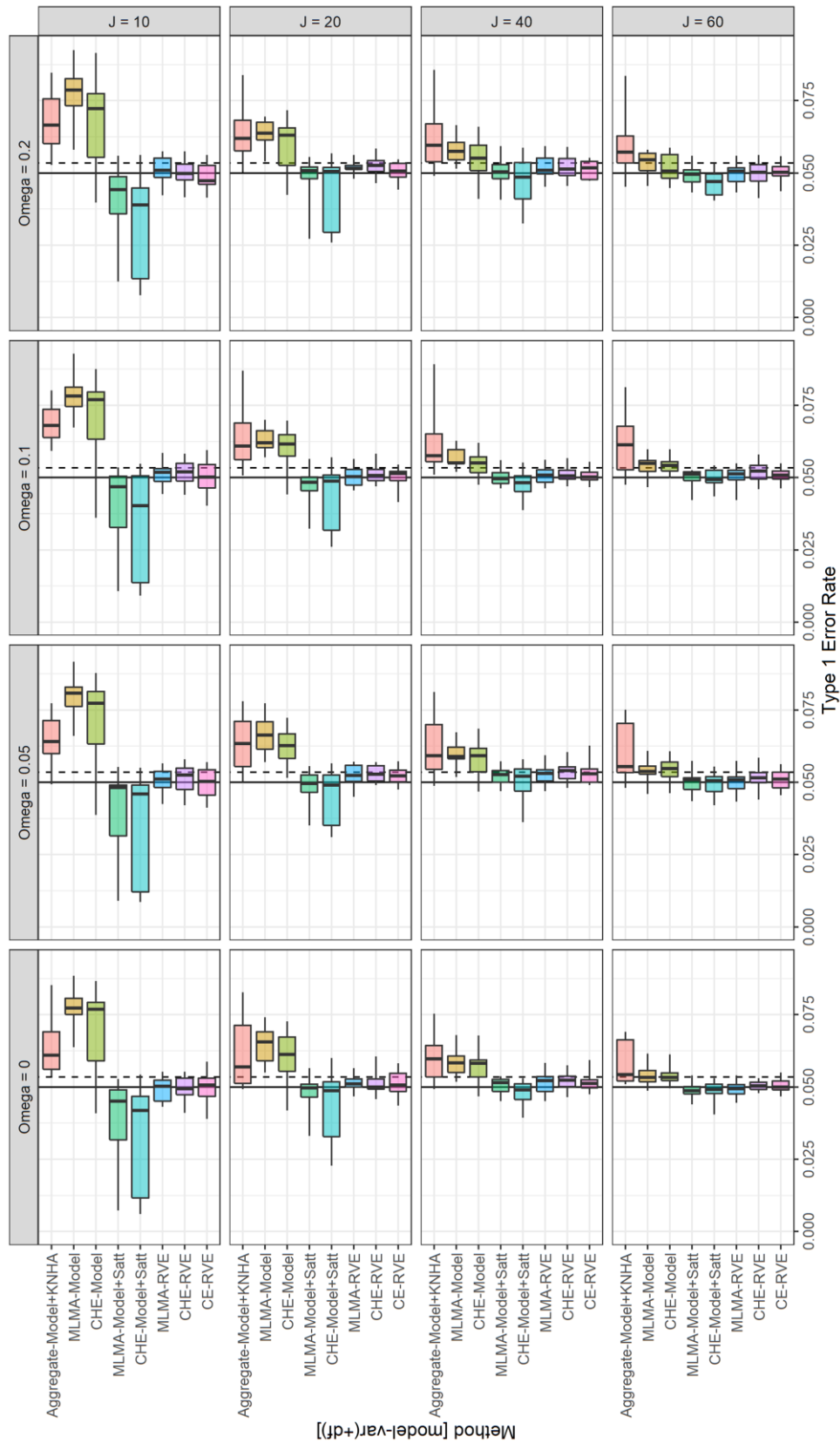


FIGURE S3. Type I error rate for $\alpha = .05$ of all estimated models by number of studies, J , and varying within-study heterogeneity values, ω . Solid lines indicate the .05 α -level and dashed lines indicate bounds for simulation error.

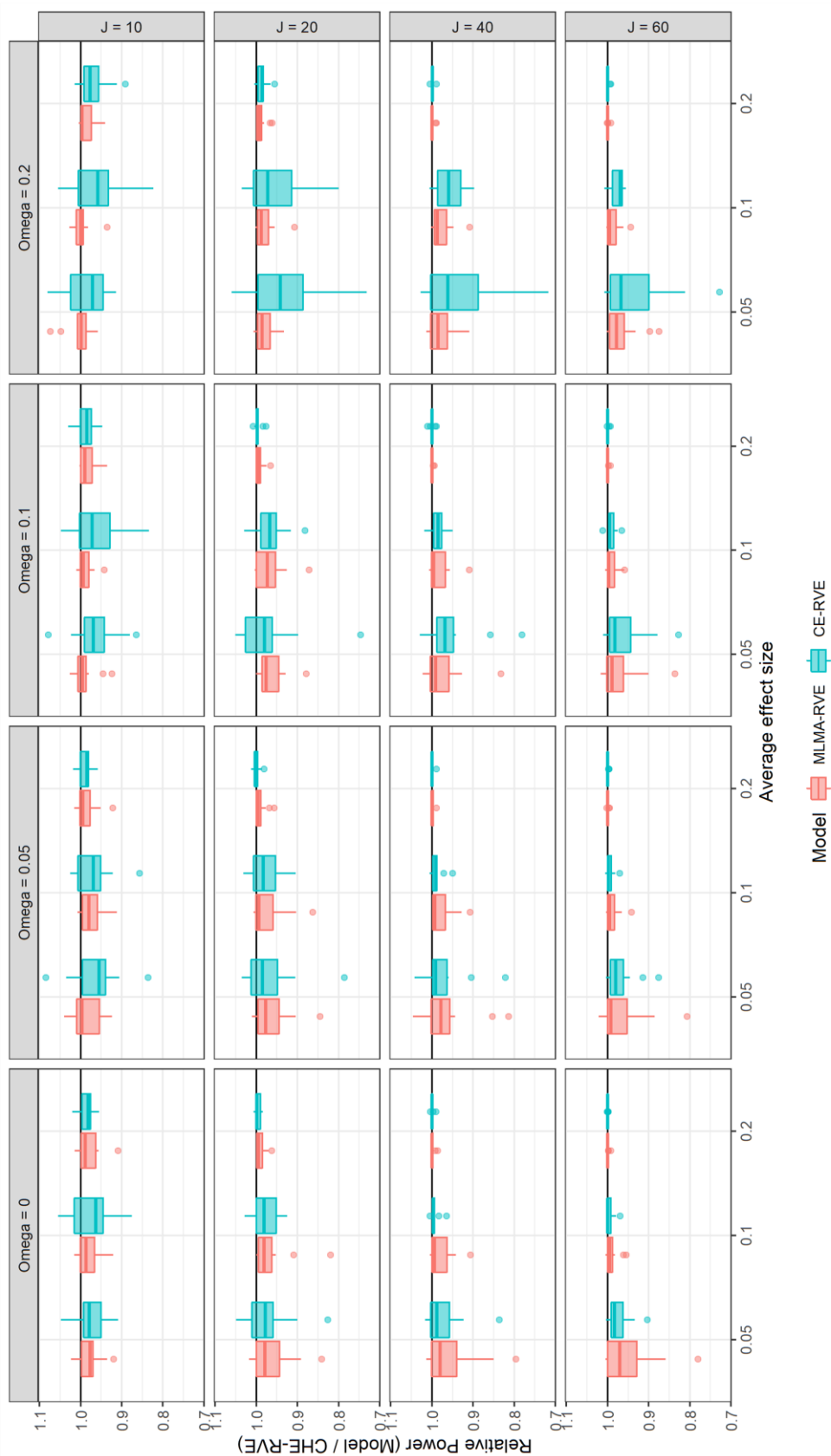


FIGURE S4. Relative power between the simulated power for the CHE-RVE model and the CE-RVE and MLMA-RVE models across different values of within-study heterogeneity ω , total number of studies, J , and average effect sizes, μ , respectively. Values less than 1 indicate loss of power relative to CHE-RVE.

Performance of the univariate approximation formula

FIGURE S5. Univariate power approximation performance for approximating the true simulated power of the RE and the RVE models, respectively, across various approaches for obtaining the sample distributions

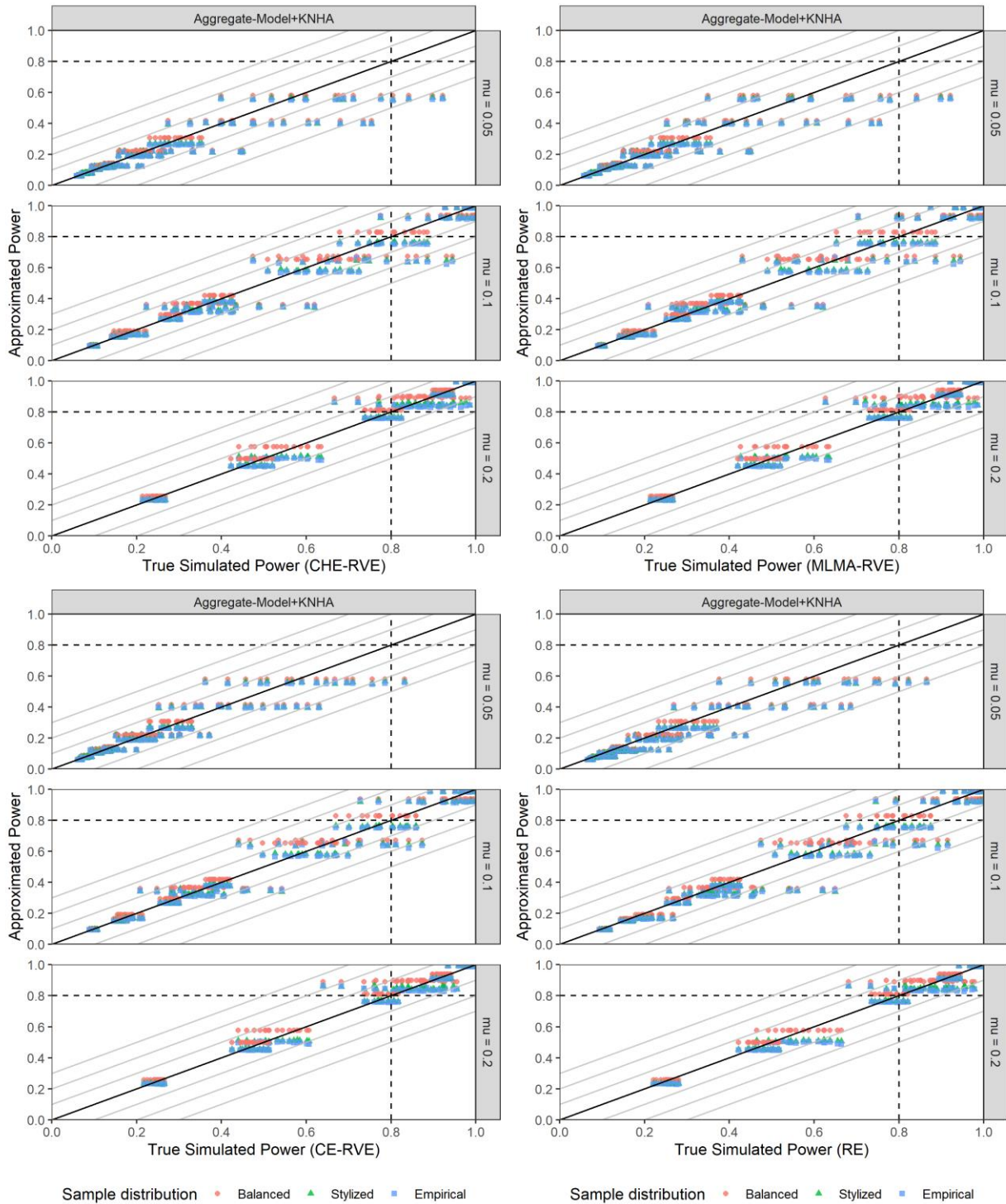


FIGURE S6. Performance of the univariate power approximation to estimate the true simulated power for RE and RVE across varying numbers of studies using the empirical approach to obtain sample distributions

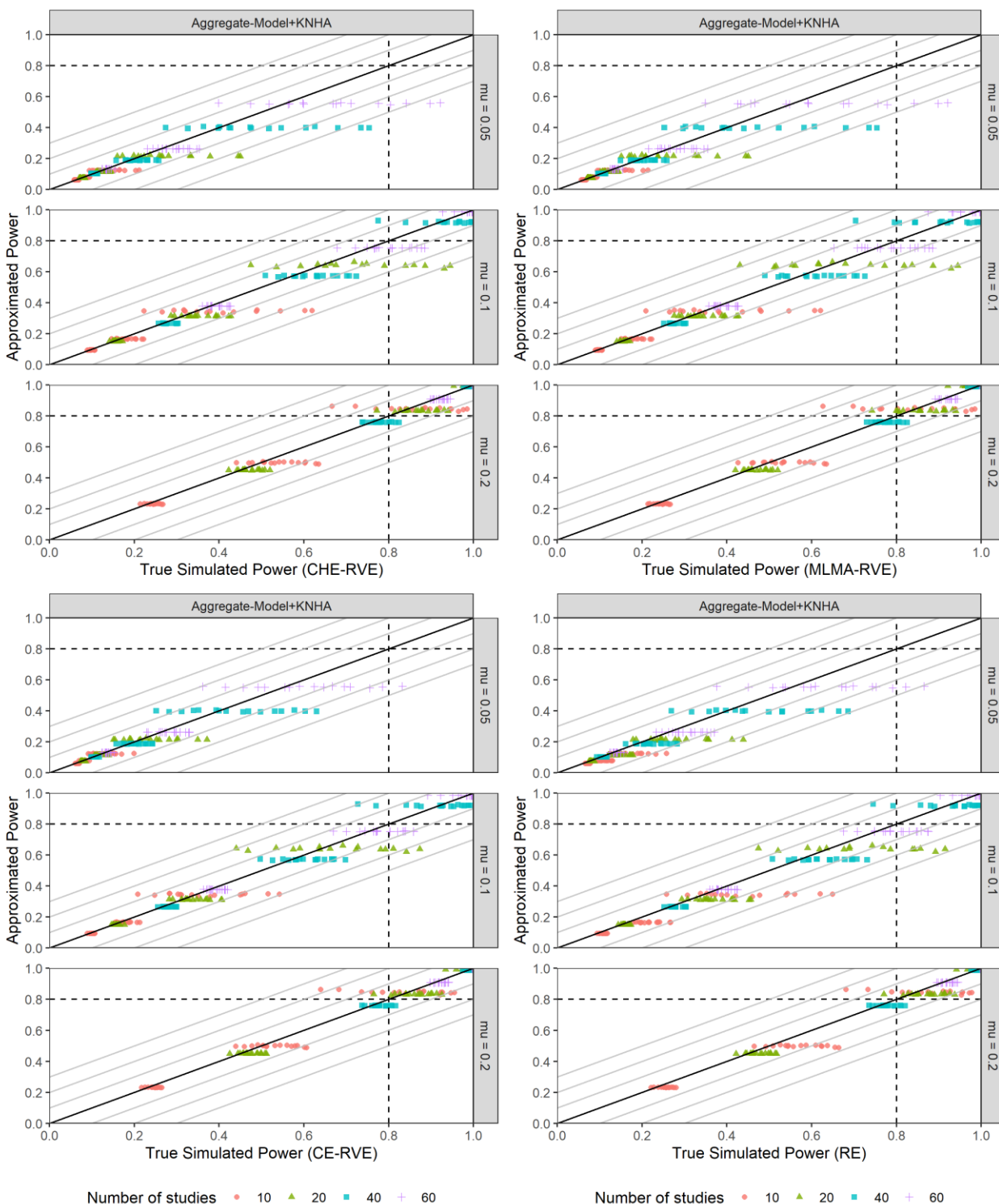
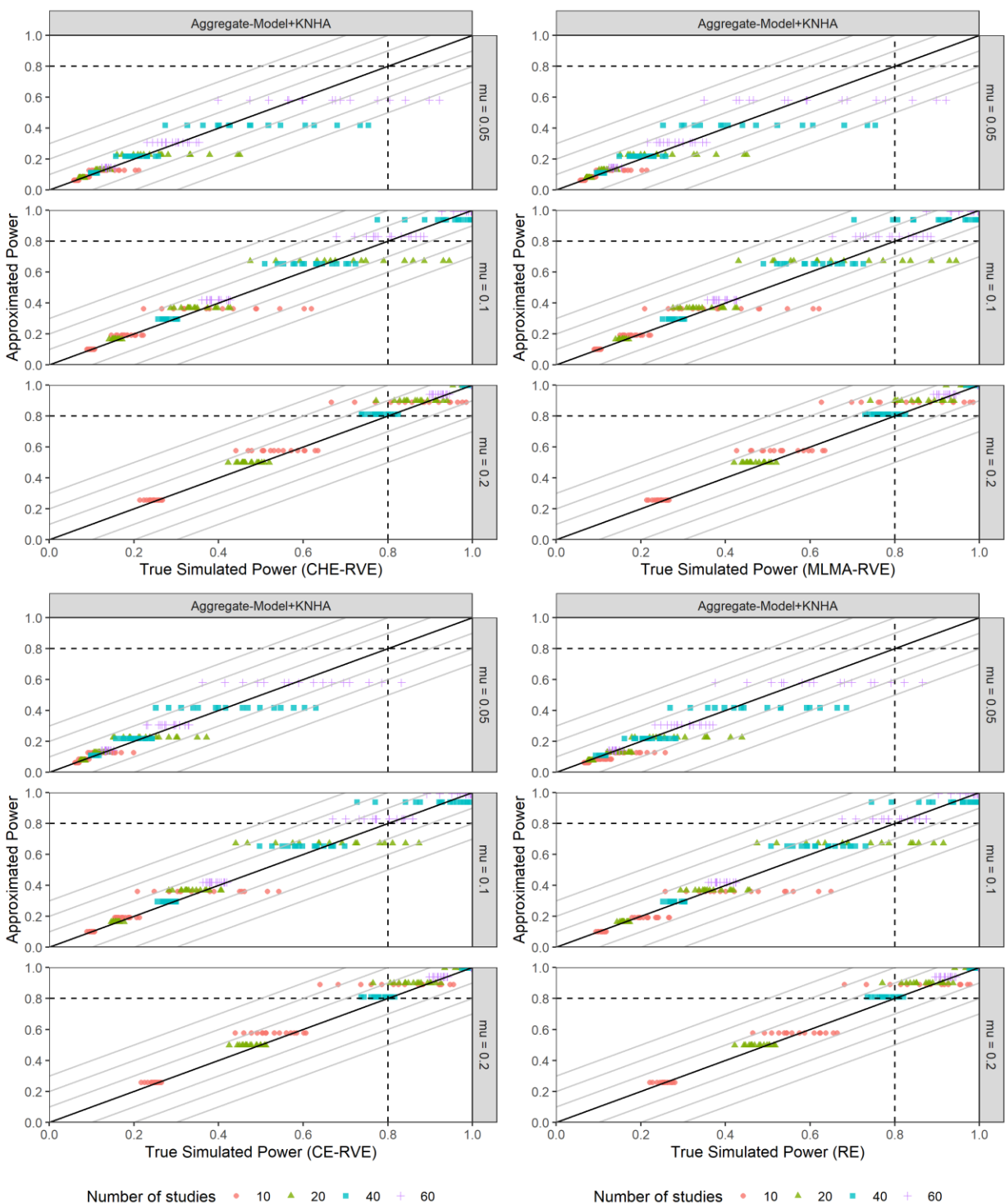


FIGURE S7. Performance of the univariate power approximation to estimate the true simulated power for RE and RVE models across varying numbers of studies assuming complete balance



Chapter IV

Conducting Power Analyses for Meta-Analysis of Dependent Effect Sizes: Common Guidelines and an Introduction to the POMADE R Package

Mikkel H. Vembye^{†,‡}

[†] Reference style: American Medical Association 11th edition.

[‡] The current version of this paper has been statistically supervised by James E. Pustejovsky and Terri D. Pigott.

ABSTRACT

In a recent paper, Vembye, Pustejovsky, & Pigott developed power approximation formulas for meta-analysis of dependent effect sizes across the multi-level meta-analysis (MLMA), the correlated effects (CE), and the correlated-hierarchical effects (CHE) models. However, the paper mainly focused on the statistical accuracy and quality assurance of the performance of the newly developed methods and less on the practical challenges encountered in applying the methods. The goal of this paper is to support applied reviewers by making these rather complex power approximation formulas practically accessible and providing guidance about obtaining the relevant quantities required to conduct reliable power approximations for meta-analyses involving statistically dependent effect sizes. In this paper, we introduce guidelines for conducting power approximations for meta-analyses of dependent effect sizes and introduce the *POMADE* R package for this purpose. We also present an overview of resources where reviewers can find information regarding parameters and quantities for making reasonable assumptions for the power approximations introduced here. We then provide R codes and examples for how to execute power analyses of the CHE model when guarding against misspecification via robust variance estimation (CHE-RVE). Hereto, we also show how to approximate the number of studies required to detect a given effect size considered to be of practical concern and how to approximate the minimum detectable effect size under fixed data and model conditions as well as with prespecified levels of statistical significance and power. Finally, we introduce new graphical tools, including the traffic light power plot for presenting power analyses across a range of plausible scenarios.

KEYWORDS: *power, meta-analysis, dependent effect sizes, CHE-RVE model, POMADE R package, traffic light power plot*

HIGHLIGHTS

What is already known

- Power of meta-analysis models for handling statistically dependent effect sizes can be approximated but is restricted by no common guidelines for how to be conducted reliably.
- Power approximations for meta-analysis of dependent effect sizes perform reliably when based on true empirical assumptions.
- Power approximations generally overestimate the true power by more than 10 percent when based on balanced assumptions, and these do not hold empirically.
- Power approximation involving robust variance estimation (RVE) outperforms other power approximation methods.

What is new

- General guidelines for the conduct of power analysis involving statistically dependent effect sizes.
- The *POMADE* R package for conducting **power** analyses of **meta-analysis** of **dependent effects**.
- Graphical tools for presenting *a priori* power analyses across a range of possible assumptions.

Potential impact

- Makes power analysis for meta-analysis of dependent effect sizes easily accessible for applied reviewers and a widely used practice in systematic reviews involving meta-analysis.
- Expansion of open science and open data practices.

1 INTRODUCTION

In a recent paper, Vembye, Pustejovsky, and Pigott¹ developed power approximation formulas for meta-analysis of dependent effect sizes across the multi-level meta-analysis (MLMA), the correlated effects (CE), and the correlated-hierarchical effects (CHE) models. However, the focus of that work was on the technical development of power formulas and assessing the accuracy of the proposed approximations rather than on the general use of the developed methods among applied reviewers and meta-analysts. There remains a need to consider the practical challenges encountered by reviewers in obtaining the relevant quantities required to conduct reliable power approximations for meta-analyses of dependent effect sizes. In this article, we provide guidelines for conducting approximations for the power of meta-analyses of dependent effect sizes and introduce the *POMADE* R package for this purpose. The ultimate goal is to make power approximation formulas for meta-analysis of dependent effect sizes accessible for applied reviewers.

The paper has four major aims. First, we present an overview of resources where reviewers can find information regarding the parameters and quantities needed to reliably execute power approximations. Second, we provide R codes and examples for how to conduct power analyses of the CHE model guarding against misspecification via robust variance estimation (CHE-RVE) since we believe that this model most adequately takes into account the dependency structures encountered in the social and behavioral sciences. Third, we introduce R codes for how to approximate the number of studies required to detect a given effect size considered to be of practical concern and how to approximate the minimum detectable effect size under fixed data and model conditions as well as with prespecified levels of statistical significance and power. Fourth, we introduce new graphical tools for presenting power analyses, including the *traffic light power plot* for presenting power analyses across a range of plausible scenarios of design features and model parameters.

2 BRIEF HISTORY OF POWER FOR META-ANALYSIS AND DEPENDENT EFFECT SIZES

Until recently, all power approximation techniques for meta-analysis^{2–5} were restricted by the assumption of independence among effect sizes, i.e., that all studies yield one effect size only. These have been shown to perform inadequately when used for approximating power for meta-analysis

models handling dependent effect sizes.¹ Furthermore, the assumption of independent effect sizes is rarely fulfilled in the social and behavioral sciences, where it is common for studies to report multiple effect sizes, producing various types of dependency structures in the meta-analysis data. Often studies report multiple eligible results for the same sample of participants (e.g., across different time points or types of measurements), creating correlated sample errors, also known as a *correlated effects dependency structure*. Yet, it is also common to find studies that report multiple results across non-overlapping samples (e.g., primary and secondary students, respectively), creating a multi-level or *hierarchical effects dependency structure* with effect sizes nested in samples and samples nested within studies. Albeit the results are deduced from non-overlapping samples, the fact the researchers apply the same estimation techniques, implementation strategies, measurement, etc., creates dependency among mean effects coming from the same study. Most often, both dependency structures appear simultaneously in social science reviews.

Various statistical techniques have been developed to handle dependencies among effect sizes. Originally, Hedges & Olkin⁶ and Raudenbush, Becker, & Kalaian⁷ suggested using multivariate effect sizes models, but these models were/are rarely used in practice⁸ because they require knowledge of the true dependency structures among effect sizes, and such information is rarely reported or retrievable from primary studies. A decade ago, methods^{9–11} based on robust variance estimation (RVE) or multi-level modeling (MLMA) were concurrently developed to handle dependency among effect sizes when the true dependency is partly or fully unknown. However, these methods are limited to either making correlated or hierarchical assumptions about the dependency structure, which in turn restricts the precision of these models when the dependency structure is misspecified, i.e., when the model substantially diverges from the true dependency structure(s)¹².

More recently, new statistical methods¹², defined as the correlated-hierarchical effects (CHE) model, have been developed in which multi-level modeling and RVE are combined while simultaneously accounting for the correlated and hierarchical effects dependency structures (therefore also defined as the CHE-RVE model). These CHE-RVE models more closely approximate the true dependency structures commonly found in meta-analysis applications in the social sciences and slightly increase the statistical power to find small effects in the circumstance when multiple dependency structures are inherent in meta-analytical datasets. Since the CHE-RVE model most adequately resembles the dependency structures found in social and behavioral science

meta-analyses, this paper concentrates on power approximations for the CHE-RVE model only. A further advantage of concentrating on the CHE-RVE model is that the common alternative models for handling dependent effect sizes can be seen as special cases of this model. For example, the MLMA model guarding against misspecification via RVE¹³ is a special case of the CHE-RVE model, assuming that the sample correlation $\rho = 0$. However, it is important to note that in cases when either no within-study heterogeneity or no correlation between effect sizes are expected, the CE or the MLMA¹³ models are the preferable models to be used. Power approximation functions for all of the common models for handling dependent effect sizes are available in the *POMADE* R package, and examples of how to use these methods will be incorporated in the accompanying vignette to the package. Power approximation formulas were also developed for the CHE and MLMA models, not guarding against misspecification via RVE or Satterthwaite degrees of freedom¹⁴, but we do not recommend using these models since they do not control the nominal Type-I error rate when the number of studies is small (i.e., less than 40 studies).

3 A PRIORI POWER APPROXIMATION FOR THE CHE-RVE MODEL

To illustrate the conduct of power analysis for meta-analysis of dependent effect sizes, we first describe the power approximation for a hypothesis test for an overall average effect based on standardized mean differences¹⁵ in which the assumed data-generating process follows that of the correlated-and-hierarchical effects (CHE) model as described by Pustejovsky and Tipton.¹²

The CHE model can be applied for meta-analyzing a set of studies where some or all included studies contribute multiple, statistically dependent effect size estimates. Suppose that we have a collection of J studies to be included in a meta-analysis, where study j includes $k_j \geq 1$ effect size estimates, for $j = 1, \dots, J$. Let T_{ij} denote effect size estimate i from study j , with corresponding standard error σ_{ij} , for $i = 1, \dots, k_j$ and $j = 1, \dots, J$. For simplicity, we assume that the sampling variances are constant within each study, so $\sigma_{1j}^2 = \sigma_{2j}^2 = \dots = \sigma_{k_j j}^2 = \sigma_j^2$.

As usual in meta-analysis, the CHE model makes the assumptions that each T_{ij} is an unbiased estimator of an effect size parameter θ_{ij} and that σ_{ij} is fixed and known. These assumptions can be expressed as

$$T_{ij} = \theta_{ij} + e_{ij}, \quad (1)$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, which has expectation zero and variance $\text{Var}(e_{ij}) = \sigma_j^2$. Effect size estimates from different studies are assumed to be uncorrelated, so $\text{cor}(e_{hj}, e_{il}) = 0$ when $j \neq l$, but effect size estimates from the same study may be correlated. Because information about the sampling correlation between effect sizes is often not available from included studies, analysts will typically need to make a more-or-less arbitrary assumption about the degree of dependence. With the CHE model, the correlations between sampling errors within a given study are all assumed to be equal to a known constant, $\text{cor}(e_{hj}, e_{ij}) = \rho$, specified by the analyst. This feature of the CHE model captures the “correlated effects” structure of the data.

The other component of the CHE model captures the “hierarchical effects” structure. Here, it is assumed that effect size parameters represent a sample from an underlying population of effects that has a hierarchical structure, according to

$$\theta_{ij} = \mu + u_j + v_{ij}, \quad (2)$$

where the study-level error term u_j has mean zero and variance τ^2 and the effect size-level error term v_{ij} has mean zero and variance ω^2 . The main parameters of the CHE model are the overall average effect size μ ; the between-study heterogeneity τ^2 ; the within-study heterogeneity ω^2 ; and the sampling correlation ρ . Under this model, we consider power approximations for tests of the null hypothesis $H_0: \mu = d$ versus a two-sided alternative, with specified Type-I error level α .

3.1 Estimation of CHE

Estimation of the overall average effect size μ entails first estimating the variance components and then using the estimated variance components to take an inverse-variance weighted average of the effect size estimates. Let $\hat{\tau}^2$ and $\hat{\omega}^2$ denote full or restricted maximum likelihood estimators of the variance components, which are calculated given an assumed sampling correlation ρ . Given values of these estimators, the overall average effect size estimate is a weighted average of the study-specific average effect size estimates, with weights given by

$$w_j = \frac{k_j}{k_j \hat{\tau}^2 + k_j \rho \sigma_j^2 + \hat{\omega}^2 + (1 - \rho) \sigma_j^2} \quad (3)$$

The overall average effect size is estimated as

$$\hat{\mu} = \frac{1}{W} \sum_{j=1}^J w_j \bar{T}_j, \quad (4)$$

where $\bar{T}_j = \frac{1}{k_j} \sum_{i=1}^{k_j} T_{ij}$ and $W = \sum_{j=1}^J w_j$. If the CHE model is correctly specified, then

$$\text{Var}(\hat{\mu}) \approx \frac{1}{W}. \quad (5)$$

Hypothesis tests or confidence intervals based on (5) will perform properly if the assumptions of the CHE model are good approximations to the true data-generating process.

In light of the lack of information about the sampling correlations between effect size estimates, meta-analysts may prefer to use tests based on robust variance estimation (RVE) methods, which maintain close-to-correct Type I error calibration even if the CHE model is misspecified. With the CHE working model, a robust estimator for the variance of $\hat{\mu}$ is given by

$$V^R = \frac{1}{W^2} \sum_{j=1}^J \frac{w_j^2 (\bar{T}_j - \hat{\mu})^2}{\left(1 - \frac{w_j}{W}\right)}. \quad (6)$$

When the working model is correctly specified, then V^R is an exactly unbiased estimator of $\text{Var}(\hat{\mu})$. However, even if the assumptions of the working model do not hold, V^R remains close to unbiased.

A robust test of the hypothesis $H_0: \mu = d$ is based on the robust Wald test statistic

$$t^R = \frac{\hat{\mu} - d}{\sqrt{V^R}}. \quad (7)$$

Tipton (2015) proposed approximating the distribution of t^R under the null hypothesis by a Student-t distribution with ξ degrees of freedom, where ξ is derived based on a Satterthwaite approximation under the assumption that the working model is correct. Specifically, the Satterthwaite degrees of freedom are calculated as

$$\xi = \left[\sum_{j=1}^J \frac{w_j^2}{(W - w_j)^2} - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{(W - w_j)^2} + \frac{1}{W^2} \left(\sum_{j=1}^J \frac{w_j^2}{W - w_j} \right)^2 \right]^{-1}. \quad (8)$$

The robust Wald test rejects the null hypothesis if $|t^R| > c_{\alpha/2, \xi}$, where $c_{\alpha/2, \xi}$ is the $\alpha/2$ critical value from a Student t distribution with ξ degrees of freedom.

3.2 Power approximation

Vembye, Pustejovsky, and Pigott¹ proposed to approximate the power of the Wald robust test using a non-central Student-t distribution, with non-centrality parameter given by

$$\lambda = \sqrt{W}(\mu - d) \quad (9)$$

and degrees of freedom as given in Equation (8). The power of the robust Wald test against a two-sided alternative is then approximated as

$$F_t(-c_{\alpha/2, \xi} | \xi, \lambda) + 1 - F_t(c_{\alpha/2, \xi} | \xi, \lambda), \quad (10)$$

where $F_t(x | \xi, \lambda)$ is the cumulative distribution function of a non-central Student-t distribution, and $c_{\alpha, \xi}$ is the upper α -level critical value for the central Student-t distribution with ξ degrees of freedom, so $F_t(c_{\alpha/2, \xi} | \xi, 0) = 1 - \alpha/2$. This approximation assumes that the CHE model is correctly specified.

The power of the test based on CHE-RVE depends on several parameters: the true average effect size μ , the between-study variance τ^2 , the within-study variance ω^2 , and the assumed correlation between sampling errors ρ . In the next section, we discuss strategies for making assumptions regarding these parameters for purposes of prospective power analysis and sample size planning.

The power of the test also depends on the number of studies in the meta-analysis (J), the magnitude of their sampling variances ($\sigma_1^2, \sigma_2^2, \dots, \sigma_J^2$), and the number of effect sizes contributed by each included study ($k_1, k_2, \dots, k_J$). Prior to completing a systematic review, the sampling variances and number of effect sizes per study will not be known precisely. For prospective power analysis, Vembye, Pustejovsky, and Pigott¹ proposed treating these quantities as random variables that follow some distribution. The distribution might be based on empirical data from an initial scoping review or a previous meta-analysis on a similar topic, or it might be based on more stylized assumptions involving a parametric distribution. With this approach, the power of the test is calculated by taking the expected value of Equation (10) over the distribution of sampling variances and effect sizes per study. Practically, the expectation is approximated by drawing a random sample of J sets of study characteristics (σ_j^2, k_j) from specified distributions, calculating λ and ξ based on the sample of study characteristics, and then calculating power with Equation (10). This process is repeated several times, with the expected power level calculated as the overall average power across repeated samples. In the *POMADE* package presented below, this process is by default repeated 100 times.

3.3 Sample size planning

The proposed methods provide a means of approximating the power of a test of the null hypothesis $H_0: \mu = d$ versus a two-sided alternative, given assumptions about the true overall average effect size, for a meta-analysis with a specified number of studies. Researchers in the planning stage of a meta-analysis might use the methods directly to answer the question “What is the power of this test?” However, they might also find it useful to frame the question somewhat differently. Two alternative framings are common: one that centers on a target sample size and one that centers on minimum meaningful effect sizes.

One alternative framing is to pose the question, “*How big a sample is needed to achieve a specified power level?*” To answer this question, we would first specify a desired power level P , such as the conventional level of $P = 0.8$, a minimum effect size of interest (μ), and a distribution of primary study sample sizes and effect sizes per study. Given these quantities, the number of included studies J affects power through the total weight W , which in turn determines the non-centrality parameter λ , and through the degrees of freedom ξ . Therefore, the target sample size is the smallest value of J that satisfies the equation

$$P = E \left[F_t \left(-c_{\alpha/2, \xi} \middle| \xi, \sqrt{W}(\mu - d) \right) + 1 - F_t \left(c_{\alpha/2, \xi} \middle| \xi, \sqrt{W}(\mu - d) \right) \right], \quad (11)$$

where the expectation is taken over the distribution of primary study sample sizes and effect sizes per study. The solution can be found through a direct grid search over a range of possible values for J . This feature is integrated in the `find_J_*` functions presented below.

Another alternative framing is to pose the question, “*How small an average effect size can be detected with a given sample size with a specified power level?*” To answer this question, we would again need to specify a desired power level P and a distribution of primary study sample sizes (or variance estimates) and effect sizes per study. We would also need to specify an anticipated sample size, J . Given these assumptions, we can find the average effect size μ that satisfies Equation (11). Just as with the previous question, the solution can be found through a direct grid search over a range of possible values for μ , and is integrated in the `MDES_*` functions presented below.

4 SUGGESTIONS FOR HOW TO OBTAIN RELEVANT EMPIRICAL PARAMETERS AND QUANTITIES NEEDED FOR POWER APPROXIMATION

As is apparent from the above-presented power approximation formula and procedure, reviewers must put forward a range of assumptions to conduct reliable power analyses. This is, of course, a clear limitation of the methods. To mitigate this limitation, this section presents guidelines for how we think researchers could make plausible and empirically informed assumptions needed to execute reasonable power approximation. We discuss each parameter and quantity needed for the power approximation one by one. In this regard, we do not consider the choice of the α -level, since

we will use the conventional $\alpha = .05$ for all the presented power calculations below. Researchers should, of course, change the α -level based on their research context¹⁶.

4.1 Smallest effect size of practical concern, μ

The first thing reviewers need to determine to conduct power analysis of meta-analysis is the smallest effect of practical concern, μ . Importantly, the determination of the smallest effect size of practical importance exclusively hinges on the specific topic of the review literature. Although common practice in the social and behavioral sciences, we do not recommend using general effect size conventions for small, medium, and large effect sizes such as Cohen's¹⁷ or Hattie's¹⁸ standards. As others have argued, relying on such decontextualized standards amounts to “characterizing a child's height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals”¹⁹.

The smallest effect size deemed to be of practical importance should be determined in relation to a range of factors such as the cost, complexity, and scalability of the intervention. Furthermore, μ should be determined by comparing the review intervention(s) to structurally related and/or similarly resource-intensive interventions from previous syntheses on similar research topics. Therefore, the smallest effect size of practical importance should ideally be deduced from relevant content sources related to the given discipline(s) and topic(s) under review.

In education, researchers interested in the effects of field experiments/interventions on student achievement could profitably apply Kraft's²⁰ empirical benchmarks for interpreting the smallest effect size of practical significance of educational interventions on standardized achievement outcomes. If reviewers are concerned with grade-specific effect sizes, they can also consult Lipsey and colleagues's¹⁹ overview of effect sizes of annual achievement gains. In psychology, reviewers could consult Schäfer & Schwartz²¹ to understand meaningful effect sizes across sub-disciplines.

4.2 Expected number of studies, J

A major aim of conducting power analysis for meta-analysis is to gain knowledge about how many studies, J , are needed to find the smallest effect size of practical concern. The number of studies expected to be found will often be based on the reviewers' content-specific knowledge of the given review topic. However, reviewers should conduct power analyses across a range of assumptions about the expected number of studies to be found to allow for the possibility that the literature

search and author solicitation reveal further studies unknown to reviewers. If reviewers are uncertain about the anticipated number of included studies, they could consult previous syntheses and reviews on similar research topics and/or from similar disciplines⁸. In education, reviewers could consult Hattie¹⁸ and Ahn et al.’s²² overviews of meta-analyses across various topics. Across education, psychology, and medicine, reviewers could look into Tipton, Pustejovsky, and Ahmadi⁸ for an overview of the average number of studies included in meta-analyses in these disciplines. Another source for retrieving empirical meta-analytical data, including J , is the *metadat* R package²³, in which a large number of datasets of previously conducted meta-analyses are stored.

4.3 Number of effect sizes per study, k_j

Making assumptions about the number of effect sizes per study, k_j , in Equation (3) can be done in various ways. Ideally, reviewers should obtain this information from pilot data of previous reviews on related topics. In practice, however, this advice might be difficult to compile because it is still not a common practice for systematic reviews and meta-analyses to open source their data. Reviewers could, of course, contact previous review authors to gain access to the relevant data. However, this might be a complicated route since author responses are generally low²⁴. If relevant data from previous systematic reviews is not available, the *metadat* R package²³ could again be used. Alternatively, reviewers could simulate k_j around the average k_j previously found in education, psychology, or medicine^{8,22}. We have made this simulation function available in the below-presented *POMADE* R package.

Researchers might be inclined to make the simplifying assumption that all studies in the synthesis will include the same number of effect sizes (i.e., a “balanced” design where $k_1 = k_2 = \dots = k_j = k$). Except when this assumption is true by design of the review, we recommend against using such an assumption because it rarely holds in practice and because, if the true k_j varies from study to study, then the power approximations will systematically overestimate the true power of the model.¹

4.4 Study sample sizes, N_j , or sampling variances, σ_j^2

To conduct reliable power approximations, reviewers must further put forward assumptions about the distribution of sampling variances, σ_j^2 , in the included studies. Such information might be

difficult to retrieve in practice, but we generally suggest that reviewers obtain this information either from pilot data of previously conducted reviews on similar research topics or from relevant meta-analytical datasets from the *metadat* package.

For a given effect size metric, the distribution of sampling variances can often be approximated from information about the distribution of sample sizes, N_j . For example, for the standardized mean difference effect size metric involving comparison of two groups of independent observations, the sampling variance of the effect size estimate is approximately

$$\sigma_j^2 \approx \left(\frac{4}{N_j} + \frac{\mu^2}{2(N_j - 2)} \right) \quad (12)$$

where μ^2 denotes the anticipated overall average effect size. As with k_j , we do not recommend the assumption of complete balance about N_j or σ_j^2 (i.e., assuming $N_1 = N_2 = \dots = N_j = N$ and $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_j^2 = \sigma^2$), because it is rarely experienced in practice, and if the true N_j and σ_j^2 vary, the power approximations will overestimate the true power of the model.¹ The *POMADE* package also includes functions from which N_j can be simulated in cases where pilot data is inaccessible.

4.4.1 Clustering

For the power approximations to work properly, reviewers must account for clustering in the included effect sizes^{25,26}. Otherwise, the power approximation will heavily overestimate the true power of the given model. Therefore, if clustered studies are expected to be included in the review, or the intervention(s) is/are provided at the cluster level²⁷, as is often the case in education²⁸, it is pivotal that reviewers either apply *effective sample sizes*⁴ (ESS) or sampling variances that both include variation from the individual and the cluster levels²⁹ for the power approximation functions to work properly. If reviewers have a vector of raw sample sizes, N_j , from clustered studies, these can be corrected for one level of clustering by roughly approximating ESS_j via

$$ESS_j = \frac{N_j}{DE} \quad (13)$$

where DE is the design effect of a two-stage sample given by

$$DE = 1 + (n - 1)\rho_{ICC} \quad (14)$$

with n being the average cluster size and ρ_{ICC} the intraclass correlation coefficient (ICC) for the cluster level. Relevant compendiums of ICC in education can be found in Hedges & Hedberg³⁰, in medicine from Gulliford, Ukoumunne, & Chinn³¹ and Verma & Lee³², and in psychology from Murray & Blitstein³³. The `effective_sample_sizes()` function from the *POMADE* package can be used to correct the raw sample size from cluster studies. If reviewers have pilot data containing a vector of sampling variances not including cluster-level variation, these can roughly be adjusted for cluster bias by multiplying DE to each sample variance component. The `cluster_bias_adjustment()` function from the *POMADE* package can be used for this purpose. Ideally, reviewers should strive to obtain pilot data, including sampling variances estimated from multi-level models or cluster robust standard errors or alternatively sampling variance components that have been cluster-bias corrected as for examples done in Tanner-Smith & Lipsey³⁴ and Dietrichson et al.³⁵

4.5 Between-study variance (study-level variance), τ^2

When making assumptions about a plausible value for the between-study variance, τ^2 , reviewers could, as with the other assumptions, consult previous reviews of similar topics. Alternatively, reviewers could follow the guideline suggested by Pigott³⁶ in which $\tau^2 = (1/3)\sigma^2$ is considered as a low degree of heterogeneity, $\tau^2 = \sigma^2$ is considered as a moderate degree of heterogeneity, and $\tau^2 = 3\sigma^2$ is considered as a large degree of heterogeneity, where σ^2 can be obtained from a simplified version of Equation (12):

$$\sigma^2 \approx \left(\frac{4}{N} + \frac{\mu^2}{2(N-2)} \right)$$

Where N is the ‘typical’ sample size or effective sample size expected to be found in the given literature. Reviewers could consult Fraley & Vazire³⁷ to gain an overview of common study sample sizes in psychology journals. To make these calculations accessible to reviewers, we have made this procedure available via the `tau2_approximation()` function from the *POMADE* package. To

recognize the uncertainty of the τ^2 estimation, we highly recommend that power approximations are conducted across a range of possible values of τ^2 . To make more intuitive estimates of τ^2 , it can be an advantage to think of the study-level heterogeneity in terms of between-study standard deviation (SD) units since these are at the same scale as the mean effect size, μ .

4.6 Within-study variance (effect size level variance), ω

As with the τ^2 estimate, the true within-study variance, ω^2 , could be obtained from result sections of previous reviews of similar research topics or estimated from relevant pilot data with dependent effect sizes. Similarly, we suggest that reviewers think of the effect-level heterogeneity in terms of within-study SD since it allows for a more intuitive interpretation of this variance component. It might also be helpful to think of ω relative to τ or *vice versa*. Say for example that reviewers expect one-third of the total true variance to come from within-study heterogeneity, then $\omega^2 = \tau^2 \times 0.5$. As with τ^2 , we think it is good practice to conduct power analyses across a range of within-study SD estimates to accommodate the uncertainty of the made assumption and then highlight the most likely scenario. We elaborate more thoroughly on this procedure in Section 5.

4.7 Assumed sample correlation, ρ

Finally, reviewers have to make assumptions about the expected sampling correlation among outcomes coming from the same study. This is indeed a tricky part of the power approximation of the CHE-RVE model. However, there are certain ways that reviewers can make reliable estimates of ρ . First, reviewers could search for literature in relevant disciplines for common sample correlations among the outcome measures relevant for the review. Second, if raw primary data containing multiple eligible outcomes measures are available to the reviewers, ρ could be estimated from this data. For example, Vembye, Weiss, & Bhat³⁸ used data from the Project STAR to estimate ρ and inform the choice of ρ in their systematic review regarding the effects of collaborative models of instruction on student achievement. Third, if reviewers have access to relevant meta-analytical pilot data containing studies reporting two outcome measures, then ρ could be obtained by simply estimating the correlation between the pairs of effect sizes estimates from those studies that provide both types of outcomes measures³⁹. Notice, however, that it is recommended, in this case, to have at least ten of such studies to be able to obtain a reliable estimate of ρ .³⁹ Independently of the

used methods to obtain ρ , we suggest that reviewers conduct power analyses across a range of different assumptions about ρ to inspect the impact of ρ on the power estimate.

5 EMPIRICAL EXAMPLE

5.1 Replication materials

All R codes for replicating the below-presented power approximation examples are available on the Open Science Framework at <https://bit.ly/3uuinTz>. For plot generation, the *POMADE* package draws on the *ggplot2* R package⁴⁰.

5.2 Power example of the CHE-RVE model using relevant pilot data

To illustrate the procedure of power analysis for meta-analysis of dependent effect sizes, suppose that we want to conduct a meta-analysis about the effects of increased instruction time by increasing the length of the school day on student achievement. To compute power for this analysis, we use pilot data from Vembye, Weiss, & Bhat's³⁸ (henceforth VWB22) meta-analysis regarding the effects of collaborative models of instruction on student achievement. We use VWB22 as pilot data because these interventions are related to increased instruction time interventions by representing true alternatives to increasing the length of the school day. From this systematic review and pilot data[§], we can find all of the relevant parameters and quantities needed to conduct power approximation except for the smallest effect size of practical concern. As previously emphasized, the smallest effect size of substantial concern must be deduced from theoretical and practical considerations.

The VWB22 study found a total of 76 studies eligible for meta-analysis, of which 82% of the effect sizes were adjusted for pretest measures, and the study data contain both correlated and hierarchical effects dependence structures, supporting the use of the CHE-RVE model. Based on this information, we assume that we will find $76 \text{ studies} \pm 10$, which might be a realistic expectation since this number of studies falls within the average number of studies found in education and applied psychology⁸. The VWB22 study further found a substantial amount of heterogeneity with variance components (reported as SDs) of $0.25 \text{ SD at the effect size level } (\omega)$ and $0.1 \text{ SD at the study level } (\tau)$. VWB22 then estimated $\rho \approx 0.7$ from paired effect size estimates for studies both

[§] Find data and background material for this study at <https://bit.ly/3nhVX3H>.

reporting STEM and Language Arts outcomes, as suggested by Kirkham et al³⁹. The pilot data of VWB22 further makes it possible to obtain a vector of k_j s, with $\bar{k}_j = 3.8$, ranging from 1 to 27, and a vector of cluster bias corrected σ_j^2 s aggregated to the study level. Cluster bias correction was needed in this case since 67 out of 76 did not adequately account for nesting of students within classes and schools. Furthermore, since both collaborative models of instruction and increased instruction time are provided at the class level, it is important to account for clustering in such reviews²⁷.

We define the smallest effect size of practical importance relative to the overall effect size of similar cost and resource-intensive interventions such as co-teaching and class size reduction, which both appear to have an overall average effect of approximately $0.1 SD$ ^{38,41}. Therefore, we here consider an overall average effect size falling below 0.1 to be practical uninteresting compared to these related interventions. With all the needed assumptions in place for power approximation of the mean effect size of the CHE-RVE model, power can be approximated from the `power_CHE()` function from the *POMADE* package as presented below.

```
# install.packages("devtools")
devtools::install_github("MikkelVembye/POMADE")

library(POMADE)
library(dplyr)

# Check information about pilot data
?VWB22_pilot

# Make a dataset with two variables, including a vector of the number of
# effect sizes per study (named kj) and a vector of the sampling variance
# components (named sigma2j).

dat_kjsigma2j <- select(VWB22_pilot, kj, sigma2j = vg_ms)

power_CHE(
  J = 76,                                # Expected number of studies
  tau2 = 0.1^2,                          # Between-study variance (from VWB22)
  omega2 = 0.25^2,                      # Within-study variance (from VWB22)
  beta = 0.1,                           # Smallest ES of practical relevance
  rho = .7,                             # Sample correlation (from VWB22)
  var_df = "RVE",                       # Type of variance and df
  sigma2_method = "empirical",          # Specifies how sigma2js are obtained
  pilot_data_kjsigma2 = dat_kjsigma2j, # Pilot data
  seed = 10052510                       # Set seed to ensure reproducibility
)

# A tibble: 1 x 7
  samp_method      method    es  var_b    df power_sig05 iterations
  <chr>          <chr>    <dbl> <dbl> <dbl>    <dbl>      <dbl>
1 empirical sigma2js CHE-RVE  0.1 0.00134 40.9      0.761      100
```

From these results, it appears that we would have 76.1% power to find $\mu = 0.1$ with 76 studies and with similar model parameters and study characteristics as found in VWB22.

5.3 Number of Studies Needed to Find the Smallest Effect of Interest

Another feature embedded in the *POMADE* package is functions that allow researchers to answer questions concerning how many studies are needed to obtain a given effect size considered to be of practical interest when the levels of statistical significance and power are prespecified. For the CHE-RVE model, this can be investigated via the `find_J_CHE()` function presented below.

```
Find_J_CHE(
  mu = 0.1,
  tau2 = 0.1^2, omega2 = 0.25^2, rho = 0.7,
  alpha = .05, target_power = .8, # Default settings
  pilot_data_kjsigma2 = dat_kjsigma2j,
  seed = 10052510
)
# A tibble: 1 x 7
  samp_method      method      es alpha target_power J_needed iterations
  <chr>          <chr>    <dbl> <dbl>      <dbl>    <dbl>      <dbl>
1 empirical sigma2s CHE-RVE    0.1  0.05        0.8      84      100
```

From these results, we can see that it would require 84 studies to have 80% power to detect $\mu = 0.1$ under conditions similar to those presented and found in VWB22.

5.4 Minimum Detectable Effect Size (MDES)

To answer the questions about the *minimum detectable effect size* (MDES) with preset levels of statistical significance and power as well as fixed study parameters and study characteristics, we have developed the MDES functions. For example, the minimum detectable effect size for the CHE-RVE model can be estimated from the `MDES_CHE()` function, as presented below.

```
MDES_CHE(
  J = 76,
  var_df = "RVE",
  tau2 = 0.1^2, omega2 = 0.25^2, rho = 0.7,
  alpha = .05, target_power = .8, # Default settings
  pilot_data_kjsigma2 = dat_kjsigma2j,
  seed = 10052510
)
# A tibble: 1 x 7
  samp_method      method N_studies alpha target_power MDES iterations
  <chr>          <chr>    <dbl> <dbl>      <dbl>    <dbl>      <dbl>
1 empirical sigma2s CHE-RVE    76  0.05        0.8  0.105      100
```

From here, it can be found that the smallest effect size detectable with 80% power under the given conditions is 0.105, i.e., very close to the smallest effect size considered to be of practical relevance, clearly underlining the importance of meta-analysis.

5.5 Plotting

5.5.1 Power

We acknowledge that it can be rather difficult to guess/approximate the true model parameters and sample characteristics, including the final number of studies *a priori*. Making only one power approximation can easily be misleading even if the true model and data structure slightly diverge from the yielded data and model assumptions. To maximize the informativeness of the power approximations, we suggest accommodating the uncertainty of the power approximations by reporting or plotting power estimates across a range of possible scenarios. Figure 1 depicts such a plot in which power estimates are approximated across varying assumptions of τ , ω , and ρ and J . In the function, reviewers can also specify and illustrate the interval in which they expect the final number of studies to fall. This provides a means for reviewers to assess the consequences of the assumptions for the power estimate and determine under which scenarios the model power exceeds 80%. Here, we follow the convention of setting 80% power as the minimum acceptable power estimate reasonable for model fitting. This means that the Type I error is considered four times as serious as making a TYPE II error, i.e., $.20/.05$.¹⁷ Reviewers can make the power plot (Figure 1) by using the `power_plot()` function in the *POMADE* package, as presented below.

```
# Black and white power plot for the CHE-RVE model with 100 iterations
power_CHE_RVE_plot <-
  power_plot(
    J = seq(50, 100, 10),          # Range of expected studies to be found
    tau2 = c(0, 0.05, 0.1, 0.2)^2, # Between-study var (reported as SD)
    omega2 = c(0.05, 0.15, 0.25, 0.35)^2, # within-study var (reported as SD)
    beta = 0.1,                   # Smallest ES of practical concern
    rho = c(.2, .4, .7, .9),      # Potential sample correlation estimates
    model = "CHE",                # Model specification (Default)
    var_df = "RVE",              # Var and df specification (Default)
    pilot_data_kjsigma2 = dat_kjsigma2j, # Pilot data (VWB22)
    expected_studies = c(66, 86),  # Expected J-interval (gray shades)
    seed = 10052510              # Set seed to ensure reproducibility
  )

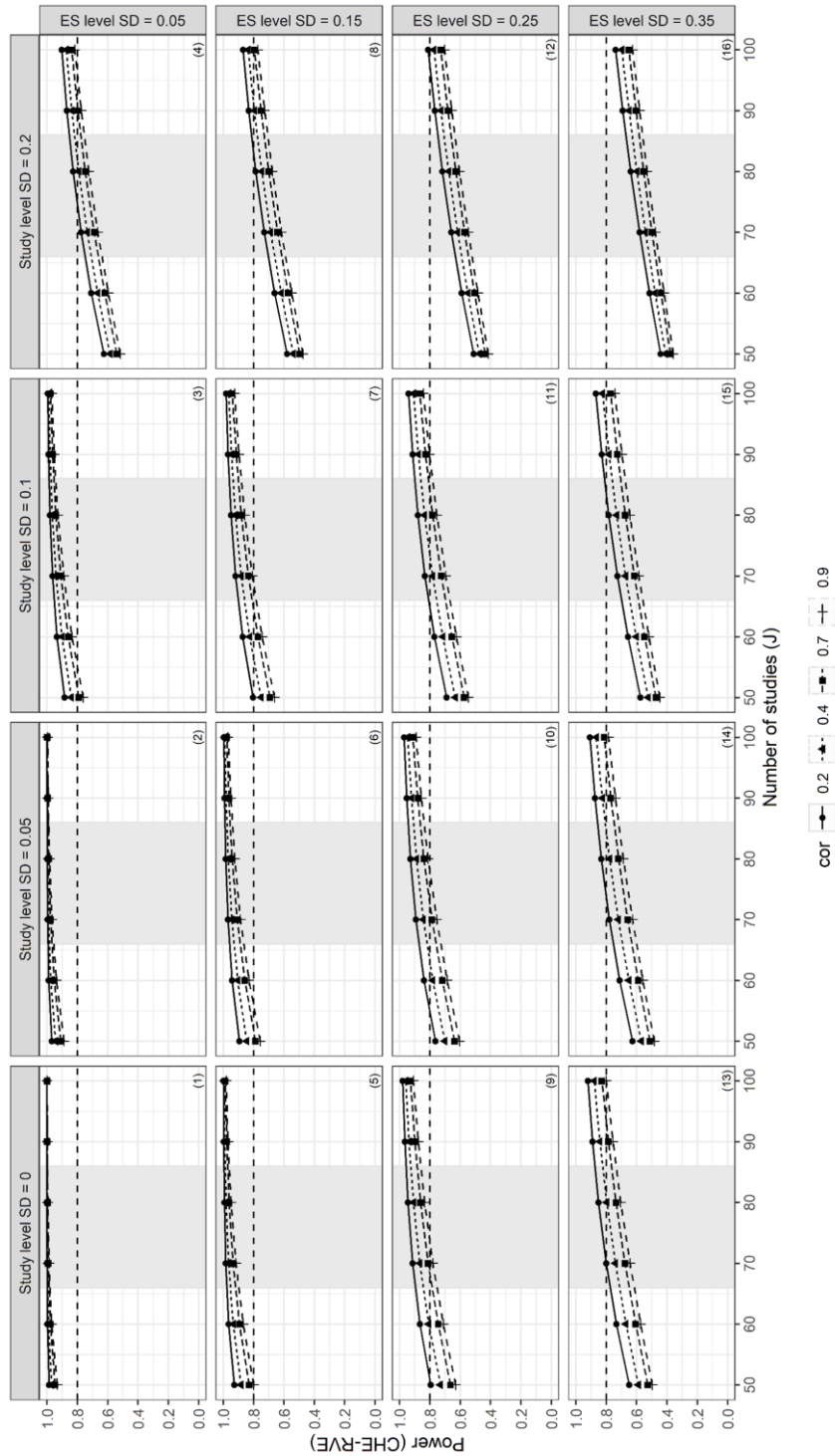
power_CHE_RVE_plot
```

The plot can be saved via the `ggsave` function from the *ggplot2* package⁴⁰.

```
# Save plot
library(ggplot2)

ggsave(power_CHE_RVE_plot, file = "/file_path/power plot.png",
  dpi = 600, height = 7, width = 12)
```

FIGURE 1. Power for meta-analysis of dependent effect sizes plot (CHE-RVE)



Note: Dashed lines indicate power of 80 percent. Shaded gray areas mark the range of studies expected to be found by the reviewer.

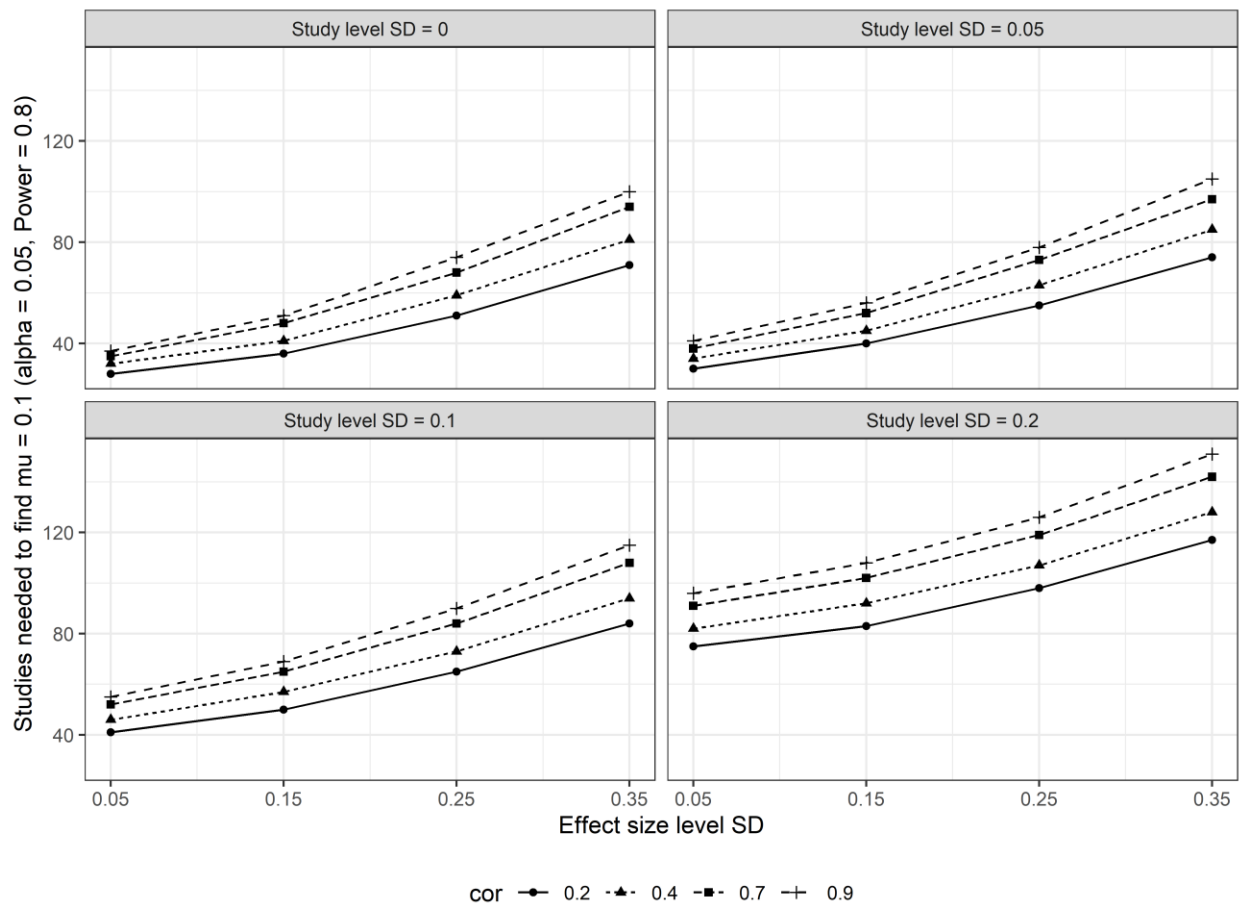
5.5.2 Number of studies (J)

To investigate the question of how many studies are needed to detect a given effect size of practical concern across varying assumptions about τ , ω , and ρ (Figure 2), the `find_J_plot()` function can be used for this purpose, as presented below.

```
J_plot <-
  find_J_plot(
    mu = 0.1,
    tau2 = c(0, 0.05, 0.1, 0.2)^2,
    omega2 = c(0.05, 0.15, 0.25, 0.35)^2,
    rho = c(.2, .4, .7, .9),
    pilot_data_kjsigma2 = dat_kjsigma2j,
    seed = 10052510, alpha = .05, target_power = .8
  )

J_plot1 <- J_plot +
  ggplot2::scale_y_continuous(breaks = seq(30, 150, 20))
J_plot1
```

FIGURE 2. Studies needed to find $\mu = 0.1$ across varying values of τ^2 and ω^2 (CHE-RVE)



From plots like Figure 2, researchers can, thereby, gain knowledge about the target range of the number of studies needed to detect the smallest effect size of practical concern.

Furthermore, by adding multiple values of μ to the `find_J_plot()` function, reviewers can also investigate how the number of studies needed changes as a function of the smallest effect size of interest. This analysis can be conducted via the codes presented below. Find the output results in Figure 3.**

```
find_J_plot(
  mu = c(0.05, 0.1, 0.15, 0.2),
  tau2 = c(0, 0.05, 0.1, 0.2)^2,
  omega2 = c(0.05, 0.15, 0.25, 0.35)^2,
  rho = c(.2, .4, .7, .9),
  pilot_data_kjsigma2 = dat_kjsigma2j,
  seed = 10052510, alpha = .05, target_power = .8
)
```

5.5.3 Minimum detectable effect size

To understand how the minimum detectable effect size varies across the number of included studies and various model parameters, the `MDES_plot()` function can be used, as presented below. Find the output results in Figure 4.

```
MDES_plot(
  J = seq(50, 100, 10),
  tau2 = c(0, 0.05, 0.1, 0.2)^2,
  omega2 = c(0.05, 0.15, 0.25, 0.35)^2,
  rho = c(.2, .4, .7, .9),
  pilot_data_kjsigma2 = dat_kjsigma2j,
  seed = 13042022, alpha = .05, target_power = .8,
  expected_studies = c(66, 86)
)
```

Figure 4 provides a means for reviewers to understand what effect sizes can actually be detected under a range of different data and model assumptions. From Figure 4, it can, for instance, be seen that across all the different scenarios, reviewers can at minimum detect a moderate²⁰ effect, clearly justifying meta-analysis.

** Note that running this plot can last more than 30 minutes when $\mu < 0.1$.

FIGURE 3. Number of studies needed as function of μ (CHE-RVE)

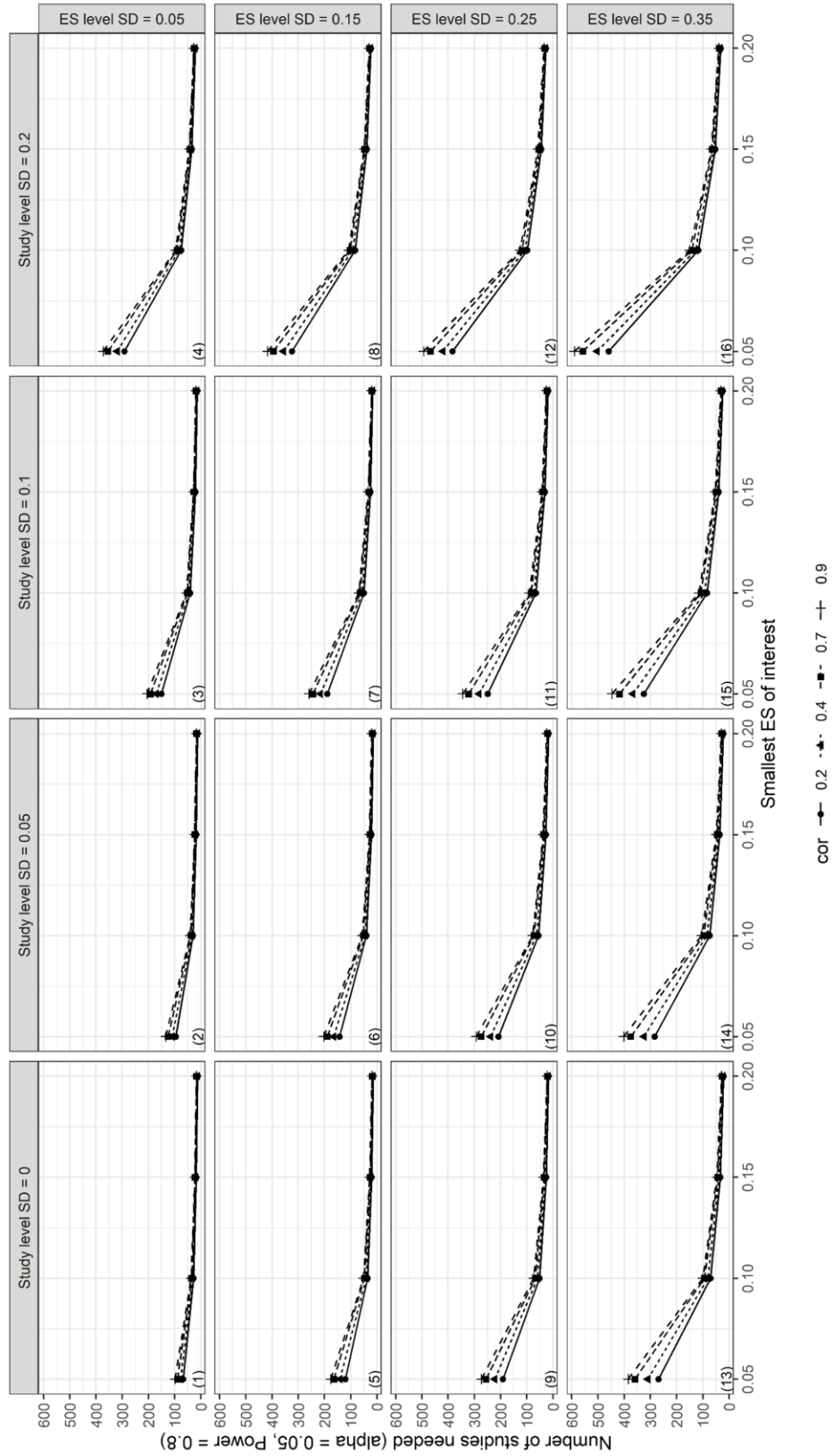
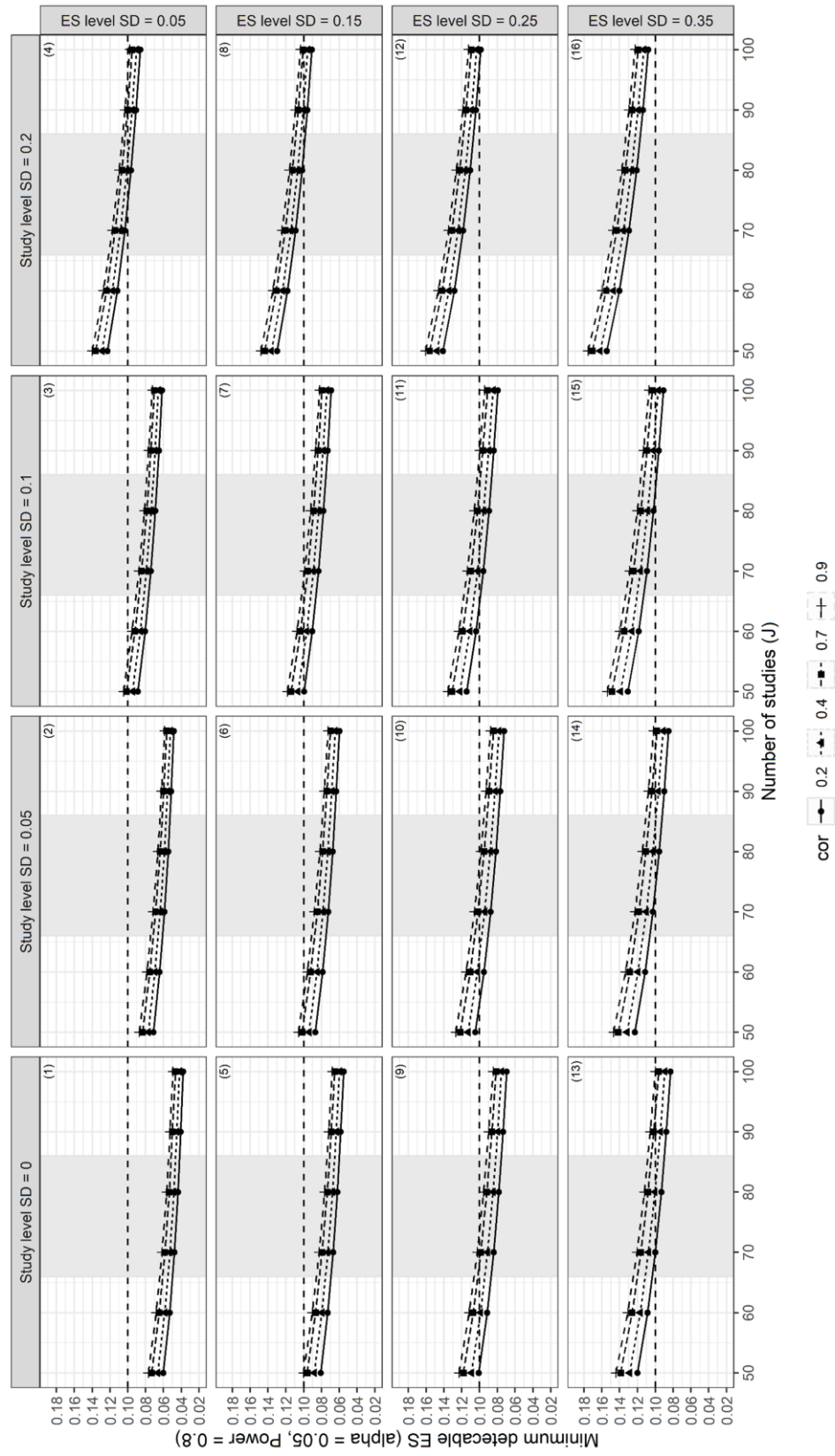


FIGURE 4. Minimum detectable effect size plot (CHE-RVE)



Note: Dashed lines indicate the smallest effect size of practical concern. Shaded gray areas mark the range of studies expected to be found by the reviewer.

5.6 Traffic light power plot

To further augment and more clearly illustrate the assumptions put forward by the reviewers, we suggest that reviewers use what we have coined as a *traffic light power plot*. Figure 5 shows a traffic light power plot in which the likelihood of the reviewers' assumptions is fleshed out by coloring the strips of the facet grid plots with green-colored parameters indicating the expected scenario, yellow-colored parameters indicating likely scenarios, and red-colored parameters indicating unlikely scenarios based on previous knowledge of the research topic. This way, it is clear to others, including funders, what they can expect in terms of power while also acknowledging the uncertainty in these estimates. We suggest approximating no more than four unlikely scenarios to keep the yielded assumptions down to a fair number. The traffic light power plot (Figure 5) can be made by using the `traffic_light_power_plot()` function in the *POMADE* package presented below.

```
power_CHE_RVE_color_plot <-
  power_plot(
    J = seq(50, 100, 10),
    tau2 = c(0, 0.05, 0.1, 0.2)^2,
    omega2 = c(0.05, 0.15, 0.25, 0.35)^2,
    beta = 0.1,
    rho = c(.2, .4, .7, .9),
    model = "CHE ",
    var_df = "RVE ",
    pilot_data_kjsigma2 = dat_kjsigma2j,
    expected_studies = c(66, 86),
    color = TRUE,          # indicate if colored lines and points should be used
    color_brewer = TRUE,  # use the "qual" palette = 2 (can be omitted)
    seed = 10052510
  )

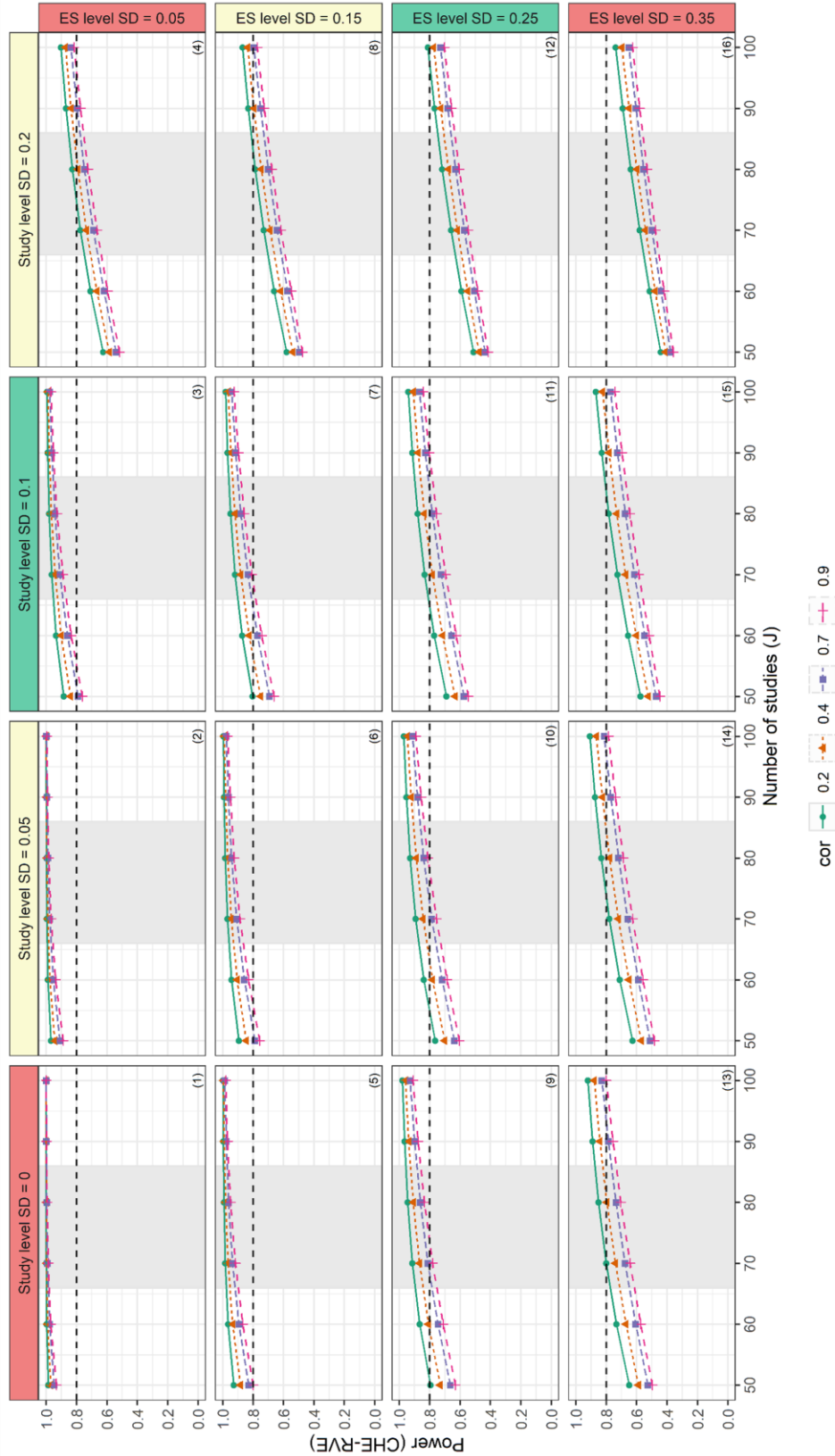
power_CHE_RVE_color_plot

# Traffic light power plot
# Coloring from upper-left strip to lower-right strip
# Remove below hashtags (#) and mark all codes to save the traffic light power plot

#png("traffic_light_power_plot.png", height = 7, width = 12, units = "in", res = 600)
traffic_light_power_plot(

  power_plot = power_CHE_RVE_color_plot,
  assumptions = c("unlikely", "likely", "expected", "likely", # Tau assumptions from left to right
                  "unlikely", "likely", "expected", "unlikely") # Omega assumptions from top to bottom
)
#dev.off()
```

Figure 5. Traffic light power plot (CHE-RVE)



Note: Dashed lines indicate power of 80 percent. Shaded gray areas mark the range of studies expected to be found by the reviewers. The colors of the strips indicate the reviewers' expectation of the given scenarios appearing in the dataset and analysis.

Figure 5 illustrates the power of the CHE-RVE model to find $\mu = 0.1$ across the assumptions we made regarding the effect of increased instruction time on student achievement with green color strips indicating our expectation to find $\tau = 0.1$ and $\omega = 0.25$ based on previous findings from VWB22. τ and ω are here reported as SDs so that they can be interpreted in the same unit as μ . Furthermore, the gray shades in Figure 5 depict our expectation to find between 66-86 studies in the given body of literature, i.e., we expect to find ± 10 studies of what was found in VWB22 and which is also the mean number of studies reported in the Review of Education Research and Applied Psychology journals⁸. The four lines in the traffic light power plot indicate various assumptions about the common sample correlation among effect sizes coming from the same study, ρ . We assumed $\rho = .7$, and under the expected (green) scenario in plot (11) in Figure 5, power estimates range from ~70% power with 66 studies to ~80% power with 86 studies. Though power does not exceed 80% in all scenarios, we would still suggest proceeding with meta-analysis, since only a minor reduction of the within-study variance would yield power above 80%. As can be seen in plot (7) in Figure 5, reducing ω with 0.1 SD would increase power by 10% or more and thus produces power above 80% across all numbers of expected studies (J). Therefore, these results indicate, in this case, that reviewers should do everything they can to reduce within-study variation, for example, by averaging results across subscale and subgroup results irrelevant to the main (subgroup) analyses of the given review.

6 UTILITY OF PROSPECTIVE POWER ANALYSIS FOR META-ANALYSIS

One of the major aims of *a priori* power analysis for meta-analysis is that it can shed light on the utility of a planned systematic review. Ultimately, it can inform reviewers and funders if enough studies are available to find the smallest effect size of practical/substantial concern and thus whether the literature is mature enough for a meta-analysis. In this regard, we must emphasize that reviewers should be careful abandoning meta-analysis based on power analyses conducted before the full literature search, partly because the power approximations require an extensive amount of assumptions that can be empirically error-prone (and thereby misleading) and partly because an unexpected number of eligible studies might be revealed to the reviewers during the literature search, e.g., through searches of gray literature databases⁴. As anecdotal evidence to support this advice, the first author was a part of a review³⁸ in which the authors only expected to find 20 eligible studies but ended up finding 128 studies, with approximately 100 studies coming from

gray literature searches. Furthermore, the true effect size can potentially diverge strongly from the smallest effect size considered to be of practical concern. This will substantially increase the power of the model. However, after finalizing the study collection, reviewers might reconsider if meta-analysis with low power should be conducted based on the detected number of studies. Hereto, it is pivotal to stress that if reviewers decide based on the power analysis not to proceed with meta-analysis—e.g., due to a small number of studies—this does not simply justify narrative synthesis as the alternative. In this case, reviewers should carefully look into relevant quantitative alternatives.^{4,42} Furthermore, power analysis can help inform reviewers about how many studies are needed to estimate the smallest effect size of interest but also what the minimum detectable effect size is under preset levels of significance and power as well as fixed study characteristics and parameters expected to be found in the literature under review.

In the social and behavioral sciences, it is common to find a large proportion of small studies that contribute with a large number of effect sizes to the common pool of effect sizes. In such cases, prospective power analyses can provide vital information about the impact of including a large proportion of such studies on the within-study variance estimation in random-effects models. This information can indicate whether reviewers should consider averaging within-study results reported across subgroups and/or sub-scales irrelevant to the main analyses of the given review. Alone by reducing the number of imprecise effect sizes, reviewers can avoid artificially inflating the within-study variance estimation and thereby gain power for their models.

Finally, one further benefit of conducting *a priori* power analysis is that it requires the reviewers to plan for and think carefully about the likely structure of their meta-analysis dataset and the smallest effect size of practical interest. This might naturally yield a deeper understanding of a meta-analysis dataset as well as the topic under review and thus provide more fine-grained and content-relevant interpretations of the final meta-analysis results. However, it is important to note that prospective power analyses should not be compared to the final results since they, *by definition*, do not add any further information to the final results.⁴

7 CONCLUSION

In this article, we have developed common guidelines for conducting power analysis for meta-analysis of dependent effect sizes and introduced the *POMADE* package for this purpose.

Moreover, we have introduced new graphical tools for illustrating power approximations across a range of plausible scenarios. As is apparent from the above illustration, power approximations for meta-analysis will be more informative when based on pilot data from previous syntheses on a similar research topic. Consequently, this makes demands on the entire meta-analysis community to embrace and follow open science and open data⁴³ policies if prospective power analyses should become common practice in meta-analysis.

Since power analysis is exclusively devoted to statistical significance testing and, thus, to some degree, based on arbitrarily selected cutpoints for determining statistical significance and relevant power, we recommend that reviewers are careful in decisions about conducting a meta-analysis based on *a priori* power analyses unless the evidence is decisive. Future research could profitably concentrate on developing methods for conducting precision analysis⁴⁴ for meta-analysis of dependent effect sizes to complement power analysis, i.e., an analysis that aims to approximate the number of studies needed to obtain a certain width of the confidence interval with a given probability. Thereby, reviewers would not need to premise the conduct of meta-analysis on a dichotomized choice of either having or not having enough power to find the smallest effect of practical concern. Nevertheless, we still believe power analysis for meta-analysis of dependent effect sizes provides a means for reviewers to make *a priori* understandings of the given stage and maturity of the literature in point for review, which can provide key guidance for the meta-analysis. With this paper, we hope to have provided the needed guidance for the power approximation to be widely disseminated among applied reviewers and used as common practice in future systematic reviews involving meta-analysis.

8 REFERENCE

1. Vembye MH, Pustejovsky JE, Pigott TD. Power approximations for overall average effects in meta-analysis with dependent effect sizes. 2022. doi:10.31222/osf.io/6tp9y
2. Hedges L V., Pigott TD. The power of statistical tests in meta-analysis. *Psychol Methods*. 2001;6(3):203-217. doi:10.1037/1082-989X.6.3.203
3. Hedges L V., Pigott TD. The power of statistical tests for moderators in meta-analysis. *Psychol Methods*. 2004;9(4):426-445. doi:10.1037/1082-989X.9.4.426

4. Valentine JC, Pigott TD, Rothstein HR. How many studies do you need?: A primer on statistical power for meta-analysis. *J Educ Behav Stat.* 2010;35(2):215-247. doi:10.3102/1076998609346961
5. Jackson D, Turner R. Power analysis for random-effects meta-analysis. *Res Synth Methods.* 2017;8(3):290-302. doi:10.1002/jrsm.1240
6. Hedges L V., Olkin I. *Statistical Methods for Meta-Analysis*. London: Academic Press; 1985.
7. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychol Bull.* 1988;103(1):111-120. doi:10.1037/0033-2909.103.1.111
8. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Res Synth Methods.* 2019;10(2):180-194. doi:10.1002/jrsm.1339
9. Hedges L V., Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods.* 2010;1(1):39-65. doi:10.1002/jrsm.5
10. Van den Noortgate W, López-López J, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. *Behav Res Methods.* 2013;45(2):576-594. doi:10.3758/s13428-012-0261-6
11. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Meta-analysis of multiple outcomes: A multilevel approach. *Behav Res Methods.* 2014;47(4):1274-1294. doi:10.3758/s13428-014-0527-2
12. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: Expanding the range of working models. *Prev Sci.* 2021;23(1):425–438. doi:10.1007/s11121-021-01246-3
13. Fernández-Castilla B, Aloe AM, Declercq L, et al. Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behav Res Methods.* 2020;53(1):702-717. doi:10.3758/s13428-020-01459-4
14. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bull.* 1946;2(6):110-114.

15. Hedges L V. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat.* 1981;6(2):107-128. doi:10.2307/1164588
16. Lakens D, Adolphi FG, Albers CJ, et al. Justify your alpha. *Nat Hum Behav.* 2018;2(3):168-171. doi:10.1038/s41562-018-0311-x
17. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge; 1988. doi:10.4324/9780203771587
18. Hattie J. *Visible Learning – A Synthesis of over 800 Meta-Analysis Relating to Achievement*. New York: Routledge; 2009.
19. Lipsey MW, Puzio K, Yun C, et al. Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *Natl Cent Spec Educ Res.* 2012.
20. Kraft MA. Interpreting effect sizes of education interventions. *Educ Res.* 2020;49(4):241-253. doi:10.3102/0013189X20912798
21. Schäfer T, Schwarz MA. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Front Psychol.* 2019;10(813):1-13. doi:10.3389/fpsyg.2019.00813
22. Ahn S, Ames AJ, Myers ND. A review of meta-analyses in education: Methodological strengths and weaknesses. *Rev Educ Res.* 2012;82(4):436-476. doi:10.3102/0034654312458162
23. White T, Noble D, Senior A, Hamilton KW, Viechtbauer W. metadat: Meta-analysis datasets. R package version 1.0-0. 2021. <https://cran.r-project.org/package=metadat>.
24. Polanin JR, Espelage DL, Grotperter JK, et al. Locating unregistered and unreported data for use in a social science systematic review and meta-analysis. *Syst Rev.* 2020;9(1):116. doi:10.1186/s13643-020-01376-9
25. Hedges L V. Effect sizes in cluster-randomized designs. *J Educ Behav Stat.* 2007;32(4):341-370. doi:10.3102/1076998606298043
26. Hedges L V. Effect sizes in three-level cluster-randomized experiments. *J Educ Behav*

- Stat.* 2011;36(3):346-380. doi:10.3102/1076998610376617
27. Higgins JPT, Eldridge S, Li T. Including variants on randomized trials. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley Online Library; 2019:569-593. doi:10.1002/9781119536604
 28. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol 1. 2nd ed. Sage; 2002.
 29. Taylor JA, Pigott T, Williams R. Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educ Res.* 2021;51(1):72-80. doi:10.3102/0013189X211051319
 30. Hedges L V., Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal.* 2007;29(1):60-87. doi:10.3102/0162373707299706
 31. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *Am J Epidemiol.* 1999;149(9):876-883. doi:10.1093/oxfordjournals.aje.a009904
 32. Verma V, Lee T. An analysis of sampling errors for the demographic and health surveys. *Int Stat Rev.* 1996;64(3):265-294. doi:10.2307/1403786
 33. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev.* 2003;27(1):79-103. doi:10.1177/0193841X02239019
 34. Tanner-Smith EE, Lipsey MW. Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *J Subst Abuse Treat.* 2015;51(1):1-18. doi:10.1016/j.jsat.2014.09.001
 35. Dietrichson J, Bøg M, Filges T, Klint Jørgensen A-M. Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Rev Educ Res.* 2017;87(2):243-282. doi:10.3102/0034654316687036

36. Pigott TD. *Advances in Meta-Analysis*. New York: Springer; 2012.
37. Fraley RC, Vazire S. The N-Pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One*. 2014;9(10):e109019. doi:10.1371/journal.pone.0109019
38. Vembye MH, Weiss F, Bhat BH. The effects of co-teaching and collaborative models of instruction on student achievement: A systematic review and meta-analysis. <https://osf.io/fby7w/>.
39. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Stat Med*. 2012;31(20):2179-2195. doi:10.1002/sim.5356
40. Wickham H. ggplot2: Elegant graphics for data analysis. 2016. <https://cran.r-project.org/web/packages/ggplot2/index.html>.
41. Filges T, Sonne-Schmidt CS, Nielsen BCV. Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Syst Rev*. 2018;14(1):1-107. doi:10.4073/csr.2018.10
42. McKenzie JE, Brennan SE. Synthesizing and presenting findings using other methods. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley Online Library; 2019:321-347.
43. Moreau D, Gamble B. Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychol Methods*. 2020. doi:http://dx.doi.org/10.1037/met0000351
44. Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology*. 2018;29(5):599-603. doi:10.1097/EDE.0000000000000876

SUPPORTING INFORMATION

R codes for replication of all examples provided in this paper are available on the Open Science Framework at <https://bit.ly/3uuinTz>.