Meta-Analysis of Educational RCTs with Follow-up (MERF)

Documentation on the Creation Process and Coding

Emma R. Hart, Drew H. Bailey, Tyler W. Watts

This document details the creation of the MERF dataset. The first portion of the document includes high-level details about the steps involved in creating the dataset. The second portion of the document includes the guidelines used by the team when conducting the inclusion/exclusion process and coding itself.

Hyperlinked documents documenting the work involved in various stages of the process are available upon request (email Emma Hart at erh2169@tc.columbia.edu or Tyler Watts at tww2108@tc.columbia.edu)

Table of Contents

Meta-Analysis Creation Process

Step 1: Identifying Articles for Inclusion

Goal: Identify studies and papers for inclusion and exclusion in the meta-analytic dataset.

1. Drew Bailey and Tyler Watts identified the following meta-analyses to start with: Li et al., 2020; Protzko, 2015; Protzko, 2017; Suggate, 2016; Kraft et al., 2018; Burns et al., 2016; Bailey et al., 2020; Taylor et al., 2017.
2. During the summer of 2020, two RAs (research assistants) did a preliminary review of the articles included in the meta-analyses and marked the ones that they thought did and did not meet inclusion criteria (details on inclusion criteria provided later in this document). They also conducted a follow-up search for interventions that they thought met inclusion criteria.
3. Emma Hart began a more formal inclusion/exclusion process. First, she ensured that all of the studies included in the 8 original meta-analyses were compiled for review. Any odd cases (where studies listed as included in the reference section didn't match that which was listed in the paper itself or vice versa) were noted in the "Important Decisions and Notes" document. Emma erred on the side of including papers.
4. Two RAs searched for any papers that Emma was unable to find from the 8 meta-analyses. Only papers with full PDFs in English were reviewed. Details on papers that could not be located were documented in the "Important Decisions and Notes" document.
5. First, each paper was reviewed by two people to determine whether it met the first two inclusion criteria: 1) that the study is a randomized control trial targeting children or adolescents and, 2) that treatment impacts on social-emotional and/or cognitive outcomes were reported. Emma reviewed each paper and formed an inclusion determination. 1 RA also reviewed each paper and formed a determination (5 RAs were involved). After this independent review, inclusion determinations were compared and discrepancies were documented and reconciled through group discussion. This process is documented here.
6. Next, RAs reviewed each study that met the first two inclusion criteria to determine whether it had usable follow-up data. For studies that were ultimately categorized as meeting inclusion criteria for being an RCT with child/adolescent measures of social-emotional/cognitive outcomes, RAs conducted a follow-up search process to determine whether each study had usable follow-up data (at least 6 months after intervention, following the same students) and to locate all possible follow-up and preceding papers through Google Scholar. This process is documented here.
   a. More details on this process are provided later in this document under step 3 of the inclusion/exclusion process. In brief, at least two RAs conducted an independent Google Scholar follow-up search and made a final determination about whether the paper had viable follow-up. Emma and the team discussed all decisions and resolved discrepancies between student decisions and Google Scholar search findings.
   b. For a handful of papers that were deemed as particularly complicated (i.e., there were several papers that reported findings), Emma, Tyler, and Drew made determinations together about what papers to include for a particular study, with the priority of maximizing papers with treatment impacts on as many follow-up assessments as possible. Details on these decisions are documented here.

7. The next step of the process involved ensuring that the information necessary to calculate at least one effect size for each study was included in the paper.
8. Of note: While studies were generally included or excluded prior to coding, there were some rare cases in which exclusion decisions were made during or after the coding process:
    a. Through the coding process, Emma and the second coder identified some papers that did not meet the aforementioned inclusion criteria. These decisions are documented here. When exclusion determinations were made, the paper was tracked as being excluded for the appropriate reason in consort diagrams (e.g., if during the coding process it was determined that the study should have been excluded at an earlier stage for not actually using random assignment, this study would be labeled as having been excluded due to lack of random assignment in consort diagrams).
    b. During the post-coding data cleaning phase, it was sometimes determined that none of the information provided by the paper could actually be used to calculate effect sizes (e.g., available statistics came from models with interaction terms). If this were the case for all outcomes within a study, then this study was excluded. In this case, the study would be labeled in consort diagrams as having been excluded due to not having the information necessary to calculate effects.

Step 2: Training

Goals: Train new coders on the process of coding through frequent meetings to review new coders' work on their first 6 papers.

1. Emma identified codable papers and several un-codable papers to use for training purposes (a representative mix of different paper difficulties that illustrated the important things to focus on while coding) and coded these. Emma checked in with Tyler to clear up any coding-related confusion.
2. Twice a week throughout the spring of 2021 semester, coders were assigned a mixture of codable and non-codable papers and asked to identify which paper was codable, why the other wasn't, and to code the one that was codable. Emma created a PowerPoint where week-to-week problems with coding were tracked. Tyler joined several meetings throughout the semester to discuss coding issues.
3. Tyler and Emma ran several "pilot" reliability tests along the way to make sure coders were becoming more reliable. Having reached a point at which there were no longer major problems in coding, the team proceeded to the reliability phase.

Step 3: Reliability

Goals: To ensure that coders are reliable before coding independently.

1. Emma randomly selected 10 papers from those she had documented as meeting inclusion criteria to use for reliability checks, and coded these (see here for details). Emma sent these 10 papers to Tyler to look over.

2. New coders independently coded the 10 papers (not consulting anyone during the process).
3. The reliability between Emma and the coders was calculated.
4. Reliability was acceptable between Emma and coder 1 (89.20%) and Emma and coder 2 (84.92%).

Step 4: Ongoing Coding/Reliability Checks

Goals: Coders independently code.

1. 2 coders double coded a small portion of the papers (one coder graduated/got a job soon after we started formal coding). Emma and coder 1 coded the rest of the papers (the majority).
2. 3 RAs were trained on determining discrepancies in Emma and coder 1's codes (discrepancies in notes and page numbers were not discrepancy checked; coder notes were discussed in the discrepancy-check process, as needed).
3. One RA checked for discrepancies between the codes. All discrepancies were documented here.
4. Meetings between the discrepancy checker and both coders were held regularly to reach consensus codes for all papers. This process led to this final dataset (excel version).
5. Confusing cases or questions that prevented the coders from reaching consensus were documented for discussion with Drew and Tyler.

Step 5: Post-Coding Data Cleaning

Goals: Prepare final dataset.

1. Emma/Tyler/Drew met to discuss all documented confusing items that came up during the coding process (see "Papers in Need of Discussion" tab on this tracking sheet). These were mostly cases in which the coders were unclear on whether reported effect sizes were viable due to the use of less-typical estimation techniques. Decisions were made about how to proceed for each case and adjustments to the data were made as appropriate (e.g., removal of effect sizes that should not have been coded in).
2. Emma and an RA assigned a study and paper ID to each study (see here for reference list).
3. 2 RAs identified all typos in the dataset and Emma resolved these.
4. Valence checking
    a. Unfortunately, we failed to code for effect size valence in the primary coding process. Thus, a post-coding process was initiated to identify the valence of each effect size included in the meta-analysis.
    b. For each effect size, Emma and Tyler independently determined whether the effect should be multiplied by 1 or -1 indicating that a higher score on the construct is positive (e.g., math scores) or negative (e.g., depressive symptoms) respectively. With the addition of Drew, the team reviewed all discrepant cases and Drew resolved discrepancies. This process was documented here.

   c. For the effect sizes that the team couldn't reach resolution on, at least 2 RAs (4 involved) reviewed each case at the paper level and gathered evidence for a valence determination. Emma reviewed these cases and made final determinations. Drew and Tyler were consulted for particularly complicated cases.
   d. Valence determinations were multiplied by effect sizes in the cases that the effect size use was an "us-calculated" effect sizes. In the case that a paper-reported effect size was used, however, an additional round of valence coding was required to identify whether the reported effect sizes were already re-valanced (i.e., reporting a reduction in behavioral problems, a positive outcome, as a positive treatment impact), or whether effect sizes were presented as expected given their measure valence (i.e., reporting a reduction in behavioral problems as a negative treatment impact). 3 RAs reviewed all of the reported effect sizes that were suspected to have a high likelihood of valence-related issues (e.g., social-emotional outcomes). 2 RAs reviewed all of the reported effect sizes that were not likely to have valence-related issues (e.g., academic outcomes). Emma reviewed the RAs work and resolved discrepancies to land at a final determination.
   e. We additionally double-checked the valence of outcomes for which the post-test effect size was negative and statistically significant after valence adjustments were made. Given the unlikelihood that treatments produce a negative, statistically significant effect, we hoped that this check would catch errors in valence coding. There were 57 cases of statistically significant, negative post-test effects. 3 RAs, or Emma and 1 RA, reviewed these cases. For each case, the reviewers indicated cases where the valence should be re-coded. Emma reviewed their determinations and resolved discrepancies as needed. 7 cases were identified as needing valence re-coded and were re-coded.
5. Emma consolidated the study-related details for each study across information reported in different papers so that there was one set of study information for each study, consistent across papers (i.e., information on intervention length, treatment targets, etc.). This consolidated information reflected the codes from the paper reporting initial impacts (the paper for which we coded in all of the study details) with updates to codes if there was a discrepancy or contrary information reported in future follow-up papers. An RA double checked this consolidation.
   a. In the case of discrepancies on intervention features that were not a simple yes/no answer (e.g., intervention length), reporting from the most recent paper was favored. In the case of yes/no codes for intervention features, the most generous and inclusive approach was used such that if an earlier or more recent paper indicated "yes" for some intervention feature, and another paper indicated "no" (suggesting no mention that the intervention held this feature), this intervention feature was coded as being present (e.g. if in initial impact paper authors said that parents were involved in the intervention, but in later paper they do not mention this, would still go with a "yes" for parents).
6. Impact estimate/SE calculations. Emma calculated impact estimates according to the documented equations (see here for line by line calculations). 1 RA checked all of these calculations and issues were discussed. Final calculations using the formulas were

computed in Stata. Documentation of these decisions and the process are detailed later in this document (see "Effect Size Calculations").

   a. Note: In the coding process, *f* statistics, *t* statistics, and unstandardized beta coefficients were not coded in (a mistake). In calculating effect sizes, if it was not possible to calculate an effect size for an outcome that was coded in, then Emma returned to the paper and searched for any of these statistics that could be used to calculate an effect size. In one case, this process revealed a construct that could be included.

7. Pre-Stata cleaning that produced new variables
   a. Analytic Sample
      i. A dummy variable was created to indicate whether each outcome that was coded during the original coding process should be included in the meta-analytic sample. A one was indicated for cases that should be included and a zero was indicated for cases that should not be included. An additional variable was created to note the reason why a variable was assigned a zero in these cases. Zeros were indicated when it was not possible to calculate an effect size for the coded outcome.
   b. Construct Groupings
      i. Emma, Tyler, and Drew reviewed the constructs and derived categories that conceptually captured the key constructs present in the data. See tracking of this process here.
      ii. Emma and Tyler independently categorized each construct according to the following options: achievement composite, attendance, general cognition, criminality, educational attainment, externalizing, grades/GPA, internalizing, language and literacy, learning skills, math, mixed composite (i.e., a measure that combined cognitive and social-emotional skills), other academic ability, retention, social-emotional skills, special education designation, and substance use.
      iii. Discrepancies in coding were resolved by Emma, Tyler, and Drew.
   c. Measure Type
      i. Emma, Tyler, and Drew reviewed the constructs and derived categories that captured the type of outcome for each outcome See tracking of this process here.
      ii. Emma and Tyler independently categorized each outcome according to the following options: tasks and tests, behavioral measures, scales and ratings.
      iii. Discrepancies in coding were resolved by Emma, Tyler, and Drew.
   d. Measure Variable Clean-up
      i. Emma made a list of all of the measures that were coded and generated a final "clean" name for the same measures that were called slightly different things, used abbreviations, etc. and created a new clean measure variable that included the appropriate cleaned measure for each outcome.
      ii. 2 RAs independently reviewed all of the original and cleaned outcomes to identify any additional issues for resolution with the ultimate goal that the final clean measure variable could be relied upon in forming analytic groupings for analysis. A new variable was created to capture subscale details that were otherwise included in the original measure variable. An

additional measure was also created to capture other important information like the measure version that included the original measure variable.

   iii. Emma reviewed RA's suggested edits and incorporated edits as appropriate. See tracking of this process here.

e. Construct Variable Clean-up
   i. 2 RAs identified issues in the original construct variables (typos, small inconsistencies) so that cleaned construct name could be used in forming analytic groupings.
   ii. Emma reviewed RA's suggested edits and incorporated edits as appropriate. See tracking of this process here.

f. Sample size and time of test issues
   i. Emma identified cases in which sample size and/or time of test was not coded for a particular study or outcome within study. Sample size and time of test were only coded by the coders if this information was explicitly provided by the study authors (inferences about these variables were not coded in; see coding protocol entries for these variables for more details on what was coded in).
   ii. For these cases, 2 RAs independently returned to the original papers and gathered any available information related to the missing information and suggested their "best estimate" of sample size and time of test. Emma then did the same review, taking into account the suggestions by the two RAs and made final estimations of sample size and time of test. These estimations took into account information from other papers on the same study, and any available information provided in the paper. Notes on these decisions were tracked here.
   iii. Emma consulted with Tyler on cases that were particularly complicated and in cases when there appeared to be no information to estimate off of. These cases were left as missing in the data. Otherwise, this estimated information was used to replace missing codes for sample size and time of test.
   iv. Note that there were a few cases in which conducting these reviews revealed that the original codes should be updated. Updates were made accordingly.

g. Demographics
   i. Race and ethnicity demographics were originally qualitatively coded (i.e., quotes from the paper were coded). At least 2 RAs independently reviewed race and ethnicity information from all included studies to create indicators for the percentage of participants of a particular race and ethnicity in each racial/ethnic group (e.g., % white, % black, % hispanic, etc). Discrepancies were reviewed by Emma. This process was tracked here.

h. Stata integration. Final dataset was imported to Stata and variables were cleaned for analysis. See Stata data cleaning syntax for data cleaning steps.
   i. Note that at the beginning of this integration, when small problems were found in the coding throughout this whole process, corrections were made

in the original dataset. Eventually, we transitioned to writing syntax in the cleaning file to make necessary idiosyncratic updates to the data.

8. Post-Stata integration cleaning and variable generation
    a. After the initial integration described above, some additional steps occurred to clean up some of the "qualitative" measures in the dataset and to generate variables that were not initially coded, or required a "second look" due to high missingness.
        i. Duration and Intensity
            1. Duration and intensity of intervention were both qualitative codes (i.e., quotes from paper were coded).
            2. At least 2 RAs independently reviewed existing codes for how many treatments were given, how many hours of treatment were given, and over how many weeks, months, or years treatments were distributed.
            3. RAs attempted to standardize across studies to produce one variable for duration and one for intensity:
                a. Duration - over how many months the treatments were administered.
                b. Intensity - how many hours children spent in treatment.
            4. Intensity required more inference because it relied on the number of sessions that were supposed to have occurred, the time of each session, and over what period of time sessions took place. Creating the duration variable required less inference and is a more solid variable for use.
            5. At least two RAs worked on each case and Emma resolved discrepancies. This process was tracked here.
        ii. Year of Intervention
            1. At least two RAs independently searched within papers from each study for the year in which the intervention was administered.
            2. When this information was not found in any of the papers, RAs searched for relevant information from grant-tracking sources.
                a. Since these are longitudinal studies, grants often lasted many years, in these cases, year of intervention was taken as the first year of funding.
            3. At least two RAs worked on each study and Emma resolved discrepancies. This process was tracked here.
        iii. Country
            1. At least two RAs returned to each of the studies to determine the country in which the study took place. Emma resolved discrepancies. This process was tracked here.
        iv. Time in school
            1. Time in school was a code with a generous amount of missingness because codes were only assigned if no inference was required to assign a code during the original coding process. To reduce missingness, Emma and an RA returned to papers with missingness to determine if a reasonable inference could be made

about whether the intervention added time (determinations were made independently, discrepancies were discussed). These determinations were tracked here. This updated time in school variable is demarcated with “coarse” in the dataset.

v. Baseline age

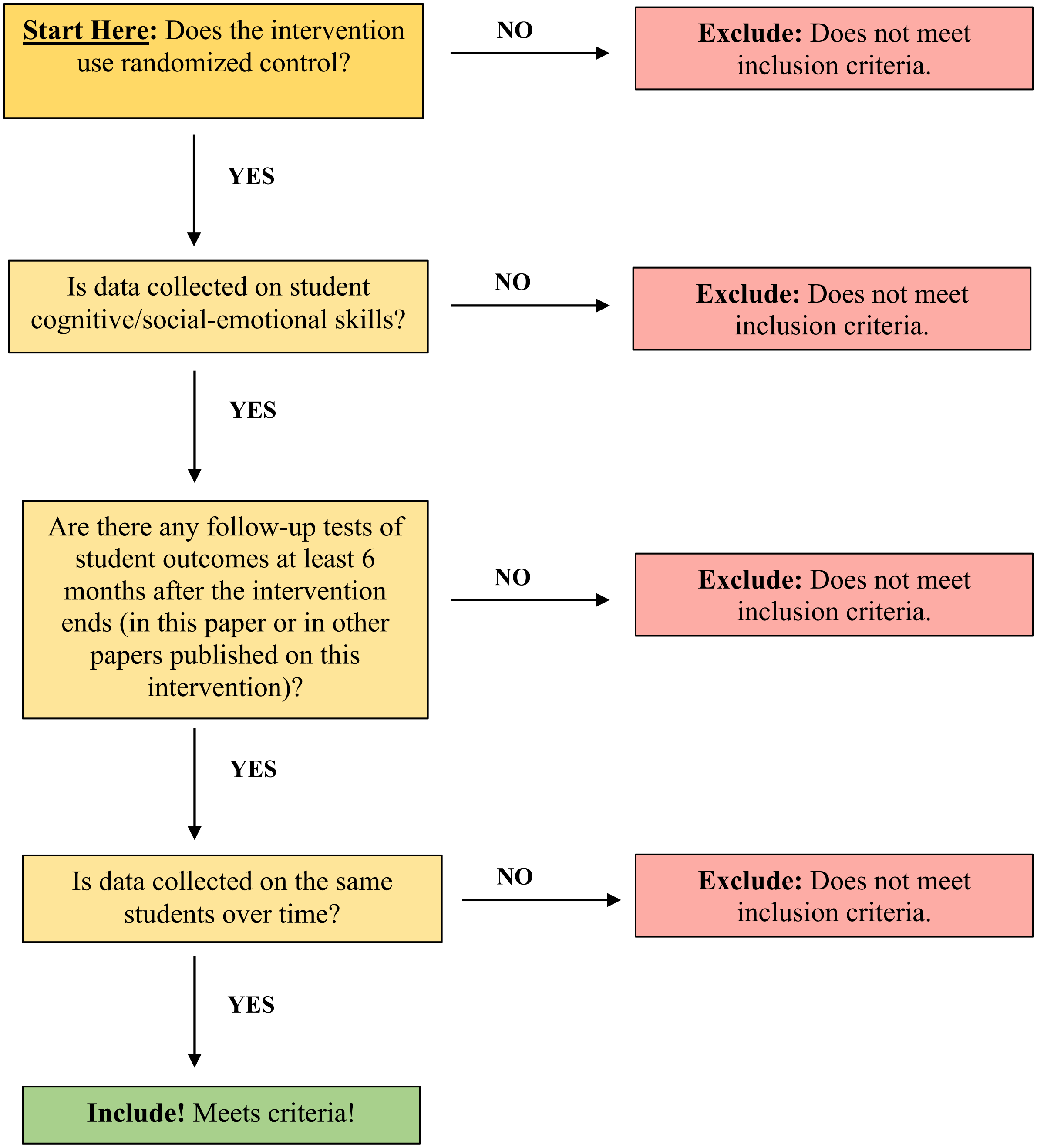
1. Baseline age of treatment and intervention participants at baseline was extracted from all papers as part of the initial coding process. An RA converted baseline age to baseline age in months for all interventions.
2. In cases where age was provided by grade level, the RA determined average age for a given grade level. For example, if a paper indicates students were in seventh grade, they were assigned a baseline age of 144 months (12 years).
3. In cases where the average age varied between treatment and control group, the RA determined a weighted average age for children in the intervention.

Step 6: Generate Final Number of Included and Excluded Interventions

Goal: Create a count of the number of interventions documented in the eight original meta-analyses

1. An RA returned to a list of all papers (whether included or excluded in the MERF sample) included in the eight original meta-analyses
    a. The RA documented intervention and sample details in order to match papers to studies to produce a unique count of the number of studies represented across the papers.
    b. Each study was assigned an ID (multiple papers on the same study would share the same ID)
        i. Papers that reported impacts on a subsample of the larger sample reported in another paper were counted as the same study
        ii. The same intervention implemented with a different sample(s) was considered separately
        iii. 3 other RAs reviewed 46 papers that seemed like they may be overlapping to check these decisions
        iv. The final count was 298 unique interventions from the eight original meta-analyses
2. Emma and the RA took this information, in combination with tracked details of which studies were included/excluded at various stages to arrive at the final number of interventions and papers – see Stata cleaning code for details

Inclusion Criteria Flow-Chart


Start Here: Does the intervention use randomized control?
NO
Exclude: Does not meet inclusion criteria.
YES
Is data collected on student cognitive/social-emotional skills?
NO
Exclude: Does not meet inclusion criteria.
YES
Are there any follow-up tests of student outcomes at least 6 months after the intervention ends (in this paper or in other papers published on this intervention)?
NO
Exclude: Does not meet inclusion criteria.
YES
Is data collected on the same students over time?
NO
Exclude: Does not meet inclusion criteria.
YES
Include! Meets criteria!

Detailed Inclusion/Exclusion Instructions (Steps 1 - 4)

These instructions were created to guide the team in steps 1 through 4 of the inclusion/exclusion process. The team referred to these guidelines throughout the process. Decisions related to inclusion/exclusion decision making were documented here. Note "behavioral" is used interchangeably with "social-emotional" in detailing the sample inclusion/exclusion procedure.

General tips

1. Start by reading the abstract to get a sense of what the intervention is.
2. Then, use control+F to search for "random" to try to discern whether the study is an RCT. Look specifically in the methods section. Read through the control/treatment assignment process and discern whether true randomization was used. If it's not an RCT, you can exclude and move on to the next paper at this point.
3. If it is an RCT, move on to read the rest of the methods section to get a sense of what outcomes were assessed and whether these count. If cognitive/behavioral outcomes were coded, then mark to "include." If not, then exclude and move on to the next paper.
4. Always proceed in this order: 1) Is it an RCT?; 2) Does it have behavioral/cognitive outcomes?

Step 1: Check for whether the paper is an RCT

5. The paper must be a randomized control trial.
    a. RCT: involves randomly assigning students/schools/blocks to intervention and control groups.
    b. Things that are okay (still considered an RCT):
        i. Lotteries.
        ii. Matching/stratification/clustering/blocking *prior to* randomization.
        iii. Author indicates that the study is "quasi-experimental", but methods section description indicates RCT.
        iv. Sometimes there will be multiple experimental and/or control groups, some of which were created through random assignment, and some were not. As long as 2 groups were created through random assignment, the study is codable (make a note of what groups were not created through random assignment and, thus, should not be coded).
        v. Random assignments can take place at multiple levels (e.g., schools, classrooms, students). Any level, as long as it undergoes random assignment, is okay.
    c. Things to look out for:
        i. "Randomly selected" – be sure not to confuse random selection with random assignment. To be an RCT, random assignment must occur. Random selection of students/schools may also occur as a way of recruiting participants, but is not a substitute for random assignment.
            1. Note: if the term "random selection" is used in referring to a randomized process by which treatment and control groups were formed, then this may be okay as long as there is equal probability that participants were assigned to either group.

   ii. "Random effects"- random effects is a statistical term that is not related to random assignment.

Step 2: Check for whether the paper has behavioral/cognitive outcomes

1. Second inclusion criteria that must be met in order to proceed is that behavioral and/or cognitive outcomes must be reported. We have taken a very liberal approach to these definitions.
   a. What counts as behavioral: self-regulation, externalizing/internalizing behaviors, pro-social/anti-social behaviors, anxiety/depression symptoms, drug/alcohol use, school suspensions/arrests, social competence, coping strategies, personality traits, etc.
   b. What counts as cognitive: vocabulary, rhyming ability, EF, IQ, GPA, educational attainment, etc.
   c. What doesn't count: sensorimotor/motor development, teacher/parent/other adult outcomes (must be *child or adolescent* outcomes).

Step 3: Check for whether the paper has follow-up at least 6 months following the end of the intervention on the same students

1. To be included, there must be a follow-up test of child/adolescent cognitive or behavioral outcomes on the same students at least 6 months following the end of the intervention.
2. Things to watch out for regarding follow-up:
   a. When some participants in the intervention receive additional intervention. This may look like: additional "booster" sessions provided after the intervention ends, participants allowed to "opt-in" to additional intervention following post-test, random selection of participants provided additional intervention, some participants receiving additional intervention on the basis of their performance following initial intervention. This is all okay (and can be conceptualized as an extension of intervention), but follow-up must occur at least 6 months after the time when the last participant finishes last intervention. Think of this as extended intervention.
   b. When control group participants receive additional intervention: For example, if some schools continue to provide students (both TX and CTRL) in the original sample with additional intervention following the end of the treatment, then this is a break of randomization and we would code: "Not an RCT."
3. Things to watch out for regarding following the same students:
   a. School level data that is not matched with the students who actually got the intervention: Sometimes outcome data will be presented for the whole school or grade (e.g., test-scores for all 4th graders). If the follow-up paper does not present data for the specific cohort of students who received intervention (and instead presents data for a new cohort of students or the whole school, some students in which were not at all part of the study), then this follow-up cannot be used.
   b. Follow-up paper presents data from multiple samples/studies in a combined way: This is okay, just make note of it.

Follow-up search process (To come to a determination for step 3):

1. Determine whether the current paper includes initial impacts (i.e., does this paper present initial impacts or is this a follow-up paper). We want to be sure to include all papers that present results for the intervention within the current sample. Indicate whether this is an initial impact paper or not. If there are lot of papers that report and re-report results from various time points, make note and we can discuss as a group which ones to include. For the Google Scholar search, we'll use the paper that feels like the best initial impacts paper.
2. If the paper does not present initial impacts, search within the paper to see if there is any indication/reference to previous papers published on the same intervention with the same sample (we are not interested in the same intervention with a different sample). If there are any previous papers referenced that you think could be included, input their references.
3. If you found references, search for the full paper, download it, and upload to the Google Drive in the appropriate folder within the "New PDFS" "Initial Impact Papers" folders and indicate that you have uploaded it. Save the paper with the reference as the name.
4. If there are multiple initial impacts, determine whether they present different information or estimate impacts in different ways. If different outcomes and/or estimates are provided, then both papers should be coded and follow-up searched. If not, identify which paper should be used for follow-up searching. If deciding between a shorter academic paper or a longer policy paper, opt for the academic paper. If one paper has significantly more citations and is an academic paper, opt for this.
5. Determine whether there is follow-up on the same students at least 6 months following intervention within the current paper and indicate this.
6. Regardless of whether there is or isn't follow-up within the paper, proceed to check within the master spreadsheet to see if there are any papers within any of our meta-analyses that present follow-up data to the current intervention with the current sample.
    a. Use "control+F" to search within each meta-analysis tab for the name of the intervention (if there is one). If there is no intervention name, search for a keyword associated with the intervention.
    b. Then search again through all tabs using the first author's last name.
    c. Finally, search once more through all tabs using the second author's last name.
    d. For any papers that seem like they might be follow-ups to the current intervention/sample, pull up the associated PDF and search to see whether there is a match. Be sure to carefully look for whether the exact sample lines up for any given intervention (there may be cases where there are multiple papers on the same intervention, but different samples/studies within our meta-analyses).
    e. If there are any matches (i.e., there is a follow-up paper to the current paper), then list the reference and associated meta-analysis you found it in.
7. Regardless of whether you find follow-up in any of the meta-analyses, proceed to conduct a Google Scholar search to identify if there are any follow-up papers on the current intervention/sample.
    a. Locate the current paper on Google Scholar, then navigate to the "cited by" link. Note: if this paper was not the initial impact paper, then use the reference for the

initial impacts paper instead of the current paper when conducting the following Google Scholar search.

b. Check "search within cited articles" to limit the following search.

c. Then use the advanced search (upper left corner drop down, three horizontal bars) to search:
    - i. "follow-up," "followup," "longitudinal" "long-term" "long term" "long-run "long run" using "with at least one of the words" advanced search.
    - ii. If there is a name for the intervention itself, also include this as a "with the exact phrase" search term on advanced search.

d. Carefully sort through the first 12 pages (10 papers per page) for papers that look like they could be follow-up to the current paper.

e. If there is a paper that looks like it may be follow-up, read the abstract. If it continues to look promising, download it and identify whether it is, indeed, follow-up.
    - i. Double check that this reference wasn't already identified as a paper within our collection of papers from all of the meta-analyses (if there are any, these would be listed under column G- "Meta-Analyses Reference").
    - ii. As was the case with the meta-analysis search, be sure to carefully look for whether the exact sample lines up for any given intervention (there may be cases where there are multiple papers on the same intervention, but different samples/studies within our meta-analyses).
    - iii. Use discretion! Look for: author names, intervention name, terms like "follow-up" "longitudinal" in title when deciding what papers to click on and read more into.
    - iv. You can use "control+F" to help highlight key words related to the intervention/author names, etc.
    - v. Please don't restrict the search by year at all.

f. If you identify any follow-ups, then list the full reference and upload the PDF to the appropriate folder on Google Drive (New Papers > Follow-up Papers > Respective Meta-Analysis). Name these PDFs the full reference name.

g. Make a final decision about whether the current paper has follow-up at least 6 months following the end of the interventions on the same students. Choose from the drop-down menu to indicate your decision.

h. Importantly: to be a viable follow-up, the paper must meet all other inclusion criteria (see steps 1, 2, and 3). Pay close attention to the inclusion criteria outlined in step 3 regarding what counts as a follow-up that is at least 6 months post-intervention on child outcomes.

i. Use columns E, F, (whether there is follow-up within the paper itself, whether there is a follow-up paper within our sample of papers from all meta-analyses) and whether you found papers on Google Scholar to determine what drop-down to choose in column L.

j. For any confusing cases, make notes in the "Step 3 Note" instead of within the other columns (stick to the drop-down options or "N/A" in these columns).

Step 4: Check if the data is reported in a codable way

1. Check to make sure that behavioral/cognitive outcomes are actually reported in codable ways.
    a. Sometimes behavioral/cognitive outcomes will be included in the methods section but will not be reported in the results section (used as covariates not DVs, just not reported). If data on behavioral/cognitive outcomes is not provided, then we should exclude as "no cognitive/behavioral outcomes."
    b. To meet this inclusion criteria, at least one of the following must be reported for at least one behavioral/cognitive outcome: means and SDs after intervention, difference score between control/experimental group outcome, effect size for intervention and/or any sort of treatment impact estimate, p value (comparing control and experimental outcome).
    c. Note that a substantial portion of this process happened throughout the coding process (coders identified reasons why a paper could not be included) and during the data cleaning process (when Emma met with Tyler and Drew to determine whether we could use impact estimates produced by confusing analytic processes, as well as when Emma attempted to calculate impact estimates for all coded outcomes).

Detailed Coding Instructions

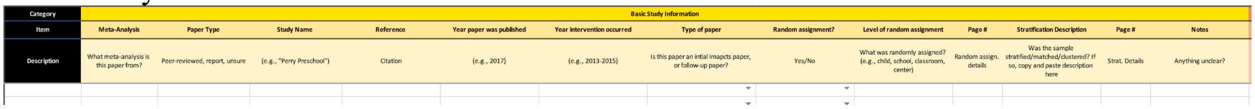
These instructions were created to guide the team in coding. All coding-related decisions were documented here for future reference and consistency.

General things to keep in mind:

1. Start by skimming the introduction, methods, and results sections of a paper. Try to gather a sense of the focus of the intervention, how the intervention was administered, timing of pre-/post-/and follow-up tests, and what child/adolescent outcomes were measured.
2. As you are coding, it is very important to keep the article inclusion criteria in mind. All of the studies that you are coding should meet inclusion criteria. In the case that you think the study you are coding does not meet inclusion criteria, contact Emma with your concern (this should happen very rarely if at all).
3. For all of the following coding steps, please do the following:
    a. Code very literally- focus on what the paper says and avoid making inferences beyond what is explicitly reported. Take what the authors write at face value.
    b. Make notes in the "notes" sections regarding any confusing or special cases as you are coding. Always not relevant page number(s). Make "general notes" that apply to most of the data in the top row for the study, put specific notes relevant to a particular row in that row.
    c. As you code, highlight the information you are coding (numbers in table, text describing relevant information) in the PDF. You should also highlight the information used to fill in the "notes" section.
    d. For any columns where you do not have the information to fill in the cells, enter "NA" (there should always be *something* entered, whether information or "NA" for every cell).
    e. Whenever possible, copy and paste descriptive information from the study itself to the template (with quotation marks and page number). Avoid using your own words to describe the intervention/any other pertinent information. Use quotation marks when copying/pasting from papers.
    f. If there is a dropdown, choose from the options provided.
    g. Triple check all inputted numbers. Check both the value of the number and whether it has been coded in the appropriate cell.
    h. Double check your spelling. Also double check that copy/paste formatting, page numbers, etc. are correct. We want the coding to be understandable/readable to people who were not involved in the project directly!
    i. When inputting multiple copy/pasted treatment descriptions, put spaces between these for readability ("return + option" keys on a mac).
    j. Always spell out abbreviations (this is very important in the measures section).
    k. Make sure to always input the most accurate numbers. Code in details from the text (versus table) when in doubt.
    l. To avoid copy/paste disasters, do not use formulas for adding cells together (e.g., sample size).

Coding process:

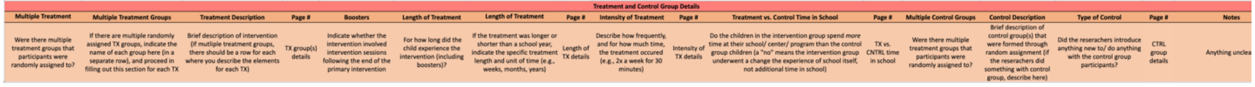
1. The following process should be followed for initial impact papers. For follow-up papers only need to code: Basic Study Information, Data Details, Data Collection, Treatment vs. Control Group Data, and Impact Estimates. Please read through the methods of follow-up papers carefully, however, and make note of any codes for which there are discrepancies between what is presented in the initial paper and in the follow-up. Note these discrepancies both in the specific, relevant cell and at the end of the coding sheet in the "Things to clear up at coding meetings" column.
2. Basic Study Information



Category	Basic Study Information												
Item	Meta-Analysis	Paper Type	Study Name	Reference	Year paper was published	Year intervention occurred	Type of paper	Random assignment?	Level of random assignment	Page #	Stratification Description	Page #	Notes
Description	What meta-analysis is this paper from?	Peer-reviewed, report, unsure	(e.g., "Perry Preschool")	Citation	(e.g., 2017)	(e.g., 2013-2015)	Is this paper an intial imapcts paper, or follow-up paper?	Yes/No	What was randomly assigned? (e.g., child, school, classroom, center)	Random assign. details	Was the sample stratified/matched/clustered? If so, copy and paste description here	Strat. Details	Anything unclear?

   a. Enter what meta-analysis the paper was drawn from- This will be indicated when Emma assigns the papers for you to code.
   b. Enter paper type- Whether the paper was peer-reviewed, a policy report, something else or whether you are unsure.
   c. Enter study name- Enter the name of the intervention as labelled by the researcher. If intervention does not have a name, Emma will have provided one that you can use.
   d. Enter reference- Enter the full citation.
   e. Year paper was published- Enter in the year the paper was published.
   f. Year intervention was conducted- If provided, enter in the year(s) that the intervention actually took place.
   g. Type of paper- Indicate using the drop-down menu whether the study was an initial impact study or a follow-up paper. Generally, if the study includes an original pre-/post-test and follow-up soon after, we consider this to be an initial impacts paper. If the study is reporting primarily on follow-up measures, then we would consider it a "follow-up" paper.
   h. Random assignment- All studies must utilize random assignment to be included in this meta-analysis. All of the papers have been screened and should use random assignment, but if you are concerned that a paper did not use random assignment, contact Emma and note it in the coding sheet.
   i. Level of random assignment- Studies can be randomly assigned at different levels (e.g., child, classroom, school). This information should be provided in the methods section. Read carefully as sometimes this is confusing (when in doubt, make a note).
   j. Page number- Indicate the page number where the study discusses random assignment procedure.
   k. Stratification description- Some studies will use matching, stratification, or clustering prior to randomization. If this was the case, copy and paste a description of how this was conducted in this cell (include how the groups were then randomized). Code "No" if there is no stratification.
   l. Page number- Indicate the page number where the study discusses stratification/matching/clustering procedures, if applicable.

3. Treatment and Control Group Details



Treatment and Control Group Details																
Multiple Treatment	Multiple Treatment Groups	Treatment Description	Page #	Boosters	Length of Treatment	Length of Treatment	Page #	Intensity of Treatment	Page #	Treatment vs. Control Time in School	Page #	Multiple Control Groups	Control Description	Type of Control	Page #	Notes
Were there multiple treatment groups that participants were randomly assigned to?	If there are multiple randomly assigned TX groups, indicate the name of each group here (in a separate row), and proceed in filling out this section for each TX	Brief description of intervention (If mutliple treatment groups, there should be a row for each where you describe the elements for each TX)	TX group(s) details	Indicate whether the intervention involved intervention sessions following the end of the primary intervention	For how long did the child experience the intervention (including boosters)?	If the treatment was longer or shorter than a school year, indicate the specific treatment length and unit of time (e.g., weeks, months, years)	Length of TX details	Describe how frequently, and for how much time, the treatment occured (e.g., 2x a week for 30 minutes)	Intensity of TX details	Do the children in the intervention group spend *more* time at their school/ center/ program than the control group children (a "no" means the intervention group underwent a change the experience of school itself, not additional time in school)	TX vs. CNTRL time in school	Were there multiple treatment groups that participants were randomly assigned to?	Brief description of control group(s) that were formed through random assignment (if the reserachers did something with control group, describe here)	Did the reserachers introduce anything new to/ do anything with the control group participants?	CTRL group details	Anything unclea

a. Multiple Treatment Groups- Indicate using the dropdown whether the study included multiple treatment groups that were formed through random assignment. Only code “yes” if more than one treatment group was created through random assignment. If there are multiple treatment groups, but these were not formed through random assignment, then make a note in the notes section about these non-randomized groups.

b. Treatment Group Names- If there are multiple treatment groups, indicate here the name of each and proceed in filling out the remaining cells for this section for each group. Only list treatment groups that were formed through random assignment. If there are not multiple treatment groups code “NA”.

c. Treatment Description- Copy and paste a description of the intervention(s). There will often be a description towards the end of the introduction or in the methods section. If there are multiple treatment groups formed through random assignment, include descriptions for each. This should capture the nature of the intervention. If the intervention involves teacher training, include information about this. If the intervention involves boosters, indicate information about this too.

d. Page Number- Indicate the page number(s) from which you copied information regarding the treatment group(s).

e. Boosters- Indicate using the dropdown whether the intervention included booster sessions (additional intervention sessions that occur after the end of the primary intervention). If there are booster sessions, code “yes.” If there are not booster sessions, code “no.” If there are any unique details about who in the sample received boosters (e.g., if not all TX participants received boosters), be sure to note this in the notes section and for the internal validity code.

    1. Defining a booster: Often will look like an initial treatment followed by a post-test assessment, then additional intervention (booster). Think about: is there a definitive end of intervention, followed by additional treatment at a later point in time?

        1. If the additional treatment feels like a continuation of or substantial component of the initial treatment, then typically do not code as booster.

f. Length of Treatment- Indicate using the dropdown whether the intervention was less than one school year, one school year, or was longer than one school year. Note- if the treatment is a change in curriculum, this should be coded as “one school year” unless explicitly indicated as otherwise.

    1. Boosters- Include boosters in your calculation of how long the intervention lasted (e.g., if primary intervention was one school year, but boosters occurred in the following school year, you would code in “more than one school year”).
    2. If the authors specify the months of intervention and they are approximately a year (e.g., 7 to 9 months), can code as “one school

year" but make a note about the exact length in the length of treatment description code.

g. Length of Treatment Description- If you indicated that the treatment was less than or more than one school year in length, describe exactly how long it was. Be sure to note the unit of time (e.g., weeks, months, years). If the treatment was one school year long, just mark "NA". In the case that the intervention is less than one school year or more than one school year, opt to code in the most specific description of the length of treatment (copy and paste information on the number of weeks/months/years the intervention lasted).
    1. Boosters- make note of boosters in this descriptive code.

h. Page Number- Indicate the page number where length of description is noted.

i. Intensity of Treatment- Describe the intensity of the intervention such as how often the treatment was administered and for how much time (e.g., 1 hour a week for 3 months). If there are multiple aspects of an intervention and information on intensity for many of these, copy/paste in all of the necessary information for each intervention component. If intervention fidelity information and "intended" intensity details are provided, just input intended intensity details. If there is no information about intensity, and ONLY information about fidelity, then code this in, but make a note in the "notes" section that it was fidelity, not pre-intervention intensity.

j. Page Number- Indicate the page number where the intensity of treatment was noted.

k. Treatment vs. Control Time in School- Using the dropdown menu, indicate "yes" if children in the intervention group spend more time in a school/center/program than the control group.
    1. Indicating "yes" means that the children experienced additional/more time in school/center/program than the control group as part of the intervention. For example, code "yes" if children in the TX group spend additional time in the school/center/etc. than CTRL group (e.g., TX goes to pre-k, CTRL does not; after-school intervention for TX, not CTRL).
    2. Code "no" if the intervention group does not spend *additional* time in school/center/program. Coding "no" means that the intervention group experienced some *change* in their school/center/program experience/environment that did not lead to additional time spent in formal instruction. For example, code "no" if children in the TX group and CTRL group spend the same amount of time in the school/center/etc. (e.g., TX gets new curriculum, CTRL gets "business-as-usual" curriculum).
    3. Only use "NA" if you absolutely cannot indicate "yes" or "no" because there is zero information provided about the control group.
    4. If there is a case where an intervention overwhelmingly took place in school and involved no additional time in school, but there were a couple after school sessions that child/parent attended, code this as a "no" but make an explicit note that there were those additional sessions that happened outside of school.

5. If there are multiple control groups formed through random assignment and your code would be different between these control groups, then make a note in the notes section about which control group you focused on for this code, and what the code would be for the other control group.

l. Page Number- Indicate page number where the authors described whether the intervention involved more time in school or a change in the school experience itself.

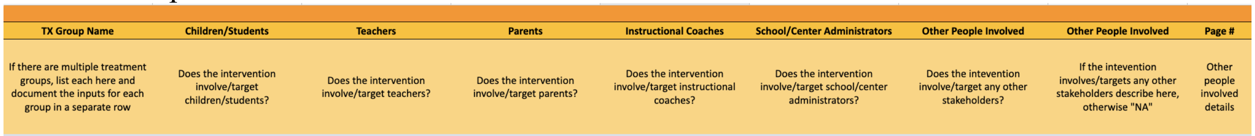
m. Multiple Control Groups- Indicate using the dropdown whether the study included multiple control groups. Indicate using the dropdown whether the study included multiple control groups that were formed through random assignment. Only code "yes" if more than one control group was created through random assignment. If there are multiple control groups, but these were not formed through random assignment, then make a note in the notes section about these non-randomized groups.

n. Control Description- Copy and paste in a description of the control group(s). If the control group(s) is just, "business as usual" describe what exactly business as usual is (e.g., control group receives no preschool, control group receives XYZ standard school curriculum). If the control group(s) receives any sort of intervention (an "active control") then be sure to describe this (e.g., control group does XYZ while the intervention receives intervention). If there is not a great quote to use to describe the control group, you can write in your own words what is happening. In other words, do not leave "NA" unless you really have no idea what happened in the control. If there are multiple control groups formed through random assignment, then be sure to copy and paste information about each.

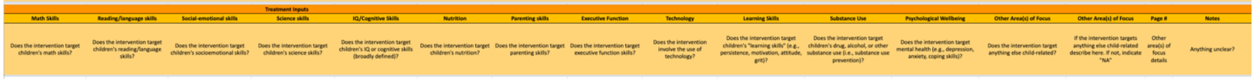
o. Type of Control- Using the dropdown to indicate whether the researchers introduced anything new within the control group (indicate "yes") or whether the researchers did not (indicate "no"). If the control group was simply "business as usual" where the researchers introduced nothing to control group, then indicate "no." Note- if the experimenters DO introduce anything to the control group, even if small, then this is a "yes" (e.g., placebo, access to some services, etc.). Simply participating in the research process (e.g., child assessments, classroom observations) does NOT qualify as an active control group. If there are multiple control groups formed through random assignment and in one control group researchers introduce something and in the other group researchers do not, make a note in the notes section about which control group you coded for and what the code would be for the other control group.

p. Page Number- Indicate the page number(s) for the information you gathered on the control group characteristics.

4. Treatment Inputs



TX Group Name	Children/Students	Teachers	Parents	Instructional Coaches	School/Center Administrators	Other People Involved	Other People Involved	Page #
If there are multiple treatment groups, list each here and document the inputs for each group in a separate row	Does the intervention involve/target children/students?	Does the intervention involve/target teachers?	Does the intervention involve/target parents?	Does the intervention involve/target instructional coaches?	Does the intervention involve/target school/center administrators?	Does the intevention involve/target any other stakeholders?	If the intevention involves/targets any other stakeholders describe here, otherwise "NA"	Other people involved details



Treatment Inputs

Math Skills	Reading/language skills	Social-emotional skills	Science skills	IQ/Cognitive Skills	Nutrition	Parenting skills	Executive Function	Technology	Learning Skills	Substance Use	Psychological Wellbeing	Other Area(s) of Focus	Other Area(s) of Focus	Page #	Notes
Does the intervention target children's math skills?	Does the intervention target children's reading/language skills?	Does the intervention target children's socioemotional skills?	Does the intervention target children's science skills?	Does the intervention target children's IQ or cognitive skills (broadly defined)?	Does the intervention target children's nutrition?	Does the intervention target parenting skills?	Does the intervention target executive function skills?	Does the intervention involve the use of technology?	Does the intervention target children's "learning skills" (e.g., persistence, motivation, attitude, grit)?	Does the intervention target children's drug, alcohol, or other substance use (i.e., substance use prevention)?	Does the intervention target mental health (e.g., depression, anxiety, coping skills)?	Does the intervention target anything else child-related?	If the intervention targets anything else child-related describe here. If not, indicate "NA"	Other area(s) of focus details	Anything unclear?

a. If there are multiple treatment groups, fill out this section using a different row for each intervention group. If this is the case, indicate the treatment group name in the first column of this section. Follow the same order of treatment group names as was used in the previous section. If there are not multiple treatment groups, indicate "NA" for this code.

b. For this section, it's very important to only consider the description of the intervention itself and not the outcomes that are measured. It's often the case that the authors of a paper will say that they are interested in multiple outcomes, but the intervention itself may not directly target some/all of these specific outcomes. This section is really about the characteristics of the intervention. Study outcomes are coded elsewhere.

    1. Another way to think about this is to consider that researchers almost always have theories of change about how their intervention is going to affect various outcomes. These theoretical standings are experimental questions in and of themselves. For example, researchers may expect a nutrition intervention to affect cognitive outcomes, but this is an empirical question up for investigation. Meanwhile, the intervention itself is just a nutrition intervention, not a cognitive one. Think carefully about what the intervention actually targets, not the outcomes that are up for empirical investigation.

c. For each of the types of people listed (i.e., children/students, teachers, parents, instructional coaches, school/center administrators) indicate using the dropdown menu whether the intervention involved/targeted this individual.

    1. Key questions for consideration:
        1. Who does the intervention interact with and/or involve?
        2. Does this individual directly receive instruction/intervention efforts that targets their skills/behaviors/characteristics? Do researchers attempt to change some adults' skills/behaviors/characteristics in order to change child/student skills/behaviors/characteristics?
        3. Are the researchers trying to *change something about this individual* through their efforts or are they using this individual to facilitate the intervention?
    2. Examples of direct targets vs. facilitators:
        1. Direct targets of interventions (code "yes"): teachers who are trained to provide a particular curriculum (regardless of the intensity of the curriculum implementation/whether the authors explicitly state that they aim to change teachers behaviors); parents who are provided 1:1 coaching, materials pertinent to parenting behaviors and/or information about child development; children who are directly provided a treatment; school district coaches who are trained to provide a new form of coaching to teachers.
        2. Facilitators of intervention (code "no"): parents who are involved in providing their children vitamins on weekends when teachers cannot do so as part of a vitamin intervention; teachers/coaches hired by the research team to provide intervention content; teachers at a researcher-affiliated child care center trained to provide

curriculum; parents/community members who are involved in sessions of the intervention for the sole purpose of providing child the opportunity to learn new skills (no materials/training provided to the adult themselves).

d. Other People Involved: If there were other people involved, describe these people.
e. Page Number: Indicate the page number where the other people involved are described/detailed, if applicable.
f. For each of the skills/component types (i.e., math, reading/language, socioemotional/behavioral, science, IQ/cognitive skills, nutrition, parenting, executive function, technology, learning skills, psychological wellbeing, substance use prevention) indicate using the dropdown menu whether the intervention targeted each skill/component.
    1. Socioemotional and behavioral skills are combined. The definition of a socioemotional skill is broad. The following would be included: internalizing/externalizing behaviors, problem solving, self-confidence, communication skills, stress management, managing emotions in positive ways, relating to peers, peer pressure, friendships, self-esteem, ethical dilemmas, conflict resolution, self-awareness, empathy. Note, communication and/or listening skills, within the context of a socioemotional intervention, should not be coded as "language/reading."
    2. IQ/cognitive skills are combined, too. This does not include "cognitive behavioral therapy" or "cognition" defined within the guise of a socioemotional/psychological wellbeing intervention.
    3. Parenting Skills- If an intervention describes targeting parenting skills, code "yes." If it is also indicated that these parenting skills are geared towards some other targeted skill (e.g., math, reading, etc.), code "yes" for these treatment inputs as well. In other words, an intervention that describes targeting parenting skills may very well simultaneously target other child skills of relevance.
    4. Executive function is defined as working memory, inhibitory control, and set/attention shifting.
    5. Technology- Only code "yes" if technology is a major component of intervention delivery.
    6. Psychological wellbeing has to do with interventions aimed at improving depression or anxiety. If an intervention is clearly aimed at improving psychological outcomes (it is explicitly stated in the title, abstract, and/or "current study" section), but this aim is not clearly stated in the methods/intervention description along with socioemotional skill targets, then code "no" for "for psychological wellbeing," and "yes" "socioemotional skills" (if socioemotional skills are explicitly listed, that is). In these cases, please make an explicit note that psychological wellbeing was addressed as a distal target in the title/abstract/etc., but that it was not directly targeted.
    7. Substance use prevention includes drug, alcohol, and smoking prevention. Similar to process for psychological wellbeing, if an intervention is clearly aimed to prevent substance use (as stated in the

title, abstract, current study section), but this is not explicitly mentioned among socioemotional skill targets in the methods section, then code in "no" for "substance use prevention" and make an explicit note in the notes section about how this is listed as a more distal outcome of interest.

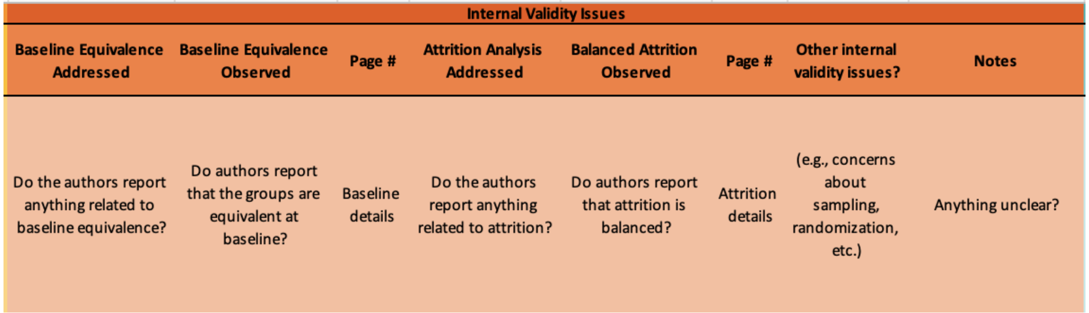
8. Learning skills involves skills such as persistence, motivation, students' attitude, and grit.

g. Other Area(s) of Focus- If there were other child-related things that the intervention targeted, indicate these (e.g., beliefs about gender relations, possible selves, critical thinking skills, a specific risky behavior, academic achievement (if given as "academic achievement" and not more specific outcome e.g., reading or math)

h. Page Number- Indicate the page number where other intervention areas of focus were detailed, if applicable.

i. For this section as with all others, if you're ever unsure of whether you properly coded a particular skill, make a note about your uncertainty.

5. Internal Validity Issues



Internal Validity Issues							
Baseline Equivalence Addressed	Baseline Equivalence Observed	Page #	Attrition Analysis Addressed	Balanced Attrition Observed	Page #	Other internal validity issues?	Notes
Do the authors report anything related to baseline equivalence?	Do authors report that the groups are equivalent at baseline?	Baseline details	Do the authors report anything related to attrition?	Do authors report that attrition is balanced?	Attrition details	(e.g., concerns about sampling, randomization, etc.)	Anything unclear?

a. General Note- For all these codes, rely on the conclusions that the researchers make. If they say that there were differences in balance and/or attrition, but ultimately conclude and state that there were not significant differences, then conclude that there was balance. As such, we are trying to understand whether the experimenters perceived whether each of these codes was a threat to internal validity. In these cases where a lack of equivalence was noted but the authors conclude that it is insignificant, make an explicit note that there were differences between the groups, but that the authors concluded there was overall balance.

b. Baseline Equivalence Addressed- Indicate using the dropdown menu whether there was any discussion/mention of equivalence between groups at baseline (may be in a table or described in the text). If there was a discussion of baseline equivalence or a table including baseline information about participants (for a table this would be means for measures), code "yes". If not, code "no."

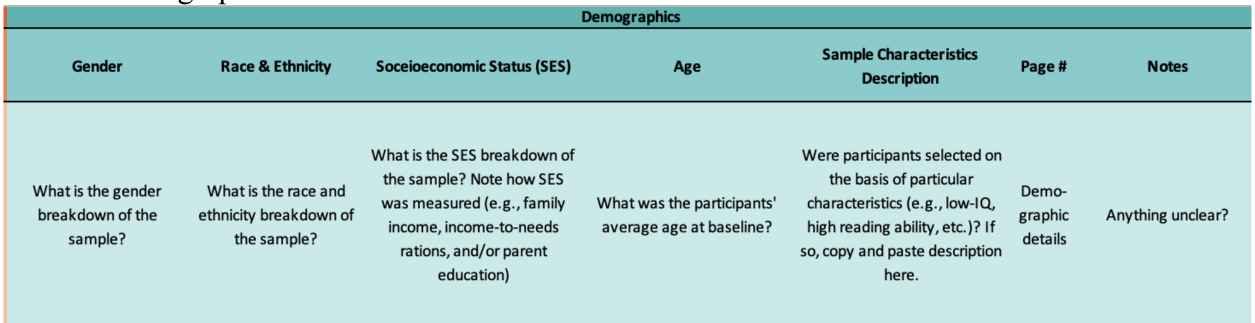
c. Page Number- Indicate the page(s) where you found information on baseline data/ equivalence.

d. Attrition Analysis Addressed- Indicate using the dropdown whether the authors report any information about attrition. This usually looks like a discussion about who dropped out of the study, *and differences in the people who remained in the control and treatment groups.* If either of these issues are mentioned, code "yes." If there is no information or data related to people dropping out, code "no." If missing data information is provided, but this information is not explicitly about attrition (e.g., missing data on baseline measures), then code "no" because missing data could be due to attrition or something else.

e. Other internal validity issues? - Here you can indicate any other pertinent internal validity issues.
    1. For example, this is where you would indicate if there were a difference in how much time children spent in the intervention (e.g., some children received more time in the treatment than others or some children started the intervention earlier than other children).
    2. There may also be interventions where some children in the TX group receive more intervention on the basis for their performance. This is okay (as long as CTRL and TX group assignment is maintained) but make a note.
    3. There are also some cases where students are added midway through the year, this is also okay, but make a note. This involves students who were not initially involved in the randomization process (either at the individual level or at their school level) being added to the study later. (e.g., they are new to an intervention school).
    4. Another great thing to code here is if there is anything notable about booster sessions (i.e., booster sessions were randomly assigned after the primary intervention ended) or if data is only reported for a subset of the sample (e.g., only non-booster participants). Note that assignment to additional boosters, if not following the original randomization, must be random. If this is not the case, tell Emma.
    5. If the authors describe something about who they included in their analytical sample based on attrition, this would be a good place to note it (e.g., all analyses only included the children who had data for all assessments).

f. Page Number- Indicate the page number(s) where you found information about attrition and attrition balance.

6. Demographics

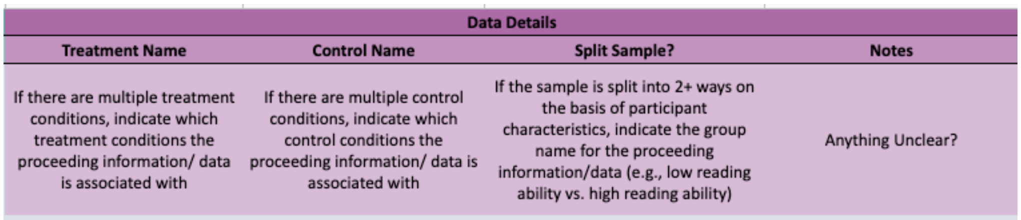


Demographics						
Gender	Race & Ethnicity	Socioeconomic Status (SES)	Age	Sample Characteristics Description	Page #	Notes
What is the gender breakdown of the sample?	What is the race and ethnicity breakdown of the sample?	What is the SES breakdown of the sample? Note how SES was measured (e.g., family income, income-to-needs rations, and/or parent education)	What was the participants' average age at baseline?	Were participants selected on the basis of particular characteristics (e.g., low-IQ, high reading ability, etc.)? If so, copy and paste description here.	Demo-graphic details	Anything unclear?

a. For this whole section, we only need to code in information about the child/family, not teachers or other people involved in the intervention. If there is an option to code in more or less specific sample characteristics, opt for the more specific ones.
    1. Generally, if SDs are provided for a demographic measure, input these.

b. Gender- Indicate the gender breakdown of the sample using descriptive information from the participants section (copy and paste). If there is no information provided in the text, information may be provided in a table that you can use instead. If this information is reported separately for the control and treatment groups, code in information for both (and note which information is for which group).
c. Race and ethnicity- Similar to gender, copy and paste in the description of the sample's race and ethnicity breakdown. Language spoken does not count for this code. If there is no information provided descriptively, look for information provided in a descriptive table that you can code in instead. If this information is reported separately for the control and treatment groups, code in information for both (and note which information is associated with which group).
d. Socioeconomic status (SES)- Copy and paste in a qualitative description of the sample's socioeconomic status breakdown (this may be reported as free and reduced lunch, income, income to needs ratios, parental education, parent occupational prestige, etc.). If this information is not provided descriptively, it may be provided in a table. If so, input data from the table, being sure to indicate how SES was measured (e.g., income vs. income-to-needs, maternal education vs. average parental education, etc.). If this information is reported separately for the control and treatment groups, code in information for both (and note which information is for which group).
e. Age at baseline- Average age of the sample at baseline in whatever unit is presented in the text (copy and paste if possible). If grade, but not age, is provided, code this.
f. Age at assessments- Average age of the sample at post-test and follow-up assessments in whatever unit is presented in the text (copy and paste if possible). If grade, but not age, is provided, code this.
g. Sample characteristics description- Did the study recruit/and or screen participants so that all participants in the study had a specific characteristics or skill levels? For example, do all the children in the sample have high IQ, low working memory, poor reading skills, etc.? If yes, copy and paste a description of details about how the participants were screened/their particular characteristics. This code would not include factors such as Head Start enrollment, belonging to a low-income family, or being a certain age. This category only involves characteristics specific to the child/adolescent that were explicitly used for screening and targeting the intervention. If there is an option to code a more lengthy/detailed description or a briefer one that communicates the main sample characteristics criteria, opt for the briefer description.
h. Page number- Indicate the page(s) where you found the demographic data.

7. Data Details



Data Details			
Treatment Name	Control Name	Split Sample?	Notes
If there are multiple treatment conditions, indicate which treatment conditions the proceeding information/ data is associated with	If there are multiple control conditions, indicate which control conditions the proceeding information/ data is associated with	If the sample is split into 2+ ways on the basis of participant characteristics, indicate the group name for the proceeding information/data (e.g., low reading ability vs. high reading ability)	Anything Unclear?

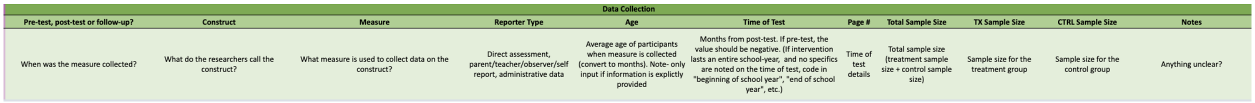
a. Treatment Name- If there are multiple treatment groups, code in each name in correspondence with the data entered in subsequent sections so we know what TX group corresponds with this data. Importantly, only code data for treatment groups that were formed through random assignment. If there are two treatment groups and one was not created through random assignment, do not code this data. Be sure the multiple treatment groups were indicated in the “Treatment and Control Group Details” section and that treatment inputs were coded for each treatment. If there are not multiple intervention groups indicate “NA”. Be sure that the name you use to define the treatment can be easily understood by someone who has not read the paper, and that that the chosen treatment names distinguish between the multiple groups. If the paper provides names for the multiple treatment groups, use these.

b. Control Name- If there are multiple treatment groups, code in the name for each of these. In subsequent rows, input information that is in accordance with the listed control group name. Only input data for control groups that were formed through random assignment. Be sure that the multiple control groups were indicated in the “Treatment and Control Group Details” section and notes were made pertaining to differences in control groups. If there is only one control group, indicate “NA”. If there are multiple treatment and control groups, input data for each combination (e.g., TX 1 and CTRL 1, TX 2 and CTRL 1, TX1 and CTRL 2, TX 2 and CTRL 2).

c. Split sample? - If there are 2+ groups of participants (e.g., low birthweight < 2000g vs. low birthweight > 200g, children at risk for reading delay vs. typically developing children, etc.) indicate which group the proceeding data is associated with. If the sample is not split into multiple samples indicate “NA”. Note- sometimes studies will report data for the whole sample as well as a specific subsample. Whenever the whole sample is reported, just code this in. In other words, only code in “split samples” or results for different sample groups if this is the only option.

d. In the case that there are two experimental groups (without a clear distinction of one being “typical”, in other words, both are true treatment groups) functioning as treatment and control, use the “treatment name” and “control name” columns to specify what treatment groups are being coded under treatment vs. control.

8. Data Collection



Data Collection										
Pre-test, post-test or follow-up?	Construct	Measure	Reporter Type	Age	Time of Test	Page #	Total Sample Size	TX Sample Size	CTRL Sample Size	Notes
When was the measure collected?	What do the researchers call the construct?	What measure is used to collect data on the construct?	Direct assessment, parent/teacher/observer/self report, administrative data	Average age of participants when measure is collected (convert to months). Note- only input if information is explictly provided	Months from post-test. If pre-test, the value should be negative. (If intervention lasts an entire school-year, and no specifics are noted on the time of test, code in "beginning of school year", "end of school year", etc.)	Time of test details	Total sample size (treatment sample size + control sample size)	Sample size for the treatment group	Sample size for the control group	Anything unclear?

a. General organizational guidelines for this section:

1. Enter constructs in the order they are reported in the paper.
2. Enter pre-test/post-test/follow-up together for each construct.
3. Enter all of the constructs for one treatment/control condition (if there are multiple) and/or part of the sample (if split) before the next.
4. We are just interested in pre-test, post-test, follow-up outcomes. If there are multiple outcomes reported that were collected while the intervention was still ongoing, we are not so interested in these (i.e., do not code these).
5. When to/when not to code in latent variables/composites:

        1. If many measures are combined into a composite and means/SDs are presented for these composites, code this data in INSTEAD of non-composite outcomes for each measure.
        2. If many measures are combined into a composite and means/SDs are NOT presented for these composites, we cannot code this data in and must use the non-composite outcomes.
            a. IMPORTANT- In this case, if impact estimates are just reported for the composite (and M/SD are **not** reported for the composite), do not code this in, but be sure to make a note of this (can copy and paste relevant information). We should not mix composite and non-composite outcomes.

b. Construct- Using data tables and/the results section, identify what child/adolescent cognitive/behavioral outcomes were collected before the intervention, after the intervention, and at follow-ups at least 6 months following the end of the intervention. We are only including studies with follow-ups of at least 6 months. Sometimes follow-up tests will be reported in additional papers. If it seems that this inclusion criteria are not met, contact Emma. List all the "constructs" that were measured using the language that the authors used (e.g., vocabulary, reading, behavior problems, etc.). Sometimes an overarching "construct" will be reported in data tables, under which specific constructs will be reported. In this case, code the construct name that is associated with the reported data (e.g., if under the label "language," "vocabulary" is reported with associated values, code "vocabulary" as the construct, not "language"). Alternatively, sometimes composite constructs will be created where several measures that are combined and reported as a larger construct (e.g., data is reported for "language" which includes several measures for constructs like "vocabulary", "speech", etc.). Report these composite constructs if they are provided (see note above about when to and when not to code composite constructs). Studies will often collect data on non-student related information such as classroom quality and teacher outcomes. Do not code this data. Only code child/adolescent data on behavioral/cognitive constructs (note that teacher-reported student/teacher relationship quality counts as a behavioral construct). Sometimes studies will also include data on outcomes such as child motor development that are not cognitive or behavioral. Do not code this data either. If it seems that there are no appropriate outcomes to code, contact Emma. We are embracing a broad definition of "behavioral"- in most cases you should code all child outcomes unless an outcome clearly not behavioral and/or cognitive.

    1. In trying to decide what to name a construct, think about what skill/characteristic/domain the researchers say they are measuring. Often this is reported in the measures section. Think about what you would say to a friend when describing what the researchers are measuring. This is the kind of label we want.
    2. When there is not a description of what the task measures, then you can code in the name of the measure (e.g., a measure is called "figure design" and there is no indication of what this really means other than

how the task was performed, so code that the construct name is “figure design”).

3. If “reverse codes” are provided (e.g., % of students who are smokers and who are not smokers), opt to code in the more “rare” or novel behavior of interest (e.g., code in % who are smokers over % who are not smokers).

c. Measure- Enter what measure was used to collect data on each construct. This should be the name of the actual scale/test/questionnaire that was used to measure the construct of interest. Using information from the methods section, identify what measure was used to collect information about the construct of interest. This should be the name of the specific measure used (e.g., Peabody Picture Vocabulary Test III, Woodcock-Johnson Test of Achievement-Revised, Bayley Scales of Infant Development, etc.). If several measures were used for a construct, be sure to list them all. If a subtest of a larger measure was used, code in the name of this subtest.

1. If a measure was created for the purposes of the current study and there is not a measure name, then indicate that the measure was created for the study, and any other information that seems relevant (e.g., if parts of the measure were derived from a standardized scale). No need to copy and paste a full description of how the measure was collected or what it was.
2. If subtests were used, be sure to indicate these.

d. Reporter Type- Indicate how each measure was collected. Was it a direct measure of children’s skills (e.g., a vocabulary test, achievement test, child performance on some task)? Was this information gathered through someone’s report (e.g., child report, teacher report, parent report)? Was it administrative data that comes from an outside source that collected this data for their own purposes which the researchers are now using for the purposes of their analyses (e.g., report card, curriculum-based assessment, district records, college attendance, special education, holding students back a grade, state achievement tests)? Or was it an observation by a researcher (e.g., observer report)? This information should be explicitly documented in the methods section (for the case of child, teacher, parent, observer report), or may be implied in the type of measure (an achievement test is a direct assessment, though the authors may not explicitly note this).

1. Self-report vs. direct assessment:
    1. Self-report- Child reports about their skills, behaviors, etc. and their perceptions are at play (e.g., attitudes towards alcohol use, frequency of drinking, Beck Anxiety Inventory).
    2. Direct assessment- A measure that directly measures children’s skills, behaviors, knowledge, etc. (e.g., spelling test, EF task, cognitive measure).

e. Pre-test, post-test, or follow-up- Use the dropdown measure to indicate when data was collected for each construct/measure. Identify whether the construct was measured before the intervention (pre-test), directly after the intervention (post-test), or at follow-up (at least 6 months after the intervention). Code in a row for each testing time point. You will likely have multiple rows for each construct (e.g., if a study collected child vocabulary at pre-test, post-test, and follow-up, there should be 3 rows, with each row corresponding to each test time point), and sometimes there will be multiple follow-up time points. In the case that the intervention is a multi-year program with multiple "pre-" and "post-" tests, code in the initial pre-test (pre-intervention) and the first post-test after the intervention actually ended. We do not need to code intermediary measures (e.g., a measure after pre-test and before post-test that occurred at some point during the intervention). Generally, this information is provided in the methods section. It is best to organize your coding by construct (i.e., organize all the pre-test, post-test, and follow-up test rows for one construct together and then go on to the next construct).

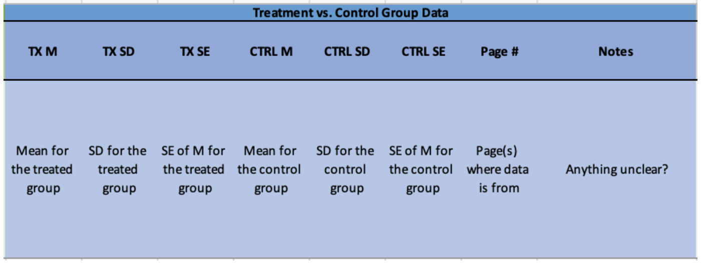
f. Time of test- Document the time when the test occurred in months and in reference to the post-test (post-test should be coded as "0" months). For example, if the pre-test occurred 6 months before the post-test, code "-6" for the pre-test, and "0" for the time of the post-test. If a follow-up test occurred 12 months after the post-test, code "12" for the follow-up test. Often this information is hidden within the methods section. Ideally a pre-test happens before the intervention, a post-test happens after an intervention ends, and a follow-up happens at least 6 months after the post-test. In the case that the pre-test did not actually occur before the intervention started, or that the post-test did not happen after the study ended (but rather slightly before), make note of this in the notes section.

    1. This can be a tricky column to code because often this information is reported somewhat ambiguously. In the case that the intervention lasts a school year, and there are no more specific details on testing, code in "beginning of school year" for pre-test and "end of school year" for post-test.
    2. In general, you should not make any assumptions about timing. Instead, code more descriptively the timing of the test and be sure to do so in a way that is interpretable.
    3. In the case that there is some information provided on exact time of testing for some measures and not for others, do not make assumptions and instead input "end of school year". If there are multiple follow-up assessments, specify which school year this refers to (e.g., "end of 1st grade" "end of 2nd grade"). When this is the case, indicate in the notes section roughly how many months after the post-test the follow-up occurred so that whoever is reading in the future knows (e.g., "end of 1st grade was approximately 1 year after post-test, end of second grade, then, was approximately 2 years after post-test").
    4. If there is no information provided about pre-test timing, but there is a note that the pre-test happened directly prior to/around the time of the start of the intervention, then you can calculate the time of pre-test using the length of the intervention. If the length of the intervention or this statement is not provided, code descriptively the start of the intervention. When something is unclear, make a note!

5. If qualitative information is provided for one code (e.g., pre-test), make a note about any information provided about post-test so that the “0” has meaning.

g. Page number- Indicate the page where you found details on the timing of the tests.
h. Total Sample Size- Indicate the total sample size for each of the measures at each of the time points. Calculate the total sample size by adding the treatment and control sample sizes. Sometimes specific information on sample size at each time point/for each measure is not provided. In this case, code in the most specific information you can find, and make note if specific information was not provided. Be sure not to use formulas when calculating this (to avoid future copy/paste disasters) and instead use a calculator.
i. TX Sample Size- Indicate the sample size for the treatment group for each of the measures at each time point. Like with total sample size, code in the most specific information the authors report.
j. CTRL Sample Size- Indicate the sample size for the control group for each of the measures at each of the time points, coding in the most specific information you can.

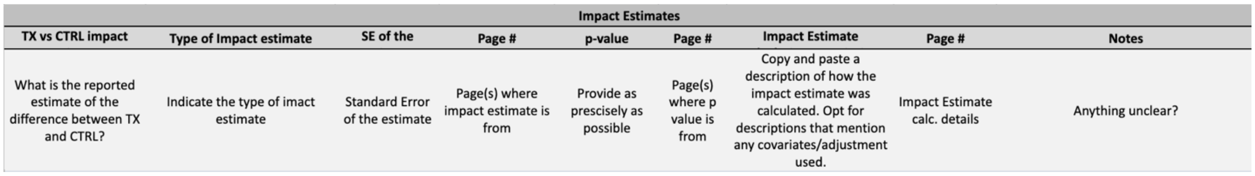
9. Treatment vs. Control Group Data



Treatment vs. Control Group Data							
TX M	TX SD	TX SE	CTRL M	CTRL SD	CTRL SE	Page #	Notes
Mean for the treated group	SD for the treated group	SE of M for the treated group	Mean for the control group	SD for the control group	SE of M for the control group	Page(s) where data is from	Anything unclear?

a. TX M/TX SD/TX SE- Enter treatment mean, and standard deviation or standard error (typically standard deviations are reported, you will almost never see BOTH SE and SD reported) for each construct/measure at the appropriate time point, and for the treatment group/split sample that was indicated in the “Data Details” section. <u>Be sure to double check the numbers you input twice to check for errors.</u>
b. CNTRL M/CNTRL SD/CNTRL SE- Same as above. If there are multiple control groups (created through random assignment), enter data for the control group that was indicated in the “Data Details” section.
c. Page number- Indicate the page(s) from which you pulled the data.
d. Note- We can only code data that reports outcomes for the intervention and control groups separately (if data is presented for the sample in an aggregated way, then this is not useful for our purposes).
e. Also note that we can’t code in change scores.
f. If proportions/percentiles of students meeting some cut-off on a given measure are presented in addition to means/standard deviations, there is no need to code these (make a note that they’re available). If this is the only data provided, then code it.
g. When different mean/standard deviations are reported in more than one place/differently, choose to code in the data from the main text.

10. Impact Estimates



Impact Estimates								
TX vs CTRL impact	Type of Impact estimate	SE of the	Page #	p-value	Page #	Impact Estimate	Page #	Notes
What is the reported estimate of the difference between TX and CTRL?	Indicate the type of imact estimate	Standard Error of the estimate	Page(s) where impact estimate is from	Provide as presciesly as possible	Page(s) where p value is from	Copy and paste a description of how the impact estimate was calculated. Opt for descriptions that mention any covariates/adjustment used.	Impact Estimate calc. details	Anything unclear?

a. TX vs. CTRL impact- Indicate the treatment impacts for post-test/follow-up time points.

1. Often this data will be provided in the data tables. However, in many cases when it is provided, the authors will not indicate the type of impact estimate. Be sure to check the notes in the bottom of the tables for information about estimates.
2. Sometimes it is hard to determine which impact estimate to use as data tables/text can have many. We want to use the impact estimate that is simply estimating the effects of the intervention on its own, **as we are NOT interested in mediation/moderation estimates**. In cases where there is the option to use an impact estimate that contains adjustment/controls for attrition, use this option (if there is a standardized beta coefficient or a Cohen's d value, for example, go with the standardized beta coefficient – this often incorporates in control variables that ultimately provide a better estimate of the effects of the intervention). Importantly, look out for whether any of these "controls" were from time-points other than the pre-test. If so, this may be indication that a moderation model was at play (where the researchers are estimating the effect of the intervention based on post-test performance, for example).
3. Note that sometimes "main" impact estimates incorporate interactions (moderation). We must be very careful to make note of these. If there is any chance that an impact estimate was calculated in a model that included an interaction term, then you can still code in the "main" estimate but should make an explicit note that this may have involved an interaction. Some signs that this may be the case: a table may present main and interaction effects, suggesting that these were estimated in the same model; a line in the analysis section that suggests that interaction terms were estimated in the same model. Be sure to make note of either of these things.
4. Code all outcomes for the full sample. If additional/different outcomes are provided for a split sample, code these in too (using the split sample column). Make note of split sample data that is presented, but that you opt not to code in when possible.
5. When in doubt, make a note about your uncertainty and which value you recorded. Sometimes impact estimates will not be provided. If this is the case, code "NA".
6. To emphasize- If mediation/moderation is involved in a model, we cannot use this (i.e., mechanisms that are explaining the relation between IV and DV which obscure the extent to which we can observe the simple effect of the intervention (IV) on the outcome). If you're unsure (as may be the case in a situation like that described in #3), then make an explicit note.

    7. If there is a binary/categorical outcome (e.g., enrollment in college) you may see these impact estimates in various forms, and can code them in- odds ratio, log odds coefficient, predicted change in probability. **Be sure to note which type of impact estimate was used in the "type" column.**

b. Type of Impact Estimate- Indicate the type of impact estimate. This information may be provided in the text, table, or text under the table. In many cases it is not provided. If the impact estimate you input does not match one of the following categories, then choose "other" and copy/paste information about the impact estimate in the notes section. If you choose other, it is helpful to indicate in your note whether the outcome of interest was continuous (e.g., test score performance) or categorical (e.g., graduated or did not graduate).

    1. Potential impact estimates in the drop down to choose from: Standardized Beta Coefficient, Cohen's D, Hedge's G, Tukey, Mean Difference Score, Odds Ratio, Log Odds Coefficient, or Predicted Change in Probability

c. SE of the estimate- Standard error of the estimate is often not provided, but in the case that it is provided, code this.

d. Page Number- Indicate the page numbers(s) where the impact estimates were reported.

e. *p* value- *p* values are often reported in data tables, but sometimes also in the results section. Code in the most accurate value. For example, if the authors provide an exact number in the text, but in table they provide a threshold value (e.g., "<. 05"), code in the exact number. If they only provide the threshold (e.g., "< .05"), code in the threshold. If they only provide *p* value estimates in the data table through "*" indications, code in "NS" (not significant) for values not marked with a star(s) but note that you assumed "NS" based on the stars in the notes section.

f. Page Number- Indicate the page number(s) where the *p* values were reported.

g. Impact Estimate Calculation Description- Copy and paste information provided by the authors about how the impact estimates were calculated. This may be provided at the bottom of data tables and/or in the statistical approach section of the methods, or in the results. No need to copy/paste this information in if there is no impact estimate provided.

h. Page number- Indicate the page(s) where the information about the impact estimate calculation information was found.

i. Note- If confidence intervals are presented for an estimate and *p* values/SEs are not provided, then make note of these in the notes section.

Effect Size and Standard Error Calculations

Emma worked closely with Drew and Tyler to determine effect sizes. An additional RA (Ph.D. level) checked all calculations. The figure pasted at the end of this section details the formulas used to calculate effect sizes based on the available, reported results.

The ultimate goal of the effect size calculation process was to identify one effect size for each coded outcome. While the standard protocol was to calculate effect sizes according to the formula detailed in the manuscript, or to use a viable author-reported effect sizes when these were available, there were many cases in which additional decision criteria were used to determine which effect size to use, or to calculate the effect size.

Adjustments for Effect Sizes Calculated using SEs, *t* statistics, and *f* statistics

In cases when standard deviations were not provided and viable reported effect sizes were not available, reported standard errors, *t* statistics, and *f* statistics were used to derive effect sizes (see figure below). In the case that any of these statistics were used to calculate effect sizes for a given outcome, the first author returned to the original paper to check whether the statistic appeared to have been calculated in a model with the inclusion of the pre-test control. In these cases, an adjustment was made when calculating the effect size given the likelihood that standard errors may have been reduced as a result of the inclusion of this control, thus biasing the effect sizes calculated using these estimates. In the cases that this control was included, the standard errors calculated from the available statistics were divided by the square root of 1 minus R^2 (assuming an R^2 between pre- and post-test measures of .50) in the effect size calculation process (using the formulas outlined in Figure S1). Thus, adjustments were made by dividing standard errors by .87 in these cases to ensure that the standard errors were not inaccurately small in the effect size calculation process.

Importantly, in many cases, these adjusted effect sizes were then used to estimate an accompanying standard error for use in our models (i.e., to weight more heavily studies with greater precision). To ensure that these estimated standard errors used were not inaccurately large in our meta-analytic models due to the .87 effect size adjustment, estimated standard errors were multiplied by .87.

Calculating Effect Sizes using *p* values

In the case that no alternative statistics were available to use in calculating effect sizes, the last resort was to estimate an effect size using reported *p* values. If precise *p* values were reported (e.g., ".002"), then *t* statistics were calculated from these *p* values and the formulas detailed in Figure S1 were then used to convert *t* values to effect sizes.

If relatively precise *p* values were reported (e.g., "< .05"), we found the smallest difference between means for each measure within a given study and assumed this *p* value was the largest possible associated *p* value (e.g., .05). For these cases, we then converted the *p* value to a *t* value using the "invt" function in Stata, assuming a two-tailed test (i.e., we divided the *p* value by 2). Next, we calculated the effect size from this *t* value (as described above), and recovered a SD from this calculated effect size. For the cases in which the same measure was available within a study but did not qualify as having the smallest difference between means, the recovered SD was then used to calculate these effect sizes.

In the case that *p*-values were only reported to be statistically non-significant, with no precise value associated, we found the largest difference between means for each measure within a given study and assumed that this *p*-value was .10. We then converted the *p* value to a *t* value using the same procedure described above for relatively precise *p* values and recovered a SD that was then used to calculate the effect size for the other cases within a study that had smaller differences between the means for each measure.

In the cases where treatment and control group means were not provided for an outcome, and the treatment impact was noted to be statistically non-significant, *p* values were assumed to be .10 and *t* statistics were calculated from these *p* values. Because means were not available, an alternate equation was used to convert *t* values to effect sizes (see next section).

For all of these aforementioned processes, we made the .87 pre-test covariate adjustment when it appeared that the *p* value came from a model including a pre-test control (see previous section for more details).

Calculating ES from *f* and *t* statistics when Means were Not Reported.

When treatment and control group means were not provided, and effect sizes were estimated using *t* statistics (only in the case of *p* value conversions) or *f* statistics (in the case of one study), the following equations were used (Higgins et al., 2023):

Choosing between Using Author-Reported or Calculated Effect Sizes

In cases when both author-reported effect sizes and calculated effect sizes were available for an outcome, we opted for consistency in using either reported or calculated effect sizes for all outcomes in a paper, if possible. For example, if a particular paper reported means and standard deviations for 20 outcomes that we used to calculate effect sizes, and also reported viable effect size estimates for 10 of those outcomes, we opted to use our calculated effect sizes for all outcomes because these were available consistently.

In cases when within-paper consistency was not an issue, we then checked for differences in reported effect sizes and calculated effect sizes. If the difference in estimates was less than 1 *SE* for all effect sizes within a paper, we opted to use the reported effect size because this estimate was, presumably, more precise if authors incorporated controls for baseline covariates or other relevant covariates in their estimations. If the difference in estimates was greater than 1 SE for any outcome within a paper, the first author checked whether issues related to valence (see next section) may have driven differences in the final reported and calculated effects. The first author also determined whether there were any issues (e.g., longitudinal effects were modeled linearly in a growth curve model, interaction terms were included in the model, etc.) in the estimation strategies used to calculate the author-reported effect size that the coders missed in the coding process (i.e., only "viable" effect sizes should have been coded). The first author reviewed decisions with Tyler to arrive at final determinations about whether to use the reported or calculated effect sizes. So long as there were no estimation issues with the reported effect sizes, these were used with the assumption that such effects should be more precise due to the inclusion of covariates when modeling the estimates.

Calculating Standard Errors for Odds Ratios, Log Odds Ratios, Proportions, and Percentages

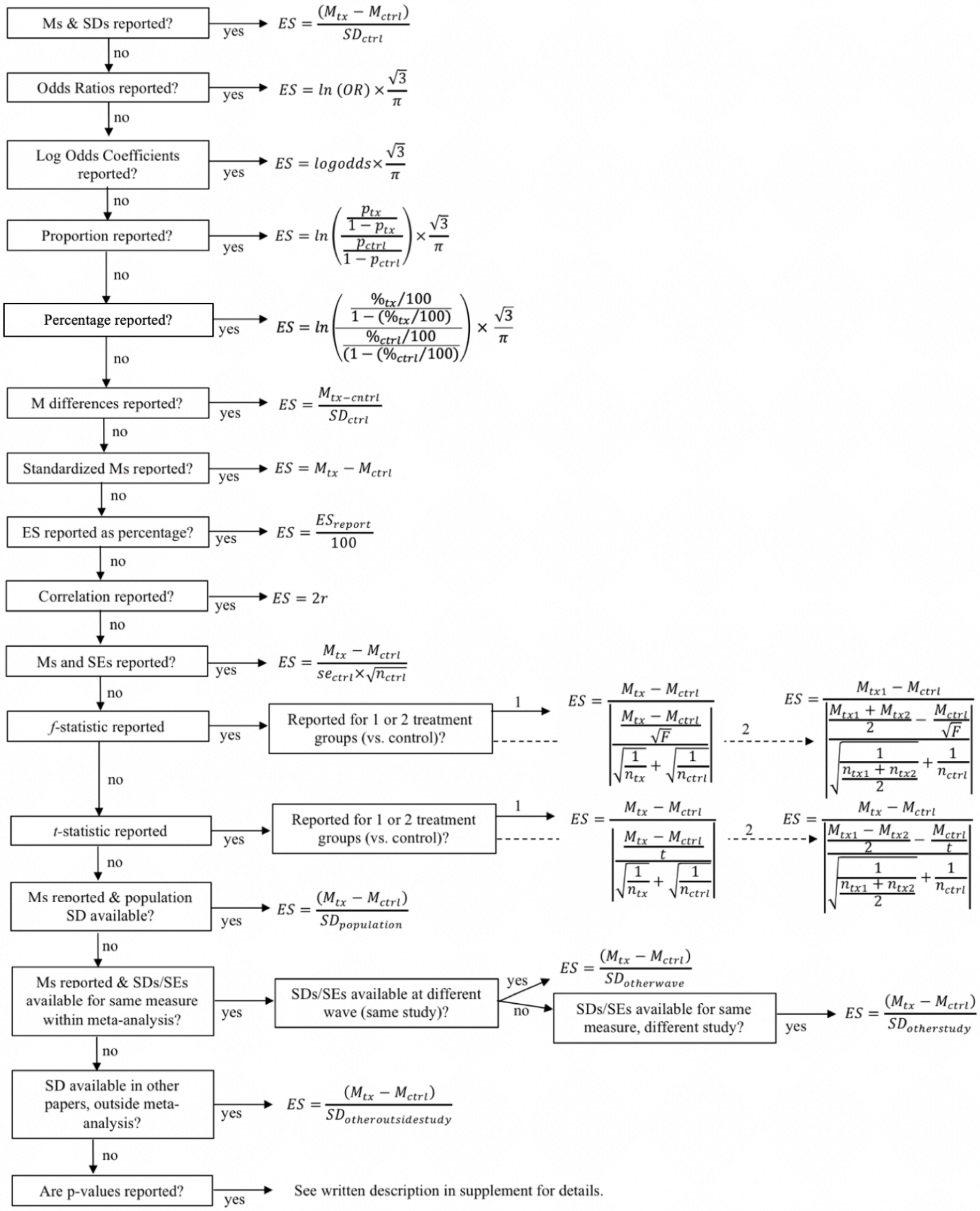
To calculate standard errors for effect sizes derived from odds ratios, log odds ratios, proportions, and percentages, we used the standard error equation presented in the manuscript, plugging in the effect size calculated using the methods detailed in Figure S1. These standard error estimates are likely slightly downwardly biased (we estimate by ~13-14%) as we were unable to use the variance associated with the original author-reported statistics (as suggested here by Hasselblad & Hedges, 1995) as this variance information was not consistently reported.

Results Presented for Subsamples & Multiple Treatment Groups

Notably, there were cases when data were reported separately for different sub-samples within a study (e.g., for boys and girls, for "low-risk" and "high-risk" participants, etc.). For these cases, we derived a main treatment effect by averaging the effect size estimates for each group, weighted by the group sample size. The same weighted averaging was used for standard errors and *p*-values. Critically, if the treatment effect was only reported for one sub-sample (e.g., only boys, only "low-risk" participants, etc.), then the effects were dropped from the meta-analysis so that each estimate represented a main treatment impact of the original random assignment to treatment or control.

Results were also commonly reported for multiple treatment groups formed via random assignment within a study. We opted to leave effect sizes presented separately by treatment group when possible since the effects reflected experimental treatment impacts. However, there were some instances when effect sizes were reported for each treatment group separately at earlier assessment waves (e.g., pre-test, post-test, 6-12-month follow-up), and in aggregated form at later assessment waves (e.g., 3-year follow-up). In these cases, treatment-specific effect sizes, standard errors, and *p*-values were averaged to form an average treatment effect that could be investigated in alignment with the effect sizes from later assessment waves.

Effect Size Calculation Flow Chart



Ms & SDs reported?
yes
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{ctrl}}$
no
Odds Ratios reported?
yes
$ES = ln\ (OR) \times \frac{\sqrt{3}}{\pi}$
no
Log Odds Coefficients reported?
yes
$ES = logodds \times \frac{\sqrt{3}}{\pi}$
no
Proportion reported?
yes
$ES = ln\left(\frac{\frac{p_{tx}}{1 - p_{tx}}}{\frac{p_{ctrl}}{1 - p_{ctrl}}}\right) \times \frac{\sqrt{3}}{\pi}$
no
Percentage reported?
yes
$ES = ln\left(\frac{\frac{\%_{tx}/100}{1 - (\%_{tx}/100)}}{\frac{\%_{ctrl}/100}{(1 - (\%_{ctrl}/100)}}\right) \times \frac{\sqrt{3}}{\pi}$
no
M differences reported?
yes
$ES = \frac{M_{tx-cntrl}}{SD_{ctrl}}$
no
Standardized Ms reported?
yes
$ES = M_{tx} - M_{ctrl}$
no
ES reported as percentage?
yes
$ES = \frac{ES_{report}}{100}$
no
Correlation reported?
yes
$ES = 2r$
no
Ms and SEs reported?
yes
$ES = \frac{M_{tx} - M_{ctrl}}{se_{ctrl} \times \sqrt{n_{ctrl}}}$
no
f-statistic reported
yes
Reported for 1 or 2 treatment groups (vs. control)?
1
$ES = \frac{M_{tx} - M_{ctrl}}{\left|\frac{\frac{M_{tx} - M_{ctrl}}{\sqrt{F}}}{\sqrt{\frac{1}{n_{tx}}} + \sqrt{\frac{1}{n_{ctrl}}}}\right|}$
2
$ES = \frac{M_{tx1} - M_{ctrl}}{\left|\frac{\frac{M_{tx1} + M_{tx2}}{2} - \frac{M_{ctrl}}{\sqrt{F}}}{\sqrt{\frac{1}{\frac{n_{tx1} + n_{tx2}}{2}} + \frac{1}{n_{ctrl}}}}\right|}$
no
t-statistic reported
yes
Reported for 1 or 2 treatment groups (vs. control)?
1
$ES = \frac{M_{tx} - M_{ctrl}}{\left|\frac{\frac{M_{tx} - M_{ctrl}}{t}}{\sqrt{\frac{1}{n_{tx}}} + \sqrt{\frac{1}{n_{ctrl}}}}\right|}$
2
$ES = \frac{M_{tx} - M_{ctrl}}{\left|\frac{\frac{M_{tx1} - M_{tx2}}{2} - \frac{M_{ctrl}}{t}}{\sqrt{\frac{1}{\frac{n_{tx1} + n_{tx2}}{2}} + \frac{1}{n_{ctrl}}}}\right|}$
no
Ms reported & population SD available?
yes
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{population}}$
no
Ms reported & SDs/SEs available for same measure within meta-analysis?
yes
SDs/SEs available at different wave (same study)?
yes
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{otherwave}}$
no
SDs/SEs available for same measure, different study?
yes
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{otherstudy}}$
no
SD available in other papers, outside meta-analysis?
yes
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{otheroutsidestudy}}$
no
Are p-values reported?
yes
See written description in supplement for details.

Note: This flow chart details the formulas used to calculate effect sizes and the decision-making process for deciding which calculation to use. Additional details relevant to this process, such as adjustments to standard errors when these were estimated in models controlling for pre-test scores, and the procedure used to calculate effect sizes from p-values (if no other information was provided) are included in the supplemental text.