

Updated February 2018

Underlying all What Works Clearinghouse (WWC) products are WWC Study Review Guides, which are intended for use by WWC certified reviewers to assess studies against the WWC evidence standards.

As part of an ongoing effort to increase transparency, promote collaboration, and encourage widespread use of the WWC standards, the Institute of Education Sciences provides external users with access to a web-based WWC Study Review Guide for conducting reviews of group design studies, including randomized controlled trials and quasi-experimental designs. Reviewers use the Study Review Guide to document the characteristics of studies, including features that pertain to a study's eligibility under a WWC protocol. The Study Review Guide assists the reviewer in assessing the study design and implementation against the WWC standards, and coding the study findings in a systematic manner consistent with WWC reporting guidelines. The WWC provides a separate Excel-based tool for conducting reviews of single-case design studies

The WWC Study Review Guide for group design studies is intended to be used by individuals trained and certified in WWC review policies and procedures, in conjunction with WWC review protocols and the **WWC Procedures and Standards Handbooks**. The Study Review Guide supports reviews of group design studies against the Version 3.0 and Version 4.0 group design standards.

This document guides users through the public-use version of the WWC Study Review Guide, available at <https://ies.ed.gov/ncee/wwc/wwcsrgpublic>. Members of the public are invited to use the public-use version of the SRG to understand how WWC reviewers document the characteristics of a study, determine its eligibility, and assess its design and implementation against the WWC standards. However, official WWC study reviews are conducted by trained and certified WWC reviewers using the full version of the SRG. The public-use SRG and the full SRG used for official reviews differ in some functions. They both apply the same WWC procedures and standards and result in the same rating for a given study, but the public-use SRG cannot be used to conduct official WWC reviews.

This page has been left blank for double-sided copying

The SRG walks users through the steps of the review process, assists reviewers in applying the WWC group design standards to determine a study rating, and applies WWC procedures for reporting findings from studies that meet WWC design standards. The review consists of five sections: screen, measures, rating, context, and narrative. You can save your work at any time in any section by clicking “Save” at the bottom of the page.

You can document general notes about the review at any time by clicking “Notes” on the bottom right of the screen on most pages of the SRG. In the notes, you can document any concerns or questions you have about the study that are not documented elsewhere in the SRG, and you can note any assumptions you made during your review.

The SRG supports reviews under both the version 3.0 and version 4.0 group design standards. However, all descriptions below pertain to the version 4.0 standards.

A. Beginning a review

Each WWC review assesses a *study* of an *intervention*, reviewed under a *protocol*, using the WWC *standards*. To begin a review, select “Start new review” from the public SRG landing page and complete the information for each of these components (see Figure 1):

- **Intervention.** Click “Select intervention” to choose the intervention that will be the focus of this review. Search for the intervention by name or create a new intervention by entering the intervention name and program description.
- **Citation.** Enter the study title, author, publication date, and publication.
- **Standards Version.** Select the version of the WWC Standards that the study should be reviewed against.
- **Protocol.** Select the review protocol to be used for the review. All WWC review protocols can be found at: <https://ies.ed.gov/ncee/wwc/Protocols>
- **Protocol Version.** Select the version of the review protocol to be used for the review.

Figure 1. Begin a new review

The screenshot shows a web form titled "Begin a new review". It contains several input fields and dropdown menus. At the top is a button labeled "Select intervention". Below it is a text input field labeled "Citation". Underneath the citation field is a dropdown menu labeled "Standards Version" with the text "Select standards version" and a downward arrow. Below the standards version dropdown are two more dropdown menus: "Protocol" with the text "Select protocol" and "Protocol Version" with the text "Select version", both with downward arrows. At the bottom left of the form is a link that says "Procedures and Standards Handbooks and Review Protocols". At the bottom right are two buttons: a green "Save" button and an orange "Cancel" button.

Click “Save” to proceed with the review. If you wish, enter your email address in the pop-up window that appears to have your PIN sent to you. Your PIN is the only way to return to an in-progress review once you have left the SRG or begun a new review. Your PIN will also appear at the top of your review when you continue, and you may also send the PIN for an active review to yourself or a collaborator by email by clicking the “Email my PIN” button in the upper right-hand corner of the SRG.

B. Screening

This section of the review consists of preliminary questions to confirm the study is eligible for review under the protocol and is relevant to the purpose of the review. Select “Yes” or “No” to move through the questions. If the answer to all screening questions is “Yes,” the review is eligible and you can continue to the next section by clicking “Save and continue.” If you answer “No” to any question, the system will display the reason why the study is ineligible. You can add any applicable notes for why the study is ineligible by clicking “Notes” at the bottom right. The system displays relevant text in italics under each question from the review protocol.

Review topics and questions include the following:

1. Is it a study?

Select “Yes” if the citation denotes a study and not a newspaper article, blog, or other publication.

2. Effectiveness.

Does the study contain primary analysis examining the effectiveness of an intervention?

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.

3. Design.

Does the study use an eligible design? *Italicized text under this question will indicate which types of study designs are eligible under this protocol (randomized controlled trial, quasi-experimental design, regression discontinuity design, or single-case design). Select “Yes” if the study uses an eligible design. Select “No” if the study uses an ineligible design, such as one without a comparison group or condition (pre-post design), meta-analysis, or literature review.*

4. Time.

Was the study published within the time frame relevant to the review protocol?

Select “Yes” if the study falls within the time frame outlined in the protocol; select “No” otherwise.

5. Outcomes.

Does the study address at least one outcome in a domain relevant for the review protocol? *Select “Yes” if the study estimates the impacts of the intervention on at least one outcome that falls into one of the domains specified in the protocol; select “No” otherwise.*

6. Age or grade range.

Does the study examine students in the age or grade range specified in the review protocol?

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.

7. Sample alignment.

Does the study meet the requirements for sample characteristics specified in the review protocol?

Select “Yes” if the intervention meets the criteria for inclusion specified in the protocol under “Types of Populations to be Included” (for example, percentage English learner students, percentage general education); select “No” otherwise.

8. Setting.

Does the study occur within a setting specified in the review protocol?

Select “Yes” if the intervention meets the criteria for the setting specified in the protocol; select “No” otherwise.

9. Location.

Does the study occur within a geographic area specified in the review protocol?

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.

10. Relevant.

Is the intervention aligned with the focus of the review?

Select “Yes” if the study is relevant to the intervention of interest. Review the intervention description and any other available information. Select “No” otherwise.

11. Other issues.

Is the study free of other issues preventing eligibility?

C. Measures

In the measures section of the SRG, reviewers identify and describe the measures used in the study and enter data to evaluate and report on findings. The SRG defines a measure as a unique combination of the instrument or method used to assess an outcome, the timing of the follow-up, and the sample used to measure the intervention’s impact. Each measure the reviewer enters is associated with a single finding within the study.

1. Adding a measure

Add a measure for each study finding by clicking “Add new outcome” under the Measures heading. You will be taken to a new page to enter information about this measure for the review. First, click “Select a measure” and enter the measure name in the search box. Always search the system for an existing measure before adding a new one. Search with multiple individual keywords in the measure name to confirm the measure does not already exist. The system will

search for measures linked to domains that are relevant to the review protocol. Carefully review the results for the correct measure and domain. If the measure already exists in the system, click “Select” next to the measure name to add the measure to the review (see Figure 2).

Figure 2. Select a measure

The screenshot shows a web interface titled "Select a measure". At the top, there is a search bar labeled "Search by measure name:" with the text "DIBELS" entered. Below the search bar is a table with two columns: "Measure" and "Domain". Each row in the table has a blue "Select" button to the left of the measure name. The table contains three rows of results:

Measure	Domain
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)	Reading achievement
Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Curriculum-Based Measurement of Reading (CBM-R) passages	Reading fluency
Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Initial Sound Fluency subtest	Phonological processing

If the measure search does not return a relevant result, you can create a new measure (see Figure 3). To add a new measure, click “Add a new measure” at the bottom of the measure search page and complete the following fields:

- **Domain.** Select the domain from the dropdown menu that is relevant to the selected review protocol.

- Measure name.** Enter the full name of the measure. Include the measure's relevant acronym, if applicable, and include the full name of the measure as well. Note: The measure should appear with the specific measure name. However, study- or sample-specific information, such as grade, cohort, student characteristics, or the follow-up period (such as Year 2 follow-up) should not appear in the measure name.
- Measure description.** Enter a brief description of the measure, without regard to the study in question.
- Is the measure a standardized test?** Select "Yes" if the test is a standardized test. A standardized test is one in which the same test is given in the same manner to all test takers with established administration and scoring procedures, often documented in a technical manual. To be considered a standardized test, the score should originate from the full test or established subscale, using the documented scoring procedures.
- Is the measure dichotomous?** A measure is dichotomous (or binary) 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." Select "Yes" if the outcome is measured as a yes/no answer for each sample member. Select "No" otherwise.
- Does the measure have evidence of face validity?** To show evidence of face validity, a sufficient description of the outcome measure must be provided to determine that the measure is clearly defined, has a direct interpretation, and measures the construct it was designed to measure. For example, a count of spoken words during a time period has face validity for measuring reading fluency, and the percentage of students who complete high school would be an outcome with face validity as a graduation rate. Select "Yes" if the measure appears to be a reasonable measure; select "No" if you see an obvious problem with the measure.
- Test-retest reliability.** Test-retest reliability or temporal stability refers to the degree to which test results are consistent over time. Enter the test-retest reliability of the outcome measure, if reported.
- Internal consistency.** Internal consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar

Figure 3. Add a measure

Add measure.

Domain
- Select domain -

Measure name

Measure description

Is this measure a standardized test? ?

Yes No

Is this measure dichotomous? ?

Yes No

Does the measure have evidence of face validity? ?

Yes No

Test-retest reliability ?

Internal consistency ?

Inter-rater reliability ?

Save Cancel

scores. For example, if a respondent expressed agreement with the statements “I like to ride bicycles” and “I’ve enjoyed riding bicycles in the past,” and disagreement with the statement “I hate bicycles,” this would indicate good internal consistency for the measure. Enter the internal consistency of the outcome, if reported.

- **Inter-rater reliability.** Inter-rater reliability is the degree of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges. If a student’s work is scored differently by different judges, the measure may not provide a clear signal about the quality of the work. Enter the inter-rater reliability of the outcome, if reported.

Click “Save” at the bottom of the page at any point to save your work.

2. Completing measure information

You will arrive at the Measure page after adding a measure to the review. The Measure page is divided into different sections by headings: Measure, Design, Sample Description, Sample Sizes, Baseline Measures, Analysis, and Measure Notes. Answer the questions in each section using information specific to the finding in the study.

The Measure Notes section is for the reviewer to document notes about a specific measure, as described below, and can be edited only on the page for the measure where you entered the notes. These notes will appear with the measure on the rating page. You can also add to the general review notes at any time by clicking “Notes” at the bottom right of the page. As a reminder, use complete sentences in the notes and include page references where needed.

The SRG asks questions about the measure that could affect the rating for the finding or how the finding will be reported by the WWC. Below, we outline some key considerations when answering these questions. For additional details and examples on these concepts, see the *WWC Standards Handbook*.

Measure

Answer the following questions about the measure and analysis (see Figure 4):

- **Is the measure 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. Select “Yes” if you have concerns that the measure might be overaligned with the intervention. Select “No” otherwise.
- **Was the measure collected in the same manner for both the intervention and comparison groups?** When outcome data are collected differently for the intervention and comparison groups, study-reported impact estimates will confound differences due to the intervention with those due to differences in the data collection methods. For example, measuring dropout rates based on program records for the intervention group and school administrative records for the comparison group will result in unreliable impact estimates.

This is because it will not be possible to disentangle the true impact of the intervention from differences in the dropout rates that are due to the particular measure used. Select “Yes” if the study collected the same measure in a similar manner for the intervention and comparison groups. Select “No” if it is clear the study collected outcome data differently for the intervention and comparison groups, potentially in a way that could lead to differences in average outcomes between groups. If the study collected outcome data differently across the intervention and comparison groups, the outcome does not meet review requirements and all findings based on the measure are rated *Does Not Meet WWC Group Design Standards*. Explain any concerns in the “Notes” field.

- **Did the analysis control for any endogenous covariates?** If a regression analysis includes a covariate that was potentially influenced by group status (that is, endogenous), the impact analysis will produce a biased test of the effect of the intervention. On the other hand, if a covariate is obtained after baseline and is unlikely to have been influenced by group status or is considered time invariant (for example, demographics such as gender and race), then there is no concern that the variable has been influenced by the intervention. For example, a study that examines the impact of an intervention on student achievement outcomes may collect data on (1) student attendance during the intervention or (2) the quality of teacher–student interactions. These variables associated with intervention dose, quality, or fidelity may have been affected by the intervention. If the impact analysis includes either of these measures as covariates, the correlation between the intervention indicator and these variables will produce bias in the impact estimate. Therefore, the WWC cannot use the results of the regression model as a credible source of information about an intervention’s effects. Select “Yes” if the analysis controls for covariates that are potentially influenced by group status. Select “No” otherwise.
- **Follow-up period.** The follow-up period is the time elapsed between the end of the intervention and the collection of the post-intervention measure. Enter the follow-up period during which the study collected outcome data and select the appropriate units from the dropdown menu: Days, Weeks, Months, Years, or Semesters. Enter “0 days” for immediate posttests. For example, if an intervention lasted the entire duration of the school year and at the conclusion of the school year the outcome measure was administered, that would be considered an immediate post-test and “0 days” should be entered for the follow-up period.

Figure 4. Adding measure to review

Measure

Phonology
Dynamic Indicators of Basic Early Literacy Skills (DIBELS) - Initial Sounds Fluency subtest
[Change to a different measure](#)

Is this measure overligned with the intervention? [?](#)
 ✖

Was the measure collected in the same manner for both the intervention and comparison groups? [?](#)
 ✖

Did the analysis control for any endogenous covariates?
 ✖

Follow-up period [?](#)
 - Select unit - ▼

Design

Answer the following questions about the study design associated with this measure:

- **Design.** Use the dropdown menu to select the appropriate design (randomized controlled trial or quasi-experimental design).
- **(If random assignment is selected as the design) Was the random assignment compromised?** There are four ways in which a randomized controlled trial that assigns individual subjects to the intervention or comparison condition can be compromised.
 - The randomized controlled trial is compromised when it includes subjects in the sample used to estimate findings (analytic sample) who were not randomly assigned. Joiners in cluster-level assignment studies are addressed separately in the SRG and reviewers should not indicate that the randomized controlled trial is compromised because of the presence of joiners.
 - The randomized controlled trial is compromised if subjects are randomly assigned to a group with different probabilities, but the findings are based on an analysis that does not account for the different assignment probabilities.
 - The randomized controlled trial is compromised when the investigator changes a subject's group membership after random assignment.
 - The randomized controlled trial is compromised when a study author manipulates the analytic sample to exclude certain subjects based on events that occurred after the introduction of the intervention when there is a clear link between the intervention and the reason for the exclusion. A clear link is present when the exclusion is based on a measure the intervention may have affected.

Select "Yes" if random assignment was compromised; select "No" otherwise.

- **Are there any confounding factors in the analysis of this measure?** In some studies, a component of the study design or the circumstances under which the intervention was implemented are perfectly aligned, or confounded, with either the intervention or comparison group. That is, some factor is present for members of only one group and absent for all members in the other group. In these cases, it is not possible to tell whether the intervention or the confounding factor is responsible for the difference in outcomes. For example, a study may have one intervention school and a different comparison school. In this case, it is impossible to separate how much of the observed effect was due to the intervention and how much was due to the particular school in which the intervention was used. Select “Yes” if there is a confounding factor and provide an explanation in the “Notes” field.
- **Comparison.** From the dropdown list, select the option that describes the services the comparison group received. Options include an intervention (that is, an alternative intervention, such as a different math curriculum from the intervention of interest for this review); business as usual; none; or unknown. If you select intervention, choose the intervention the comparison group received from the dropdown list.

Sample description

In the Sample Description section, select “Yes” if the sample for this measure is for the full sample (rather than a subgroup, such as female students or a cohort). If the sample for this measure is a subsample, select “No.” By selecting “No,” a set of options for common types of subsamples will appear (see Figure 5). Select the nature of the subsample. The system will auto-fill the sample description text field based on your selection. If the subsample comprises another characteristic that does not appear in the provided options, type the sample description in the text field. Use the auto-fill description whenever possible; only edit the sample description field if additional text is necessary to distinguish the sample from others in the study.

It is possible that no finding in a study uses the study’s full sample. For example, a study may examine multiple subgroups separately, but never analyze the combined sample. The SRG can aggregate these findings into a full sample finding, as discussed below. In this case, use the sample description that best describes the sample for the subgroup finding.

Sample sizes

Begin by selecting “Yes” or “No” in response to the question “Is this an analysis with cluster-level assignment?” Select “Yes” if the study satisfies two conditions: (1) individuals were assigned to the intervention or comparison condition as groups, and (2) outcomes were measured for individuals within those clusters (but may be analyzed as individual-level data or as cluster-level averages).

Figure 5. Sample description

Sample description

Does this data represent the full sample? ⓘ

☐ Yes ☒ No ✕

Select each category by which this subsample is differentiated from the full sample.

☐ Gender

☐ Grade

☐ Race

☐ Ethnicity

☐ Financial position

Sample description ⓘ

If “Yes” is selected to denote that the analysis contains cluster-level assignment, fields will appear to define the individuals and clusters for the analysis, and enter key sample sizes (see Figure 6). The individuals in the analytic sample are the units that contribute outcome data to the analysis. For example, in a study that aggregates student-level achievement data to school-level averages for analysis, the students are the individuals. The clusters in the analytic sample are the units that were assigned to conditions and contain the individuals who contribute outcome data. For both individuals and clusters, select the type of unit from the dropdown menu: Student, Teacher, Class, School, District, Campus, or Center.

Figure 6. Sample sizes for cluster studies

The screenshot shows the 'Sample sizes' form. At the top, it asks 'Is this an analysis with cluster level assignment?' with 'Yes' selected. Below this, there are two main sections: 'Individuals' and 'Clusters'. Each section has a dropdown menu for 'Unit' (set to 'Student' and 'Class' respectively) and a text input for 'Analytic sample size'. There are also checkboxes for 'Not reported' for each. Below these, there are two columns for 'Total present in clusters at time pre-intervention data were collected (reference sample for baseline representativeness)' and 'Total present in clusters at time outcome data were collected (reference sample for outcome representativeness)'. Each column has inputs for 'Intervention' and 'Comparison', with 'Not reported' checkboxes. At the bottom, it asks 'Did the analysis account for clusters?' with 'No' selected. There is also a field for 'Intracluster correlation coefficient' and a dropdown for 'Are the reported outcome standard deviations measured using data for individuals or clusters?' set to 'Select Level'. A 'Notes' button is in the bottom right corner.

Enter the analytic sample sizes for the intervention group and the comparison group for both the number of individuals and the number of clusters. If the study is a randomized controlled trial, also enter the reference samples for calculating individual non-response and cluster-level attrition. Finally, also enter the samples present in clusters when pre-intervention and follow-up data were collected. These samples represent the reference samples for assessing the representativeness of the sample at baseline and follow-up. Check the “Not reported” checkbox if a sample size for individuals or clusters is not reported in the study.

Select “Yes” or “No” in response to the question “Did the analysis account for clusters?” A “mismatch” problem occurs when assignment is carried out at the cluster level (for example, classroom or school level), while the analysis is conducted at the individual level (for example, student level). When the analysis ignores the correlation between outcomes among individuals within the same clusters when computing the standard errors of the impact estimates, this approach leads to underestimated standard errors and overestimated statistical significance. To assess the statistical significance of an intervention’s effects in cases where study authors have not corrected for the clustering, the WWC computes clustering-corrected statistical significance estimates. An example of an analysis that accounts for clusters and does not require the WWC correction is hierarchical linear modeling. Similarly, analyses of cluster-level averages do not require the correction.

If “No” is selected because the analysis did not account for clusters, enter the intraclass correlation coefficient that the WWC should use to perform the adjustment into the box provided. Check the review protocol for guidance on the appropriate intraclass correlation coefficient. As defaults, the WWC uses the intraclass correlation coefficient values of 0.20 for achievement outcomes and 0.10 for all other outcomes, but will use study-reported intraclass correlation coefficient values when available.

Select “Yes” or “No” in response to the question “Does the analytic sample include any joiners who pose a risk of bias according to the review protocol?” The analytic samples for cluster-level assignment studies may include two types of individuals: *stayers*, who were in the sample when the clusters were assigned to condition; and *joiners*, who entered the sample afterwards. For example, schools may be assigned in the fall, but the analytic sample might include new students who transferred into the schools later in the school year. The new students are joiners. Review protocols specify which joiners may pose a risk of bias based on when they entered clusters. Joiners who do not pose a risk of bias according to the review protocol can be treated as stayers in WWC reviews. Select “Yes” if the analytic sample included joiners; select “No” otherwise.

Finally, select from the dropdown list whether the standard deviations for the reported outcome were measured using data for individuals or clusters (select either Individual or Cluster in the dropdown). Standard deviations based on cluster-level data will be smaller than those based on individual-level data, leading to a larger effect size. Because they are not directly comparable to individual-level effect sizes, effect sizes based on cluster-level standard deviations will not be reported in most WWC products. (However, cluster-level means can be used to calculate effect sizes that are comparable to student-level effect sizes, so long as the calculation uses a standard deviation based on individual-level data.) The statistical significance and direction (positive or negative) of cluster-level effect sizes is taken into account in determining the characterization of study findings.

If “No” is selected to denote that the analysis does not contain cluster-level assignment, answer the two questions about missing and imputed data that appear. First, was any outcome data imputed to estimate this finding? Imputed outcome data can affect how the WWC calculates attrition and influence the study rating in other ways. Select “Yes” if the analytic sample includes any imputed outcome data; select “No” otherwise. In particular, select “No” if the authors addressed missing outcome data by excluding all cases with missing outcome data from the analytic sample (that is, complete case analysis).

Second, does the study use an acceptable approach to address all missing data in the analytic sample? Assess the description of the approach used to address any missing baseline or outcome data and compare to the list of acceptable approaches in Table II.6 of the *WWC Standards Handbook*. Consider all baseline measures included in the analysis, whether or not the measure is required for assessing baseline equivalence. The common approach of excluding all cases with missing baseline and outcome data (that is, complete case analysis) is an acceptable approach. Select “Yes” if an acceptable approach was used; select “No” otherwise.

Enter the requested sample sizes for the intervention group and comparison group, including the analytic sample sizes (see Figure 7). The analytic sample sizes include any records with

imputed data included in the analysis. If the study is a randomized controlled trial, record the number of subjects randomly assigned to conditions. If the study has imputed outcome data, also record the number of subjects with observed outcome data in both the intervention and comparison groups (that is, the analytic sample sizes minus any records with imputed outcome data that were analyzed). Because the WWC counts imputed outcome data as attrition, these sample sizes will be used as the numerators in attrition calculations. Check the “Not reported” checkbox if the sample size for individuals is not reported in the study. Also, select the type of unit from the dropdown menu: Student, Teacher, Class, School, District, Campus, or Center.

Figure 7. Sample sizes for non-cluster studies

The screenshot shows a web form titled "Sample sizes". It contains three questions, each with "Yes" and "No" radio buttons and a close icon (X):

- Is this an analysis with cluster level assignment? (The "No" button is selected.)
- Was any outcome data imputed to estimate this finding?
- Does the study use an acceptable approach to address all missing data in the analytic sample?

Below the questions are two main sections for data entry:

- Analytic sample size:** Includes a "Unit" dropdown menu (set to "Student"), and input fields for "Intervention" and "Comparison". Each input field has a "Not reported" checkbox below it.
- Number of subjects with observed outcome data:** Includes input fields for "Intervention" and "Comparison", each with a "Not reported" checkbox below it.

Baseline measures

A baseline measure is an assessment of skills or characteristics prior to the intervention. The WWC uses baseline data to assess the baseline equivalence of the intervention and comparison groups, which can affect the rating of quasi-experimental designs, high-attrition randomized controlled trials, and compromised randomized controlled trials. If there is baseline data for the sample relevant to a measure, click “Add a baseline measure.” A new page titled “Baseline” will appear. A reviewer should enter all available baseline data for the sample in the SRG, whether or not the data are required to assess baseline equivalence. Also, a reviewer should enter the baseline measure when the baseline data are reported for the analytic sample, or a smaller sample as a result of some missing baseline data.

Click “Select measure” to search for the measure in the system. If the measure exists, click “Select” next to the measure name to select that measure and domain. If the measure does not exist, create the measure in the system by clicking “Add a new measure” at the bottom of the window and complete the same fields required for new measures discussed on pp. 17-18.

After selecting or adding the baseline measure, you will return to the baseline page. **Indicate whether the measure is one that is required for assessing baseline equivalence under the review protocol.** Only measures that are indicated as required will influence a study rating.

Then, if the study is a cluster-level assignment study:

- **Indicate whether the baseline standard deviations reported are based on individual- or cluster-level data.** Standard deviations based on cluster-level data will be smaller than those based on individual-level data, leading to a larger baseline effect size. Cluster-level standard deviations may not be used to establish baseline equivalence of individuals (see Step 4 of the

review process for cluster-level assignment studies in the *WWC Standards Handbook*), but may be used to establish baseline equivalence of clusters (see Step 7 of the review process for cluster-level assignment studies in the *WWC Standards Handbook*). The baseline equivalence requirement for clusters (that is, Step 7) can be satisfied using individual- or cluster-level means and individual- or cluster-level standard deviations (in any combination), as long as the weighting of the means is consistent with the weighting used in the analysis. The WWC will use individual-level standard deviations when possible, so enter these into the SRG if available.

- **Is the baseline analytic sample comprised of the same individuals in the outcome analytic sample?** Select “Yes” if the baseline sample includes the exact analytic sample of individuals used to measure outcomes at follow-up; select “No” otherwise. You will select “No” if the baseline data are based on the same clusters as the analytic sample, but do not include the same individuals used to measure outcomes at follow-up. If you select “No,” the baseline measure can be used to establish equivalence of clusters only. Check the review protocol to confirm that the sample meets any requirements for establishing equivalence of clusters.

For all study designs: Was the baseline characteristic measured using the same units as the outcome? For example, answer “Yes” if the researchers administered the same test, using the same scoring procedures, as a pretest and posttest. Answer “No” if (a) the researchers administered different assessments at baseline and follow-up, or (b) the measures were the same, but different subscales or scoring procedures were used to score the tests.

If the study is an individual-level assignment study you will also be asked two questions about missing baseline data in the analytic sample:

- **Was any baseline data imputed to estimate this finding?** Select “Yes” if the baseline sample includes any imputed data; select “No” otherwise.
- **Is there missing baseline data for the analytic sample that was not imputed?** Select “Yes” if the same subjects in the analytic sample for the measure were excluded from the calculation of means for the baseline sample because of missing data; select “No” otherwise.

Enter the baseline sample sizes. First, record the number of intervention and comparison group subjects used to calculate the means to assess baseline equivalence (including any imputed baseline data). Then, if the study has imputed baseline data, record the number of subjects with observed baseline data used to calculate the baseline means in both the intervention and comparison groups (that is, excluding any records with imputed baseline data). These sample sizes will be used to assess baseline equivalence using the approach described in section II.C.4 of the *WWC Standards Handbook*. Check the “Not reported” checkbox if a requested sample size is not reported.

Enter the baseline means. First, record the means for the intervention and comparison group subjects used to assess baseline equivalence (including any imputed baseline data). If the study has imputed baseline data, also record the means for the samples with observed baseline data only. If the measure is dichotomous, use the numeric percentage value (0 to 100) for the mean. For example, a 79 percent graduation rate should be entered as 79. Check the “Not

reported” checkbox if the baseline means for the subjects used to assess baseline equivalence are not reported.

Enter the standard deviations for the subjects with observed baseline data. Check the “Not reported” checkbox if the standard deviations for subjects with observed baseline data are not reported.

If the study has imputed baseline data, enter the outcome means for the intervention and comparison group samples with both observed baseline and observed outcome data. These means will be used to assess baseline equivalence using the approach described in section II.C.4 of the *WWC Standards Handbook*.

Indicate whether the study analysis adjusted for this baseline measure and, if so, select the method of adjustment. The analysis may include a statistical adjustment for differences in the intervention and comparison groups at baseline, which is required for satisfying the baseline equivalence requirement when the baseline effect size falls between 0.05 and 0.25 standard deviations. For example, authors may include the baseline measure in the regression analysis. Select “Yes” if the authors report an analysis that statistically controls for the baseline difference; select “No” otherwise. If you select “Yes,” then also select the method from the dropdown method. The options are Regression or ANCOVA and Gain Scores, Difference-in-Differences, or Fixed Effects. If you select the second of these two options, the SRG will consider the adjustment acceptable for satisfying the baseline equivalence requirement only if (1) the baseline and outcome measures are measured using the same units (based on your response to an earlier question) and (2) the correlation between the baseline and outcome measures exceeds 0.6.

Enter the correlation between the baseline and outcome measures, if reported. This correlation may be used by the SRG for two purposes. First, the correlation is used to assess baseline equivalence when some data are missing or imputed. Second, the correlation is used to assess whether gain scores, difference-in-differences (including the WWC-calculated difference-in-difference adjustment), or fixed effects can be used to satisfy the baseline equivalence requirement when a statistical adjustment for the baseline measure is required.

Finally, select the type of effect size computation used to calculate baseline differences. To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect size. The effect size can be calculated in a number of ways, depending on the information provided in the study. The options for the baseline effect size computations are below:

- **Unadjusted mean and standard deviation.** Select this option if the measure is continuous and the authors report an unadjusted pre-intervention mean and standard deviation.
- **t-stat.** Select this option 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.** Select this option if the authors provide unadjusted pre-intervention means reported for both groups using a dichotomous measure. You do not need to enter standard deviations.

Select “Save and continue” to return to the Measure page, which will now show the baseline measure and whether baseline equivalence is established. Repeat these steps if more than one baseline measure exists for the outcome.

Analysis

Enter the unadjusted and/or adjusted means for the analytic sample in their respective fields. If the study includes missing data, also include the means for the sample of individuals with observed data. If the measure is dichotomous, use the numeric percentage value (0 to 100) for the mean. For example, a 79 percent graduation rate should be entered as 79. Check the “Not reported” checkbox if the unadjusted or adjusted means for the analytic sample are not reported. Enter the unadjusted standard deviations for the analytic sample. Check the “Not reported” checkbox if the unadjusted standard deviations for the analytic sample are not reported.

Select “Yes” or “No” to specify whether a negative result for this measure indicates a favorable outcome. For example, consider an outcome that measures the number of negative behavior incidents a student exhibited. An observed decrease would be a favorable outcome; the student would be exhibiting fewer negative behavior incidents. For other measures, including standardized test scores measuring student achievement, select “No” because a positive result indicates a favorable outcome.

In the Effect Size Computation dropdown list, select the difference computation analysis used in the study. To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect size. The effect size can be calculated in a number of ways, depending on the information provided in the study. When multiple calculation methods are possible for a finding, please select the calculation method appropriate for the most rigorous analysis conducted for that contrast. For example, if both adjusted and unadjusted means are reported, use the adjusted means to calculate the effect size. The options for the effect size computations are below:

- **Unadjusted mean and standard deviation:** Select this option if the authors report unadjusted post-intervention means and standard deviations for both groups. Enter the unadjusted means and standard deviations for the intervention and comparison groups.
- ***t*-stat:** Select this option if the authors report the summary *t*-statistic from a *t*-test comparison of means. Enter the *t*-statistic. 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:** Select this option if the authors report post-intervention means for both groups using a dichotomous measure. Enter the means for the intervention and comparison groups as percentages. You do not need to enter standard deviations.
- **ANOVA *F*-test:** Select this option if the authors report the summary *F*-statistic from a one-way (one-factor) ANOVA. Enter the *F*-statistic. 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.

- **ANCOVA adjusted post-intervention:** Select this option if the authors report adjusted post-intervention means and standard deviations for both groups. Enter the adjusted means and unadjusted standard deviations for the intervention and comparison groups.
- **ANCOVA *F*-test and correlation:** Select this option if the authors report the summary *F*-statistic for the test of the intervention effect from an ANCOVA along with the pre-post correlation. Enter the *F*-statistic and correlation.
- **OLS:** Select this option if the authors report results from an OLS regression. Enter the regression coefficient.
- **HLM level-2 coefficient:** Select this option if the authors report results from an HLM regression that examines impacts at a particular point in time (that is, not a growth-curve analysis). Enter the regression coefficient.
- **Favorable/unfavorable designation only:** Select this option if the study only indicates that the effect is favorable or unfavorable and then select Favorable or Unfavorable from the dropdown menu that appears, as appropriate.

Enter the study-reported effect size, *p*-value, and significance for this outcome. Leave these fields blank if the study does not report an effect size and/or *p*-value or if the study reports the *p*-value as a range (for example, <0.05 or >0.05). The study could report the significance as categorical or quantitative. If only a categorical *p*-value is provided, select whether it was statistically significant.

Select “Yes” or “No” to indicate whether the WWC should use the study-reported effect size, study-reported *p*-value, and study-reported significance to characterize the findings in place of the WWC-calculated values for these items (see Figure 8). In some rare circumstances a study will report a study-reported effect size calculated using the WWC formula for Hedges’ *g* (including the small sample size adjustment). If so, the WWC will use and report the study-reported effect size. The WWC generally accepts the *p*-values and statistical significance for a finding reported by the author(s) of the study. However, there are two common circumstances in which the WWC will compute the statistical significance levels:

- The study does not include statistical significance estimates or there is a known problem with the study calculations.
- The statistical significance levels reported in the study do not account for clustering when there is a mismatch between the unit of assignment and unit of analysis.

The WWC will make an adjustment to the statistical significance of a finding when the study reports multiple estimates of impacts within a single domain, but the reported statistical significance levels do not account for the multiple comparisons. In this scenario, the WWC will apply the multiple comparisons adjustment to the study-reported *p*-values if neither of the above two issues require the WWC to use its own calculation. In that case, the reviewer should indicate that the study-reported *p*-values should be used (the SRG will perform the multiple comparisons adjustment to these study-reported values).

Even when the WWC must calculate its own values to form its official characterization of the study's findings, please include the study-reported data in the SRG. The WWC sometimes uses this information when reporting on the study.

Figure 8. Enter analysis data

Does a negative result indicate a favorable outcome?

✕

Effect size computation ?

- Select calculation -

▼

	Study reported value	Use in place of calculated value? ?
Effect size		Yes No ✕
p-value		Yes No ✕
Is finding significant? ?	Yes No ✕	Yes No ✕

Measure notes

Enter any notes about the measure in the Measure Notes textbox. Be sure to note the table or page on which you found the sample size, analysis, and effect size information. Include a note if the review requires an author query specific to this measure or if you have concerns about the measure. Document any response to the author query that affects this measure. Document whether any calculations for this measure were done outside of the SRG and if so, describe those calculations.

After you have entered all data for the specific measure, click “Save and continue” to return to the review measure page. Click “Save” to save your work at any time. Click “Cancel” to return to the review measures page. The review measures page will now include the measure you just entered in a table with the domain, measure name, comparison, sample, period, and initial rating for the measure.

Follow the above steps to enter all of the measures for the review or edit the information on a measure you have already added. After completing all measures, click “Save and continue” to move to the next section.

3. Aggregating findings

When a study presents findings separately for several groups of sample members without presenting an aggregate result, the WWC will query authors to determine whether they conducted an analysis on the full sample. If the WWC is unable to obtain aggregate results from the authors, the WWC uses an average of subgroups within a study as

the primary finding and presents the subgroup results as supplemental findings. To aggregate findings in the review, select “Aggregate findings” above the Save and Continue button. Then, click the check box for the findings to aggregate. The findings should comprise the same measure, domain, and time period with different, non-overlapping samples. After selecting the findings, click “Save and continue” (see Figure 9). A new finding will be created in the measures table with the sample name “Aggregated sample.” The findings that were aggregated will now be marked as supplemental findings. You can edit the aggregated sample by clicking “Edit” in the finding row. If you need to re-run the aggregate findings, delete the initial aggregate sample by clicking “Edit” in the finding row and then click “Delete” at the bottom of the measures page. You can then re-aggregate the findings.

Figure 9. Select aggregate findings

Select	Domain	Measure
☒	Alphabetics	Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Initial Sound Fluency subtest
☒	Alphabetics	Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Initial Sound Fluency subtest
☐	Reading achievement	Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

4. Main or supplemental findings

After completing the data entry for all of the measures and aggregating measures as necessary, select whether the findings from each measure represent main or supplemental findings. Selecting “Main” will signal that the finding should contribute to the overall study rating. Supplemental findings, such as those using secondary measures, subsamples, subtests, or different follow-up period, will not factor into the rating of the study. For each outcome measure, and among those findings that meet WWC design standards, the WWC uses the following criteria to designate one finding or set of findings as the main finding: (1) includes the full sample; (2) uses the most aggregate measure of the outcome measure (rather than individual subscales); and (3) is measured at a time specified by the protocol (for example, latest follow-up period, earliest follow-up period after conclusion of the intervention, or after one year of exposure). However, not all studies will report a single finding or set of findings that meets these criteria that the WWC can designate as the main finding. When applying these rules is not straightforward because of incomplete information about findings, overlapping samples, or other complications, reviewers have discretion for a study or group of studies under review to identify main and supplemental findings (among those that meet WWC design standards) in a way that best balances the goals of characterizing each study’s findings based on the criteria above and presenting the findings in a clear and straightforward manner, while avoiding overlap in the samples and subscales in the main findings.

C. Rating

The next section of the review is the rating. The system will provide a preliminary rating based on the information entered about the study. A summary by outcome domain appears at the top. Under that, the information for each outcome measure appears, grouped by domain. Adjustments performed by the SRG, such as a difference-in-differences adjustment, will be noted with a footnote. Confirm (1) the rating is correct and the disposition accurately describes the reason for the rating, (2) the information summarized on this page matches the information reported in the study, (3) the correct effect size and significance data are listed in the “official” column for each measure, and (4) any adjustments performed by the SRG appear to be appropriate. You can use the navigation bar across the top to get to the previous sections to correct any information.

D. Context

The SRG proceeds to the Context section after calculating an initial rating. Reviewers will report key sample and study information in the Context section. This information includes the analytic sample size for the main findings; grades, race, ethnicity, and gender of the students in the WWC-reviewed sample; other key sample characteristics, such as English learner status, disability, and free and reduced-price lunch eligibility; characteristics of the sample classroom and school; and location of the study (see Figure 10). Go through each section of the Context menu to confirm you have coded all applicable information reported in the study.

Figure 10. Context menu

Context

For each category, provide the values that match this study. These values should describe the analytic sample.

Sample size	Main analytic sample size 1200
Grade	
Race	Main unit of analysis Student
Ethnicity	
Gender	
Language	
Disability	
Financial position	
Class type	
School type	
Urbanicity	
Region/State	

← Previous Save Save and continue →

For race, the options follow the U.S. Census racial categories. Report what the study reported. Then, enter the remaining percent in the “Unspecified” category to have the coded race equal 100 percent. Ethnicity and gender should equal 100 percent. When available, use demographic data from the analysis sample. Use data from a related sample (the randomized sample or overall school or district) if that is all the study provides. Calculate the combined sample demographics if the study reports demographics separately for the intervention and comparison groups.

After entering all of the applicable information, click “Save and continue” to move to the next section.

E. Narrative

The Narrative section allows the reviewer to describe the setting of the study, study design, sample sizes, sample characteristics, intervention, comparison, and implementation in narrative

form. Each narrative question has a separate Notes field for you to document where in the study you obtained the information or to provide any additional notes.

When you have completed your review of each page and confirmed the information you entered is correct, click “Complete” to finalize your review. At any point in the review, you can save and come back to the review later.