

Study Design and Implementation Assessment Device (version 1.1)

The Study Design and Implementation Assessment Device (Study DIAD) is a system for assessing the degree to which the design and implementation of an individual evaluation permits conclusions about the causal effects of an intervention.

The fact that the Study DIAD focuses on research pertaining to the causal effects of educational interventions does not mean we believe that research designs meant to uncover causal relationships are the *only* tools that should be used by social scientists. Nor does it mean we believe that to be truly “scientific,” social science must be limited to randomized trials. To the contrary, we believe that (a) no single method can be used to address all interesting and important questions about educational interventions and (b) even when causal relationships are of primary interest, qualitative studies and quantitative surveys, among other types of research, yield important information about when, why, and how interventions work, and for whom. However, our central focus and the focus of the Study DIAD are on research designs—such as randomized trials, certain quasi-experiments, regression discontinuity designs—that have as their primary purpose uncovering causal relationships¹ (Shadish, Cook, & Campbell 2002).

Assessment of Study Design and Implementation

As Boruch (1997) writes, “Estimating the effect of ... [an] educational program requires comparing the condition of the individuals who have received the new service against the condition they would have been in had they not received the service” (p. 1). However, determining this latter condition is a difficult task. It involves taking into account ordinary growth, random fluctuations, and changes in the environment, among other variables. Design and implementation assessments are meant to evaluate the extent to which studies that claim to estimate the relative effects of an intervention take these issues into account.

Assessments of the design and implementation of studies on the causal effects of educational interventions can be used for multiple purposes. One purpose is to establish criteria for including and excluding studies from a research synthesis. Specifically, some studies may be so poorly designed and implemented that we would not want to include them in our evidence reports. Their results are so suspect that we would not know what to conclude from them. Another purpose for design and implementation assessments is to see if different research designs lead to different results. Finally, design and implementation assessments can be used to help draw conclusions about the cumulative strength of an entire set of studies. Each of these different purposes can play a role in how the Study DIAD is used.

Other Attempts to Assess Study Design and Implementation

Problems with existing scales. Before setting out to develop our own tool to assess study design and implementation, we examined tools that had been developed by others. We found several, and also found reasons to be skeptical about their validity. Specifically, empirical evidence suggests that existing quality scales disagree about what quality is. In a demonstration of this disagreement in medicine (a field often thought to have greater consensus about research quality than education), Jüni,

¹ We are currently exploring the feasibility of incorporating standards for single case experimental research into the Study DIAD.

Witschi, Bloch, & Egger (1999) applied 25 different quality scales to 17 studies reporting on trials comparing the effects of low-molecular weight heparin (LMWH) to standard heparin on post-operative deep vein thrombosis. The authors applied the 25 quality scales to the 17 trials, and then performed 25 different meta-analyses examining in each case the relationship between study quality and the effect of LMWH (relative to standard heparin). Studies were divided into “high quality” and “low quality” categories, with the high and low categories defined by a quality threshold given by the original authors of the quality scales or by median split when such a threshold was not provided. Then, the authors examined the conclusions of the meta-analyses separately for “high” and “low” quality trials. For six of the quality scales, the “high quality” studies suggested no difference between LMWH and standard heparin, while the “low quality” studies suggested a significant positive effect for LMWH. For seven other quality scales, this pattern was reversed. That is, the “high quality” studies suggested a positive effect for LMWH, while the “low quality” studies suggested no difference between the two conditions. The remaining 12 quality scales resulted in conclusions that did not differ between “high” and “low” quality trials. In addition, there was no association in these studies between the overall quality score and effect size using any of the 25 quality scales.

Lipsey & Wilson (1993) examined the results of over 300 *meta-analyses* in clinical psychology and education (broadly defined). They found 27 meta-analyses that involved an explicit comparison of studies rated on their methodological quality. Like Jüni et al (1999), they found that there was no relationship between the quality rating and the magnitude of the effect observed in their studies.

In both cases, the quality assessments seem to have been at best useless and at worst misleading. We do not believe this result occurred because study quality doesn't matter. Rather, we believe it is likely that the study quality assessments were so poor that they made it impossible to detect the real relation between study quality and study results. Most problematically, the instruments share a reliance on a single summary score to represent a study's quality. Especially when scales focus on more than one aspect of validity, the single score approach results in a score that is summed from very different aspects of study design and implementation, many of which are not necessarily related to one another. For example, there is no necessary relation between the validity of outcome measures and the mechanism used to allocate participants to groups. Thus, when scales combine disparate elements of study design into a single score, it is likely that important considerations of design are being obscured. For example, hypothetical Scale A might give Study X low marks because a fair comparison was not used (say 20 points) but high marks because a variety of participant characteristics were represented in the sample (say 60 points) to get a total score of 80 points. Scale A might then give hypothetical Study Y high marks because a fair comparison was used (60) but low marks because a very limited range of participant characteristics were represented in the sample (20). This would result in these two very different studies receiving the same total score of 80. When multiple dimensions are combined, it makes it very difficult to understand what those scores mean.

Lessons learned from existing quality scales. What lessons can be learned from this review of study quality scales? We think there are at least four. First, study design and implementation needs to be assessed on multiple dimensions. Internal validity is an important aspect of study quality, but it is clearly not the only one. Thus, we believe that a thoughtful approach to assessing design and implementation requires recognition of the importance of all four general classes of validity. Second, we believe that it is a mistake for scales that do focus on more than one dimension of study quality to attempt to summarize those dimensions using a single score. Doing so obscures important differences between studies and results in a number that is both useless and uninterpretable. Third, there is little justification or even agreement for complex schemes that weight items on quality scales (once we

abandon the single score approach, the value of this exercise is greatly diminished anyway). Fourth, assessments should be tied to explicit and transparent rules for relating the operational characteristics of studies to the judgments of quality. This way, the interjudge reliability of the scale will be enhanced and when disagreements about quality do arise, the source of the disagreement can be identified.

A note about context. As mentioned earlier, one potential use for assessments of study design and implementation is to set criteria for inclusion and exclusion from a research synthesis. It is therefore important to identify the critical contextual elements of the synthesis question. For example, one cannot evaluate the design and implementation of a study without knowing specifically what the intervention is, how it should be implemented, and what outcomes it should affect.

This assertion may seem trivial, but in practice it requires substantial input from individuals with significant substantive expertise to address these questions adequately. The list of questions that must be answered before undertaking an assessment of study design and implementation is presented below (see “Definitions of important terms” below).

Structure of the Study DIAD

With the lessons from our review of existing quality scales in mind, we developed the What Works Clearinghouse’s Study Design and Implementation Assessment Device, or Study DIAD. We attempted to create one instrument that can be used to answer questions at four different levels of specificity. We wanted the most general level to be understandable to an audience of nonresearchers and the most detailed level to be specific enough to satisfy researchers’ desire for comprehensiveness and explicitness. Thus, the Study DIAD results in four levels of assessment of design and implementation that are hierarchically related. The people who complete the Study DIAD answer specific, low-inference questions about study design and implementation. The answers to these questions are then compared to algorithms, or patterns of responses, that result in the answers to more than thirty questions about design and implementation. These questions feed into eight more general composite questions about design and implementation and finally, the answers to the eight questions combine yet again to answer four even broader questions.

Put somewhat differently, the Study DIAD is arranged as a hierarchy of related questions (and answers) starting with specific study characteristics that are used to build global assessments of the *confidence* we have that a study has uncovered the effects of the intervention. The questions are arranged so that answers at one level—the most specific level—feed into a set of design and implementation questions at a second level, a set of composite questions at a third level, and then into a fourth level of even more general questions. In essence, anybody could look at the tool at four different levels of abstractness depending on the uses they intended for its assessments. But, no matter which level served their purpose, users could also see the other levels and know how the assessments at each level related to or depended on the level below it. Each of the four levels is described below, starting with the most specific.

Coding Level Questions

The Study DIAD rests on a foundation of highly specific questions that are answered about each study. The number of questions varies by topic, but most topics will have from eighty to 100 questions. These questions are designed to require as little inference as possible to answer them. Examples of these questions include the sample sizes, specific sample characteristics (e.g., student age), and the reliability of scores on the outcome measures.

Design and Implementation Questions

The coding-level questions feed into thirty-eight more general questions that are answered about each study. For example, at this level you will find a question that reads “Was the assumption of statistical independence met, or could dependence (including dependence arising from clustering) be accounted for in estimates of effect sizes and their standard errors?” Another question that appears at the most specific level is “Were intervention conditions known to study participants, providers, data collectors, and/or other authorities (e.g., parents, teachers, case managers)?” This question relates to issues concerning how participants in the comparison group might have reacted to not receiving the intervention.

Eight Composite Questions

The answers to specific questions are then compared to a set of algorithms used to generate answers to eight composite questions about design and implementation. These eight questions are more general. For example, one question at this level is “Were the participants in the intervention group comparable to the participants in the comparison group?” The answer to the more specific question mentioned above on random assignment clearly feeds into answering this question, but other questions are relevant as well. The eight composite questions are listed below:

1. Was the intervention relevant to the review?
2. Were the outcome measures relevant to the review and properly aligned to the intervention?
3. Were the participants (e.g., students, schools) in the group receiving the intervention comparable to the participants in the comparison group? (Note: The meaning of the term “comparability” differs somewhat depending on the research design employed in the study.)
4. Was the study free of events that happened at the time as the intervention that confused the intervention’s effect?
5. Were targeted participants, settings, outcomes, and occasions used in the study?
6. Was the intervention tested for its effect within important subgroups of participants, settings, outcomes, occasions, and intervention variations?
7. Could accurate effect sizes be estimated?
8. Were statistical tests adequately reported?

Four Global Questions

After the eight composite questions are answered, they can be used to answer four more global questions about a study. Again, explicit algorithms are used to turn the composite eight questions into the global four questions. The questions at the most global level deal with what Cook and Campbell (1979), and Shadish, Cook, and Campbell (2002) referred to as the four most general sets of threats to the validity of a study: construct validity, internal validity, external validity, and statistical validity. The four questions that the Study DIAD answers at its most global level are related to each of these four sets of threats. In order, they are the following:

1. Were the intervention and outcome relevant to the review?
2. Was the intervention the cause of the change in the outcome?
3. Was the intervention tested on relevant participants (for example, students and schools) and environments (for example, classrooms and occasions)?
4. Could accurate effect sizes be derived from the study report?

Missing Data

Unfortunately, study reports often omit critical aspects of study design, implementation, analysis, and results that are of interest to later readers. This can occur for several reasons, including: it is sometimes difficult for study authors to predict in advance what information later users of the report will find interesting; (b) journal editors, in the interests of saving space, sometimes edit important details out of the manuscripts they publish; (c) reporting conventions in a topic area don't require consistent presentation of some information; (d) the information later readers need may not be central to the goals of the study; and (e) simple carelessness or poor training.

Some missing data can be recovered by contacting study authors, and we recommend that reviewers do so whenever feasible. In addition, studies are occasionally available in multiple formats (e.g., both as a dissertation and as a journal article). We believe that good practice involves scouring all available information for data relevant to the review.

The Study DIAD anticipates the problem of missing data by providing advice to reviewers about what they should assume about a study's design and implementation if the study report does not mention some characteristics. For instance, we suggest that reviewers assume that scores on an outcome measure are *not* reliable if the study authors give no indication of score reliability. Alternatively, we suggest that reviewers assume that there were no events confounded with the intervention unless there was explicit reason to believe so. We do not expect that reviewers adopt our conventions uncritically. Rather, we hope to influence the practice of literature reviewing by encouraging that missing data problems be thoughtfully considered before the reviewers start examining their data.

Definitions of Important Terms

Finally, we need to point out two issues related to how different design features are defined. First, many terms we use are defined in a glossary that accompanies the Study DIAD. New terms may be added in the future (see Appendix A). Please, let us know if you have suggestions regarding these definitions.

Second, we noted above that a good design and implementation assessment device needs to be flexible. In particular, certain terms and contextual issues will be specific to each topic area. Therefore, you will see in the Study DIAD some terms that appear quite vague or ambiguous. In most instances, these are terms that have to be given meaning by the individuals conducting the review. For example, the What Works Clearinghouse asks the experts assigned to a topic to provide answer these questions prior to reviewing any studies. In addition, independent peer reviewers are given the opportunity to comment on them. These contextual questions are as follows:

1. What commonly shared and/or theoretically derived characteristics of the intervention should be present in its definition and implementation?
 - i. Which of these characteristics are necessary to define interventions that “fully,” “largely,” and “somewhat” reflect commonly shared and/or theoretically derived characteristics?
 - ii. What variations in the intervention are important to examine as potential moderators of effect size?
2. What important characteristics of the intervention would we need to know in order to reliably replicate it with different participants, in other settings, at other times?

3. What are the important classes of outcomes?
 - i. What classes of outcomes are needed to conclude that a reasonable range of operations and/or methods have been included and tested?
4. Does the evidence report team have a minimum level of score reliability for outcomes to be considered in the review? If so, what are the specific minimum reliability coefficients for internal consistency, temporal stability, and/or inter-rater reliability (as appropriate)?
5. During what interval of time should studies have been conducted to be appropriate for the evidence report?
6. What characteristics must a sample have in order to be eligible for this review?
7. In addition to a pretest of the outcome, what are the important characteristics of participants that might be related to the intervention's effect and must be equated if a study does not employ random assignment?
8. What characteristics of subgroups of participants are important (a) to have variation on and (b) to test within a study to determine whether an intervention is effective within these groups? What levels or labels capture this variation?
 - i. Which of these characteristics of subgroups of participants are needed to conclude that a reasonable range of characteristics have been included and tested?
9. What characteristics of settings are important to test within a study to determine whether an intervention is effective within these groups?
 - i. Which of these characteristics and settings are needed to conclude that a reasonable range of tests have been conducted?
10. What is the appropriate interval for measuring the intervention's effect relative to the end of the intervention?
11. For purposes of sampling, what constitutes the local pool of participants?
 - i. If students are drawn from the same local pool, which groups of individuals (e.g., students, teachers, parents, administrators, case workers) might have been able to interfere with the fidelity of the comparison if they had known who was in the intervention and comparison groups?
12. For research on this topic, how would you define differential attrition from the intervention and control groups?
13. For research on this topic, how would you define severe overall attrition?
14. What constitutes a minimal sample size that would permit a sufficiently precise estimate of the effect size?
15. Are there statistical properties of the data that the team wished to record and explore during data analysis as potential moderators of the effect size? If so, what are they?

16. What percentage of important statistical information (i.e., sample size, direction of effect, effect size) is needed for the results of this study to be “fully”, “largely”, and “rarely” reported?

Other Important Study Variables

The questions on the Study DIAD do not represent the totality of information that should be extracted from studies included in a review. The review teams should also extract more specific information about (a) the intervention and how it was implemented, (b) the research design and how it was implemented, (c) the individuals in the study, and (d) the outcomes of the study. To assist with this, we have provided a list of other desirable characteristics of study design and implementation that are not part of the Study DIAD (see Appendix B). Other characteristics will undoubtedly be needed for particular topic areas. These characteristics may be desirable aspects of methodology or may be of particular relevance to particular research questions. As an example, even when random assignment is not possible, it is often desirable to allocate participants to study groups on a basis that is *not* related to the outcome variable. As noted, these characteristics of studies will be coded and, if possible, examined as potential moderators of effect size.

Updating the Study DIAD

We have labeled this version of the Study DIAD as Version 1.1 because we consider it to be an evolving document that will incorporate changes that arise as a function of putting it to use or through recommendations made by others. Our development process has led to an impressive amount of convergence regarding what design and implementation features should be included on the Study DIAD, but we also hope to collect information on the experience of reviewers as they apply the Study DIAD to specific studies in diverse topic areas, on the reliability of coding, on differences between Study DIAD assessments of study design and implementation and other instruments, and on end-user reactions to its output, among other indices of its validity and utility.

References

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage Publications.

Chalmers, I., Hedges, L. V., & Cooper, H. (2001). A brief history of research synthesis. *Evaluation and the Health Professions, 25*, 12-37.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges, (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Jüni, P., Witschi, A., Bloch, R., and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*, 1054-1060.

Kenny, D. A. & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Methods, 99*, 422-431.

Lipsey, M. W. & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2d ed.). New York: McGraw Hill.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

For citation purposes, please refer to this document as follows:

Valentine, J. C. and Cooper, H. (2004). *What Works Clearinghouse Study Design and Implementation Assessment Device* (Version 1.1). Washington, DC: U.S. Department of Education.

Study DIAD (Version 1.1)

The Study DIAD (Version 1.1) is presented below. Because the instrument is complex, we have divided it into two sections. First, we show how the design and implementation questions relate to the eight composite questions. Then, we demonstrate how the eight composite questions are translated into four global questions.

Composite Question 1. Intervention’s Relevance to the Review

Was the intervention relevant to the review?

- Yes, the intervention was adequately described, and it fully reflected ideas about what the intervention should be.
- Maybe yes, the intervention was adequately described, and it at least largely reflected ideas about what the intervention should be.
- Maybe no, there were important details missing from the description of the intervention and/or possible problems with its implementation.
- No, the intervention did not reflect ideas about what it should be and/or there were known problems with its implementation.

| | Response Pattern (Read down to determine the answer to the question.) | | | |
|--|--|------------------|-----------------------------|------------|
| | Fully | Largely | Fully, Largely, or Somewhat | Not at All |
| 1.1 To what extent does the intervention reflect commonly held or theoretically derived characteristics about what it should contain? | Yes | Yes | Yes or No | Yes or No |
| 1.2 Was the intervention described at a level of detail that would allow its replication by other implementers? | No | No | Yes | Yes or No |
| 1.3 Was there evidence that the group receiving the intervention might also have experienced a changed expectancy, novelty, and/or disruption effect not also experienced by the comparison group? | Yes | Yes | No | Yes or No |
| 1.4 Was there evidence that the intervention was implemented in a manner similar to the way it was defined? | Yes | Maybe Yes | No | Yes or No |
| Answer to Composite Question 1 associated with this response pattern: | Yes | Maybe Yes | No | No |

Note: If unable to determine from the report, the answer to questions 1.2, 1.3, and 1.4 is “no,” and the answer to question 1.1 is “not at all.”

Note: Any pattern of answers to questions 1.1, 1.2, 1.3, and 1.4 not identified above results in a “maybe no” answer to composite question 1.

Composite Question 2. Outcome Measure’s Relevance to the Review

Were the outcome measures relevant to the review and properly aligned to the intervention?

- Yes, the report presented evidence that the outcome measure was properly defined and aligned to the intervention.
- [There is no “maybe yes” answer for this question.]
- Maybe no, there was evidence that the measure had face validity and was properly aligned to the intervention. However, evidence suggested the measure might not be reliable.
- No, it is unclear what the outcome was.

| | Response Pattern (Read down columns to determine the answer to the question.) | | |
|---|--|-----------------|-----------|
| | Yes | Yes | Yes or No |
| 2.1 Do items on the outcome measure appear to represent the content of interest to this Evidence Report (i.e., have face validity)? | Yes | Yes | Yes or No |
| 2.2 Was there evidence that the outcome measure was measured with acceptable score reliability? | Yes | No | Yes or No |
| 2.3 Was the outcome measure properly aligned to the intervention condition? | Yes | Yes | No |
| Answer to Composite Question 2 associated with this response pattern: | Yes | Maybe No | No |

Note: If unable to determine from the report, the answer to questions 2.1, 2.2, is “no” and 2.3 is “proper.”

Composite Question 3. Clarity of Causal Inference: Fair Comparison (for Randomized Designs)

Were the participants (e.g., students, schools) in the group receiving the intervention comparable to the participants in the comparison group?

- Yes, participants were randomly assigned to conditions and few participants dropped out during the study.
- Maybe yes, random assignment was used but there was severe dropping out by participants.
- Maybe no, random assignment was used but there was differential dropping out of participants across conditions.
- No, although random assignment was used, participants dropping out during the study probably led to the groups not being comparable.

| | Response Pattern (Read down to determine the answer to the question.) | | | |
|--|---|------------|--------------|-----------|
| 3.1 Was there differential attrition between intervention and comparison groups? | No | No | Yes | Yes |
| 3.2 Was there severe attrition overall? | No | Yes | No | Yes |
| Answer to Composite Question 3 associated with this response pattern: | Yes | Yes | Maybe | No |

Note: If unable to determine from the report, the answer to questions 3.1 and 3.2 is “no.”

Note: A failed randomized experiment (i.e., one receiving a “maybe no” or a “no”) may still qualify for inclusion as a quasi-experimental design.

Composite Question 3. Clarity of Causal Inference: Fair Comparison (for Quasi-Experimental Designs)

Were the participants (e.g., students, schools) in the group receiving the intervention comparable to the participants in the comparison group?

- [There is no “yes” answer for these types of designs.]
- Maybe yes, reasonable steps were taken to make the groups comparable.
- Maybe no, although steps were taken to make the groups comparable, the steps may not have been adequate.
- No, it is unlikely that the participants in the groups were comparable.

| | Response Pattern (Read down to determine the answer to the question.) | | | | | |
|--|---|---------------------|---------------------|-----------|-----------|-----------|
| | Yes | Yes | Yes | No | No | No |
| 3.3 Were adequate equating procedures used to recreate the selection model? | Yes | Yes | Yes | No | No | No |
| 3.4 Was there differential attrition between intervention and comparison groups after equating occurred? | No | Yes | No | Yes | No | Yes |
| 3.5 Was there severe overall attrition after equating occurred? | No | No | Yes | No | No | Yes |
| Answer to Composite Question 3 associated with this response pattern: | Maybe Yes | Maybe No | Maybe No | No | No | No |

Note: In rare cases, groups may be demonstrably equivalent (see glossary). When accompanied by an attrition problem (questions 3.4, 3.5a, and 3.5b above), groups must be demonstrably equivalent after attrition has occurred. When demonstrable equivalence has been established, the answer to composite question #3 is “maybe yes.”

Note: Any pattern of answers to questions 3.3, 3.4, and 3.5 not identified above results in a “maybe no” answer to Composite Question 3.

Note: If unable to determine from the report, the answer to questions 3.3, 3.4, and 3.5 is “no.”

Composite Question 3. Clarity of Causal Inference: Fair Comparison (for Regression Discontinuity Designs)

Were the participants (e.g., students, schools) in the group receiving the intervention comparable to the participants in the comparison group (that is the slopes of regression lines were similar on the assignment variable)?

- Yes, an assignment variable with specified cutoffs was used to place participants into groups and there was no attrition problem.
- Maybe yes, an assignment variable with specified cutoffs was used to place participants into groups but severe attrition may have affected study results.
- Maybe no, an assignment variable with specified cutoffs was used to place participants into groups, but differential attrition may have affected study results.
- No, an assignment variable with specified cutoffs was not used to place participants into groups.

| Internal Validity— Selection | Response Pattern (Read down to determine the answer to the question.) | | | |
|--|---|------------------|-----------------|-----------|
| | No | No | Yes | Yes |
| 3.6 Was there differential attrition between intervention and comparison groups? | No | No | Yes | Yes |
| 3.7 Was there severe attrition overall? | No | Yes | No | Yes |
| 3.8 Could all observations have received the intervention had the cutoff point been set differently? | Yes | Yes | Yes | Yes or No |
| Answer to Composite Question 3 associated with this response pattern: | Yes | Maybe Yes | Maybe No | No |

Note: If unable to determine from the report, the answer to questions 3.6, 3.7 and, 3.8 is “no.”

Composite Question 4. Clarity of Causal Inference: Lack of Contamination

- Was the study free of events that happened at the time as the intervention that confused the intervention’s effect?
- Yes, other events that might be alternative explanations to the intervention’s effect have been ruled out.
- Maybe yes, there were no other identified events that could be alternative explanations, but some alternative explanations remain plausible.
- [There is no “maybe no” answer for this question.]
- No, other events happening at the same time as the intervention may have caused the effect.

| | Response Pattern (Read down to determine the answer to the question.) | | | | | |
|---|---|------------|------------------|-----------|-----------|-----------|
| | No | No | No | Yes | No | Yes |
| 4.1 Was there evidence of a local history event? | No | No | No | Yes | No | Yes |
| 4.2a Were the intervention and comparison groups drawn from the same local pool? | No | Yes | Yes | Yes or No | Yes or No | Yes or No |
| 4.2b If yes, were intervention conditions known to study participants, providers, data collectors, and/or other authorities (e.g., parents, teachers, case managers)? | n/a | No | Yes | Yes or No | Yes or No | Yes or No |
| 4.3 Did the description of the study give any other indication of the strong plausibility of other intervention contaminants? | No | No | No | No | Yes | Yes |
| Answer to Composite Question 4 associated with this response pattern: | Yes | Yes | Maybe Yes | No | No | No |

Note: If unable to determine from the report, the answer to questions 4.1, 4.2a, 4.2b, and 4.3 is “no.”

Composite Question 5. Generality of Findings: Inclusive Sampling

Were targeted participants, settings, outcomes, and occasions included in the study?

- Yes, the targets are represented in the sample.
- Maybe yes, most important characteristics of the targets are represented in the sample.
- Maybe no, although some important characteristics of targets are represented in the sample, many important targets are not.
- No, the sampled participants were not part of the target population.

| | Response Pattern (Read down to determine the answer to the question.) | |
|--|---|---------------------------|
| | Yes | Yes |
| 5.1 Did the sample contain participants with the necessary characteristics to be considered part of the target population (for purposes of this Evidence Report)? | Yes | No |
| 5.2 To what extent did the sample capture variation among participants on important characteristics of the target population (for purposes of this Evidence Report)? | Fully | Fully or Reasonable Range |
| 5.3 To what extent did the study include variation on important characteristics of the target setting (for purposes of this Evidence Report)? | Fully | Fully or Reasonable Range |
| 5.4 To what extent were important classes of outcome measures included in the study? | Fully | Fully or Reasonable Range |
| 5.5 Did the study measure the outcome at a time appropriate for capturing the intervention's effect? | Yes | Yes |
| 5.6 Was the study conducted during the time frame appropriate for the Evidence Report? | Yes | Yes |
| Answer to Composite Question 5 associated with this response pattern: | Yes | Maybe Yes |
| | | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Note: If unable to determine from the report, the answer to question 5.1 is “no,” the answer to questions 5.2, 5.3, and 5.4 is “limited,” and the answer to questions 5.5 and 5.6 is “yes.”

Composite Question 6. Generality of Findings: Effects Tested within Subgroups

Was the intervention tested for its effectiveness within important subgroups of target participants, settings, outcomes, occasions, and intervention variations?

- Yes, the intervention was tested for its effectiveness on targeted variations.
- Maybe yes, the intervention was tested for its effectiveness within most important subgroups of the participants and settings.
- Maybe no, although the intervention was tested for its effectiveness within some important subgroups of the participants and settings, many were left out.
- No, at best the intervention was only tested for its effectiveness within limited important subgroups of the participants, settings, outcomes, occasions, and intervention variations.

| | | Response Pattern (Read down to determine the answer to the question.) | | | | |
|--|---|---|------------------|-------------------------------|------------------|------------------------|
| | | Fully | Reasonable Range | All Important Characteristics | Reasonable Range | Somewhat or Not at All |
| 6.1 | To what extent was the intervention tested for effectiveness within important subgroups of participants (for purposes of this Evidence Report)? | Fully | Fully | Reasonable Range | Reasonable Range | Somewhat or Not at All |
| 6.2 | To what extent was the intervention tested for effectiveness within important subgroups of settings (for purposes of this Evidence Report)? | Yes | Yes or No | Yes or No | Yes or No | No |
| 6.3 | Was the intervention tested for its effectiveness across important classes of outcomes? | Yes | Yes or No | Yes or No | Yes or No | No |
| 6.4 | Was the time of measurement (relative to the end of the intervention) tested as an influence on the intervention's effect? | Yes | Yes or No | Yes or No | Yes or No | No |
| 6.5 | Was the intervention tested for its effectiveness across important variations in intervention implementation? | Yes | Yes or No | Yes or No | Yes or No | No |
| Answer to Composite Question 6 associated with this response pattern: | | Yes | Maybe Yes | Maybe Yes | Maybe Yes | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Note: If unable to determine from the report, the answer to questions 6.1 and 6.2 is “not at all” and the answer to questions 6.3, 6.4, and 6.5 is “no.”

Composite Question 7. Precision of Outcome: Effect Size Estimation

Could accurate effect sizes be estimated?

- Yes, accurate effect sizes appear could be estimated.
- Maybe yes, there was some evidence of statistical issues that may have caused the effect sizes to be inaccurately estimated, but the likely impact on inferences was minimal.
- Maybe no, there was evidence that statistical issues may have caused the effect sizes to be inaccurately estimated.
- No, the assumption of statistical independence was not met, and dependence was not accounted for in the effect sizes.

| | Response Pattern (Read down to determine the answer to the question.) | | | |
|---|---|------------------|------------------|-----------|
| | Yes | Yes | Yes | No |
| 7.1 Was the assumption of independence met, or could dependence (including dependence arising from clustering) be accounted for in estimates of effect sizes and their standard errors? | Yes | Yes | Yes | No |
| 7.2 Did the statistical properties of the data (e.g., distributional and variance assumptions, if any, presence of outliers) allow for valid estimates of the effect sizes? | Yes | No | Yes | Yes or No |
| 7.3 Were the sample sizes adequate to provide sufficiently precise estimates of effect sizes? | Yes | Yes | No | Yes or No |
| 7.4 Were the outcome measures sufficiently reliable to allow adequately precise estimates of the effect sizes? | Yes | Yes or No | Yes or No | Yes or No |
| Answer to Composite Question 7 associated with this response pattern: | Yes | Maybe Yes | Maybe Yes | No |

Note: If unable to determine from the report, the answer to questions 7.1 and 7.3 is “no,” and the answer to questions 7.2 and 7.4 is “yes.”

Composite Question 8. Precision of Outcome: Statistical Reporting

Were the statistical tests adequately reported?

- Yes, the statistical tests were adequately reported.
- Maybe yes, sufficient statistical information was reported to allow, at a minimum, imprecise effect sizes to be calculated for most measured outcomes.
- Maybe no, effect sizes could not be calculated for most outcome measures.
- No, sample sizes were not reported for most measured outcomes, and/or neither the magnitude nor the direction of the effects could be discerned for most outcome measures.

| | Response Pattern (Read down to determine the answer to the question.) | | | | |
|--|---|---------------------------|---------------------------|---------------------------|---------------------------|
| 8.1 To what extent were sample sizes reported (or estimable) from statistical information presented? | Fully | Fully | Fully, Largely, or Rarely | Fully, Largely, or Rarely | Rarely |
| 8.2 To what extent could directions of effects be identified for important measured outcomes? | Fully | Fully or Largely | Fully, Largely, or Rarely | Rarely | Fully, Largely, or Rarely |
| 8.3a To what extent could effect sizes be estimated for important measured outcomes? | Fully | Fully or Largely | Rarely | Rarely | Fully, Largely, or Rarely |
| 8.3b If yes, could estimates of effect sizes be computed using a standard formula (or its algebraic equivalent)? | Fully | Fully, Largely, or Rarely |
| Answer to Composite Question 8 associated with this response pattern: | Yes | Maybe Yes | Maybe No | No | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Global Question 1. Relevance to the Review

Were the intervention and outcome relevant to the review?

- Yes, the intervention and the outcome measures were properly defined.
- Maybe yes, at a minimum the intervention at least largely reflected ideas about what it should be, and the outcome measures appeared to measure the content of interest.
- Maybe no, the intervention and/or the outcome measures were described only as members of broader classes (across which significant variation in content is to be expected).
- No, it is unclear what was done in the study.

| | Response Pattern (Read down to determine the answer to the question.) | | | | | |
|---|--|----------------------|----------------------|----------------------|--------------|--------------|
| Was the intervention properly defined? | Yes | Yes | Maybe Yes | Maybe Yes | No | Yes or No |
| Was the outcome properly defined? | Yes | Maybe Yes | Yes | Maybe Yes | Yes or No | No |
| Answer to Global Question 1 associated with this response pattern: | Yes | Maybe Yes | Maybe Yes | Maybe Yes | No | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Global Question 2. Clarity of Causal Inference

Was the intervention the cause of the change in the outcome?

- Yes, the major alternative explanations (selection and contamination) have been ruled out.
- Maybe yes, although steps were taken to make the groups comparable, it is possible that dropping out or a lack of randomization caused them to differ somewhat.
- Maybe no, random assignment was not used to make groups comparable, and it seems likely that any steps taken to make them comparable were inadequate.
- No, it is unclear what might have caused the difference.

| | Response Pattern (Read down to determine the answer to the question.) | | | | |
|---|--|----------------------|----------------------|----------------------|-----------|
| Were the participants in the group receiving the intervention comparable to the participants in the comparison group? | Yes | Maybe Yes | Yes | Maybe Yes | No |
| Was the study free of events that happened concurrently with the intervention that confused its effect? | Yes | Yes | Maybe Yes | Maybe Yes | No |
| Answer to Global Question 2 associated with this response pattern: | Yes | Maybe Yes | Maybe Yes | Maybe Yes | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Global Question 3. Generality of Findings

Was the intervention tested on relevant participants (for example, students, schools) and environments (for example, classrooms, occasions)?

- Yes, the proper targets were included and the effect of the intervention was tested within targets.
- Maybe yes, at least some important targets were included in the study, and the intervention was tested within some of these targets.
- Maybe no, many important targets were not included in the study, and/or the intervention was rarely tested within targets.
- No, the accessed sample was not part of the target population.

| | Response Pattern (Read down to determine the answer to the question.) | | | | |
|--|--|----------------------|----------------------|----------------------|-----------------|
| Were targeted participants, settings, outcomes, and occasions included in the study? | Yes | Maybe Yes | Yes | Maybe Yes | No |
| Was the intervention tested for its effectiveness within important subgroups of participants, settings, outcomes, and occasions? | Yes | Yes | Maybe Yes | Maybe Yes | Yes or No |
| Answer to Global Question 3 associated with this response pattern: | Yes | Maybe Yes | Maybe Yes | Maybe Yes | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Global Question 4. Precision of Outcomes

Could accurate effect sizes be derived from the study report?

- Yes, the statistical results were adequately reported and the effect sizes accurately estimated.
- Maybe yes, either the statistical results were only imprecisely reported, or there was evidence that the effect sizes may have been inaccurately estimated.
- Maybe no, there were important problems with the reporting of the statistical results and/or the estimation of the effect size.
- No, statistical results were not reported for most measured outcomes, and/or there was evidence that the effect sizes may have been inaccurately estimated.

| | Response Pattern (Read down to determine the answer to the question.) | | | | |
|---|--|----------------------|----------------------|--------------|--------------|
| Were the effect sizes accurately estimated? | Yes | Maybe Yes | Yes | Yes or No | No |
| Were the statistical tests adequately reported? | Yes | Yes | Maybe Yes | No | Yes or No |
| Answer to Global Question 4 associated with this response pattern: | Yes | Maybe Yes | Maybe Yes | No | No |

Note: A pattern of answers to these questions that is not specifically identified results in a “maybe no.”

Appendix A

GLOSSARY OF TERMS

This glossary includes the basic terms and phrases that are used in the introduction to the Study Design and Implementation Assessment Device (Study DIAD) and other materials. The glossary depends on efforts in the health, education, and other sciences to clarify language, and will be updated as standards of evidence and the language for explaining the standards improve. References to earlier definitions in scientific publications, and on which we depend here, are given where it seems sensible to do so.

Alignment of outcome measures to the intervention: The extent to which material presented on an outcome measure, such as an academic achievement test, is drawn from the same universe of content represented in (meant to be influenced by) the intervention, such as an academic program or practice that is purported to enhance achievement. Misalignment can occur because of over alignment (i.e., the materials used in the intervention are identical to material on the outcome measure) or because of under alignment (i.e., the outcome measure is irrelevant to the intervention).

Appropriate time for capturing the intervention’s effect: Occasionally, programs may be evaluated at an inappropriate time. For example, the program might be evaluated before the intervention might reasonably be expected to have an effect on participants.

Assumption of statistical independence: See independence assumption.

Attrition: Loss of participants that occurs after participants’ assignment to the intervention and control groups has taken place (also called *mortality*) (Shadish, Cook, & Campbell, 2002, p. 505). See also differential attrition.

Cluster randomized trial: A randomized trial in which organizations or other entities (other than individuals themselves) are assigned to one of two or more intervention conditions. See also randomized trial.

Comparison group: In a randomized trial or a quasi-experiment, a group that is compared with an intervention group and that may receive either an alternate intervention or no intervention (Shadish, Cook, & Campbell 2002, p. 506). In a randomized trial, this is often called a control group.

Construct: A theoretical variable that can be measured in a variety of ways, such as “value added,” “mathematics ability,” and so on.

Construct validity: The degree to which inferences are warranted from observed characteristics of participants, the intervention, and outcomes sampled within a study to the constructs these samples represent. For example, an outcome measure may demonstrate construct validity by establishing convergent and/or discriminant validity, or by behaving in theoretically predicted

ways in experimental settings, among others. For achievement outcomes, construct validity generally includes the degree to which outcome measures sample material not directly taught.

Contamination: As defined here, an event that (a) occurs at the same time as the intervention and (b) plausibly affects the outcome, such that it might be confused for an effect of the intervention.

Control Group: In a randomized trial, a control group consists of participants who were randomly assigned to normal or ambient conditions, were not randomly assigned to the intervention, and are being compared to the control group's performance.

Convergent validity: The idea that two measures of the same construct should correlate with each other (Shadish, Cook, & Campbell 2002, p. 506).

Crossover: The switching of intervention conditions by one or more participants during a study. The important factors relevant in identifying the likely effects of crossovers include: the direction of the movement (i.e., intervention to comparison group, comparison to intervention group, or both), how the crossovers were handled in the statistical analysis (intent to treat vs. treated as a member of the new group vs. dropped from analysis), when the crossovers occurred (e.g., before the intervention began/ early in the intervention vs. later in the intervention), and why the crossovers occurred. Some of the effects of crossover are addressed in the Study DIAD under contamination and attrition.

Differential attrition: In a randomized trial or a quasi-experiment, the individuals or entities assigned to one condition attrit at a rate that is different from the individuals or entities that were assigned to the alternative condition. Attrition may be influenced by faulty follow up strategies in a study, or may be caused by other factors.

Direction of effect: The direction (positive or negative) that the outcome measure has (relative to the comparison group) on the intervention group's standing on the outcome variable.

Discriminant validity: The idea that a measure of construct A can be discriminated from a measure of construct B, when B is thought to be different from A; discriminant validity correlations should be lower than convergent validity correlations (Shadish, Cook, & Campbell 2002, p. 507).

Effect size: The strength (or magnitude) of the relationship in the population, or the degree of departure from the null hypothesis (Rosenthal & Rosnow 1991, p. 42). In randomized trials, estimated effect size depends on the difference between mean outcomes in the interventions and control conditions, the variability in the groups, and the sample sizes in each group. Standardized effect size formulae are given in Rosenthal & Rosnow (1991).

Equate: Procedures used to make groups more comparable at the study design stage (e.g., matching, blocking, stratifying), or the analysis stage statistical controls (propensity score matching, covariance analysis), or both.

Face validity: The extent to which a test of academic achievement, or other characteristic of the target population “looks like” it measures what it is intended to measure (Nunnally 1967, p. 99).

Group randomized trial: See cluster randomized trial.

Implementation: The activities, both intended and unintended, that did and did not occur as part of the intervention conditions; includes intervention delivery, intervention receipt, and intervention adherence (Shadish, Cook, & Campbell 2002, p. 508, 316).

Independence assumption: The assumption that the random variation in participants’ responses to an intervention (stochastic error) in a randomized trial or a quasi-experiment are independent of one another (Rosenthal & Rosnow 1991, p. 315). Put less technically, the assumption of the independence of observations means that knowing one person’s score on an outcome measure give you no information about another person’s score on that same measure. Although the assumption of independence of observations is critical to the validity of statistical tests used to judge the effectiveness of interventions (Kenny & Judd, 1986), it is often the case that the assumption is violated in educational research, because many groupings are present in the design of the study (e.g., small work groups, classrooms, schools, etc.).

Intervention: A policy, program, or practice, that is, an assembly of activities and processes that is given to some students, schools, or classrooms, and not given to others.

Local history event: An event occurring between the beginning of the intervention and the posttest within the context of the intervention, outcome, time, setting, and persons studied that could have produced the observed effect in the absence of the intervention (Shadish, Cook, & Campbell 2002, p. 508).

Local pool: Persons having a geographical or other contextual relationship with the sampled participants. For example, students in the same classroom or school might be considered part of the same local pool.

Meta-analysis: The statistical analysis of a collection of study results (Cooper & Hedges, 1994). See also: research synthesis.

Multiple occasions: Outcomes are measured at multiple points in time, relative to the end of the intervention.

Observational study: See Quasi-experiment.

Other intervention contaminants: In the Study DIAD, events or processes other than those listed individually that might be confused with an intervention effect. These would include interactions between contamination threats and group membership, for example, when multiple interventions occur and the intervention effect only appears when other interventions are present.

Participants: Individuals, classrooms, schools, or other entities that are the targets of observations of a study on the effects of an intervention.

Place-based randomized trial: See cluster randomized trial.

Quasi-experiment: An experiment in which participants were not randomly assigned to groups (Shadish, Cook, & Campbell, 2002, p. 511). Also called observational studies in the Cochrane statistical tradition. See Rosenbaum (2002).

Random assignment: In a randomized trial, any procedure for assigning participants to conditions based on chance, with every participant having the same fixed and known probability of being randomly assigned to the intervention condition. This probability may be .50 or some other value between 0 and 1.

Random selection: A procedure in which each member of the population has the same fixed and known probability of being included in a sample. This probability may be .50 or some other value between 0 and 1.

Randomized trial: A study in which individuals, classrooms, or entities such as schools, are randomly assigned to different intervention and control conditions so as to produce a statistically unbiased estimates of the relative effect of the interventions and a legitimate statement of one's confidence in results (Boruch, 1997).

Regression discontinuity design: A study design in which participants are assigned to the intervention and control conditions on the basis of a cutoff score on a pre-intervention measure of need or merit whose statistical relation to an outcome measure is known or assumed. This is as opposed to a random assignment process being used to assign participants to groups (Shadish, Cook, & Campbell, 2002, p. 208).

Regression to the mean: Also known as statistical regression, the tendency of extreme scores on one measure (e.g., a pretest) to be less extreme on another measure (e.g., a post-test). Regression to the mean can be confused for or mask a real effect of an intervention. Regression to the mean is most likely to be problematic when participants have been selected because they had scores that were higher or lower than average (Shadish, Cook, & Campbell, 2002, p. 57).

Research synthesis: Research synthesis attempts to integrate empirical research for the purpose of creating generalizations in a way that is (a) intended to be exhaustive in the coverage of the database and (b) initially nonjudgmental with regard to the outcomes of the synthesis (Chalmers, Hedges, & Cooper, 2001). See also: meta-analysis.

Sample: The subset of the population for whom we have obtained observations (Rosenthal & Rosnow 1991, p. 628).

Sample size: The number of observations upon which an effect size is based.

Selection effect: A bias in estimating the effect of an intervention that is caused by the fact that eligible people or entities in a quasi-experiment choose to participate in either the intervention or

the control condition, as opposed to eligible people or entities being randomly assigned to either condition.

Self-selection: When participants decide the condition they will be in (Shadish, Cook, & Campbell 2002, p. 512). For example, program volunteers are compared to non-volunteers or people completing the intervention are compared to people not completing the intervention.

Significance levels: In statistical convention, the threshold condition for probability of a Type I error. The level of alpha; also called *p value* (Rosenthal & Rosnow 1991, p. 629).

Single case research: Research in which the main target of observation is a single unit. The unit is most commonly an individual, but can also be a classroom, a school, a community, etc. Within this unit, multiple observations may be made over time, over the individuals contained in the unit, and so on.

Systematic review: See research synthesis.

Target population: The complete collection of individuals for whom the intervention is designed.

Unit of analysis: The unit at which statistical analyses were performed. This may differ from the unit of assignment, the unit of intervention delivery, or the unit of observation.

Unit of assignment: The level at which participants (e.g., students, schools) were assigned to intervention and comparison or control conditions.

Appendix B

Other Characteristics to Code from Studies and Examine as Potential Moderators of Effect

Additional Descriptors for Report Characteristics

- Source of report (e.g., published vs. not)
- Source of published reports (e.g., peer reviewed vs. not)

Additional Descriptors for Research Design

- Specific research design (e.g., regression discontinuity, interrupted time series)
- Type of randomization procedure, if any (e.g., random vs. stratified random assignment)
- Information on the specifics of how the randomization procedure was carried out (e.g., randomization mechanism, whether masked allocation was used, etc.)
- Type of equating procedure, if any, for randomized designs
- If random assignment was not used, how did participants get into groups (i.e., self-selection vs. put in groups on a basis related to outcome variable vs. put in groups on a basis not related to outcome variable)
- Type of comparison condition (e.g., waiting list control vs. no treatment control)
- Relationship between evaluator and intervention (e.g., product developer vs. independent evaluator)

Additional Descriptors of the Intervention

- Unit in which intervention was delivered (e.g., individual vs. small group)
- Length of time the intervention had been in operation prior to evaluation
- Incentives for participation given to intervention and comparison groups
- Implementation of the comparison condition
- Ratio of participants to staff

Fidelity of Implementation of the Intervention Relative to its Definition

- Implementation staff (e.g., teachers vs. clinicians)
- Staff training
- Access and/or dosage
- Proper use

Additional Descriptors of the Sample

- Age of sample (mean age, range)
- SES of sample
- Percent of each sex
- Percent speaking language other than English in the home
- Student characteristics (e.g., average vs. gifted)
- Population density of school district (e.g., large urban vs. small urban)
- State or region in which school district resides

Additional Descriptors of the Outcome Measures

- Data type (e.g., dichotomous vs. discrete)

- Source of measure (e.g., self-report vs. teacher report)
- Type of reliability estimate (e.g., internal consistency vs. test-retest)
- Reliability estimate
- Range restriction

Additional Descriptors of the Data Analysis

- How attrition was handled in analyses (e.g., intent to treat vs. completers)
- Original analysis or reanalysis by independent party
- Evidence of systematic data exclusion