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.