WHAT WORKS CLEARINGHOUSE
EVIDENCE STANDARDS FOR REVIEWING STUDIES

REVISED MAY 2008

INTRODUCTION

The Institute of Education Sciences (IES) and the What Works Clearinghouse (WWC) have identified topic areas that present a wide range of our nation's most pressing issues in education (e.g., middle school math, beginning reading, and character education). Within each selected topic area, the WWC collects studies of interventions (i.e., programs, products, practices, and policies) that are potentially relevant to the topic area through comprehensive and systematic literature searches. The studies collected are then subjected to a three-stage review process.[1]

First, the WWC screens studies based on their relevance to the particular topic area, the quality of the outcome measures, and the adequacy of data reported. Studies that do not pass one or more of these screens are identified as *Does Not Meet Evidence Screens* and hence excluded from the WWC review.

Second, for each study that meets these initial screens, the WWC assesses the strength of the evidence that the study provides for the effectiveness of the intervention being tested. Studies that provide strong evidence for an intervention's effectiveness are characterized as *Meet Evidence Standards*. Studies that offer weaker evidence *Meet Evidence Standards with Reservations*. Studies that provide insufficient evidence are characterized as *Does Not Meet Evidence Screens*. In order to meet evidence standards (either with or without reservations), a study has to be a randomized controlled trial or a quasi-experiment with one of the following three designs: quasi-experiment with equating, regression discontinuity designs, or single-case designs.[2] The rules for determining the specific evidence category that a study falls under depends on the design of the study, as will be detailed later in the document.

At the third stage, studies that are rated as meeting evidence standards (either with or without reservations) during the second stage are reviewed further to assure consistent interpretation of

[1] The WWC regularly updates WWC technical standards and their application to take account of new considerations brought forth by experts and users. Such changes may result in re-appraisals of studies and/or interventions previously reviewed and rated. Current WWC standards offer guidance for those planning or carrying out studies, not only in the design considerations but the analysis and reporting stages as well. WWC standards, however, may not pertain to every situation, context, or purpose of a study and will evolve.

[2] Randomized controlled trials are studies in which participants are randomly assigned to an intervention group that receives or is eligible to receive the intervention and a control group that does not receive the intervention. Quasi-experimental designs are primarily designs in which participants are not randomly assigned to the intervention and comparison groups, but the groups are equated. Quasi-experimental designs also include regression discontinuity designs and single case designs. Regression discontinuity designs are designs in which participants are assigned to the intervention and the control conditions based on a cutoff score on a pre-intervention measure that typically assesses need or merit. This measure should be one that has a known functional relationship with the outcome of interest over the range relevant for the study sample. Single-case designs are designs that involve repeated measurement of a single subject (e.g., a student or a classroom) in different conditions or phases over time.

study findings and allow comparisons of findings across studies. During this stage, WWC gathers information about variations in participants, study settings, outcomes, and other study characteristics that provide important information about the studies and study findings. Note that the information collected from the third review stage is for consistency in presenting findings from different studies and other descriptive purposes. The information does not affect the rating of the strength of the study determined during the second review stage.

Based on studies that *Meet Evidence Standards* and *Meet Evidence Standards with Reservations*, the WWC produces two types of reports: WWC intervention reports and WWC topic reports. Intervention reports summarize evidence from studies on a specific intervention. Similarly, topic reports summarize evidence from all interventions that qualify for a WWC intervention report in a specific topic area.

Neither the WWC nor the U.S. Department of Education endorses any interventions.

STAGE 1: DETERMINING THE RELEVANCE OF A STUDY TO A WWC REVIEW

OVERVIEW

In each topic area identified by the IES and the WWC, the WWC collects both published and unpublished impact studies that are potentially relevant to the topic. The WWC review team then screens all collected studies to ensure that the studies to be included in a WWC review are eligible for the review based on WWC screening standards and criteria specified in the WWC review protocol developed for each topic area. The main considerations are whether a study was conducted within a relevant timeframe, was focused on an intervention that meets the protocol criteria, included a sample that meets the protocol criteria, used appropriate measures for relevant outcomes, and reported findings adequately.

SCREENING STANDARDS

- **Relevant Timeframe:** The study must have been conducted during a timeframe relevant to the WWC review. For example, according to the WWC review protocol for the topic area of middle school math, only studies conducted after 1983 are eligible for inclusion in the WWC review.
- **Relevant Intervention:** The intervention must be relevant to the WWC review. An intervention designed to improve students' writing skills, for example, is not a relevant intervention for the topic area of beginning reading. In contrast, a study of an intervention designed to improve vocabulary would be.
- **Relevant Sample:** The study's sample must be relevant to the WWC review. In the topic area of beginning reading, for example, a relevant study sample has to consist of students in grades K–3.
- **Relevant Outcome:** The study must report on at least one outcome relevant to the WWC review. Student engagement, for example, is not considered a relevant outcome for interventions in middle school math, which focuses on achievement outcomes.

- **Adequate Outcome Measure:** The measure used must be able to reliably measure a relevant outcome that it is intended to measure.[3] For example, a nationally normed, validated test of math computation skills would be an adequate measure of math skills. In contrast, a self-report of math competency would not be considered a reliable measure of math competency.

- **Adequate Reporting:** It must be possible to calculate the effect size for at least one adequate measure of a relevant outcome. In the simplest randomized controlled trial, for example, this requires the study report means and standard deviations of the outcomes for the intervention and comparison groups respectively, and usually the sample sizes for the intervention and comparison groups.

    - By default, the WWC calculates effect sizes using the pooled standard deviation. If the pooled standard deviation is not available, the standard deviation for the comparison group, if available, will be used to calculate the effect sizes.

    - For studies that report effect sizes but do not provide data for computing the effect sizes, the WWC will report the effect sizes presented in the study unless there is reason to cast them in doubt (e.g., unusually large effect sizes).

[3] The study author must provide the title of the test and one or more of the following: (1) documentation that the test items are relevant to the topic, (2) a description of the test items that is sufficient to demonstrate that the items are relevant to the topic, or (3) evidence of test reliability.

STAGE 2: ASSESSING THE STRENGTH OF THE EVIDENCE THAT A STUDY PROVIDES FOR THE INTERVENTION'S EFFECTIVENESS

OVERVIEW

The WWC reviews each study that passes the preceding screens to determine whether the study provides strong evidence (*Meets Evidence Standards*), weaker evidence (*Meets Evidence Standards with Reservations),* or insufficient evidence (*Does Not Meet Evidence Screens*) for an intervention's effectiveness. Studies that *Meet Evidence Standards* are well-designed and implemented randomized controlled trials. Studies that *Meet Evidence Standards with Reservations* are quasi-experiments with equating[4] and no severe design or implementation problems, or randomized controlled trials with severe design or implementation problems. The evidence standards for two special cases of quasi-experimental designs, regression discontinuity designs and single-case studies, are under development as of September 2006.

EVIDENCE STANDARDS

Study Design: In order for a study to be rated as meeting evidence standards (with or without reservations), it must employ one of the following types of research designs: a randomized controlled trial or a quasi-experiment (including quasi-experiments with equating, regression discontinuity designs, and single-case designs).

[4] Equating may be done either through matching to make the study groups comparable in terms of important pre-intervention characteristics, or through statistical controls during the analysis stage to adjust for pre-intervention difference between the study groups, or both.

*If the study appears to be a **randomized controlled trial (RCT)**, the following rules are used to determine whether the study Meets Evidence Standards or Meets Evidence Standards with Reservations.*

- **Randomization:** For an RCT to *Meet Evidence Standards*, the study participants (e.g., students, teachers/classrooms, or schools) should have been placed to each study condition through random assignment or a process that was haphazard and functionally random.
    - For studies received by the WWC prior to December 31, 2006: If the study authors used the term "random assignment" but gave no other indication of how the assignment procedure was carried out, the label is assumed to have been properly applied unless there is reason to doubt this claim.
    - For studies received by the WWC beginning January 1, 2007: For the sample allocation to be considered "random assignment," the study authors must report specifics about the randomization procedure, including: (a) details about how the assignment sequence was generated, (b) information about the role of the person who generated the sequence, and (c) methods used to conceal the sequence until participants were assigned to conditions.
    - Examples of haphazard assignment that *might* be functionally random include: alternating by date of birth (e.g., January 5 is placed into group A, January 7 is placed into group B, and January 13 is placed into group A); and alternating by the last digit of an identification code (e.g., "evens" are placed into group A, "odds" are placed into group B). Examples of haphazard assignment that are *unlikely* to be functionally random include: placing birth months January–June into group A, birth months July–December into group B; and using scheduling software to assign students to conditions.

If the assignment process in an RCT is truly random or functionally random as described above, the RCT *Meets Evidence Standards.* If the study has high levels of overall or differential attrition, it cannot receive the top rating.

- **Overall Attrition:** Attrition is defined as a failure to measure the outcome variable on all the participants initially assigned to the intervention and comparison groups. High overall attrition generally makes the results of a study suspect, although there may be rare exceptions.
- **Differential Attrition:** Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the comparison groups. Severe differential attrition makes the results of a study suspect because it may compromise the comparability of the study groups.

If the study has high levels of overall or differential attrition, it should demonstrate baseline equivalence of the post-attrition analysis samples to receive the *Meets Evidence Standards with Reservations* rating.

- **Baseline Equivalence:** The groups should have been equated on a pretest (or a proxy of the pretest) of the outcome measure and across any other characteristics identified in the WWC review protocol for the topic area.

If the study has high levels of overall or differential attrition and does not demonstrate baseline equivalence, it *Does Not Meet Evidence Standards.* However, if statistical adjustment was used to account for these differences in the analysis, the Principal Investigator for the topic area has discretion to determine whether the study *Meets Evidence Standards with Reservations*.

- **Statistical Adjustment:** The use of statistical procedures (e.g., covariate adjustment in an ANCOVA) to equate groups on pretest may address baseline incomparability in the impact analysis.

- **Intervention Contamination:** Intervention contamination occurs when something happens after the beginning of the intervention and affects the outcome for the intervention or the comparison group, but not both. For an RCT to *Meet Evidence Standards,* there should be no evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants.[5]

    - If there is evidence of intervention contamination, the study *Meets Evidence Standards with Reservations*.

- **Teacher-Intervention Confound:** A teacher-intervention confound occurs when only one teacher is assigned to each condition.[6] For an RCT to *Meet Evidence Standards*, there should be more than one teacher assigned to each condition or, if there is only one teacher per condition, there should be strong evidence that teacher confound problem is negligible.[7]

    - If there is only one teacher per condition and there is no evidence that teacher effects are negligible, the study *Does Not Meet Evidence Screens*.

    - If there is only one teacher per condition and there is evidence that teacher effects are minimal but not negligible, the study *Meets Evidence Standards with Reservations*.

- **Mismatch Between Unit of Assignment and Unit of Analysis:** Some RCTs may be designed and implemented well, but the analysis of data may be incorrect. A common problem is that the units of random assignment may not match up with the units of analysis and this feature of the study design is ignored in the analysis. Ignoring this fact may lead to inflated estimates of the statistical significance of study findings.

---

[5] Intervention contamination poses a threat to the validity of the evidence for an intervention's effects in that the observed difference between the intervention and the comparison groups may not be entirely attributable to the intervention, but may reflect the effect of the contaminant.

[6] This standard also applies to studies with assignment at the level of other aggregated units, such as classrooms, schools or districts, in which only one aggregated unit is assigned to each condition.

[7] See technical guidance on teacher-intervention confound for more details.

Mismatch does not affect the rating given to a study; that is, it does not affect the statement about meeting evidence standards because the standards rely solely on the design rather than the data analysis of the study. Nevertheless, WWC reports need to recognize the mismatch problem and adjust the estimates of statistical significance when it occurs.

*If the study appears to use a **quasi-experimental design (QED) with equating**, use the following rules to determine whether the study Meets Evidence Standards with Reservations or Does Not Meet Evidence Screens.*

- **Group Assignment:** Studies in which participants were placed into groups using procedures other than random assignment or a cutoff score on a pre-intervention measure are assumed to *Meet Evidence Standards with Reservations*, unless one or more of the following conditions is violated:
- **Equating and Baseline Equivalence:** The groups should have been equated on a pretest (or a proxy of the pretest) of the outcome measure and across any other characteristics identified in the WWC review protocol for each topic area through matching and/or statistical adjustment to establish baseline equivalence.
    - Equating accomplished through matching involves creating or identifying intervention and comparison groups that "look" similar on a pretest of the outcome measure.
    - Equating accomplished through statistical adjustment involves using statistical procedures (e.g., covariate adjustment in an ANCOVA) to equate groups on pretest and address baseline incomparability in the impact analysis. If there was baseline incomparability that was not accounted for in the analysis, the study *Does Not Meet Evidence Screens*.
    - If the groups appeared to be patently incomparable at baseline,[8] and the incomparability was unlikely to be adequately addressed through statistical adjustment, the study *Does Not Meet Evidence Screens.*
- **Overall Attrition:** For a QED to Meet Evidence Standards with Reservations, there should not be a severe overall attrition problem or, if there was, it should have been accounted for in the analysis.
    - Severe overall attrition (if not too extreme) can be addressed by demonstrating post-attrition equivalence of the groups. If addressed in this way, the study is not downgraded.
    - Random attrition (e.g., random selection of several students from a class to test) is not considered a threat to internal validity, and does not contribute to severe overall attrition.

[8] The PI and the Review Team for a given topic area have the discretion to determine whether the baseline incomparability in a study was too substantial to be adequately adjusted. The decision rules for handling such studies will be documented and justified.

- If there was severe overall attrition that cannot be discounted on the basis of evidence, the study *Does Not Meet Evidence Screens*.

- **Differential Attrition:** For a QED to *Meet Evidence Standards with Reservations,* there should not have been a severe differential attrition problem or, if there was, it should have been accounted for in the analysis.

  - Severe differential attrition (if not too extreme) can be addressed by demonstrating post attrition equivalence of the groups. If addressed in this way the study is not downgraded.

  - If there was severe differential attrition that cannot be discounted on the basis of evidence, the study *Does Not Meet Evidence Screens*.

- **Intervention Contamination:** There should be no evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants.

  - If there is evidence of an intervention contamination, the study *Does Not Meet Evidence Screens*.

- **Teacher-Intervention Confound:** A teacher-intervention confound occurs when only one teacher is assigned to each condition.[9] For a QED to *Meet Evidence Standards with Reservations*, there should be more than one teacher assigned to each condition or, if there is only one teacher per condition, there should be strong evidence that teacher effects on the findings are negligible.[10]

  - If there is only one teacher per condition and there is no evidence that teacher effects are negligible, the study *Does Not Meet Evidence Screens*.

- **Mismatch Between Unit of Assignment and Unit of Analysis:** Some QEDs may be designed and implemented well but the analysis of data may be incorrect. A common problem is that the units of random assignment may not match up with the units of analysis and this feature of the study design is ignored in the analysis. Ignoring this fact leads to inflated estimates of the statistical significance of study findings.

  Mismatch does not affect the rating given to a study; that is, it does not affect the statement about meeting evidence standards because the standards rely solely on the design rather than the data analysis of the study. Nevertheless, WWC reports need to recognize the mismatch problem and correct the estimates of statistical significance when it occurs.

[9] This standard also applies to studies with assignment at the level of other aggregated units, such as classrooms, schools or districts, in which only one aggregated unit is assigned to each condition.

[10] See technical guidance on teacher-intervention confound for more details.

STAGE 3: IDENTIFYING OTHER IMPORTANT CHARACTERISTICS OF A STUDY THAT MEETS EVIDENCE STANDARDS (WITH OR WITHOUT RESERVATIONS)

OVERVIEW

All studies that pass the evidence standards and are rated as either *Meets Evidence Standards* or *Meets Evidence Standards with Reservations* during the second review stage are further reviewed to describe other important study characteristics. The purpose of the Stage 3 review is to collect contextual information about the studies that provide evidence for the effectiveness of the interventions being tested, and to aid the interpretation of the findings presented in the WWC intervention and topic reports. The additional information collected during the third review stage does not affect the ratings of the studies on the evidence standards (i.e., *Meets Evidence Standards*, *Meets Evidence Standards with Reservations*, or *Does Not Meet Evidence Screens),* which are determined during the second review stage.

OTHER STUDY CHARACTERISTICS

- **Variations in People, Settings, and Outcomes**[11]
  - Subgroup Variation: What subgroups were included in the study?
  - Setting Variation: In what settings did the study take place?
  - Outcome Variation: What outcomes were measured in the study? Which outcome domains did the outcome measures pertain to according to the outcome domain classification specified in the WWC review protocol for each topic area?
- **Analysis of Intervention's Effects on Different Subgroups, Settings, and Outcomes**
  - Analysis by Subgroups: For what subgroups were effects estimated?
  - Analysis by Setting: For what settings were effects estimated?
  - Analysis by Outcome Measures: For what outcome measures and outcome domains were effects estimated?

[11] Information about the variations in people, settings, and outcomes of the studies as well as information about analysis within subgroups will help to assess the generalizability of the study findings.

- **Statistical Reporting**
  - Complete Reporting: Are findings reported for most of the important measured outcomes?[12]
  - Relevant Statistics Reported: Are the following statistics reported: intervention and comparison group posttest means and standard deviations, posttest mean differences, sample sizes, pretest means, statistical significance levels of the posttest mean differences?
  - Covariate Adjustments: Are outcome measures adjusted for differences in pretest or other important pre-intervention differences between the intervention and comparison group?

[12] The purpose of this question is to assess the extent to which the study findings are biased by potential selective reporting, as reported findings are more likely to demonstrate positive intervention effects than findings from the same study that are not reported by the study authors.

What Works Clearinghouse
Study Design Classification

Revised September 2006

To be eligible for WWC review, a study must be a randomized controlled trial or a quasi-experiment. An eligible quasi-experiment must be one of the following three designs: quasi-experiment with equating on pretest, regression discontinuity design, or single-case design with multiple changes of condition. The questions and examples below are meant to help WWC staff to classify properly the design of each study potentially relevant to WWC review.[1]

Is this study a randomized controlled trial?

1. Was random assignment used to place participants into different study groups?

    - For studies received by the WWC prior to December 31, 2006: If the study authors used the term "random assignment" but gave no other indication of how the assignment procedure was carried out, the label is assumed to have been properly applied unless there is reason to doubt this claim.[2]

    - For studies received by the WWC beginning January 1, 2007: For the sample allocation to be considered "random assignment," the study authors must report specifics about the randomization procedure, including: (a) details about how the assignment sequence was generated (e.g., use of a random number table or generator, coin flip, roll of a die), (b) information about the role of the person who generated the sequence, and (c) methods used to conceal the sequence until participants were assigned to conditions.

    - Occasionally, researchers will use the term "random assignment" when they really mean "random selection." Alternatively, they may use the term "random selection" to mean "random assignment." Coders should examine closely the context of the language used in the report for evidence of these types of confusion.

    - Occasionally, researchers will use matching, blocking, or stratifying *before* randomization in order to minimize group differences on a variable or set of variables. Coders should closely examine studies to ensure that these are classified properly as randomized controlled trials.

[1] The WWC regularly updates WWC technical standards and their application to take account of new considerations brought forth by experts and users. Such changes may result in re-appraisals of studies and/or interventions previously reviewed and rated. Current WWC standards offer guidance for those planning or carrying out studies, not only in the design considerations but the analysis and reporting stages as well. WWC standards, however, may not pertain to every situation, context, or purpose of a study and will evolve.

[2] Reasons to doubt the claim of randomization include the following: (1) the assignment procedure was described and it resembles one of the strategies identified as "not functionally random" (see below) or (2) the sample sizes for the intervention and comparison conditions are markedly different at the level of assignment.

2. If a randomization procedure was not used, were participants placed into intervention groups using a process that was haphazard and functionally random?

    - Examples of haphazard assignment that *might* be functionally random include: (a) alternating by date of birth (e.g., January 5 is placed into group A, January 7 is placed into group B, and January 13 is placed into group A); (b) alternating alphabetically by last name (e.g., Acosta is placed into group A, and Aguilera is placed into group B); and (c) alternating by the last digit of an identification code (e.g., "evens" are placed into group A, and "odds" are placed into group B).

    - Examples of haphazard assignment that are *unlikely* to be functionally random include: (a) placing birth months January – June into group A, birth months July – December into group B; (b) placing participants with a last name beginning with A-M into group A, and last names beginning with N-Z into group B; (c) placing the first 20 arrivals into group A, and the last 20 arrivals into group B, and (d) using scheduling software to assign students to groups.[3]

    - Because it is often difficult to determine what is functionally random and what is not, the WWC's Principal Investigators (PIs) and Technical Review Team (TRT) should weigh in whenever this decision is not clear cut.

An answer of "yes" to either of these questions leads to a categorization of the study as a randomized controlled trial. If the categorization is based on haphazard assignment, it will be noted in the write-up of the intervention report.

Is this study a quasi-experiment with equating on pretest?

1. Were participants placed into different study groups on a non-random basis?

2. Were the groups equated on a pretest (or a proxy of the pretest) of the outcome measure and across any other characteristics identified in the WWC review protocol for each topic area?

[3] For the WWC to consider student assignment based on scheduling software functionally random, the study author would need to demonstrate that the assignment of students to conditions was independent of students' other interests and course selections. For example, class scheduling software might be used to produce random samples in these two situations: (1) The scheduling system is used with no pre-specified conditions (e.g., no classes or students with certain characteristics were entered into the system before other students were assigned to groups) OR (2) The sample was limited to students who were not affected by scheduling parameters or constraints (e.g., if gym, band, and art classes were already set in the scheduling system, the random assignment of students who did not take gym, band, or art classes may produce a functionally random sample).

- Equating can be accomplished through:

  — Matching. This involves creating or identifying intervention and comparison groups that "look" similar on a pretest of the outcome measure and across any other characteristics identified in the WWC Review Protocol for each topic area. Because adequate matching may not be easy to accomplish, the WWC's PIs and TRT should be consulted to determine whether the matching method used in a particular study resulted in adequate equating on pretest.

  — Statistical adjustment. This involves using statistical procedures (e.g., covariate adjustment in an ANCOVA) to equate groups on a pretest measure of the outcome.

- Timing of equating: Groups may be identified and matched before the intervention was implemented or prior to analysis after implementation. Groups may also be statistically equated during analysis.

- Timing of pretest: The pretest may be administered at baseline, or it may be administered quite some time before the intervention was implemented (e.g., collected from achievement testing the previous year).

- Sample pretested: Under limited conditions, the pretest used in equating may come from a preceding cohort of the students that comprise a larger unit of intervention delivery. For example, for interventions where the school is the unit of intervention delivery (e.g., a school-wide math curriculum), the pretest or "baseline" data may come from achievement testing at the school during the year that preceded the intervention's implementation. In this case, the pretest data would not come from the student cohort in the study sample, but from a different cohort of students who were in a given grade in the year before the intervention was implemented. Only when the unit of intervention delivery is at the school level or higher is this approach acceptable. The timing and characteristics of the pretest should be noted during coding.

If the answer to both of these questions is "yes," then the study meets the WWC's definition of a quasi-experiment with equating.

What Works Clearinghouse Extent of Evidence Categorization

The Extent of Evidence Categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and sizes of studies. This scheme has two categories: small and moderate/large.

- The extent of evidence is moderate/large:
  - The domain includes more than one study; AND
  - The domain includes more than one school; AND
  - The domain findings are based on a total sample size of at least 350 students OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.

- The extent of evidence is small:
  - The domain includes only one study; OR
  - The domain includes only one school; OR
  - The domain findings are based on a total sample size of less than 350 students AND, assuming 25 students in a class, a total of less than 14 classrooms across studies.

Each intervention domain receives its own categorization. For example, each of the three domains in character education—behavior; knowledge, attitudes, and values, and academic achievement—receives a separate categorization.

Example:
Intervention Do Good, a character education intervention, had three studies that met WWC standards and were included in the review. All three studies reported on academic achievement. There were a total of 6 schools across the three studies. The first study reported testing on 150 students, the second study 125 students, and the third study reported testing 4 classes with 15 students in each class. The extent of evidence on academic achievement for the Do Good intervention is considered "moderate/large" – it met the condition for both the number of studies and the number of schools, and although the total number of students is less than 350 (150+125+(4*15)=335), the number of classes exceeded 14 (150/25+125/25+4=15).

A "small" extent of evidence indicates that the amount of the evidence is low. There is currently no consensus in the field on what constitutes a "large" or "small" study or database. Therefore, the WWC set the conditions above based on the following rationale:

- When there is only one study, there is the possibility that some characteristics of the study—the outcome instruments, the timing of the intervention, etc.—might have affected the findings. When there are multiple studies, especially if they differ, provide some assurance that the effects can be attributed to the intervention, and not some features of the particular place where the intervention was studied. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only one setting.
- Similarly, when there is only one school, there is a possibility that some characteristics of the school—the principal, student demographics, etc.—might have affected the findings or are

intertwined or confounded with the findings. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only a single school.

- The sample size of 350 was derived from the following assumptions:
  - a balanced sampling design that randomizes at the student level,
  - a minimum detectable effect size of 0.3,
  - the power of the test at 0.8,
  - a two-tailed test with an alpha of 0.05, and
  - the outcome was not adjusted by an appropriate pretest covariate.

The Extent of Evidence Categorization provided in recent reports, and described here, signals WWC's intent to eventually provide a rating scheme on the external validity, or the generalizability, of the findings, for which the extent of evidence is only one of the dimensions. The Extent of Evidence Categorization, in its current form, is not a rating on external validity; instead, it serves as an indicator that cautions readers when findings are drawn from studies with small size samples, a small number of school settings, or a single study.

What Works Clearinghouse Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates the effect size (see the WWC Effect Size Technical Paper) of the intervention's effect into an "improvement index." The improvement index represents the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the control group mean (i.e., 50th percentile) in the control group distribution. Alternatively, the improvement index can be interpreted as the expected change in percentile rank for an average control group student if the student had received the intervention.

As an example, if an intervention produced a positive impact on students' reading achievement with an effect size of 0.25, the effect size could be translated to an improvement index of 10 percentile points. We could then conclude that the intervention would have led to a 10% increase in percentile rank for an average student in the control group, and that 60% (10% + 50% = 60%) of the students in the intervention group scored above the control group mean.

Specifically, the improvement index is computed as follows:

1. Compute Cohen's U3 index that corresponds to the effect size estimate.

The U3 index represents the percentile rank of a control group student who performed at the level of an average intervention group student. An effect size of 0.25, for example, would correspond to a U3 of 60%, which means that an average intervention group student would rank at the 60th percentile in the control group. Equivalently, an average intervention group student would rank 10 percentile points higher than an average control group student, who, by definition, ranks at the 50th percentile.

Mechanically, the conversion of an effect size to a U3 index entails looking up on a table that lists the proportion of area under the standard normal curve for different values of z-scores, which can be found in the appendices of most statistics textbooks. For a given effect size, U3 has a value equal to the proportion of area under the normal curve below the value of the effect size—under the assumptions that the outcome is normally distributed and that the variance of the outcome is similar for the intervention group and the control group.

2. Compute the improvement index as (U3 – 50%).

Given that U3 represents the percentile rank of an average intervention group student in the control group distribution, and that the percentile rank of an average control group student is 50%, the improvement index, defined as (U3 – 50%), would represent the difference in percentile rank of an average intervention group student and an average control group student in the control group distribution.

What Works Clearinghouse Intervention Rating Scheme

Factors Determining the Rating

Explicit heuristics will be applied to support two judgments about the findings of each qualifying study about a given outcome (or outcome domain) for a given intervention:

1. The direction, magnitude, and statistical significance of the empirical effect estimate. This will be characterized as a ***statistically significant positive, substantively important positive, indeterminate***, or ***statistically significant negative*** effect.
2. The quality of the research design generating the effect estimate. This will be characterized as a ***strong*** or ***weak*** design. (See the WWC Study Review Standards for further details.)

The rating scheme based on these two factors is presented below. After that are the detailed descriptions and heuristics for making the judgments on these factors for each study and outcome.

Rating Scheme Based on These Judgments

Positive Effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant positive* effects, at least one of which met WWC evidence standards for a *strong* design.
- No studies showing *statistically significant* or substantively important *negative* effects.

***Potentially Positive Effects*:** Evidence of a positive effect with no overriding contrary evidence.

- At least one study showing a *statistically significant* or *substantively important positive* effect.
- No studies showing a *statistically significant* or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing *statistically significant* or *substantively important positive* effects.

Mixed Effects: Evidence of inconsistent effects as demonstrated through either of the following.

- At least one study showing a *statistically significant* or *substantively important positive* effect; AND at least one study showing a *statistically significant* or substantively important *negative* effect, but no more such studies than the number showing a *statistically significant* or *substantively important positive* effect.
- OR, at least one study showing a *statistically significant or substantively important* effect AND more studies showing an *indeterminate* effect than showing a *statistically significant or substantively important* effect.

No Discernible Effects: No affirmative evidence of effects.

- None of the studies shows a *statistically significant or substantively important effect, either positive or negative.*

Potentially Negative Effects: Evidence of a negative effect with no overriding contrary evidence.

- At least one study showing a *statistically significant* or substantively important *negative* effect.

- No studies showing a *statistically significant* or *substantively important positive* effect OR more studies showing *statistically significant* or substantively important *negative* effects than showing statistically *significant* or *substantively important positive* effects.


Negative Effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant negative* effects, at least one of which is based on a *strong* design.
- No studies showing *statistically significant* or *substantively important positive* effects.

Evidence Base and Heuristic Rules

Points of Evidence

For each study of the intervention and for each outcome, the following points of evidence are the basis for characterizing the empirical findings:

1. Quality of the study design: RCT (meets evidence standards) or QED (meets evidence standards with reservations) under the current WWC criteria.
2. Effect size: A single effect size or, in the case of multiple measures of the specified outcome, either (i) the mean effect size, or (ii) the effect size for each individual measure within the domain. The effect size is defined as the standardized mean difference (i.e., the difference between the student-level posttest means on an outcome variable divided by the pooled standard deviations, either calculated directly or derived from other appropriate statistics, corrected for small sample sizes).
3. Sample size: The number of units of assignment per condition and the number of students in those units per condition if students were not the units of assignment.
4. Statistical significance of the effect based on a correct ("aligned") analysis if reported. Statistical significance is assumed to mean the conventional alpha=.05, two-tailed for single measures and for mean effects within each domain. When multiple hypothesis tests are performed using the number of measures greater than one (m>1measures) within each domain, the Benjamini Hochberg procedure may be used to correct for multiple comparisons and identify statistically significant effects for individual measures (Benjamini, Y., and Y. Hochberg, *Journal of the Royal Statistical Society* 1995, Vol. 57, No.1, 289-300 [http://www.math.tau.ac.il/~ybenja/MyPapers.html]).

Characterizing the Quality of the Research Design Generating the Effect Estimates

The heuristics for categorizing the quality of the research design used in a given study are as follows:

- ***Strong design:*** designs that meet the WWC's evidence standards, which are RCTs without severe design or implementation problems.)
- ***Weak design:*** designs that meet WWC's evidence standards with reservations, which include RCTs with severe design or implementation problems, and QEDs with equating and without severe design or implementation problems.

(See the WWC Study Review Standards for further details.)

Characterizing the Direction and Magnitude of the Empirical Effect Estimate

These heuristics are applied to the outcome variable(s) identified by the Principal Investigator (PI) as relevant to the review. The PI may choose to ignore some variables if they are judged

sufficiently peripheral or unrepresentative and consider only the remaining ones. Similarly, if the PI judges that there is one core variable with all the others secondary or subsidiary, only that one may be considered.

A. Definitions and Suggested Defaults

The heuristics in the next section require that values be set for certain terms. These terms and associated procedures are defined below with suggested default values.

Minimum effect size. The smallest positive value at or above which the effect is deemed substantively important with relatively high confidence for the outcome domain at issue. Effect sizes at least this large will be taken as a qualified positive effect even though they may not reach statistical significance in a given study. The suggested default value is a student-level effect size greater than or equal to 0.25 (ES $\geq$ 0.25), corresponding to a 10 percentile point difference between the percentile rank of the average student in the comparison group (50^{th} percentile) and the percentile rank of the average student in the intervention group (60^{th} percentile) based on the comparison group distribution. The PI may set a different default if explicitly justified in terms of the nature of the intervention or the outcome domain. A similar default applies in the negative direction. The suggested default value for a minimum negative effect is a student-level effect size less than or equal to –0.25 (ES $\leq$ -0.25).

t test adjusted for clustering. A *t* test applied to the effect size (or mean effect size in cases of multiple measures of the outcome) that incorporates an adjustment for clustering. This procedure allows the reviewer to test the effect size directly in cases where a misaligned analysis is reported. (Computational details are provided in the appendix.) However, the clustering adjustment requires specifying an ICC value. The suggested default ICC value for achievement outcomes is .20. The suggested default ICC for behavioral and attitudinal outcomes is .10. The PI may set different defaults if explicitly justified in terms of the nature of the research circumstances or the outcome domain.

B. Heuristics for Characterizing Effects of a Study

(Note: The italicized terms involve default values and are defined above.)

Statistically significant positive effect: Any one of the following:

If the analysis as reported by the study author is properly aligned:

- For a single outcome measure within an outcome domain, either of the following is appropriate. (If the results differ, select the strategy which demonstrates significance.)
  - The effect is reported as positive and statistically significant.
  - The effect size is positive and statistically significant when tested using a *t test adjusted for clustering.*
- For multiple measures of outcomes within an outcome domain, any of the following are appropriate. (If the results differ, select the strategy that demonstrates statistical significance.)
  - Univariate statistical tests are reported for each outcome measure and *at least half* of the effect sizes are positive and statistically significant and *no* effect sizes are negative and statistically significant, ignoring multiple hypothesis tests.

- The omnibus effect for all the outcome measures together is reported as positive and statistically significant on the basis of a multivariate statistical test.
- Univariate statistical tests are reported for each outcome measure and the effect size for *at least one* measure within the domain is positive and statistically significant and *no* effect sizes are negative and statistically significant, *when accounting for clustering and for multiple hypothesis tests within the domain*.
- The *mean* effect size for the multiple measures of the outcome is positive and statistically significant when tested using a *t test adjusted for clustering.*[1]

If the analysis as reported by the study author is not properly aligned, either of the following is appropriate:

- The effect size or the *mean* effect size (if multiple measures of outcomes within a domain) is positive and statistically significant when tested using a *t test adjusted for clustering*.
- Univariate statistical tests are reported for each outcome measure and *at least one* effect size is positive and statistically significant and *no effect sizes* are negative and statistically significant, *accounting for clustering and multiple comparisons within the domain*.

Substantively important positive effect:

- The effect size is not statistically significant in any of the senses described above, but the student-level effect size (if there was a single student-level measure within an outcome domain) or the *mean* effect size based on multiple student-level findings (if there were multiple student-level measures within an outcome domain) is equal to or greater than the *minimum effect size*.[2]

Indeterminate effect:

- The effect size is not statistically significant and does not qualify as a substantively important positive effect as defined above (that is, the effect size or the mean effect size is less than the *minimum effect size*).

Substantively important negative effect:

- The effect size is not statistically significant in any of the senses described above, but the student-level effect size (if there was a single student-level measure within an outcome domain) or the *mean* effect size based on multiple student-level findings (if there were multiple student-level measures within an outcome domain) is equal to or less than the *minimum negative effect size*.[2]

[1] Note that this formula is still acceptable if there is no clustering, as the clustering term drops out of the equation.
[2] Note that this criterion, as well as the default *minimum effect size*, is entirely based on student-level ESs. Cluster-level ESs are ignored for the purpose of the rating scheme because they are based on a different ES metric than the student-level ESs, and therefore not comparable with student-level ESs. Moreover, cluster-level ESs are relatively rare, and there is not enough knowledge in the field yet to set a defensible *minimum effect size* for cluster-level ESs.

Statistically significant negative effect: Any one of the following where no statistically significant or substantively important positive effect has been detected (in the sense outlined above):

If the analysis as reported by the study author is properly aligned:

- For a single outcome measure within an outcome domain, either of the following is appropriate. (If the results differ, select the strategy that demonstrates significance.)
  - The effect is reported as negative and statistically significant.
  - The effect size is negative and statistically significant when tested using a *t test adjusted for clustering.*
- For multiple measures of outcomes within an outcome domain, any of the following is appropriate. (If the results differ, select the strategy which demonstrates significance.)
  - Univariate statistical tests are reported for each outcome measure and *at least half* of the effect sizes are negative and statistically significant and *no* effect sizes are positive and statistically significant, ignoring multiple hypothesis tests.
  - The omnibus effect for all the outcome measures together is reported as negative and statistically significant on the basis of a multivariate statistical test.
  - Univariate statistical tests are reported for each outcome measure, and *at least one* effect size is negative and statistically significant and *no* effect sizes are positive and statistically significant, *accounting for clustering and multiple comparisons within the domain*.
  - The *mean* effect size for the multiple measures of the outcome is negative and statistically significant when tested using a *t test adjusted for any clustering.*[3]

If the analysis as reported by the study author is not properly aligned, either of the following is appropriate:

- The effect size or the *mean* effect size (if multiple measures of outcomes within an outcome domain) is negative and statistically significant when tested using a *t test adjusted for clustering*.
- Univariate statistical tests are reported for each outcome measure and *at least one effect size* is negative and statistically significant and *no* effect sizes are positive and statistically significant, *accounting for clustering and multiple comparisons within the domain*.

[3] Note that this formula is still acceptable if there is no clustering, as the clustering term drops out of the equation.

Appendix: Computational details for the *t test adjusted for clustering*

To determine if it is plausible that the effect size in a study with a misaligned analysis is statistically significant

(1) The reviewer has:

N_T, N_C, and $N = N_T + N_C$
(student-level sample sizes for the intervention and comparison groups respectively)

$m = m_T + m_C$
(number of clusters—classrooms or schools)

$ES = (X_T - X_C)/S_p$
(effect size computed from student level means and SDs with no attention to clustering)

(2) A default rho is assumed (current default is ρ=.20)

(3) The *t* statistic is computed for the effect size ignoring clustering:

$$t = ES\sqrt{\frac{N_T N_C}{N_T + N_C}}$$

(4) The *t* value above is corrected for clustering using the default rho and assuming equal *n* in each cluster:

$$t_A = ct \quad \text{where } c = \sqrt{\frac{(N-2) - 2\left(\frac{N}{m} - 1\right)\rho}{(N-2)[1 + \left(\frac{N}{m} - 1\right)\rho]}}$$

(5) Adjusted degrees of freedom are calculated:

$$h = \frac{\left[(N-2) - 2\left(\frac{N}{m} - 1\right)\rho\right]^2}{(N-2)(1-\rho)^2 + \frac{N}{m}\left(N - 2\frac{N}{m}\right)\rho^2 + 2\left(N - 2\frac{N}{m}\right)\rho(1-\rho)}$$

(6) Significance is determined in the usual way using adjusted t_A with adjusted *df=h*

Technical Details of WWC-Conducted Computations

(4-16-2007)

To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect sizes (ES) and the improvement indices associated with study findings on outcome measures relevant to the WWC's review. In general, the WWC focuses on ESs based on student-level findings regardless of the unit of assignment or the unit of intervention. Focusing on student-level findings not only improves the comparability of ES estimates across studies, but also allows us to draw upon existing conventions among the research community to establish the criterion for "substantively important" effects for intervention rating purposes. In addition to ESs and improvement indices, the WWC also computes the levels of statistical significance of student-level findings corrected for clustering and/or multiple comparisons where necessary.

The purpose of this document is to provide the technical details about the various types of computations conducted by the WWC as part of its review process, which will allow readers to better understand the findings that we report and the conclusions that we draw regarding the effectiveness of the educational interventions reviewed by the WWC.[1] Specifically, the technical details of the following types of WWC-conducted computations are presented:

I Effect Size Computation for Continuous Outcomes
- ES as Standardized Mean Difference (Hedges's g)
- ES Computation Based on Results from Student-Level T-Tests or ANOVA
- ES Computation Based on Results from Student-Level ANCOVA
- ES Computation Based on Results from Cluster-Level Analyses
- ES Computation Based on Results from HLM Analysis in Studies with Cluster-Level Assignment

II. Effect Size Computation for Dichotomous Outcomes

III. Computation of the Improvement Index

IV Clustering Correction of the Statistical Significance of Effects Estimated with Mismatched Analyses

V Benjamini-Hochberg Correction of the Statistical Significance of Effects Estimated with Multiple Comparisons

In addition to computational procedures, this document presents the rationale for the specific computations conducted and their underlying assumptions. These procedures are currently used to compute effect sizes and make corrections for study designs and reporting practices most commonly encountered during WWC's review process. It is not meant to serve as a comprehensive compendium of an exhaustive list of ES computation methods that have ever been developed in the field.

[1] The WWC regularly updates WWC technical standards and their application to take account of new considerations brought forth by experts and users. Such changes may result in re-appraisals of studies and/or interventions previously reviewed and rated. Current WWC standards offer guidance for those planning or carrying out studies, not only in the design considerations but the analysis and reporting stages as well. WWC standards, however, may not pertain to every situation, context, or purpose of a study and will evolve.

I. Effect Size Computation for Continuous Outcomes

ES as Standardized Mean Difference (Hedges's g)

Different types of ES indices have been developed for different types of outcome measures, given their distinct statistical properties. For continuous outcomes, the WWC has adopted the most commonly-used ES index—the standardized mean difference, which is defined as the difference between the mean outcome of the intervention group and the mean outcome of the comparison group divided by the pooled within-group standard deviation (SD) on that outcome measure. Given that the WWC generally focuses on student-level findings , the default SD used in ES computation is the student-level SD.

The basic formula for computing standardized mean difference is as follows:

$$\text{Standardized mean difference} = (X_1 - X_2) / S_{\text{pooled}}, \quad (1)$$

where X_1 and X_2 are the means of the outcome for the intervention group and the comparison group respectively, and S_{pooled} is the pooled within-group SD of the outcome at the student level. Formulaically,

$$S_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}, \text{ and} \quad (2)$$

$$S_{\text{pooled}} = \text{sqrt}\{[(n_1\text{-}1)S_1\text{^}2 + (n_2\text{-}1)S_2\text{^}2]/(n_1+n_2\text{-}2)\}$$

$$\text{Standardized mean difference (g)} = \frac{X_1 - X_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}} \quad (3)$$

$$(g) = (X_1 - X_2)/\text{sqrt}\{[(n_1\text{-}1)S_1\text{^}2 + (n_2\text{-}1)S_2\text{^}2]/(n_1+n_2\text{-}2)\}$$

where n_1 and n_2 are the student sample sizes, and S_1 and S_2 the student-level SDs, for the intervention group and the comparison group respectively.

The ES index thus computed is referred to as Hedges's g.[2] This index, however, has been shown to be upwardly biased when the sample size is small. Therefore, we have applied a simple correction for this bias developed by Hedges (1981), which produces an unbiased ES estimate by multiplying the Hedges's g by a factor of [1-3/(4N-9)], with N being the total sample size. Unless otherwise noted, Hedges's g corrected for small-sample bias is the default ES measure for continuous outcomes used in the WWC's review.

[2] The Hedges' g index differs from the Cohen's d index in that Hedges's g uses the square root of degrees of freedom (sqrt(N-k) for k groups) for the denominator of the pooled within-group SD (S_{pooled}), whereas Cohen's d uses the square root of sample size (sqrt(N)) to compute S_{pooled} (Rosenthal, 1994; Rosnow, Rosenthal, & Rubin, 2000).

In certain situations, however, the WWC may present study findings using ES measures other than Hedges's g. If, for instance, the SD of the intervention group differs substantially from that of the comparison group, the PIs and review teams may choose to use the SD of the comparison group instead of the pooled within-group SD as the denominator of the standardized mean difference, and compute the ES as Glass's Δ instead of Hedges's g. The justification for doing so is that when the intervention and comparison groups have unequal variances, as in the case where the variance of the outcome is affected by the intervention, the comparison group variance is likely to be a better estimate of the population variance than the pooled within-group variance (Cooper, 1998; Lipsey & Wilson, 2001). The WWC may also use Glass's Δ, or other ES measures used by the study authors, to present study findings—if there is not enough information available for computing Hedge's g. These deviations from the default will be clearly documented in the WWC's review process.

The sections to follow focus on the WWC's default approach to computing student-level ESs for continuous outcomes. We describe procedures for computing Hedges's g based on results from different types of statistical analysis most commonly encountered in the WWC reviews.

ES Computation Based on Results from Student-Level T-Tests or ANOVA

For randomized controlled trials, study authors may assess an intervention's effects based on student-level t-tests or analyses of variance (ANOVA) without adjustment for pretest or other covariates, assuming group equivalence on pre-intervention measures achieved through random assignment. If the study authors reported posttest means and SD as well as sample sizes for both the intervention group and the comparison group, the computation of ESs will be straightforward using the standard formula for Hedges's g (see Equation (3)).

Where the study authors did not report the posttest mean, SD, or sample size for each study group, the WWC computes Hedges's g based on t-test or ANOVA F-test results, if they were reported along with sample sizes for both the intervention group (n_1) and the comparison group (n_2). For ESs based on t-test results,

$$\text{Hedges's g} = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

$$\text{Hedges's g} = t * \text{sqrt}\,[(n_1 + n_2)/n_1 n_2] \quad (4)$$

For ESs based on ANOVA F-test results,

$$\text{Hedges's g} = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}, \quad (5)$$

$$\text{Hedges's g} = \text{sqrt}\,[(F(n_1 + n_2)/n_1 n_2]$$

ES Computation Based on Results from Student-Level ANCOVA

Analysis of covariance (ANCOVA) is a commonly used analytic method for quasi-experimental designs. It assesses the effects of an intervention while controlling for important covariates,

particular pretest, that might confound the effects of the intervention. ANCOVA is also used to analyze data from randomized controlled trials so that greater statistical precision of parameter estimates can be achieved through covariate adjustment.

For study findings based on student-level ANCOVA, the WWC computes Hedges's g as *covariate adjusted mean difference* divided by *unadjusted pooled within-group SD*. The use of adjusted mean difference as the numerator of ES ensures that the ES estimate is adjusted for covariate difference between the intervention and the comparison groups that might otherwise bias the result. The use of unadjusted pooled within-group SD as the denominator of ES allows comparisons of ES estimates across studies by using a common metric to standardize group mean differences, i.e., the population SD as estimated by the unadjusted pooled within-group SD.

Specifically, when sample sizes, and adjusted means and unadjusted SDs of the posttest from an ANCOVA are available for both the intervention and the comparison groups, the WWC computes Hedges's g as follows:

$$\text{Hedges's g} = \frac{X_1' - X_2'}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}}, \qquad (6)$$

Hedges's g = $(X_1' - X_2')$/sqrt{[$(n_1$-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}

where X_1' and X_2' are adjusted posttest means, n_1 and n_2 the student sample sizes, and S_1 and S_2 the student-level unadjusted posttest SD, for the intervention group and the comparison group respectively,

It is not uncommon, however, for study authors to report unadjusted group means on both pretest and posttest, but not adjusted group means or adjusted group mean differences on the posttest. Absent information on the correlation between the pretest and the posttest, as is typically the case, the WWC's default approach is to compute the numerator of ES—the adjusted mean difference—as the difference between the pretest-posttest mean difference for the intervention group and the pretest-posttest mean difference for the comparison group. Specifically,

$$\text{Hedges's g} = \frac{(X_1 - X_{1-pre}) - (X_2 - X_{2-pre})}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}}, \qquad (7)$$

Hedges's g = [(X_1 –$X_{1\text{-}pre}$)- (X_2 –$X_{2\text{-}pre}$)]/sqrt{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}

where X_1 and X_2 are unadjusted posttest means, $X_{1\text{-pre}}$ and $X_{2\text{-pre}}$ unadjusted pretest means, n_1 and n_2 the student sample sizes, and S_1 and S_2 the student-level unadjusted posttest SD, for the intervention group and the comparison group respectively,

This "difference-in-differences" approach to estimating an intervention's effects while taking into account group difference in pretest is not necessarily optimal, as it is likely to either

overestimate or underestimate the adjusted group mean difference, depending on which group performed better on the pretest.[3] Moreover, this approach does not provide a means for adjusting the statistic significance of the adjusted mean difference to reflect the covariance between the pretest and the posttest. Nevertheless, it yields a reasonable estimate of the adjusted group mean difference, which is equivalent to what would have been obtained from a commonly used alternative to the covariate adjustment-based approach to testing an intervention's effect—the analysis of gain scores.

Another limitation of the "difference-in-differences" approach is that it assumes the pretest and the posttest are the same test. Otherwise, the means on the two types of tests might not be comparable, and hence it might not be appropriate to compute the pretest-posttest difference for each group. In cases where different pretest and posttests were used, and only unadjusted means on pretest and posttest were reported, the Principal Investigators (PIs) will need to consult with the WWC Technical Review Team to determine whether it is reasonable to use the difference-in-differences approach to compute the ESs.

The difference-in-differences approach presented above also assumes that the pretest-posttest correlation is unknown. In some areas of educational research, however, empirical data on the relationships between pretest and posttest may be available. If such data are dependable, the WWC PIs and the review team in a given topic area may choose to use the empirical relationship to estimate the adjusted group mean difference that is unavailable from the study report or study authors, rather than using the default difference-in-differences approach. The advantage of doing so is that, if indeed the empirical relationship between pretest and posttest is dependable, the covariate-adjusted estimates of the intervention's effects will be less biased than those based on the difference-in-differences (gain score) approach. If the PIs and review teams choose to compute ESs using an empirical pretest-posttest relationship, they will need to provide an explicit justification for their choice as well as evidence on the credibility of the empirical relationship.

Computationally, if the pretest and posttest has a correlation of r, then

$$\text{Hedges's g} = \frac{(X_1 - X_2) - r(X_{1-pre} - X_{2-pre})}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}, \quad (8)$$

Hedges's g = $[(X_1 - X_2) - r(X_{1\text{-}pre} - X_{2\text{-}pre})]$/sqrt{$[(n_1\text{-}1)S_1\text{^}2 + (n_2\text{-}1)S_2\text{^}2]/(n_1+n_2\text{-}2)$}

where all the other terms are the same as those in Equation (7).

A final note about ANCOVA-based ES computation is that Hedges's g cannot be computed based on the F-statistic from an ANCOVA using Equation (5). Unlike the F-statistic from an ANOVA, which is based on unadjusted within-group variance, the F-statistic from an ANCOVA is based on covariate-adjusted within-group variance. Hedges's *g*, however, requires

[3] If the intervention group had a higher average pretest score than the comparison group, the difference-in-difference approach is likely to underestimate the adjusted group mean difference. Otherwise, it is likely to overestimate the adjusted group mean difference.

the use of unadjusted within-group SD. Therefore, we cannot compute Hedges's g with the F-statistic from an ANCOVA in the same way as we compute *g* with the F-statistic from an ANOVA. If the pretest-posttest correlation is known, however, we could derive Hedges's *g* from the ANCOVA F-statistic as follows:

$$\text{Hedges's g} = \sqrt{\frac{F(n_1 + n_2)(1 - r^2)}{n_1 n_2}}, \quad (9)$$

Hedges's g = sqrt[$F(n_1 + n_2)(1-r\text{^}2)/n_1n_2$]

where r is the pretest-posttest correlation, and n_1 and n_2 are the sample sizes for the intervention group and the comparison group respectively.

ES Computation Based on Results from Cluster-Level Analyses

The ES computation methods described above are all based on student-level analyses, which are appropriate analytic approaches for studies with student-level assignment. The case is more complicated, however, for studies with assignment at the cluster level (e.g., assignment of teachers, classrooms, or schools to conditions), where data may have been analyzed at the student level, the cluster level, or through multilevel analyses. Although there has been a consensus in the field that multilevel analysis should be used to analyze clustered data (e.g., Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Flay & Collins, 2005; Murray, 1998; and Snijders & Bosker, 1999), student-level analyses and cluster-level analyses of such data still frequently appear in the research literature despite their problems.

The main problem with student-level analyses in studies with cluster-level assignment is that they violate the independence of observations assumption underlying traditional hypothesis tests and result in underestimated standard errors and inflated statistical significance (see Section IV for details about how to correct for such bias). The estimate of the group mean difference in such analyses, however, is unbiased and therefore can be appropriately used to compute the student-level ES using methods explained in the previous sections.

For studies with cluster-level assignment, analyses at the cluster level, or aggregated analyses, are also problematic. Other than the loss of power and increased Type II error, potential problems with aggregated analysis include shift of meaning and ecological fallacy (i.e., relationships between aggregated variables cannot be used to make assertions about the relationships between individual-level variables), among others (Aitkin & Longford, 1986; Snijders & Bosker, 1999), Such analyses also pose special challenges to ES computation during WWC reviews. In the remainder of this section, we discuss these challenges and describe WWC's approach to handling them during WWC reviews.

<u>How to compute student-level ESs for studies with cluster-level analyses</u>

For studies that only reported findings from cluster-level analyses, it might be tempting to compute ESs using cluster-level means and SDs. This, however, is not appropriate for the purpose of the WWC reviews for at least two reasons. First, because cluster-level SDs are

typically much smaller than student-level SDs,[4] ESs based on cluster-level SDs will be much larger than, and therefore incomparable with, student-level ESs that are the focus of WWC reviews. Second, the criterion for "substantively important" effects in the WWC Intervention Rating Scheme (ES of at least 0.25) was established specifically for student-level ESs, and does not apply to cluster-level ESs. Moreover, there is not enough knowledge in the field as yet for judging the magnitude of cluster-level effects. A criterion of "substantively important" effects for cluster-level ESs, therefore, cannot be developed for intervention rating purposes. An intervention rating of potentially positive effects based on a cluster-level ES of 0.25 or greater (i.e., the criterion for student-level ESs) would be misleading.

In order to compute the student-level ESs, we need to use the student-level means and SDs on the findings. This information, however, is often not reported in studies with cluster-level analyses. If the study authors could not provide student-level means, the review team may use cluster-level means (i.e., mean of cluster means) to compute the group mean difference for the numerator of student-level ESs if: (1) the clusters were of equal or similar sizes, (2) the cluster means were similar across clusters, or (3) it is reasonable to assume that cluster size was unrelated to cluster means. If any of the above conditions holds, then group means based on cluster-level data would be similar to group means based on student-level data, and hence could be used for computing student-level ESs. If none of the above conditions holds, however, the review team will have to obtain the group means based on student-level data in order to compute the student-level ESs.

While it is possible to compute the numerator (i.e., group mean difference) for student-level ESs based on cluster-level findings for most studies, it is generally much less feasible to compute the denominator (i.e., pooled SD) for student-level ESs based on cluster-level data. If the student-level SDs are not available, we could compute them based on the cluster-level SDs and the actual intra-class correlation (ICC) (student-level SD = (cluster-level SD)/sqrt(ICC)). Unfortunately, the actual ICCs for the data observed are rarely provided in study reports. Without knowledge about the actual ICC, one might consider using a default ICC, which, however, is not appropriate, because the resulting ES estimate would be highly sensitive to the value of the default ICC and might be seriously biased even if the difference between the default ICC and the actual ICC is not large.

Another reason that the formula for deriving student-level SDs (student-level SD = (cluster-level SD)/sqrt(ICC)) is unlikely to be useful is that the cluster-level SD required for the computation was often not reported either. Note that the cluster-level SD associated with the ICC is not exactly the same as the observed SD of cluster means that were often reported in studies with cluster-level analyses, because the latter reflects not only the true cluster-level variance, but also part of the random variance within clusters (Raudenbush & Liu, 2000; Snijder & Bosker, 1999).

It is clear from the above discussion that in most cases, requesting student-level data, particularly student-level SDs, from the study authors will be the only way that allows us to compute the student-level ESs for studies only reporting cluster-level findings. If the study authors could not provide the student-level data needed, then we would not be able to compute

[4] Cluster-level SD = (student-level SD)*sqrt(ICC).

the student-level ESs. Nevertheless, such studies will not be automatically excluded from the WWC reviews, but could still potentially contribute to intervention ratings as explained below.

How to handle studies with cluster-level analyses in intervention ratings if the student-level ESs could not be computed

A study's contribution to the effectiveness rating of an intervention depends mainly on three factors: the quality of the study design, the statistical significance of the findings, and the size of the effects. For studies that only reported cluster-level findings, the quality of their design is not affected by whether student-level ESs could be computed or not. Such studies could still meet WWC evidence standards with or without reservations and be included in intervention reports even if student-level ESs were not available.

While cluster-level ESs cannot be used in intervention ratings, the statistical significance of cluster-level findings could contribute to intervention ratings. Cluster-level analyses tend to be underpowered, hence estimates of the statistical significance of findings from such analyses tend to be conservative. Therefore, significant findings from cluster-level analyses would remain significant had the data been analyzed using appropriate multilevel models, and should be taken into account in intervention ratings. The size of the effects based on cluster-level analyses, however, could not be considered in determining "substantively important" effects in intervention ratings for reasons described above. In WWC's intervention reports, cluster-level ESs will be excluded from the computation of domain average ESs and improvement indices, both of which will be based exclusively on student-level findings.

ES Computation Based on Results from HLM Analyses in Studies with Cluster-Level Assignment

As explained in the previous section, multilevel analysis is generally considered the preferred method for analyzing data of from studies with cluster-level assignment. With recent methodological advances, multilevel analysis has gained increased popularity in education and other social science fields. More and more researchers have begun to employ the hierarchical linear modeling (HLM) method to analyze data of a nested nature (e.g., students nested within classes and classes nested within schools) (Raudenbush & Bryk, 2002)[5]. Similar to student-level ANCOVA, HLM can also adjust for important covariates such as pretest when estimating an intervention's effect. Unlike student-level ANCOVA that assumes independence of observations, however, HLM explicitly takes into account the dependence among members within the same higher-level unit (e.g., the dependence among students within the same class). Therefore, the parameter estimates, particularly the standard errors, generated from HLM are less biased than those generated from ANCOVA when the data have a multilevel structure.

Hedges's g for intervention effects estimated from HLM analyses is defined in a similar way to that based on student-level ANCOVA: adjusted group mean difference divided by unadjusted pooled within-group SD. Specifically,

[5] Multilevel analysis can also be conducted using other approaches, such as the SAS PROC MIXED procedure. Although different approaches to multilevel analysis may differ in their technical details, they are all based on similar ideas and underlying assumptions.

$$\text{Hedges's g} = \frac{\gamma}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}}, \quad (10)$$

Hedges's g = γ/sqrt{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}

Where γ is the HLM coefficient for the intervention's effect, which represents the group mean difference adjusted for both level-1 and level-2 covariates, if any;[6] n_1 and n_2 are the student sample sizes, and S_1 and S_2 the unadjusted student-level SDs for the intervention group and the comparison group respectively.

One thing to note about the denominator of Hedges's g based on HLM results is that the level-1 variance, also called "within-group variance," estimated from a typical two-level HLM analysis is not the same as the conventional unadjusted pooled within-group variance that should be used in ES computation. The within-group variance from an HLM model that incorporates level-1 covariates has been adjusted for these covariates. Even if the within-group variance is based on an HLM model that does not contain any covariates (i.e., a fully-unconditional model), it is still not appropriate for ES computation, because it does not include the variance between level-2 units within each study condition that is part of the unadjusted pooled within-group variance. Therefore, the level-1 within-group variance estimated from an HLM analysis tends to be smaller than the conventional unadjusted pooled within-group variance, and would thus lead to an overestimate of the ES if used in the denominator of the ES.

The ES computations for continuous outcomes explained above pertain to individual findings within a given outcome domain examined in a given study. If the study authors assessed the intervention's effects on multiple outcome measures within a given domain, the WWC computes a domain average ES as a simple average of the ESs across all individual findings within the domain.

II. Effect Size Computation for Dichotomous Outcomes

Although not as common as continuous outcomes, dichotomous outcomes are sometimes used in studies of educational interventions. Examples include dropout vs. stay in school; grade promotion vs. retention; and pass vs. fail a test. Group mean differences, in this case, appear as differences in proportions or differences in the probability of the occurrence of an event. The ES measure of choice for dichotomous outcomes is odds ratio, which has many statistical and practical advantages over alternative ES measures such as the difference between two probabilities, the ratio of two probabilities, and the phi coefficient (Fleiss, 1994; Lipsey & Wilson, 2001).

[6] The level-2 coefficients are adjusted for the level-1 covariates under the condition that the level-1 covariates are either uncentered or grand-mean centered, which are the most common centering options in an HLM analysis (Raudenbush & Bryk, 2002). The level-2 coefficients are not adjusted for the level-1 covariates if the level-1 covariates are group-mean centered. For simplicity purposes, the discussion here is based on a two-level framework (i.e., students nested with clusters). The idea could easily be extended to a three-level model (e.g., students nested with teachers who were in turn nested within schools).

The measure of odds ratio builds on the notion of odds. For a given study group, the odds for the occurrence of an event are defined as follows:

$$\text{Odds} = \frac{p}{1-p}, \quad (11)$$

Odds = *p/(1-p)*

where p is the probability of the occurrence of an event within the group. Odds ratio (OR) is simply the ratio between the odds for the two groups compared:

$$\text{OR} = \frac{Odds_1}{Odds_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}, \quad (12)$$

OR = $Odds_1/Odds_2 = [p_1(1-p\rho_2)]/[p_2(1- p_1)]$

where p_1 and p_2 are the probabilities of the occurrence of an event for the intervention group and the comparison group respectively.

As is the case with ES computation for continuous variables, the WWC computes ESs for dichotomous outcomes based on student-level data in preference to aggregate-level data for studies that had a multi-level data structure. The probabilities (p_1 and p_2) used in calculating the odds ratio represent the proportions of students demonstrating a certain outcome among students across all teachers/classrooms or schools in each study condition, which are likely to differ from the probabilities based on aggregate-level data (e.g., means of school-specific probabilities) unless the classrooms or schools in the sample were of similar sizes.

Following conventional practice, the WWC transforms odds ratio calculated based on Equation (12) to logged odds ratio (LOR) (i.e., the natural log of the odds ratio) to simplify statistical analyses:

$$\text{LOR} = \text{In(OR)} \quad (13)$$

The logged odds ratio has a convenient distribution form, which is approximately normal with a mean of 0 and a SD of π/sqrt(3), or 1.81.

The logged odds ratio can also be expressed as the difference between the logged odds, or logits, for the two groups compared. Equivalent to Equation (13),

$$\text{LOR} = In(Odds_1) - In(Odds_2), \quad (14)$$

LOR = $In(Odds_1) - In(Odds_2)$

which shows more clearly the connection between the logged odds ratio index and the standardized mean difference index (Hedges's g) for ESs. To make logged odds ratio comparable to standardized mean difference and thus facilitate the synthesis of research findings based on different types of outcomes, researchers have proposed a variety of methods for "standardizing" logged odds ratio. Based on a Monte Carlo simulation study of seven different types of ES indices for dichotomous outcomes, Sanchez-Meca, Marin-Martinez, and Chacon-Moscoso

(2003) concluded that the ES index proposed by Cox (1970) is the least biased estimator of the population standardized mean difference, assuming an underlying normal distribution of the outcome. The WWC, therefore, has adopted the Cox index as the default ES measure for dichotomous outcomes. The computation of the Cox index is straightforward:

$$LOR_{Cox} = LOR/1.65 \quad (15)$$

The above index yields ES values very similar to the values of Hedges's g that one would obtain if group means, SDs, and sample sizes were available—assuming that the dichotomous outcome measure is based on an underlying normal distribution. Although the assumption may not always hold, as Sanchez-Meca and his colleagues (2003) note, primary studies in social and behavioral sciences routinely apply parametric statistical tests that imply normality. Therefore, the assumption of normal distribution is a reasonable conventional default.

III. Computation of the Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates ES into an "improvement index." The improvement index represents the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the comparison group mean (i.e., 50th percentile) in the comparison group distribution. Alternatively, the improvement index can be interpreted as the expected change in percentile rank for an average comparison group student if the student had received the intervention.

As an example, if an intervention produced a positive impact on students' reading achievement with an effect size of 0.25, the effect size could be translated to an improvement index of 10 percentile points. We could then conclude that the intervention would have led to a 10% increase in percentile rank for an average student in the comparison group, and that 60% (10% + 50%=60%) of the students in the intervention group scored above the comparison group mean.

Specifically, the improvement index is computed as follows:

(1) Convert the ES (Hedges's g) to Cohen's U3 index.

The U3 index represents the percentile rank of a comparison group student who performed at the level of an average intervention group student. An effect size of 0.25, for example, would correspond to a U3 of 60%, which means that an average intervention group student would rank at the 60th percentile in the comparison group. Equivalently, an average intervention group student would rank 10 percentile points higher than an average comparison group student, who, by definition, ranks at the 50th percentile.

Mechanically, the conversion of an effect size to a U3 index entails looking up on a table that lists the proportion of area under the standard normal curve for different values of z-scores, which can be found in the appendices of most statistics textbooks. For a given effect size, U3 has a value equal to the proportion of area under the normal curve below the value of the effect

size—under the assumptions that the outcome is normally distributed and that the variance of the outcome is similar for the intervention group and the comparison group.

(2) Compute:

$$\text{Improvement index} = U3 - 50\% \quad (16)$$

Given that U3 represents the percentile rank of an average intervention group student in the comparison group distribution, and that the percentile rank of an average comparison group student is 50%, the improvement index, defined as (U3 – 50%), would represent the difference in percentile rank between an average intervention group student and an average comparison group student in the comparison group distribution.

In addition to the improvement index for each individual finding, the WWC also computes a domain average improvement index for each study as well as a domain average improvement index across studies for each outcome domain. The domain average improvement index for each study is computed based on the domain average effect size for that study rather than as the average of the improvement indices for individual findings within that study. Similarly, the domain average improvement index across studies is computed based on the domain average effect size across studies, with the latter computed as the average of the domain average effect sizes for individual studies.

IV. Clustering Correction of the Statistical Significance of Effects Estimated with Mismatched Analyses

In order to adequately assess an intervention's effects, it is important to know not only the magnitude of the effects as indicated by ES, but also the statistical significance of the effects. The correct statistical significance of findings, however, is not always readily available, particularly in studies where the unit of assignment does not match the unit of analysis. The most common "mismatch" problem occurs when assignment was carried out at the cluster level (e.g., classroom or school level), whereas the analysis was conducted at the student level, ignoring the dependence among students within the same clusters. Although the point estimates of the intervention's effects based on such mismatched analyses are unbiased, the standard errors of the effect estimates are likely to be underestimated, which would lead to inflated Type I error and overestimated statistical significance.

In order to present a fair judgment about an intervention's effects, the WWC computes clustering-corrected statistical significance for effects estimated from mismatched analyses and the corresponding domain average effects based on Hedges' (2005) most recent work. As clustering correction will decrease the statistical significance (or increase the p-value) of the findings, non-significant findings from a mismatched analysis will remain non-significant after the correction. Therefore, the WWC only applies the correction to findings reported to be statistically significant by the study authors.

The basic approach to clustering correction is to first compute the t-statistic corresponding to the ES that ignores clustering, and then correct both the t-statistic and the

associated degrees of freedom for clustering based on sample sizes, number of clusters, and the intra-class correlation. The statistic significance corrected for clustering could then be obtained from the t-distribution with the corrected t-statistic and degrees of freedom. In the remainder of this section, we detail each step of the process.

(1) Compute the t-statistic for the ES ignoring clustering

This is essentially the reverse of Equation (4) that computes Hedges's g based on t:

$$t = g\sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

$$\text{t} = g\text{*sqrt}\,[n_1 n_2/(\,n_1 + n_2)] \qquad (17)$$

where g is the ES that ignores clustering, and n_1 and n_2 are the sample sizes for the intervention group and the comparison group respectively for a given outcome. For domain average ESs, n_1 and n_2 are the average sample sizes for the intervention and comparison groups respectively across all outcomes within the domain

(2) Correct the above t-statistic for clustering

$$t_A = \text{t}\sqrt{\frac{(N-2) - 2\left(\frac{N}{m} - 1\right)\rho}{(N-2)[1 + \left(\frac{N}{m} - 1\right)\rho]}}, \qquad (18)$$

$$t_A = t\text{*sqrt}\{[(N\text{-}2)\text{-}2(N/m\text{-}1)\rho]/[(N\text{-}2)(1+(N/m\text{-}1)\rho)]\}$$

where N is the total sample size at the student level ($N = n_1 + n_2$), m is the total number of clusters in the intervention and comparison groups ($m = m_1 + m_2$, m_1 and m_2 are the number of clusters in each of the two groups), and ρ is the intra-class correlation (ICC) for a given outcome.

The value of ICC, however, is often not available from the study reports. Based on empirical literature in the field of education, the WWC has adopted a default ICC value of .20 for achievement outcomes and .10 for behavioral and attitudinal outcomes. The PIs and review teams may set different defaults with explicit justification in terms of the nature of the research circumstances or the outcome domain.

For domain average ESs, the ICC used in Equation (18) is the average ICC across all outcomes within the domain. If the number of clusters in the intervention and comparison groups differs across outcomes within a given domain, the total number of clusters (m) used for computing the corrected t-statistic will be based on the largest number of clusters in both groups across outcomes within the domain (i.e., largest m_1 and m_2 across outcomes). This gives the study the benefit of the doubt by crediting the measure with the most statistical power, so that the WWC's rating of interventions will not be unduly conservative.

(3) Compute the degrees of freedom associated with the t-statistics corrected for clustering:

$$h = \frac{\left[(N-2) - 2\left(\frac{N}{m} - 1\right)\rho\right]^2}{(N-2)(1-\rho)^2 + \frac{N}{m}\left(N - 2\frac{N}{m}\right)\rho^2 + 2\left(N - 2\frac{N}{m}\right)\rho(1-\rho)} \quad (19)$$

$h =$ [(*N*-2)-2(*N*/*m*-1)ρ]^2/[(*N*-2)(1- ρ)^2+ (*N*/*m*)(*N*-2*N*/*m*)ρ^2+2(*N*-2*N*/*m*)ρ(1- ρ)]

(4) Obtain the statistical significance of the effect corrected for clustering

The clustering-corrected statistical significance (p-value) is determined based on the t-distribution with corrected t-statistic (t_A) and the corrected degrees of freedom (h). This p-value can either be looked up in a t-distribution table that can be found in the appendices of most statistical textbooks, or computed using the t-distribution function in Excel: p = TDIST(t_A, h, 2).

Further information on this topic is available in the WWC's technical papers on the WWC Tutorial on Mismatch Between Unit of Assignment and Unit of Analysis and the WWC Intervention Rating Scheme.

V. Benjamini-Hochberg Correction of the Statistical Significance of Effects Estimated with Multiple Comparisons

In addition to clustering, another factor that may inflate Type I error and the statistical significance of findings is when study authors perform multiple hypothesis tests simultaneously. The traditional approach to addressing the problem is the Bonferroni method, which lowers the critical p-value for individual comparisons by a factor of *1/m*, with *m* being the total number of comparisons made. The Bonferroni method, however, has been shown to be unnecessarily stringent for many practical situations; therefore the WWC has adopted a more recently developed method to correct for multiple comparisons or multiplicity—the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995). The BH method adjusts for multiple comparisons by controlling false discovery rate (FDR) instead of familywise error rate (FWER). It is less conservative than the traditional Bonferroni method, yet still provides adequate protection against Type I error in a wide range of applications. Since its conception in the 1990's, there has been growing evidence showing that the FDR-based BH method may be the best solution to the multiple comparisons problem in many practical situations (Williams, Jones, & Tukey, 1999)

As is the case with clustering correction, the WWC only applies the BH correction to statistically significant findings, because non-significant findings will remain non-significant after correction. For findings based on analyses where the unit of analysis was properly aligned with the unit of assignment, we use the p-values reported in the study for the BH correction. If the exact p-values were not available, but the ESs could be computed, we will convert the ESs to t-statistics (see Equation (4)), and then obtain the corresponding p-values.[7] For findings based on

[7] The p-values corresponding to the t-statistics can either be looked up in a t-distribution table, or computed using the t-distribution function in Excel: p = TDIST(t, df, 2), where df is the degrees of freedom, or the total sample size minus 2 for findings from properly aligned analyses.

mismatched analyses, we first correct the author-reported p-values for clustering, and then use the clustering-corrected p-values for the BH correction.

Although the BH correction procedure described above was originally developed under the assumption of independent test statistics (Benjamini & Hochberg, 1995), Benjamini and Yekutieli (2001) point out that it also applies to situations where the test statistics have positive dependency, and that the condition for positive dependency is general enough to cover many problems of practical interest. For other forms of dependency, a modification of the original BH procedure could be made, which, however, is "very often not needed, and yields too conservative a procedure" (p. 1183).[8] Therefore, the WWC has chosen to use the original BH procedure rather than its more conservative modified version as the default approach to correcting for multiple comparisons. In the remainder of this section, we describe the specific procedures for applying the BH correction in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group, studies that tested a given outcome measure with multiple comparison groups, and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on Multiple Outcome Measures within the Same Outcome Domain Tested with a Single Comparison Groups

The most straightforward situation that may require the BH correction is when the study authors assessed an intervention's effect on multiple outcome measures within the same outcome domain using a single comparison group. For such studies, the review team needs to first check whether the study authors' analyses already took into account multiple comparisons (e.g., through a proper multivariate analysis). If so, obviously no further correction is necessary. If the authors did not address the multiple comparison problem in their analyses, then the review team will need to correct the statistical significance of the authors' findings using the BH method. For studies that examined measures in multiple outcome domains, the BH correction will be applied to the set of findings within the same domain rather than across different domains. Assuming that the BH correction is needed, the review team will apply the BH correction to multiple findings within a given outcome domain tested with a single comparison group as follows:

(1) Rank order statistically significant findings within the domain in ascending order of the p-values, such that: $p_1 \leq p_2 \leq p_3 \leq \ldots \leq p_m$, with m being the number of significant findings within the domain.

(2) For each p-value (p_i), compute:

$$p_i' = \frac{i * \alpha}{M}, \qquad (20)$$

$[p_i' = i * \alpha / M]$

[8] The modified version of the BH procedure uses α over the sum of the inverse of the p-value ranks across the m comparisons (i.e., $\alpha / \sum_{i=1}^{m} \frac{1}{i}$) instead of α in Equation (20).

where i is the rank for p_i, with $i = 1, 2, \ldots m$; M is the total number of findings within the domain reported by the WWC; and α is the target level of statistical significance.

Note that the M in the denominator of Equation (20) may be less than the number of outcomes that the study authors actually examined in their study for two reasons: (1) the authors may not have reported findings from the complete set of comparisons that they had made, and (2) certain outcomes assessed by the study authors may be deemed irrelevant to the WWC's review. The target level of statistical significance, α, in the numerator of Equation (20) allows us to identify findings that are significant at this level after correction for multiple comparisons. The WWC's default value of α is 0.05, although other values of α could also be specified. If, for instance, α is set at 0.01 instead of 0.05, then the results of the BH correction would indicate which individual findings are statistically significant at the 0.01 level instead of the 0.05 level after taking multiple comparisons into account.

(3) Identify the largest i—denoted by k— that satisfies the condition: $p_i \leq p_i'$. This establishes the cut-off point, and allows us to conclude that all findings with p-values smaller than or equal to p_k are statistically significant, and findings with p-values greater than p_k are not significant at the pre-specified level of significance ($\alpha = 0.05$ by default) after correction for multiple comparisons.

One thing to note is that, unlike clustering correction, which produces a new p-value for each corrected finding, the BH correction does not generate a new p-value for each finding, but rather only indicates whether the finding is significant or not at the pre-specified level of statistical significance after the correction. As an illustration, suppose a researcher compared the performance of the intervention group and the comparison group on eight measures in a given outcome domain, and reported six statistically significant effects and two non-significant effects based on properly aligned analyses. To correct the significance of the findings for multiple comparisons, we would first rank-order the p-values of the six author-reported significant findings in the first column of Table 1, and list the p-value ranks in the second column. We then compute $p_i' = i^* \alpha / M$, using Equation (20) with $M = 8$ and α=0.05, and record the values in the third column. Next, we identify k, the largest i, that meets the condition: $p_i \leq p_i'$. In this example, $k = 4$, and $p_k = 0.014$. Thus, we can claim that the four finding associated with a p-value of 0.014 or smaller are statistically significant at the 0.05 level after correction for multiple comparisons. The other two findings, although reported as being statistically significant, are no longer significant after the correction.

Table 1. An Illustration of Applying the Benjamini-Hochberg Correction for Multiple Comparisons

Author-reported or clustering-corrected p-value (p_i)	**P-value rank (i)**	$p_i' = \frac{i^*(0.05)}{8}$ $p_i' = i^*(0.05)/8$	$p_i \leq p_i'$**?**	**Statistical significance after BH correction (α = .05)**
0.002	1	0.006	Yes	significant
0.009	2	0.013	Yes	significant
0.011	3	0.019	Yes	significant
0.014	**4**	**0.025**	**Yes**	**significant**

0.034	5	0.031	No	n.s.
0.041	6	0.038	No	n.s.

Note. n.s.: not statistically significant.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on a Given Outcome Tested with Multiple Comparison Groups

The above discussion pertains to the multiple comparisons problem when the study authors tested multiple outcomes within the same domain with a single comparison group. Another type of multiple comparisons problem occurs when the study authors tested an intervention's effect on a given outcome by comparing the intervention group with multiple comparison groups. The WWC's recommendation for handling such studies is as follows:

1. In consultation with the PI and the study authors if needed, the review team selects a single comparison group that best represented the "business as usual" condition or that is considered most relevant to the WWC's review. Only findings based on comparisons between the intervention group and this particular comparison group will be included in the WWC's review. Findings involving the other comparison groups will be ignored, and the multiplicity due to one intervention group being compared with multiple comparison groups could also be ignored.

2. If the PI and the review team believe that it is appropriate to combine the multiple comparison groups, and if adequate data are available for deriving the means and SDs of the combined group, the team may present the findings based on comparisons of the intervention group and the combined comparison group instead of findings based on comparisons of the intervention group and each individual comparison group. The kind of multiplicity due to one intervention group being compared with multiple comparison groups will no longer be an issue in this approach.

   The PI and the review team may judge the appropriateness of combining multiple comparison groups by considering whether there was enough common ground among the different comparison groups that warrant such a combination; and particularly, whether the study authors themselves conducted combined analyses or indicated the appropriateness, or the lack thereof, of combined analyses. In cases where the study authors did not conduct or suggest combined analyses, it is advisable for the review team to check with the study authors before combining the data from different comparison groups.

3. If the PI and the review team believe that neither of the above two options is appropriate for a particular study, and that findings from comparisons of the intervention group and each individual comparison group should be presented, they need to make sure that the findings presented in the WWC's intervention report are corrected for multiplicity due to multiple comparison groups if necessary. The review team needs to first check the study report or check with the study authors whether the comparisons of the multiple groups were based on a proper statistical test that already took multiplicity into account (e.g., Dunnett's test (Dunnet, 1955), the Bonferroni method (Bonferroni, 1935), Scheffe's test (1953), and Tukey's HSD test (1949)). If so, then there would be no need for further corrections.

It is also advisable for the team to check with the study authors regarding the appropriateness of correcting their findings for multiplicity due to multiple comparison groups, as the authors might have theoretical or empirical reasons for considering the findings from comparisons of the intervention group and a given comparison group without consideration of other comparisons made within the same study. If the team decides that multiplicity correction is necessary, they will apply such correction using the BH method in the same way as they would apply it to findings on multiple outcomes within the same domain tested with a single comparison group as described in the previous section.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on Multiple Outcome Measures in the Same Outcome Domain Tested with Multiple Comparison Groups

A more complicated multiple comparison problem arises when a study tested an intervention's effect on multiple outcome measures in a given domain with multiple comparison groups. The multiplicity problem thus may originate from two sources. Assuming that both types of multiplicity need to be corrected, the review team will apply the BH correction in accordance with the following three scenarios.

Scenario 1: The study authors's findings did not take into account either type of multiplicity.

In this case, the BH correction will be based on the total number of comparisons made. For example, if a study compared one intervention group with two comparison groups on five outcomes in the same domain without taking multiplicity into account, then the BH correction will be applied to the 10 individual findings based on a total of 10 comparisons.

Scenario 2: The study authors's findings took into account the multiplicity *due to multiple comparisons, but not the* multiplicity *due to multiple outcomes.*

In some studies, the authors may have performed a proper multiple comparison test (e.g., Dunnett's test) on each individual outcome that took into account the multiplicity due to multiple comparison groups. For such studies, the WWC will only need to correct the findings for the multiplicity due to multiple outcomes. Specifically, separate BH corrections will be made to the findings based on comparisons involving different comparison groups. With two comparison groups, for instance, the review team will apply the BH correction to the two sets of findings separately—one set of findings (one finding for each outcome) for each comparison group.

Scenario3: The study authors's findings took into account the multiplicity *due to multiple outcomes, but not the* multiplicity *due to multiple comparison groups.*

Although this scenario may be relatively rare, it is possible that the study authors performed a proper multivariate test (e.g., MANOVA or MANCOVA) to compare the intervention group with a given comparison group that took into account the multiplicity due to multiple outcomes, and performed separate multivariate tests for different comparison groups. For such studies, the review team will only need to correct the findings for multiplicity due to

multiple comparison groups. Specifically, separate BH corrections will be made to the findings on different outcomes. With five outcomes and two comparison groups, for instance, the review team will apply the BH correction to the five sets of findings separately—one set of findings (one finding for each comparison group) for each outcome measure.

The decision rules for the three scenarios described above are summarized in the table below.

Table 2. Decision Rules for Correcting the Significance Levels of Findings from Studies That had a Multiple Comparison Problem due to Multiple Outcomes in a Given Domain and/or Multiple Comparison Groups, by Scenario

Authors' Analyses	Benjamini-Hochberg Correction
1. Did not correct for multiplicity from any source	• BH correction to all 10 individual findings
2. Corrected for multiplicity due to multiple comparison groups only	• BH correction to the 5 findings based on T vs. C1 comparisons • BH correction to the 5 findings based on T vs. C2 comparisons
3. Corrected for multiplicity due to multiple outcomes only	• BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O1 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O2 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O3 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O4 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O5

Note. T: treatment (intervention) group;
C1 and C2: comparison groups 1 and 2;
O1, O2, O3, O4, and O5: five outcome measures within a given outcome domain.

On a final note, although the BH corrections are applied in different ways to the individual study findings in different scenarios, such differences do not affect the way in which the intervention rating is determined. In all three scenarios of the above example, the 10 findings will be presented in a single outcome domain, and the characterization of the intervention's effects for this domain in this study will be based on the corrected statistical significance of each individual finding as well as the magnitude and statistical significance of the average effect size across of the 10 individual findings within the domain.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, A*(149), 1-43.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165–1188.

Bloom, H. S., Bos, J.M., & Lee, S.W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review,* 234:445- 69.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni.* Rome: Italy, pp. 13–60.

Cooper, H. (1998). *Synthesizing research: A guide for literature review*. Thousand Oaks, CA: Sage Publications.

Cox, D.R. (1970). *Analysis of binary data*. New York: Chapman & Hall/CRC.

Donner, A. and Klar, N. (2000) Design and Analysis of Cluster Randomized Trials in Health Research. London: Arnold Publishing.

Dunnett, C. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of American Statistical Association, 50*, 1096–1121.

Flay, B. R., & Collins, L. M. (2005). Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *The Annals of the American Academy of Political and Social Science, 599*, 147-175.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Russel Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Hedges, L. V. (2005). Correcting a significance test for clustering. Unpublished manuscript.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. (Vol. 27). New York: Oxford University Press.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*(6), 446–453.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomous outcomes in meta-analysis. *Psychological Methods, 8*(4), 448–467.

Scheffe, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40*, 87–104.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrika, 5*, 99–114.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of Educational and Behavioral Statistics, 24(1), 42-69.

What Works Clearinghouse Tutorial on Mismatch between Unit of Assignment and Unit of Analysis

The mismatch problem of concern here occurs when the units of assignment do not match the units of analysis in a study of an intervention and this feature of the study's design is ignored in the study's data analysis. For instance, a study may have assigned entire classrooms (the unit of assignment) to the intervention and control conditions. But the study analyzed data at the individual student level *rather* than at the classroom level or at both the classroom level and student level. Such analyses are common, but they are incorrect on statistical grounds.

This kind of mismatch leads to statistics with greater *apparent* precision than they actually have, because students are treated as independent units when they are not. By ignoring the design effect due to the clustering of students within classrooms (Kish, 1965), such analyses are likely to yield misleadingly high levels of statistical significance (*p* values that are too small) and misleadingly narrow confidence intervals for an observed difference between intervention and control conditions. In a well executed randomized trial, for example, the estimates of a difference will be statistically unbiased, but statements about statistical tests of hypotheses and about one's confidence in results may not be correct.

In particular, a difference found to be statistically significant under an incorrect mismatch analysis could, under a correct analysis, turn out to be not statistically significant. A difference found to be not statistically significant under an improper mismatch analysis, on the other hand, would generally remain non-significant under a correct analysis.

Calculating effect sizes, confidence intervals, *p* values for statistical tests, and standardized effect sizes correctly, when groups are the units of assignment, requires information that is often not available in original reports when study authors analyzed the data incorrectly. In particular, to properly analyze the data, one needs to (a) know the intraclass correlation, which represents the degree to which individuals are dependent on each other within groups, or (b) employ methods such as hierarchical linear modeling that take this relationship into account. This intraclass correlation is rarely reported in studies with the mismatch problem. And hierarchical linear modeling and related approaches usually require access to and resources for reanalyzing original micro-record data. These are often not available.

Example: Consider a study in which 10 classrooms, each containing 20 students, were randomly allocated to an intervention and control conditions. Classes were then the units of assignment, with five classes in each condition. If students were independent of one another (i.e., intraclass correlation = 0), a statistical test that used students as the units of analysis would have an actual probability of rejecting the null hypothesis of .05. If the intraclass correlation among students was .05, the actual probability of rejection would be .16. If the intraclass correlation was .10, the probability of rejection would be .26.

Ignoring the intraclass correlation, therefore, will lead to specious declarations of statistical significance. The problem was recognized in the 1980s by Wolins (1982) among others, but its importance has become clear as a consequence of more recent work. See Hedges (2005) for technical detail and discussion of contemporary work.

Bibliography:

Bryk, A. S. and Raudenbush, S. W. (1992) Hierarchical Linear Models. Thousand Oaks, Ca: Sage Publications.

Donner, A. and Klar, N. (2000) Design and Analysis of Cluster Randomized Trials in Health Research. London: Arnold Publishing.

Hedges, L. V. (2005) Correcting a Significance Test for Clustering. Northwestern University. Submitted for publication (August 2005).

Kish, L. (1965) Survey Sampling. New York: Wiley.

Murray, D. M. (1998) Design an Analysis of Group Randomized Trials. New York: Oxford University Press.

Wolins, L. (1982) Research Mistakes in the Social and Behavioral Sciences. Ames, Iowa: Iowa State University Press.

Teacher-Intervention Confound

In some studies reviewed by the WWC, only one teacher is assigned to each condition. In particular, three different kinds of studies involve only one teacher per condition. The technical guidance discusses each case in turn. Some of these cases apply only to randomized controlled trials (RCT) while others apply to both RCTs and quasi-experimental designs (QEDs). The final case applies not only to teachers, but to any aggregated units such as classrooms, schools, or districts.

1) RCTs with one teacher per condition, and students randomly assigned to teachers
2) RCTs and QEDs with one teacher teaching both conditions, and students assigned to conditions.
3) RCTs with one teacher, school, or district randomly assigned to each condition and students are not randomly assigned, and similar QEDs

Finally, this guidance focuses on one specific technical issue, the confound between teacher and intervention. The study's ultimate disposition (i.e., meets evidence standards, meets evidence standards with reservations, does not meet evidence screens) also depends on how it fares on other criteria in the WWC Study Review Standards.

RCTs with one teacher per condition and students randomly assigned to teachers

This part of the guidance focuses on RCTs only, and does not apply to QEDs.

1. In some studies, one teacher may teach curriculum A and a different teacher may teach curriculum B. Children are then randomly assigned to each teacher/curriculum combination.

This is indeed a randomized trial. But the estimate of the intervention's effect is problematic because the teacher and intervention are confounded. That is, the effect of teachers usually cannot be disentangled from the effect of the intervention; consequently, the estimate of the intervention's effect could be then subject to potentially serious bias.

2. The default for handling these studies is the following:

> Such an RCT study should generally be downgraded or even discarded if (a) the study does not demonstrate that the confounding problem is negligible and (b) the PI and Review Team regard the potential bias in estimating effect as non-trivial.

3. In certain domains and interventions, it is possible for teacher effects to be negligible. For instance, a computer instruction program may be relatively free-standing and require little teacher engagement in the actual programmatic instruction and measurement of outcomes. In a comparison of two such computer programs, teachers might have very little effect on either condition. If the PI and Review Team agree that the study author demonstrates that teacher effects and the potential bias are negligible, then the study may

be regarded as an RCT without a teacher confound problem (that is, the study is neither downgraded nor discarded).

If the teacher has some role in implementing the intervention, but that role is limited by the nature of the intervention (e.g., predominantly computer-based), it is reasonable to assume some limited teacher effects. In this case, the study might be downgraded, but not discarded.

For interventions where the extent of teacher engagement (and therefore the possible teacher effect) is unclear, the burden of proof is on the study authors to demonstrate that teacher effects are negligible, are likely to have little or no impact on the study findings, and therefore the study should not be downgraded or discarded.

RCTs and QEDs with one teacher teaching both conditions and students assigned to conditions

1. In some studies, one teacher may teach curriculum A in one class and the same teacher may teach curriculum B in a second class. Students are then randomly assigned to each class.

2. The study is a fair test of the intervention if the PI and Team believe it is reasonable to assume (a) that the teacher's ability and motivation to teach curriculum A is the same as his or her ability and motivation to teach curriculum B or (b) that effects of the teacher are negligible for this intervention (e.g., as in the example above, an intervention may require very little input on the part of a teacher). The study may provide evidence bearing on either assumption, and this should be recognized by the PI and Review Team. For instance, the study may tell the reader that the teacher is well trained in each curriculum.

This situation is analogous to some surgical trials in which the same surgeon uses two different approaches in each arm of a randomized trial. Patients are randomly assigned to each arm of the trial, but the same surgeon performs the surgery in both arms.

3. The study is not a fair test of the intervention if the PI and Review Team do not feel there is adequate basis for making any of the above assumptions. For instance, (a) the study may provide no information about the plausibility of the assumptions, and (b) the PI and Review Team regard the assumptions as implausible based on the study's contents, and (c) the PI and Review Team regard the potential bias due to teacher confound in estimating the intervention's effect as non-trivial.

4. For RCTs in which a single teacher teaches both the intervention and the control conditions, and students are randomly assigned to conditions, the WWC recommends the following default disposition.

The study should be downgraded if:

- The study author does not demonstrate equal ability and motivation of the teacher in teaching both conditions

OR

- The study author does not demonstrate negligible teacher effects for the particular intervention (if counter evidence exists)

OR

- The PI and Review Team regard the potential biases in estimating the intervention's effect as non-trivial .

The study is not downgraded if there is a strong case that teacher ability and motivation are equal in each condition or teacher effects are negligible for the particular intervention (and consequently there are no serious potential biases in the estimate of the intervention's effect). However, the PI and team should explain in the intervention report that the teacher is assumed to be equally skilled and motivated to teach in each condition.

5. QEDs are handled similarly. The study should be downgraded (i.e., discarded) if:

- The study author does not demonstrate equal ability and motivation of the teacher in teaching both conditions

OR

- The study author does not demonstrate negligible teacher effects for the particular intervention (if there is counterevidence)

OR

- The equating procedures are absent or inadequate,.

The reasons for discard should be documented and explained in the intervention report.

QEDs in which a single teacher teaches both conditions can be included in the WWC's review if the study author demonstrates that the teacher ability and motivation are equal in both conditions or teacher effects are negligible for the particular intervention. Again, this should be made explicit in the intervention report .

RCTs with one teacher, school or district randomly assigned to each condition, and students are not randomly assigned, and similar QEDs

1. A study may be based on two intact classrooms and their teachers, where one intact class and its teacher may be assigned randomly to condition A, and the second intact class and teacher assigned to condition B.

More generally, a study involving two aggregated intact units (e.g., classrooms, schools, or districts) may randomly assign one aggregated unit to the intervention condition and a second aggregated unit to a second condition. In the technical jargon, the aggregates are often called "clusters," "groups," or "places."

2. In each case, the unit of randomization is at the cluster (aggregate or place) level. In each case, only one unit (the teacher and her classroom, an entire school, etc.) out of two such units is randomly assigned to each treatment condition.

3. A correct statistical analysis at the level of the unit of randomization (schools or districts or classes) cannot be done without invoking untestable assumptions. This is because the number of degrees of freedom associated with statistical tests (such as t or F), confidence intervals, etc. is zero. Put another way, neither statistical significance tests nor confidence intervals can be calculated at the proper level of analysis (i.e. the level of randomization) if the study is viewed as a randomized trial. In addition, any estimate of the intervention's effect is confounded with the teacher's effect.

This design, with N=1 unit of randomization in each arm of a randomized trial, is not a good randomized design.

4. The study author may have analyzed the data at a level of units lower than the level of random assignment. For instance, the study may report an analysis based on data at the level of individual students within the randomly assigned classes, or students within the randomly assigned schools in an attempt to adjust for difference between students in different classes or schools. The study can be construed as a QED if the analysis was done this way.

5. The default disposition for such RCTs is as follows:

> Such a study should generally be downgraded by the PI and Team. Depending on the study design, analysis, and the assumptions, the study may have been analyzed as a QED. If it does not meet the standards for a QED, it should be discarded. The PI and Review Team should document the reason for the discard.

6. QEDs in which schools (or other entities, such as classes or districts) are confounded with interventions are problematic in that the effects of schools and effects of the intervention usually cannot be disentangled, and the assumption that the school effects are equal is usually not plausible. Further, post facto matching of students or (equivalently) statistically equating is often suspect. For instance, if the schools differ appreciably in their location and characteristics of students (New York City versus rural Iowa), no amount of matching or statistical equating is likely to assure that the groups of children that are finally selected as being comparable within schools will indeed yield a statistically unbiased estimate of the effect of the intervention.

The WWC regards the assumption as patently implausible in the study context (or regard the equating as patently inadequate) and should then downgrade the study and discard it. Reasons for the discard must be given. However, the PI and Review Team may include a study of this type if they can provide compelling evidence that the required assumptions have been met.

Bibliography

Wolins, Leroy. (1982) *Research Mistakes in the Behavioral and Social Sciences*. Ames, Iowa: Iowa State University Press.