WHAT WORKS CLEARINGHOUSE
Study Review Guide Instructions for Reviewing Randomized Controlled Trials and Quasi-Experimental Designs

Released March 3, 2014
Updated August 18, 2014

This document provides step-by-step instructions on how to complete the Study Review Guide (SRG, Version S3, V1) for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs).

For every What Works Clearinghouse (WWC) review, reviewers will be asked to complete an SRG. A completed SRG should be a reviewer's independent assessment of the study, relative to the criteria specified in the review protocol and the WWC Procedures and Standards Handbook (Handbook). At the end of the review process, a Master SRG (MRG) will represent the final assessment of the study and provide a record for the WWC, as well as serve as a key input in producing reports. For more details on the review process, see the instructions provided in the protocol for a particular review.

This guide is intended to be used by individuals trained and certified in the WWC review standards and procedures and in conjunction with the WWC Procedures and Standards Handbook. See the Inside the WWC tab at http://whatworks.ed.gov for the links to Instructions, the Handbook, and the most current version of the SRG.

General Characteristics of the SRG

The SRG uses color to indicate aspects of the review. Color is introduced by filling cells or using different colored text. Below is the color legend for this SRG.

Blue, Accent 1, Lighter 60%: Indicates a section heading in the *Main Tab*
Blue, Accent 1, Lighter 80%: Indicates a stem for a row in the *Main Tab*
Black, Text 1, Lighter 15%: Indicates text entered by reviewer
White, Background 1, Darker 50%: Indicates a drop-down menu or link to a section in the *Data Tab*
Green: Indicates a response that keeps the study moving through the review process
Red: Indicates a response that results in the study not moving forward through the review process
Light Blue: Indicates a value from a drop-down menu that does not affect study rating (design, not applicable)
Orange, Accent 6, Lighter 60%: Indicates outcomes information in the *Data Tab*
Purple, Accent 4, Lighter 60%: Indicates attrition section in the *Data Tab*
Red, Accent 2, Lighter 60%: Indicates baseline equivalence section in the *Data Tab*
Aqua, Accent 5, Lighter 60%: Indicates study reported findings section in the *Data Tab*
Olive Green, Accent 3, Lighter 60%: Indicates WWC calculated findings section in the *Data Tab*
Olive Green, Accent 3, Darker 50%: Indicates a formula is in the column (and the column is locked)
Main, Data, and *Summary Tabs*: Cells become tan when reviewer needs to complete
Data and *Summary Tabs*: Cells become yellow when in a row that needs to be completed AND the cell includes a formula

In the *Main Tab*, there are two places a reviewer may need to add rows: (1) to represent all appropriate disposition codes for why the study is identified as *Not Eligible for Review* (Row 22), and (2) to represent all appropriate disposition codes for why the study is rated *Does Not Meet WWC Group Design Standards* (Row 42).

To add additional rows to capture the disposition, click on the *Review Tab*, and click "Unprotect Sheet." Place your cursor in the row below (i.e., Cell D23 and Cell D43); from the *Home Tab*, click "Insert" and select "Insert Sheet Rows." Select Cells A–D for the row you are copying. Press Ctrl + C. Move your cursor to Cell A in the row you added. Press Ctrl + V. On the *Review Tab*, click "Protect Sheet" to ensure you do not overwrite a formula by mistake.

The *Data Tab* includes 20 rows for outcomes (Rows 7–26). If the study you are reviewing has more than 20 outcomes x sample x time period, please click on the *Review Tab*, and click "Unprotect Sheet." Highlight a blank row. Press Ctrl + C. Right click and select "Insert Copied Cells." Repeat to add as many rows as you need. On the *Review Tab*, click "Protect Sheet" to ensure you do not overwrite a formula by mistake.

Main Tab of the SRG

The *Main Tab* of the SRG captures a prose summary and assessment of the study and is structured in three separate steps, described in detail below.

Your role as a reviewer is to provide complete information for each element in a given section. In general, you will be required to enter data into Columns C, D, and E in the *Main Tab*.

- Column C will typically require a short answer to a question posed in Column B.
- Column D will require a justification for the short answer in Column C.
- Column E will require the page numbers from the study that serve as the source of the justification presented in Column D.

Stage 1: Preliminary Screening

Stage 1 of the study review assesses whether the given study is eligible for WWC review under a given review protocol. All reviews are conducted under a specific review protocol. All screening and review decisions relate to that specific protocol and could differ under another review protocol. For instance, a study that is ineligible under one protocol could be eligible under another protocol because the sample or outcome measures align more closely with the latter than the former. Similarly, a study could meet standards under one protocol, but not another, because of aspects of the research design that play out differently under different review protocols.

- **Study ID.** (Cell C2)

  Enter the Study ID for the review.
- **Reviewer Number.** (Cell C3)

  Enter your reviewer number. If a Master Review Guide (MRG), enter "MRG."
- **Review Date.** (Cell C4)

  Enter the date of your review.

- **Full Citation.** (Cell D2)

  Enter the full citation for the study.

Example: Darch, C., Eaves, R. C., Crowe, D. A., Simmons, K., & Conniff, A. (2006). Teaching spelling to students with learning disabilities: A comparison of rule-based strategies versus traditional instruction. *Journal of Direct Instruction, 6*(1), 1–16.

Overview

- **Standards and Protocol.** (Row 9)

  Enter the version of standards being used for the review (Cell C9). Enter the name and version of the protocol being used for the review (Cell D9).

- **Intervention Name.** (Row 10)

  Clearly indicate the name of the intervention(s) you are reviewing within this SRG (Cell C10). Note whether a single name refers to multiple versions of an intervention or different names refer to the same or related interventions (Cell D10).

  Please enter the exact name of the intervention and a brief description from the study (Cell D10).

Example: *The study examines the effectiveness of Reading Recovery implemented as a pull-out program for struggling readers. Students participated in 30-minute lessons 4 days a week.*

Screening Criteria

- **Effectiveness.** Does the study examine the effect of an intervention? (Row 13)

  In Column C (Cell C13), select "Yes" if the study claims to examine the effect of an intervention within the scope of the review, regardless of the quality of the design; select "No" otherwise.

- **Design.** Does the study use an eligible design (randomized controlled trial, quasi-experimental design, regression discontinuity design, or single-case design)? (Row 14)

  In Column C (Cell C14), select "Yes" if the study uses any design with a comparison group or condition, regardless of the quality of the design; select "No" if the study uses a design without a comparison group or condition, such as a pre-post design that lacks a comparison group, meta-analysis, or literature review. In Column D, enter the type of design. If the design is a regression discontinuity or single-case design, stop entering information into the RCT and QED SRG template. Provided that you are certified to conduct a review for the given design, obtain the appropriate SRG template for the regression discontinuity or single-case design, and complete the SRG in the correct template. If you are working with the WWC, please route the study to review team leadership so they can obtain a reviewer certified in that design.

- **Focus.** Is the intervention a program, product, policy, or practice with the primary focus aligned with the review protocol? (Row 15)

  In Column C (Cell C15), select "Yes" if the intervention meets the criteria for inclusion specified in the protocol under *Types of Interventions to be Included*; select "No" otherwise. If the study is rated either *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, a more detailed description of the intervention will be provided in Stage 3 of the SRG.

- **Sample Alignment.** Does the study meet the requirements for sample characteristics specified in the review protocol? (Row 16)

  In Column C (Cell C16), select "Yes" if the intervention meets the criteria for inclusion specified in the protocol under *Types of Populations to be Included* (e.g., % English language learners or % general education); select "No" otherwise. In Column D, describe the sample with respect to how it satisfies or does not satisfy this requirement. If the study is rated either *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, a more detailed description of the study participants will be provided in Stage 3 of the SRG.

Example: *The sample includes 60% students with learning disabilities.*

- **Time.** Was the study published within the time frame relevant to the review protocol? (Row 17)

  In Column C (Cell C17), select "Yes" if the study falls within the time frame outlined in the protocol; select "No" otherwise.

- **Age or Grade Range.** Does the study examine students in the age or grade range specified in the review protocol? (Row 18)

  In Column C (Cell C18), select "Yes" if the intervention meets the criteria for the age or grade range specified in the protocol under *Types of Populations to be Included*; select "No" otherwise. In Column D, please indicate (1) the grade levels of the students in the sample OR (2) the age range of the children in the sample, if the sample is described in terms of age rather than grade level. If the study is rated either *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, a more detailed description of the study participants will be provided in Stage 3 of the SRG.

Example: *Ninety percent of the intervention students were identified as ninth graders. Ninety-five percent of the comparison students were identified as ninth graders. The mean age of the intervention group is 15.5 years (SD = 1.20). The average age of the comparison group is 15.7 years (SD = 1.15).*

- **Location.** Does the study examine sample members in a location specified for the review protocol? (Row 19)

  In Column C (Cell C19), select "Yes" if the study sample was drawn from the geographic region described in the protocol under *Types of Populations to be Included*; select "No" otherwise. In Column D (Cell D19), describe the location. If the study is rated either *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, a more detailed description of the study setting will be provided in Stage 3 of the SRG.

- **Outcomes.** Does the study address at least one outcome in a domain relevant for the review protocol? (Row 20)

  In Column C (Cell C20), select "Yes" if study estimates the impacts of the intervention on at least one outcome that falls into one of the domains specified in the protocol under *Types of Outcomes to be Included,* and in Column D[Cell D20], summarize the types of outcomes included (you do not need to list each outcome here).

  For example, you might enter the names of the domains for which there are outcomes and the number of outcomes in each domain.

Example: *The study reports three math outcomes: full scores, numeracy and geometry/spatial subscales, from the research-based early mathematics assessment (REMA), developed by Clements, Sarama, and Liu (2008).*

  Select "No" if there are no outcomes that fall within the domains of the review protocol. In Column D, briefly list the outcomes that are available. Provide any additional documentation needed for a reviewer to confirm that the outcomes are not eligible under that protocol. If you are working on a WWC team and are uncertain whether any of the study's outcomes qualify under the protocol, seek additional guidance from your review leadership.

Example: *The study is not eligible for review under the Dropout Prevention review protocol. Although the study is an examination of a youth development program that aims to keep youth engaged in school, the only outcomes reported in this publication are related to sexual behavior (pregnancies, use of birth control including condoms).*

- **Does the study meet WWC and review protocol screening criteria?** (Row 22)

  If more than one disposition is appropriate, copy and paste this row and select the additional disposition code.

  From the drop-down menu ("Yes/No?") in Column C, select "Yes" if the study met all of the screening criteria in this section and warrants a full review to determine if the study design meets the WWC's standards. In Column D, select "The study is eligible for review. An explanation for that decision is below."

Select "No" in Column C if the study failed to meet any of the screening criteria in this section, and select the appropriate screening disposition code in Column D.

The Screening Disposition Codes are:

… is out of scope of the protocol. This cover scenarios in which the study did not (1) use an intervention conducted in English, (2) occur within the time frame specified in the protocol, (3) occur within a geographic area specified in the protocol, (4) occur within a setting specified in the protocol, or (5) include an outcome within a domain specified in the protocol.

… does not use a sample aligned with the protocol. This covers scenarios in which (1) the study did not use a sample within the age or grade range specified in the protocol, or (2) [the study did not show/the authors could not confirm/the WWC could not confirm] that at least 50% of the sample was classified [review specific classification].

… does not use an eligible design. This covers scenarios in which the study did not (1) examine the effectiveness of the intervention; (2) contain a primary analysis; (3) use a comparison group design, regression discontinuity design, or single-case design; or (4) provide adequate or consistent information to assess whether it was eligible for review.

- **Explanation for Screening Disposition.** (Row 23)

  If you selected "Yes" in Cell C22, provide a brief explanation of why the study is eligible for review.

Example: *The publication is eligible for review under the Adolescent Literacy review protocol; the students are in fifth and sixth grades. It is a study of Reading Mastery and reports reading scores from the SAT-11.*

If you selected "No" in Cell C22, provide a complete explanation for the screening disposition you selected in Cell D22. One option is to consider using the scenarios described above as the starting point for your explanation.

Example: *The publication is not eligible for review under the Beginning Reading review protocol; the students are in fourth and fifth grades. It may be eligible for review under the*

Adolescent Literacy review protocol, as it is a study of Reading Mastery and reports reading scores from the SAT-11.

Stage 2: Quality of Evidence (if the study passes Stage 1)

Stage 2 of the review determines if the study meets WWC standards. In this section, you will be making assessments about the level of evidence from the study. Whenever possible, use direct quotes from the study when asserting that a study does not pass a particular standard to improve transparency of the review process.

- **Design details** (Row 27) **How are the intervention and comparison groups formed?** (Row 28)

  In Column C (Cell C28), use the drop-down menu to select the appropriate design (RCT, cluster RCT, or QED). The selection of a design results in the highlighting of cells on the *Data Tab* to guide you to the relevant cells for that design regarding attrition and baseline equivalence.

  In Column D (Cell D28), describe each stage in the process by which the intervention and comparison samples were formed. Identify any stage of the process for which the study is unclear (and may warrant an author query). Please draft any author questions you have on the *Author Query & Response Tab*.

Examples:

1) *The study is a cluster randomized controlled trial (classes were assigned, but students are analyzed). Ten schools volunteered for the program. Each school provided two ninth-grade Biology sections that participated in the study. Random assignment was done within each school using a coin to assign one section to the intervention (lab-based learning) and the other section to the comparison (business-as-usual) condition. In four of the schools, the same teacher taught both sections, as there was only one Biology teacher. In the remaining six schools, different teachers taught the conditions.*

2) *The study is a clustered quasi-experimental design (classes were assigned, but students are analyzed). Twenty classrooms were identified as intervention classrooms (computer-based dissection), and twenty classrooms were identified as comparison classrooms (business-as-usual). The authors do not provide any additional information on how classrooms were recruited or identified as one condition or the other.*

3) *The design is not clear. The authors write "Ten schools were assigned to the intervention condition, and ten schools were assigned to the control condition. Pairs of schools were identified matching on size, per-pupil spending, free reduced lunch status, and percentage minority students. One member of each pair was assigned to the intervention condition (p. 242)." An author query is needed to determine whether the authors randomly assigned schools to condition, as the language "assigned" and use of matched pairs suggest a possible blocked RCT. If no response is given, or the authors report they did not randomly assign schools to condition, then the study will be assessed as a quasi-experimental design.*

- **Is the study free of factors that are confounded with either group?** (Row 29)

  Select "No" in Column C (Cell C29) and provide an explanation in Column D (Cell D29) if there is a confounding factor; select "Yes" otherwise.

  There is a confound if (1) there is an n = 1 problem (i.e., only one unit assigned in the intervention group, the comparison group, or both), OR (2) the study suffers from another confound that would lead you to expect differences in outcomes between the groups, even if the intervention group had *not* received the intervention. Please draft any author questions you have on the *Author Query & Response Tab.*

Examples:

1) *The study is rated Does Not Meet WWC Group Design Standards. There is an n = 1 confound at the interventionist level. A single teacher taught the three classes in the intervention condition. A single teacher taught the two classes in the comparison condition.*

2) *An author query is needed to determine whether there is an n = 1 confound at the classroom level. The authors report that five classes from two schools participated in the study. School A contributed three classes. School B contributed two classes. The intervention group had three classes, and the comparison group had two classes. It is not clear if School A is the only school in the intervention condition and School B is the only school in the comparison condition, or if both schools contributed classes to both conditions. Please see the Author Query & Response Tab for the proposed question. If no response is received, the study will be rated Does Not Meet WWC Group Design Standards. If the authors respond that School A was the intervention school and School*

B was the comparison school, the study will be rated Does Not Meet WWC Group Design Standards.

The authors responded (see Author Query & Response Tab), indicating that both schools contributed to both conditions. The intervention condition includes one class from School B and two classes from School A. The comparison condition includes one class from School B and one class from School A.

Complete Orange Section of the Data Tab (Row 31)

Click on Cell A31 to be taken to the beginning of the *Data Tab*. Complete the orange section (instructions begin on page 23).

- **Is there at least one relevant outcome that meets review requirements?** (Row 32)

  Select "Yes" or "No" in Column C (Cell C32). List all eligible outcomes with a brief description of the measure. Please draft any author questions you have on the *Author Query & Response Tab*.

Examples:

1) *The authors report on four eligible outcomes, although an Author Query (AQ) is needed to determine whether one has acceptable reliability. The three eligible outcomes are: Terra-Nova Reading (a standardized measure in the General Reading domain); Terra-Nova Math (a standardized measure in the General Mathematics domain); and the SAT-10 Science (a standardized measure in the General Science domain). The authors also report a score derived from 20 publicly-released TIMSS items, which is not an established subscale. An AQ is needed regarding reliability for the TIMSS score. If the authors do not have or do not provide reliability information, then the TIMSS outcome is rated Does Not Meet WWC Group Design Standards.*

2) *The authors did not respond to the query for reliability information on the TIMSS. The TIMSS outcome is rated Does Not Meet WWC Group Design Standards as it does not meet requirements. It is not an established subscale, and no reliability information is provided.*

3) *The authors report on "risky sexual behavior," which is not eligible for review, as the Dropout Prevention review protocol focuses on progress in school, not sexual behavior.*

Complete the Purple Section of the Data Tab (Row 34)

Click on Cell A34 to be taken to the beginning of the *Data Tab*. Complete the purple section (instructions begin on page 27.

Select the appropriate attrition boundary (liberal or conservative) in Cell C34.

- **Is there at least one outcome, sample, or time point with low attrition at the cluster and subcluster level?** (Row 35)

  If the study is a QED, on the *Main Tab* (Row 35) in Column C, select "NA," as attrition is not assessed for studies using a QED.

  If the study is an RCT or a cluster RCT, on the *Main Tab* (Row 35), select "No" in Column C (Cell C35) if either cluster-level or subcluster-level attrition is high. If the study is an RCT or cluster RCT with low attrition, select "Yes."

  For all studies, in Column D (Cell D35), briefly describe the units of assignment and analysis and the associated attrition as necessary. In particular, provide details on how you identified the numbers used in the *Data Tab*, if this information is not readily obtained from the page numbers listed in Column E. If attrition varies for different samples/analyses, briefly describe that as well. Please draft any author questions you have on the *Author Query & Response Tab*.

Examples:

1) *The study is a QED; attrition is not assessed by the WWC for QEDs.*

2) *The study is a clustered RCT. Ten schools were randomized to intervention (n = 5) and comparison (n = 5). The analytic sample included seven schools (intervention n = 5; comparison n = 2). Using the liberal attrition threshold, there is high attrition at the cluster level. Examining attrition at the sub-cluster level, focusing on only the seven schools in the analytic sample, the numbers of students who were randomized are 1,000 students in the intervention condition and 350 students in the comparison condition. The analytic sample included 800 students in the intervention condition and 300*

students in the comparison condition; therefore, there is low attrition at the sub-cluster level. Due to the high attrition at the cluster level (schools), the study must demonstrate baseline equivalence.

Complete Red Section of the Data Tab (Row 37)

Click on Cell A37 to be taken to the equivalence section of the *Data Tab*. For studies that provide baseline data for the analytic sample[1], complete the red section (instructions begin on page 29.

Please complete even for RCTs with low attrition, as this information may be used for a difference-in-differences adjustment.

- **Is evidence of baseline equivalence provided for at least one analytic sample, including statistical adjustment for characteristics relevant to equating the groups as given in the protocol, as needed?** (Row 38)

  The list of measures on which the groups are required to be equivalent is specified in the protocol. (Note: You will need to determine whether demonstrating baseline equivalence is required based on the design and the assessment of attrition.)

  For RCTs with low attrition, including cluster RCTs, select "NA" for "not applicable" in Column C (Cell C38).

  For RCTs with high attrition and QEDs, select "Yes" if the study established baseline equivalence for all eligible outcomes (no statistical adjustment was required or all required statistical adjustments were made); select "No" if the study did not establish baseline equivalence for any of the eligible outcomes.

  In Column D (Cell D38), list the outcomes that pass the baseline equivalence standard, and then list the outcomes that do not pass the baseline equivalence standard. If appropriate, describe the statistical adjustment used to control for baseline differences in the analyses. Please draft any author questions you have on the *Author Query & Response Tab*.

[1] The analytic sample is the set of study participants that were observed at the focal assessment period for the outcome under review.

Examples:

1) *The study is a QED. The baseline differences for the outcomes are within the adjustment range (REMA Total Score = 0.15 SD; Mental Addition Score = 0.19 SD). The authors report HLM analyses, including the REMA pretest and Mental Addition pretest as covariates. The study demonstrates baseline equivalence.*

2) *The study is an RCT with high attrition. The baseline differences for the outcomes are within the adjustment range (SAT-10 Reading = 0.23 SD; WJ-Passage Comprehension = 0.07 SD). The authors report repeated measures ANOVA analyses, which do not provide acceptable statistical control for baseline differences. The study is rated Does Not Meet WWC Group Design Standards due to a failure to demonstrate baseline equivalence for any eligible outcomes.*

3) *The study is a QED. The baseline differences for the three outcomes are all below 0.05 SD; thus, the authors do not need to adjust for baseline differences, and they did not adjust their analyses. They presented only unadjusted means and standard deviations. The WWC will apply a post-hoc difference-in-differences adjustment to improve the precision of the estimated ES.*

- **Is the study free of other data or analytic issues that would affect the rating?** (Row 40)

  In Column C (Cell C40), select "No" if there are issues about the analysis or data that were not captured in an earlier entry. Use Column D to summarize the issues. If there are no issues, select "Yes."

Examples:

1) *The authors do not adjust for the clustered nature of the data; a post-hoc clustering adjustment will be made using 0.20 as the ICC for all outcomes as per the review protocol. There are multiple outcomes in the General Mathematics domain; a post-hoc multiple adjustment comparison using the Benjamini-Hochberg method will be applied.*

2) *The degrees of freedom reported for the reported F-statistics are not consistent, suggesting there is variation in the number of students providing each outcome. For the most part, they appear to be within 10 students of the sample sizes reported in the*

article and used in the Data Tab to assess attrition. Rows 30–34 were used to determine whether the attrition rating would change if the 10 students were removed (a) equally from both groups (Row 30); (b) all from intervention (Row 31); (c) all from comparison (Row 32); (d) 3/5 from intervention and 2/5 from comparison (Row 33); or (e) 2/5 from intervention and 3/5 from comparison (Row 34). Regardless of the distribution of the sample loss, the study still has low attrition of youth.

- **What is the highest rating of an analysis in the study, given current information?** (Row 42)

  Select a rating using the drop-down menu in Column C (Cell C42).

  If you select *Meets WWC Group Design Standards Without* (or *With*) *Reservations*, in Column D (Cell D42), select "The study meets WWC Group Design Standards with (or without) reservations." In Row 43, briefly describe the contrast(s) and rating(s). You will also need to complete Stage 3.

Example: *The study is rated Meets WWC Group Design Standards Without Reservations. It is an RCT with low attrition at immediate posttest. The follow-up test is rated Meets WWC Group Design Standards With Reservations due to high attrition; however, the pretest difference for the analytic sample for the follow-up test is less than 0.05 SD.*

  If you select *Does Not Meet WWC Group Design Standards*, do not complete Stage 3.

  If the study is rated *Does Not Meet WWC Group Design Standards* (*DNMGDS*) based on the information provided, select the appropriate *DNMGDS* Disposition Code in Column D (Cell D42) and provide an explanation for the disposition code in Row 43. Copy and paste Row 42 as many times as needed to capture the exact disposition code (selected from menu in Cell D42) for each comparison that could be rated *Does Not Meet WWC Group Design Standards*.

  The *DNMGDS* Disposition Codes are:

  … the measures of effectiveness can not be attributed solely to the intervention. This covers scenarios in which (1) a group design study had only one unit assigned to one or both conditions, or (2) the effects of the intervention of interest were reported only in combination with other interventions.

… the eligible outcomes does not meet WWC requirements. This covers scenarios in which (1) outcomes were overaligned with the intervention, (2) outcomes were determined not to be sufficiently valid or reliable, or (3) inter-assessor agreement did not meet minimum thresholds.

… equivalence of the analytic intervention and comparison groups is necessary and not demonstrated. This covers scenarios in which the study was (1) a QED, (2) an RCT or RDD in which the combination of overall and differential attrition rates exceeded the WWC standards for this review, (3) an RCT or RDD in which attrition rates could not be assessed, (4) an RCT in which the groups were not generated using a random process, (5) an RCT in which there was nonrandom allocation after random assignment. Note that this occurs when equivalence (a) could not be assessed; (b) could be assessed and an adjustment was required, but not used; or (c) could be assessed and the difference was too large.

If an Author Query is needed to determine the final rating, indicate what response would be needed. The Author Query should be referenced in the appropriate row above and the question drafted on the *Author Query & Response Tab*.

If the rating differs by analysis, provide the rating for each sample, outcome, and time period combination, as necessary.

Examples:

1) *The study is rated Does Not Meet WWC Group Design Standards. It is a quasi-experimental study that does not demonstrate baseline equivalence, as all domains have one outcome with a baseline difference greater than 0.25 SD.*

2) *The study is rated Does Not Meet WWC Group Design Standards. The study is a clustered RCT with high attrition at the sub-cluster level. Baseline differences are in the adjustment zone (ranging from 0.07–0.23 SD); however, the authors only report gain score analyses, which do not adjust for the baseline differences in an acceptable manner for the WWC.*

3) *The study is rated Does Not Meet WWC Group Design Standards. The study is an RCT in which groups were formed using a nonrandom process. The intervention group included students whose last names started with A–L, while the comparison group included*

students whose last names started with M–Z. Baseline differences are in the adjustment zone (ranging from 0.05–0.20 SD); however, the authors only report unadjusted M and SD.

4) *The study is rated Does Not Meet WWC Group Design Standards. The study is a CRCT in which attrition at the cluster level cannot be assessed, as the number of teachers originally assigned by condition is not provided. The magnitude of the baseline difference cannot be calculated, as no SDs are reported. The authors report a two-level HLM with the pretest as a covariate. The study rating may be changed depending on whether the author responds. NOTE: The AQ was sent on 10/10/2012; no response was received by 11/10/2012.*

Stage 3: Study Details (if the study passes Stage 2)

Stage 3 of the SRG summarizes the key findings and a broad description of the study design and intervention. Ideally, this section should be written so that the text for each subsection can be directly pasted into the appropriate appendices for an intervention report or single study review. As such, do not use any text that was directly copied from the study in the descriptions of the study details in Stage 3.

Complete Blue and Green Sections of the Data Tab (Row 49)

Click on Cell A49 to be taken to the study-reported and WWC-computed findings sections of the *Data Tab*. Complete the blue and green sections (instructions begin on page 31 (blue) or page 33 (green)).

- **Did the authors present effect sizes? If so, how were they computed?** (Row 50)

  If the study presents effect size estimates, select "Yes" in Column C (Cell C50) and indicate how those effect sizes were computed in Cell D50 of the *Main Tab* (regardless of whether impact estimates are provided in other metrics). If the authors do not report effect sizes, select "No." In particular, note if the author-reported ES differs from the WWC-calculated ES and provide information that explains why (e.g., the authors use the comparison group SD, not the pooled SD, in their calculations).

Examples:

1) *The authors do not report effect sizes.*

2) *The authors report effect sizes; however, they use the comparison group SD in the denominator, rather than the pooled SD. This results in slight differences in magnitude from the WWC-calculated effect sizes.*

- **Are estimates presented for subgroups in protocol?** (Row 51)

  In Column C (Cell C51), select "Yes" if the study provides impact estimates for any of the subgroups outlined in the protocol under *Types of Populations to be Included*; select "No" otherwise.

  In Cell D51, briefly describe which subgroups are analyzed and whether those analyses meet WWC group design standards. In the *Data Tab*, be sure to include information from subgroup analyses in addition to the main analyses. (See instructions on completing the *Data Tab* for more information.)

Example: *The study reports impacts within low- and high-performing subgroups. Please see the rating boxes above. The low-performing contrast is rated Meets WWC Group Design Standards With Reservations, as it is an RCT with high attrition that demonstrates baseline equivalence (difference is 0.14 SD, and pretest is a covariate in regression analysis). The high-performing contrast is rated Meets WWC Group Design Standards Without Reservations. It is an RCT with low attrition.*

In summary, describe … (Row 53)

- **Setting of the study (e.g., location, classrooms, courses, schools).** (Row 54)

  Include the locations from which the sample was drawn.

Example: *The study was conducted in five states in the Northwestern region of the United States. Ten districts participated in the study. The ninth-grade Algebra classes at each high school in the district participated (a total of 20 classes, taught by eight teachers).*

- **Study design.** (Row 55)

  Summarize the study design, including how the sample was selected and the number of clusters/students that were assigned to each condition.

Example: *The sample was recruited from 20 schools in Georgia. Schools were randomized to condition using a random number algorithm. All ninth-grade history teachers in each school and their students participated in the study (n = 25 for intervention; n = 20 for comparison). Passive consent was used, as the intervention was the required history curriculum for ninth grade, and outcomes are the standard state history test. There were 1,500 students from 10 schools in the intervention condition and 1,350 students from 10 schools in the comparison condition at the beginning of the year.*

- **Sample sizes (e.g., students, classrooms, teachers, schools).** (Row 56)

  Summarize the participants, including the characteristics of the participant sample. Include a description of the number of clusters/students in the analytic sample (use a range if the sample sizes vary across outcomes).

Example: *The intervention condition included 10 schools, 25 classrooms, and 1,500 students. The analytic sample included 10 schools, 25 classrooms, and 1,000 students.*

The comparison condition included 10 schools, 20 classrooms, and 1,350 students. The analytic sample included nine schools, 18 classrooms, and 1,200 students.

There is a single outcome, so there is no variation in analytic sample sizes by outcome.

- **Sample characteristics in protocol (e.g., race, gender, free/reduced-price lunch).** (Row 57)

Example: *The analytic sample for the intervention condition included 1,000 students. The analytic sample was 51% female, 35% Black, 30% Asian, 25% Caucasian, and 10% Hispanic. Eighty percent of the students were eligible for free or reduced-price lunch. Twenty percent were receiving special education or related services, but were enrolled in general education history classes.*

The analytic sample for the comparison condition included 1,200 students. The student sample was 49% female, 35% Caucasian, 25% Black, 20% Asian, and 20% Hispanic. Seventy

percent of the students were eligible for free or reduced-price lunch. Fifteen percent were receiving special education or related services, but were enrolled in general education history classes.

- **Intervention condition as implemented in the study (including number of days/weeks/months, number of sessions, time per session).** (Row 58)

  Summarize the intervention(s) in sufficient detail to help readers understand what makes this intervention similar to or different from other interventions. The level of detail should be similar to what would be provided in an introductory section of a typical impact evaluation report. It should include the length of the intervention and the dosage, as well as information about the content, delivery, and implementation of the intervention. Note that this description should be about the intervention as used in this study, not as described by the developer or in ideal conditions.

Example: *The intervention, Making History Come Alive, is a project-based curriculum with 20 units, one per week. Each unit consists of a 30-minute lecture portion and a project-based activity to be completed in small groups. Five classrooms completed the lecture portion of all 20 units but did not complete any project-based activities. Ten classrooms completed the lecture and project-based activity for 15 of the 20 units. Five classrooms completed the lecture portion of all 20 units and the project-based activity for 15 of the 20 units. Five classrooms were able to implement as designed (20 units; both lecture and project-based activities). Teacher surveys indicated the primary reason for not completing the project-based activities included: classroom management challenges and time constraints. The primary reason for not completing the full 20 units was time constraints.*

- **Comparison condition as implemented in the study.** (Row 59)

  Indicate the consequences of being assigned to the comparison group (e.g., what comparison group members could not receive, what the study suggests they did receive, etc.). Clarify whether the counterfactual was a particular alternative intervention, and if so, name the intervention, and provide a brief description if the study provides that information.

Example: *The comparison schools implemented their standard history curriculum. The authors do not identify the particular curriculum used. Teacher surveys indicate the primary mode of instruction was lecture, used in all 18 classes. Projects were used at least*

once in 15 of the 18 classrooms. The connection between the project and lecture and nature of the project (group or individual) is not clear. This condition would be best described as "business-as–usual," although some specific details about the curriculum are available from the teacher survey.

- **Describe all eligible outcomes reported and how they were measured.** (Row 60)

  Describe all of the outcomes within relevant domains examined in the study, and identify which of those outcomes are eligible based on the criteria specified in the protocol under *Types of Outcomes to be Included*. Also, indicate how each eligible outcome was measured (if it is not self-explanatory from the name of the outcome), was collected (if relevant), and can be interpreted (the scale of the measure).

Example: *The single outcome is the Georgia State History Test. This outcome is in the General History domain. It is a state test and thus considered a standardized test by the WWC. The state assessment is given in the spring, and scaled scores are reported.*

- **Are there outcomes that do not meet review requirements? If yes, provide the domain and a brief description of the reason why.** (Row 61)

  Select "Yes" or "No" in Cell C61. In Cell D61, list all eligible outcomes and why they do not meet review requirements.

Example: *The authors did not respond to the query for reliability information on the TIMSS. The TIMSS outcome is based on 20 publicly-available items but is not an established subscale. The outcome is rated Does Not Meet WWC Group Design Standards, as it is not an established subscale, and no reliability information was provided.*

- **Are there any outcomes that are not eligible for review? If yes, provide a brief description and the reason why.** (Row 62)

  Select "Yes" or "No" in Cell C62. In Cell D62, list the outcomes that are not eligible for review, along with a reason for ineligibility.

Example: *The authors report on "risky sexual behavior," which is not eligible for review, as the Dropout Prevention review protocol focuses on progress in school, not sexual behavior.*

- **Support for implementation.** (Row 63)

  Indicate both the staff training and technical assistance conducted to support the implementation of the intervention (as evaluated in the study).

Example: *Teachers in the intervention condition participated in a 2-week summer institute. During the institute, they developed lesson plans and delivered them to a selected sample of ninth graders attending summer school at schools not participating in the study. Members of the developer's team visited each teacher at least once a semester to observe a lesson and provide comprehensive feedback. Teachers were placed into small learning communities that used a protected workspace online to share ideas and problem-solve. Teachers were able to send messages to the developer's team for support related to any particular lesson or activity.*

Data Tab of the SRG

Each row corresponds to a single outcome for a particular sample at a particular point in time. Pre- and post-intervention outcomes using the same measure and sample should be reported in a single row. Most of the column headings have a comment attached that provides a short description of the information to be captured within it.

Elements in Row 2 of the *Data Tab* will auto-fill with information from the *Main Tab*.

- **Study ID.** (Cell A2)

  Auto-fills with content in *Main Tab* Cell C2.
- **Select design.** (Cell B2)

  Auto-fills with content in *Main Tab* Cell C28.

Orange (Outcome Name, Domain, Construct, and Measure Characteristics) Section of the Data Tab

General instructions:

- Include only outcomes that fall within one of the eligible domains for the review.
- If you are uncertain whether a particular outcome measure falls under one of these domains, check for additional guidance that your team may provide on outcome measures.
- If results are presented separately for different samples or periods, you will need to have a separate row for each. Similarly, if there are multiple follow-ups, include a row for each. If reliability information differs across the samples, make sure it is accurately reflected in the table.
- Measures that are used only as a pre-intervention measure (such as age or gender) or post-intervention measure (such as dropout) are recorded on a separate row and labeled accordingly.
- When you enter an outcome name in Column A, cells that need to be completed will be shaded tan. Cells that are shaded yellow have formulas in them and are locked, which means you will not be able to place your cursor in the cell.

- **Measure.** (Column A)

  Indicate the name of the outcome or test, exactly as it was specified in the article or report.

- **Domain.** (Column B)

  Indicate which of the eligible domain(s) the outcome comes from. This should be the name of the DOMAIN only (no constructs). If a study reports subgroup analyses, composite and subtests, or outcomes at different points in time, then the domain name in the *Data Tab* must reflect this so that the domain averages and multiple comparisons (MC) are conducted correctly. In taking this approach, the main analyses will not suffer an MC penalty when the authors present subgroup, subtest, or additional follow-up examples. In the presentation of the results in an intervention report or single study report, the evidence rating for the study can be based on the main analyses (which will be presented in Appendix C of the report), and the additional analyses can be included in Appendix D for transparency (though these findings will not contribute to the evidence rating).

  For example, in subgroup analyses (which are recorded in the *Data Tab* below the full sample analysis), the "Math" domain in the full sample analysis could be named the "Math-SE" domain for a special education subgroup analysis.

  For studies that report outcomes for both a composite measure and multiple subtests, the SRG could use the "Math" domain for the composite and "Math-Subtests" for the arithmetic, fractions, and whole number subtests.

  Similarly, for an outcome in the Math domain collected at three time points, the SRG could use "Math" for the time period that will contribute to the evidence and "Math-X mo" for the time periods that will be reported in appendices.

  If an outcome or contrast is rated *does not meet WWC Group Design Standards* and will not be included in the report, the domain column should be empty to ensure the multiple comparison adjustments are correct.

- **Construct.** (Column D)

  Note any relevant constructs for the domain in this column. This is to ensure that the domain averages calculated later in the worksheet are for the full domain,

regardless of construct. Not all areas have constructs; check the protocol.

- **Binary.** (Column E)

  Select "Yes" from the drop-down menu if the outcome is a 1/0 variable for which the underlying construct is a yes/no answer, such as "ever graduated" or "retained in grade" (i.e., there is not an underlying distribution of the variable). Select "No" otherwise.

- **Standardized test.** (Column F)

  Select "Yes" from the drop-down menu if the test is a standardized test. The score should be from the full test or established subscale to be considered a standardized test.

- **Face validity.** (Column G)

  Select "Yes" from the drop-down menu if the measure appears to be a reasonable measure; select "No" if you see an obvious problem with the measure, and indicate your concerns in Column N.

- **Test-retest reliability.** (Column H)

  Enter the test-retest reliability of the outcome if it is reported; enter "NR" otherwise. Unless specified in the review protocol, the minimum acceptable value is 0.40.

- **Internal consistency.** (Column I)

  Enter the internal consistency of the outcome if it is reported; enter "NR" otherwise. Unless specified in the review protocol, the minimum acceptable value is 0.50.

- **Inter-rater reliability.** (Column J)

  Enter the inter-rater reliability of the outcome if it is reported; enter "NR" otherwise. Unless specified in the review protocol, the minimum acceptable value is 0.50.

- **Not overaligned?** (Column K)

Select "Yes" from the drop-down menu if you have no concerns that the measure may be overaligned with the intervention.

Select "No" from the drop-down menu if you have concerns that the measure may be overaligned with the intervention.

Measures that are closely aligned or tailored to the intervention are likely to demonstrate larger effect sizes than those that are less closely aligned with the intervention. An example of overalignment is if the measure includes some of the same materials (such as specific reading passages) that are used in the intervention or administered to the intervention group as part of the intervention. Explain any concerns in Column N.

- **Same measure & collection?** (Column L)

  Select "Yes" from the drop-down menu if the same measure was collected in a similar manner for the intervention and comparison groups.

  Select "No" from the drop-down menu if it is clear that outcome data were collected in a different manner for the intervention and comparison groups, potentially in a way that could lead to differences in average outcomes between groups.

  In a situation where the outcome data were collected differently across the intervention and comparison groups, this may be considered a confounding factor and documented in the *Main Tab* (Row 31). Explain any concerns in Column N.

- **Meets WWC requirements?** (Column M)

  Based on the answers in the previous columns and the standards established in the review protocol, select "Yes" from the drop-down menu if the outcome meets all of the requirements; select "No" otherwise.

  If you select "No," ensure Column N includes details on why the measure does not meet WWC requirements. (Note: The measure name in Column A will turn red to indicate it should not be reported.)

- **Notes or concerns about the measure.** (Column N)

Summarize any concerns you have about this measure. In particular, if the measure does not meet the requirements for inclusion in this review as noted in Column O, indicate why.

If a citation is provided for a standardized test or subscale, it may be helpful to note that here.

Now, go back and enter a decision in Cell C32 on the *Main Tab* (Is there at least one relevant outcome that meets review requirements?).

Purple (Attrition) Section of the Data Tab

For each outcome, enter the necessary sample information in that row as follows.

For clustered studies, complete Columns O–X.

- **Unit of assignment (if different).** (Column O)

  For clustered designs, enter the unit of assignment (teacher, student).

- **Baseline sample.** (Columns P and Q)

  In Column P, enter the number of clusters for the intervention group at the time of (random) assignment.

  In Column Q, enter the number of clusters for the comparison group at the time of (random) assignment.

- **Analytic sample.** (Columns S and T)

  In Column S, enter the number of clusters for the intervention group for the analytic sample.

  In Column T, enter the number of clusters for the comparison group for the analytic sample.

- **Attrition rates.** (Columns V and W)

  The overall attrition rate at the cluster level is calculated in Column V.

The differential attrition (in percentage points) is calculated in Column W.

- **Low?** (Column X)

  A determination of whether attrition is low at the cluster level is reported ("Yes" or "No").

For all studies, complete columns Z–AI.

- **Unit of analysis.** (Column Z)

  Enter the unit of analysis (also known as the subcluster).

- **Baseline sample.** (Columns AA and AB)

  In Column AA, enter the sample size for the intervention group at the time of (random) assignment.

  In Column AB, enter the sample size for the comparison group at the time of (random) assignment.

- **Analytic sample.** (Columns AD and AE)

  In Column AD, enter the sample size for the intervention group for the analytic sample.

  In Column AE, enter the sample size for the comparison group for the analytic sample.

- **Attrition rates.** (Columns AG and AH)

  The overall attrition rate for the unit of analysis is calculated in Column AG.

  The differential attrition for the unit of analysis (in percentage points) is calculated in Column AH.

- **Low?** (Column AI)

  A determination of whether attrition is low for the unit of analysis is reported ("Yes" or "No").

- **Is there low attrition?** (Column AJ)

  A determination of whether attrition is low is reported ("Yes" or "No").

Now, go back and enter a decision in Cell C35 on the *Main Tab* (Is there at least one outcome, sample, or time point with low attrition at the cluster and subcluster level?).

Red (Baseline Equivalence) Section of the Data Tab

- **Using data from …** (Column AK)

  Select the type of data that will be used to assess baseline equivalence.

  Unadjusted M and SD should be selected if the measure is continuous and the unadjusted pre-intervention mean and standard deviation are reported.

  "T-Stat" should be selected if the authors report a summary *t*-statistic from a *t*-test comparison of means. Note: In general, it is not appropriate to use a *t*-statistic from any analysis other than a *t*-test for group means for WWC effect size calculations.

  "Dichotomous" should be selected if the authors provide unadjusted pre-intervention means reported for both groups for variables that are true 0/1 variables (e.g., dropout, graduated) and not for proportions (e.g., % correct).

- **Intervention.** (Columns AL–AN)

  Based on the selection in Column AK, Columns AL and AM will be shaded to indicate a value should be entered.

  In Column AL, enter the pre-intervention mean for the intervention group. If the measure is binary, this should be entered as a decimal value (i.e., "0.50," not "50" for 50%).

  In Column AM, enter the pre-intervention standard deviation for the intervention group.

  Column AN will be filled based on the value in Column AD.

- **Comparison.** (Columns AP–AR)

  Based on the selection in Column AK, Columns AP and AQ will be shaded to indicate a value should be entered.

  In Column AP, enter the pre-intervention mean for the comparison group. If the measure is binary, this should be entered as a decimal value (i.e., "0.50," not "50" for 50%).

  In Column AQ, enter the pre-intervention standard deviation for the comparison group.

  Column AR will be filled based on the value in Column AE.

- **t.** (Column AT)

  Based on the selection in Column AK, Column AT will be shaded to indicate a value should be entered.

  In Column AT, enter the summary *t*-statistic.

- **g.** (Column AU)

  Based on the information entered in Columns AK–AT, the magnitude of the baseline difference will be reported as Hedges' *g.*

- **Equiv?** (Column AV)

  Based on the information calculated in Column AU, the determination of equivalence will be provided.

  "Yes," indicating the magnitude of the baseline difference is less than 0.05 SD.

  "No," indicating the magnitude of the baseline difference is greater than 0.25 SD.

  "Adj," indicating the authors must report findings from an analysis that statistically controlled for the baseline difference.

- **Was the pretest different?** (Column AX)

  Select "Yes" from the drop-down menu if the pre-intervention measure was different from the post-intervention measure; select "No" otherwise.

  If different forms of the same measure were used, it is not a different pre-intervention measure. If a different pre-intervention measure was used, consult with the review team leadership to determine if the pre-intervention measure is an acceptable proxy for the post-intervention measure.

- **Did analysis adjust for pre?** (Column AY)

  Select "Yes" from the drop-down menu if the authors report an analysis that statistically controls for the baseline difference; select "No" otherwise.

- **Should we do D-n-D with pre?** (Column AZ)

  A formula will return "Yes" or "No" to indicate whether a difference-in-differences adjustment should be conducted.

- **Okay to use in report?** (Column BA)

  A formula will return "Yes" or "No" to indicate whether the measure should be included in the report.

Now, go back and enter a decision in Cell C38 on the *Main Tab* (Is evidence of baseline equivalence provided for at least one analytic sample, including statistical adjustment for characteristics relevant to equating the groups as given in the protocol, if needed?).

Blue (Analysis and Results) Section of the Data Tab

For each outcome that is eligible and for which baseline equivalence is established or properly accounted for in the analysis:

- **Sample.** (Column BB)

  Indicate the sample for which intervention and comparison group means were computed. Possible options include "full sample" or a description of the subgroup or subsample for which means are provided (e.g., "Grade 1" or "boys").

- **Period.** (Column BC)

  Indicate any relevant information about the timing of the pre- or post-intervention measure, such as when the pre-intervention variable was measured (e.g., in the spring prior to random assignment or the school year) and whether it is a pretest only.

- **Using data from …** (Column BD)

  Select the type of analysis used to compare post-intervention differences in the intervention and comparison groups.

  Selecting the analysis type will highlight the corresponding cells that need to be completed in the remaining columns of this section.

  Enter as much information as you have. You may want to make notes for the author query if there is information that you need.

There are a number of ways to estimate an effect size and *p*-value from the information presented in a study. Some studies may provide enough information to use different types of results for WWC calculations. As such, we have created a stratified list of analysis types, from which WWC reviewers should prioritize analyses earlier in the list (provided that the study reported analysis is correct/appropriate).

Possible Analysis Types (taken from the Handbook)

Model Based Estimates of Program Impacts (using pretest or other covariates):

- **OLS:** Results from an OLS regression are reported. Enter the standard deviation for the intervention group (Column BG) and comparison group (Column BK), along with the regression coefficient (Column BP).

- **HLM Level-2 coefficient:** Results from an HLM regression are reported that examine impacts at a particular point in time (i.e., not a growth-curve analysis). Enter the unadjusted standard deviation for the intervention group (Column BG) and comparison group (Column BK), along with the regression coefficient (Column BP).

- **ANCOVA adjusted post-intervention:** Adjusted post-intervention means and

standard deviations reported for both groups. Enter the adjusted means and unadjusted standard deviations for the intervention group (Columns BF and BG) and the comparison group (Columns BJ and BK).

- **ANCOVA F-test and correlation:** Summary F-statistic for the test of the intervention effect from an ANCOVA is reported along with the pre/post correlation. Enter the F-statistic (Column BN) and correlation (Column BO).

Posttest only (though there might be an opportunity for the WWC to do a D-n-D):

- **Unadjusted post-intervention:** Unadjusted post-intervention means and standard deviations reported for both groups. Enter the means and standard deviations, for the intervention group, Columns BE and BG and for the comparison group, Columns BI and BK.

- **Dichotomous means:** Unadjusted post-intervention means reported for both groups for variables that are true 0/1 variables (e.g., dropout, graduated) and not for proportions (e.g., % correct). Enter the means in Column BE (intervention) and Column BI (comparison).

Test Statistic (no information on means):

- ***t*-stat:** Summary *t*-statistic from a *t*-test comparison of means is reported. Enter the *t*-statistic in Column BM. Note: In general, it is not appropriate to use a *t*-statistic from any analysis other than a *t*-test for group means for WWC effect size calculations.

- **ANOVA F-test:** Summary F-statistic from a one-way (one-factor) ANOVA is reported. Enter the F-statistic in Column BN. Note: In general, it is not appropriate to use an F-statistic from any analysis other than a one-way ANOVA for WWC effect size calculations.

Green (Findings) Section of the Data Tab

For each outcome that is eligible and for which baseline equivalence is established or properly accounted for in the analysis, complete Columns BQ–CN as appropriate.

- **Effect size | S** (Column BV)

A locked cell with a formula to calculate the pooled standard deviation.

- **Effect size | N_s** (Column BW)

  A locked cell with a formula to calculate the ratio of sample sizes necessary for some effect size calculations.

- **Effect size | *g*** (Column BX)

  A locked cell with a formula to calculate Hedges' *g* for continuous outcomes or Cox index for dichotomous outcomes.

- **Effect size | WWC** (Column BY)

  A locked cell with a formula to calculate the effect size with a small sample size adjustment and a difference-in-differences adjustment, if appropriate.

- **Effect size | II** (Column CA)

  A locked cell with a formula to calculate the improvement index based on the WWC-reported effect size (Column BY).

- **Study *p*-value** (Column CC)

  Enter the *p*-value reported in the study.

- **Use study *p*-value?** (Column CD)

  Select "Yes" from the drop-down menu if the WWC should use the *p*-value as reported in the study, such as when the author made the appropriate adjustments for (1) baseline equivalence and (2) clustering, and (3) conducted their analysis appropriately given their design; select "No" otherwise.

- **Cluster correction | ICC** (Column CF)

  If the unit of assignment and unit of analysis differ, enter the value of the intra-class correlation for the outcome domain as specified in the review protocol. The default intra-class correlation is 0.20 for achievement outcomes and 0.10 for behavioral and attitudinal outcomes. The review team leadership may set different defaults in the protocol if explicitly justified in terms of the nature of the research circumstances or the outcome domain. If the study reports the ICC from

an appropriate HLM analysis, include this ICC to have the WWC-calculated *p*-value better align with the study-calculated *p*-values.

- **Cluster correction | M** (Column CG)

  A locked cell with a formula to calculate the total number of clusters in the intervention and comparison groups.

- **Cluster correction | ta** (Column CH)

  A locked cell with a formula to calculate the *t*-statistic, adjusted for clustering.

- **Cluster correction | df** (Column CI)

  A locked cell with a formula to calculate the degrees of freedom for the *t*-statistic calculated in Column CH.

- **Cluster correction | WWC *p*** (Column CJ)

  A locked cell with a formula to calculate the *p*-value, including incorporating the cluster adjustment, if necessary.

- **MC rank** (Column CL)

  Enter the rank of the outcome (smallest *p*-value = 1) if a multiple comparison correction is needed within the domain; leave the cell blank otherwise.

- **Critical *p*-value** (Column CM)

  A locked cell with a formula to calculate the new critical *p*-value that will be used in determining statistical significance.

  Below are brief instructions for determining statistical significance in the SRG. Refer to the Handbook for more details on the Benjamini-Hochberg (BH) correction procedures.

  To determine whether a finding is statistically significant after applying the BH correction, you should identify the *p*-value with the *largest* value that is statistically significant relative to its new critical *p*-value. Then all findings with a *p*-value less than that new critical *p*-value are statistically significant. This means you may identify a finding as statistically significant that was not significant

relative to its own critical *p*-value.

To assess statistical significance in the example below, start at the highest rank and compare each outcome's *p*-value with the revised critical *p*-value based on its rank. In the example, the eighth- through sixth-ranked outcomes are not statistically significant. The fifth-ranked outcome is statistically significant (0.030 < 0.031); therefore, this outcome and all others ranked higher (1–4) would be designated as statistically significant as well, including the fourth-ranked finding, where the *p*-value is higher than its own revised critical *p*-value.

Author-reported or clustering corrected *p*-value (Px)	Rank (x)	New Critical *p*-value (P`x = 0.05x/X)	Px < P`x?	Statistically significant after BH correction?
0.002	1	0.006	Yes	Significant
0.009	2	0.013	Yes	Significant
0.014	3	0.019	Yes	Significant
0.027	4	0.025	No	Significant
0.030	5	0.031	Yes	Significant
0.042	6	0.038	No	Not Significant
0.052	7	0.044	No	Not Significant
0.076	8	0.050	No	Not Significant

- **Significant after MC?** (Column CN)

  Select "Yes" if the finding is statistically significant after adjusting for multiple comparisons; select "No" otherwise.

Summary Tab of the SRG

Rows 2 and 3 will auto-fill based on information entered in the *Main Tab*.

- **Study ID.** (Cell A2)
  Auto-fills with content in *Main Tab* Cell C2.

- **Full citation.** (Cell B2)
  Auto-fills with content in *Main Tab* Cell D2.

- **Review date.** (Cell S2)
  Auto-fills with content in *Main Tab* Cell C4.

- **Standards version.** (Cell A3)
  Auto-fills with content in *Main Tab* Cell C9.

- **Review protocol and version.** (Cell B3)
  Auto-fills with content in *Main Tab* Cell D2.

- **Select a design.** (Cell E3)
  Auto-fills with content in *Main Tab* Cell C28.

- **Select rating.** (Cell H3)
  Auto-fills with content in *Main Tab* Cell C42.

- **Select *DNMGDS* disposition code.** (Cell S3)
  Auto-fills with content in *Main Tab* Cell D42.

Each row corresponds to a single outcome domain.

- **Domain.** (Column A)
  Enter a domain name. It must be entered exactly as in the *Data Tab*.

- **Number of outcomes.** (Column B)

A locked cell with a formula to calculate the number of outcomes listed in the *Data Tab* that are in the domain entered in Column A.

- **Effect size.** (Column C)

  A locked cell with a formula to calculate the average effect size for all outcomes in the domain entered in Column A.

- **Max effect size.** (Column D)

  A locked cell with a formula to calculate the maximum effect size for all outcomes in the domain entered in Column A.

- **Improvement index.** (Column E)

  A locked cell with a formula to calculate the improvement index from the average effect size for the domain entered in Column A.

- **Min improvement index.** (Column F)

  A locked cell with a formula to calculate the minimum improvement index for any measure in the domain entered in Column A.

- **Max improvement index.** (Column G)

  A locked cell with a formula to calculate the maximum improvement index for any measure in the domain entered in Column A.

- **Max sample size.** (Column H)

  A locked cell with a formula to calculate the maximum sample size for any measure in the domain entered in Column A.

- ***p*-value.** (Column I)

  A locked cell with a formula to calculate the *p*-value for the average effect size for the domain entered in Column A.

  This is calculated based on the hidden *t*-statistic (Column N) which is based on the average effect size (Column C).

- **Are any outcomes significant?** (Column J)

  A locked cell with a formula to calculate whether any outcomes in the domain entered in Column A remained significant after necessary adjustments were made.

- **Characterization of finding.** (Column S)

  Select the appropriate characterization of findings for the domain entered in Column A.

  See the Handbook for more details, but in brief:

  Single, SS+: Estimated effect is positive and statistically significant after any necessary adjustments.

  Single, SI+: Estimated effect is positive and not statistically significant after any necessary adjustments, but is substantively important.

  Single, Indeterminate: Estimated effect is neither statistically significant after any necessary adjustments nor substantively important.

  Single, SI-: Estimated effect is negative and not statistically significant after any necessary adjustments, but is substantively important.

  Single, SS-: Estimated effect is negative and statistically significant after any necessary adjustments.

  Multiple, SS+ (A): Univariate statistical tests are reported for each outcome measure, at least half of the effects are positive and statistically significant, and no effects are negative and statistically significant.

  Multiple, SS+ (B): Univariate statistical tests are reported for each outcome measure, at least one measure is positive and statistically significant, and no effects are negative and statistically significant.

  Multiple, SS+ (C): Mean effect is positive and statistically significant after any necessary adjustments.

  Multiple, SS+ (D): Omnibus effect for all outcome measures together is reported as positive and statistically significant on the basis of a multivariate statistical

test in a properly-aligned analysis.

Multiple, SI+: Mean effect size is positive and not statistically significant, but is substantively important.

Multiple, Indeterminate: Mean effect reported is neither statistically significant nor substantively important.

Multiple, SI-: Mean effect size is negative and not statistically significant, but is substantively important.

Multiple, SS- (A): Univariate statistical tests are reported for each outcome measure, at least half of the effects are negative and statistically significant, and no effects are negative and statistically significant.

Multiple, SS- (B): Univariate statistical tests are reported for each outcome measure, at least one measure is negative and statistically significant, and no effects are negative and statistically significant.

Multiple, SS- (C): Mean effect is negative and statistically significant after any necessary adjustments.

Multiple, SS- (D): Omnibus effect for all outcome measures together is reported as negative and statistically significant on the basis of a multivariate statistical test in a properly-aligned analysis.

Author Query & Response Tab of the SRG

Each reviewer should draft their questions for an author query (AQ). The MRG should capture the final set of questions as sent to the author. It should also document whether (and when) a response was received and the response.

- **Date AQ sent** (Cell B3)

  Enter the date the author query was sent.

  Note: The team coordinator will send author queries from the WWC email account for reviews conducted under contract ED-IES-13-C-0010.

- **Response received? (Yes/No)** (Cell B4)

  Enter "Yes" or "No" to document whether any response was received.

- **Date of response** (Cell B5)

  Enter the date the response(s) was/were received by the team.

- **Mode of response (fax, email, etc.)** (Cell B6)

  Enter the particular mode(s) by which the response(s) was/were received by the team.