WWC SINGLE STUDY REVIEW PROTOCOL FOR GROUP DESIGN STUDIES THAT ESTIMATE AVERAGE INTERVENTION EFFECTS, VERSION 2.0

What Works Clearinghouse (WWC) Single Study Reviews (SSRs) are designed to provide education practitioners and policymakers with timely and objective assessments of the quality of the research evidence from recently released research papers and reports.

These reviews focus primarily on studies of the effectiveness of education or school-based interventions serving students in the prekindergarten through twelfth grade, as well as those in a postsecondary setting. However, in some instances, they might focus on studies of interventions serving other groups, such as teachers or other school leaders. An SSR may be created any time the WWC reviews a study that is of sufficient importance or interest to make the review available to users. Typically, only studies for which the full study details are not yet accessible in another WWC product, such as an intervention report, will be considered candidates for an SSR. The reports do not represent the results of a systematic review; instead, they provide a summary of one study.

A. Study Eligibility Criteria

All studies eligible for an SSR must examine the effectiveness of an intervention (for example, a program, policy, or practice) intended to directly or indirectly affect academic and/or nonacademic student outcomes. To be eligible for an SSR, a study also must fall in one or more of the following categories:

- Research evidence from recently released research papers and reports whose public release was reported in a major national news source which received a Quick Review
- Randomized controlled trials (RCTs) that have been completed and entered into the National Center for Education Evaluation and Regional Assistance (NCEERA) registry but have not been reviewed previously by the WWC
- New studies that relate to an existing WWC intervention report
- Studies reviewed for WWC practice guides that are influential in determining the recommendations
- Studies that the Institute of Education Sciences (IES) requests be reviewed for other reasons

B. Determination of Review Protocol

All studies identified for SSRs will be reviewed under an existing topic area protocol if the intervention and population group for the study are eligible for review under that protocol. Studies may be reviewed under a combination of protocols if outcomes or populations span more than one topic area. For example, a study may report on the impact of an intervention on both math and reading outcomes, or report on both general and special education students. If a study is not directly eligible for review under a given WWC topic area, the protocol(s) in related areas may still be used to guide selection of eligible outcomes and domain definitions.

C. Outcome Eligibility Criteria

SSRs focus only on eligible outcomes, which are those considered to be educationally relevant. These can include, but are not limited to, student achievement and student behavioral and emotional outcomes. Attitudes and teacher-reported grades/grade point averages (GPAs) are not eligible outcomes for SSRs, unless specified as eligible outcomes under the relevant topic area protocol. Teacher and administrator outcomes are reported only for interventions targeted at teachers, such as professional development or teacher compensation.

SSRs assess and summarize study outcomes by outcome domain, as findings may vary across different types of outcomes. Domains are based on topic area protocols or are developed in consultation with content experts. If a study presents information on a large number of outcomes, it may be useful to use the study's presentation of the main/primary findings as the focal outcomes to include in Appendix C, or to consult with a content expert to determine the subset of outcomes that should be included in the SSR (so as not to penalize the author with a WWC calculated MC adjustment).

A study's rating is based only on outcome measures that are not overaligned with the intervention. Overalignment occurs when the outcome measure includes some of the same materials (such as books or passages) that are used in the intervention or is administered to the intervention group as part of the intervention. In these situations, the study does not provide a good estimate of the effect of the intervention, but rather, exposure to the outcome on which students are assessed.

D. Single Study Review and Reporting Criteria

SSRs provide a summary of one study. In general, the WWC defines a study as an analysis that examines the effect of an intervention on a particular sample (e.g., a set of students, schools, or districts) and on a particular set of outcomes. In some cases, a single manuscript may report on multiple studies (e.g., a study of an intervention tested on middle school and on high school students; or a study of a regular and enhanced implementation of the intervention). In this case, WWC treats each study within the manuscript separately. Results from a single study also may unfold over multiple manuscripts (e.g., a study of a beginning reading program may include reports on immediate and on long-term effects). In this case, WWC selects one set of outcomes and, thus, one manuscript as the primary focus for the review and references others as additional sources.

SSRs are produced only when a study meets WWC standards with or without reservations. A study reviewed under the SSR protocol that does not meet WWC standards is included in the WWC studies' database and listed in a monthly summary of studies reviewed but not meeting standards. One exception is that studies identified for review due to recent media attention will be summarized in an SSR even if the study does not meet WWC standards with or without reservations.

SSRs describe the intervention, study design, and eligible outcomes; rate the study using WWC evidence standards; and summarize the study findings on all eligible outcomes within each domain. Results for eligible subgroups also are included when designs meet WWC standards (with or without reservations).

1. Describing the study design and intervention

SSRs provide a brief description of the study design, including the data sources and methodology used. The intervention description provides a brief overview of the general features of

the intervention and other information relevant for describing the intervention as implemented in the study.

The SSR contains information about why the study was selected to be reviewed in the SSR format. The front page also includes a disclaimer that indicates that the SSR does not represent the results of a systematic review of the literature on a particular intervention, but is a summary of only one study.

2. Rating the study using WWC evidence standards

The study rating is a function of the study design. This protocol pertains to group design studies that estimate average impacts of the intervention, specifically randomized controlled trials (RCTs) and quasi-experimental designs (QEDs).

Two certified reviewers assess a study against WWC evidence standards, and a third certified reviewer reconciles any discrepancies (for more information on WWC evidence standards, see the *WWC Procedures and Standards Handbook*). Reviewers of SSRs use WWC evidence standards to assign one of three possible study ratings:

- *Meets WWC evidence standards without reservations.* This rating applies to well-designed and implemented randomized controlled trials.
- *Meets WWC evidence standards with reservations.* This rating applies to RCTs with high attrition and quasi-experimental studies that establish equivalence of the analytic sample at baseline.
- *Does not meet WWC evidence standards.* This rating applies to studies that provide insufficient evidence of causal validity.

The SSR states the assigned rating and, if appropriate, discusses the study's strengths and weaknesses. When a study uses more than one design or study sample to answer a research question, the SSR focuses on the findings that are based on the strongest design and/or study sample as determined by the WWC design standards. The WWC may assign different ratings for analyses of different outcomes or for findings pertaining to different subgroups. When different ratings are assigned, SSRs provide an explanation of each.

Sample attrition is a key factor in determining the WWC rating for RCTs. Baseline equivalence on measures of the outcome variable or factors correlated with the outcome measure is a key factor in determining the WWC rating for QEDs and RCTs with high attrition.

Attrition in RCTs. The WWC considers both the overall sample attrition rate and the differential in sample attrition between the intervention and comparison groups, as both contribute to the potential bias of the estimated effect of an intervention. The WWC has established conservative and liberal standards for acceptable levels of attrition. The conservative standards are applied in cases where the lead methodologist has reason to believe that relatively more of the attrition may be endogenous to the intervention reviewed—for example, high school students choosing whether or not to participate in a drop-out prevention program. The liberal standards are applied in cases where the lead methodologist has reason to believe that relatively little of the attrition is endogenous to the intervention reviewed—for example, movement of young children in and out of school districts due to family mobility.

Attrition rates are based on the number of sample cases used in the analysis sample with measured, as opposed to imputed, values of the outcome measures. Table 1 presents the maximum difference in the attrition rate for the intervention and comparison groups that is acceptable for a given level of overall sample attrition. The empirical basis for these thresholds is described in the *WWC Procedures and Standards Handbook*.

Table 1. WWC Determinants for Establishing Baseline Equivalence

Overall Attrition	Conservative Boundary	Liberal Boundary	Overall Attrition	Conservative Boundary	Liberal Boundary
0	5.7	10.0	34	3.5	7.4
1	5.8	10.1	35	3.3	7.2
2	5.9	10.2	36	3.2	7.0
3	5.9	10.3	37	3.1	6.7
4	6.0	10.4	38	2.9	6.5
5	6.1	10.5	39	2.8	6.3
6	6.2	10.7	40	2.6	6.0
7	6.3	10.8	41	2.5	5.8
8	6.3	10.9	42	2.3	5.6
9	6.3	10.9	43	2.1	5.3
10	6.3	10.9	44	2.0	5.1
11	6.2	10.9	45	1.8	4.9
12	6.2	10.9	46	1.6	4.6
13	6.1	10.8	47	1.5	4.4
14	6.0	10.8	48	1.3	4.2
15	5.9	10.7	49	1.2	3.9
16	5.9	10.6	50	1.0	3.7
17	5.8	10.5	51	0.9	3.5
18	5.7	10.3	52	0.7	3.2
19	5.5	10.2	53	0.6	3.0
20	5.4	10.0	54	0.4	2.8
21	5.3	9.9	55	0.3	2.6
22	5.2	9.7	56	0.2	2.3
23	5.1	9.5	57	0.0	2.1
24	4.9	9.4	58	-	1.9
25	4.8	9.2	59	-	1.6
26	4.7	9.0	60	-	1.4
27	4.5	8.8	61	-	1.1
28	4.4	8.6	62	-	0.9
29	4.3	8.4	63	-	0.7
30	4.1	8.2	64	-	0.5
31	4.0	8.0	65	-	0.3
32	3.8	7.8	66	-	0.0
33	3.6	7.6	67	-	-

Studies based on cluster random assignment designs must meet attrition standards for both the study sample units that were assigned to intervention or comparison group status (e.g., schools or districts) and the study sample units for analysis (e.g., typically students). In applying the attrition standards to the subcluster level (e.g., students), the denominator for the attrition calculation includes only sample members in the clusters that remained in the study sample.

RCTs with combinations of overall and differential attrition rates that exceed the applicable threshold, based on the applicable standard, must demonstrate baseline equivalence of the analysis sample, or, if nonequivalence falls within the allowable range, statistically control for the nonequivalence, in order to receive the second-highest rating of *meets WWC evidence standards with reservations*.

Baseline Equivalence. RCTs with high attrition and all QEDs must demonstrate baseline (that is, pre-intervention) equivalence between the intervention and comparison groups in the analysis sample, or, in cases with moderate evidence of nonequivalence, statistically control for differences if the groups are nearly equivalent, in order to receive the rating of *meets WWC evidence standards with reservations*.

Baseline equivalence is examined on measures of the outcomes or baseline measures that are expected to be highly correlated with these outcomes (e.g., a test that is the same or highly correlated with the outcome measures). Rules for establishing baseline equivalence should be applied *within each outcome domain*.

Groups are considered equivalent if the reported differences in mean baseline characteristics of the groups are less than or equal to 5% of the pooled standard deviation in the sample. If this is the case, the equivalence standard is met, and the study can receive a rating of *meets WWC evidence standards with reservations* for that domain. Statistical significance of the difference in means is not considered.

Groups are considered to have moderate evidence of nonequivalence if differences are greater than 5% but less than or equal to 25% of the pooled standard deviation in the sample; in this case, the study findings must be based on analytic models that control for the individual-level baseline characteristic(s) on which the groups differ in order to receive a rating of *meets WWC evidence standards with reservations*. Otherwise, the study is rated *does not meet WWC evidence standards* for that domain.

Studies with baseline differences greater than 25% of the pooled standard deviation do not meet the baseline equivalence standard for that domain, regardless of whether or not the impacts are estimated using models that control for baseline characteristics. The study is rated *does not meet WWC evidence standards* for that domain.

Finally, when a study shows evidence of equivalence on a baseline measure of the outcome variable, but there also is evidence that the populations being compared are drawn from very different settings (such as rural versus urban, or high-SES versus low-SES), these settings may be deemed too dissimilar to provide an adequate comparison. In these cases, the study is rated *does not meet WWC evidence standards*.

Author queries. If a study does not provide enough information on key aspects of the study design, such as sample attrition or equivalence of intervention and comparison groups, the WWC submits a query to the study author(s) in order to gather information for use in determining a study

rating. Authors are asked to respond to queries within five business days. When applicable, study authors are advised of the rating the study will be given if the WWC review team does not receive the requested information within the five-day period.

Unless a study focuses its presentation of findings on specific subgroups, we will not query authors for information on subgroups, since these findings are relegated to Appendix D and do not contribute to the study rating for the SSR.

3. Summarizing study findings on key outcomes

SSRs contain a summary of the study findings by domain. They include estimates of the impacts of the intervention in natural units (e.g., means, percentages), as well as in effect size units, and *p*-values as reported by study authors. Where warranted, the SSRs may also indicate that the WWC adjusted a study's estimates of statistical significance to correct for multiple comparisons or clustering or for preexisting differences in baseline characteristics not adjusted for by the authors (see *WWC Procedures and Standards Handbook* for more details).