Analytic Technical Assistance and Development

Conducting Strong Quasi-experiments

Version 1

May 2015

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Contents

Exhibits

Appendix

Choosing a Study Design to Measure Program Effects

Researchers who plan to study the effectiveness of a policy, program, or practice should choose a study design that maximizes scientific rigor for the context and fits within cost and operational constraints.

Researchers and developers should always consider first whether it is feasible to implement a randomized controlled trial to examine program effectiveness.

Randomized controlled trials (RCTs), or "experiments," provide strong evidence of effectiveness. Random assignment ensures that treatment and control groups do not differ except for receipt of the intervention. In a well-designed and well-implemented RCT, researchers can be more confident that they are measuring program effects and not effects of something else.

Implementing an RCT is not always feasible. For example, providers may be unwilling or unable to limit participation in a program to some students when the program being studied has more seats available than applicants. Also, funders or policymakers may decide to begin a study after a program is under way. A late start to a study makes random assignment infeasible except in special circumstances. For example, some charter schools use lotteries to allocate their open spaces, and it is possible to use the lotteries after the fact to set up an experiment. But nearly all experiments are planned from the start.[1]

When conducting an RCT is not possible, a strong quasi-experimental design (QED), or quasi-experiment, can provide valuable evidence about a program's effectiveness. This brief discusses best practices and objectives in designing and implementing strong QEDs, presents answers to frequently asked questions from developers and researchers who want their studies to meet the U.S. Department of Education's standards of rigor, as defined by What Works Clearinghouse (WWC). The brief also summarizes common pitfalls that cause QEDs not to meet WWC standards for group design studies.[2]

How key terms are used in this brief

Treatment (intervention): the policy, program, practice, or strategy that will be evaluated.

Treatment group: the group of individuals, classrooms, schools, districts, or institutions participating in the study and the intervention. (Studies sometimes call this the *participant group* or the *intervention group*.)

Comparison group: the group of individuals, classrooms, schools, districts, or institutions participating in the study but not participating in the intervention. Although often used interchangeably in other studies, in this brief, we refer to this group as the "comparison group" for QED studies and "control group" for RCT studies.

Strong QED: a quasi-experimental design study that meets standards for credible evidence of effectiveness. In this brief, we focus primarily on the What Works Clearinghouse evidence standards.

[1] See Resch, Berk and Akers (2014) for guidance on recognizing and conducting opportunistic experiments in education field settings.

Factors and Guidelines in a QED Design

A key principle in designing quasi-experiments is that the more the quasi-experiment is like a true experiment, the stronger its validity (Rosenbaum 2010). In an experiment, random assignment creates both (1) a treatment group and (2) a control group that is the treatment group's mirror image. A QED will be stronger if its comparison group is as close as it can be to a mirror image of the treatment group. Because a QED is not using random assignment, it should use other approaches to create the mirror-image comparison group.

Exhibit 1. In a strong QED, the comparison group will be close to a mirror image of the treatment group.

Treatment Group **Comparison Group**

The analogy of a mirror image is a useful way to think about an ideal quasi-experiment. In practice, a comparison group for any QED is not a perfect mirror image of its treatment group. Even if what can be seen or measured (for example, gender, race/ethnicity, achievement, or previous experience) about the groups is exactly equivalent, there is no presumption that what cannot be seen will be equivalent, as there is in an experiment. And, crucially, it is not possible to test whether these characteristics differ, for the simple reason that they cannot be measured.

Unobserved variables—related or unrelated to outcomes

These unmeasured characteristics, often called "unobserved" variables, do not present issues for the study if they are unrelated to outcomes. For example, researchers evaluating a postsecondary developmental math program do not need to worry that more students in the treatment group like peanut butter and jelly sandwiches. Taste in sandwiches is not correlated with math skills.

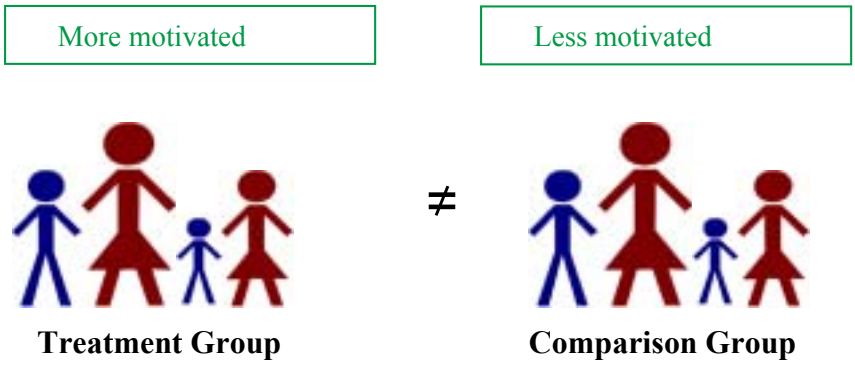
But it is easy to think about unobserved variables that are related to outcomes. Suppose that a school offers a voluntary summer catch-up program for children who are reading below grade expectations. Parents who enroll their children in this program may be different in important, hard-to-observe ways from those who do not enroll their children. For example, they may read to their children more often than parents whose children did not enroll. That difference by itself could create a difference in children's motivation to read, even for students who are the same on other dimensions such as age, gender, family income, and spring test scores. If more motivated

---

[2] The WWC standards relevant for RCTs or QEDs, and applied here, are called "group design" standards. Group design studies measure impacts by comparing outcomes for a treated set of individuals versus a comparison set of individuals, not by comparing outcomes for the same individuals over time. This brief focuses on aspects of the WWC group design standards specifically related to QEDs, hereafter referred to simply as "the WWC standards." For a comprehensive discussion of all WWC evidence standards, including group design standards, consult the What Works Clearinghouse Procedures and Standards Handbook, Version 3.0.

students with greater parent support enroll in the catch-up program, their reading skills may improve faster over the summer than those of non-enrolled students, independent of any effect the catch-up program might have had. This example illustrates a key limitation facing any quasi-experiment. When the study estimates program effects without measuring characteristics that are related to outcomes, part of the measured effect may arise because of differences in, say, motivation to read. The study cannot say how much of the measured effect is a real program effect and how much is due to differences in unmeasured characteristics. And no matter how much effort the study invests in gathering data to increase the number of measured characteristics (such as by surveying parents about their literacy practices when their child was younger), something relevant will always be unobserved and unmeasured.

Exhibit 2. The comparison group may look the same as the treatment group, but may differ in ways that researchers cannot observe or measure (like motivation), making it hard to argue that differences in outcomes are due solely to the treatment.



Matching strategies for creating comparison groups

Strategies for creating comparison groups range from conveniently identifying a group that is "like" the treatment group to carefully selecting a group by using matching techniques. For example, a convenient approach might be to select students in neighboring schools who are not implementing a particular program. These students could be used as a comparison group for a treatment group in schools that are using the program. This approach is inexpensive and straightforward to implement, but it risks creating groups that are not equivalent on important characteristics, which would be evident only after data are collected.

Careful matching strategies identify comparison group individuals that satisfy some metric of equivalence or closeness to treatment group individuals ("individual" could refer to students, teachers, schools, districts, postsecondary institutions or higher education systems). Rosenbaum (2010) surveys the vast literature and provides extensive discussion about matching and useful examples.

What Works Clearinghouse (WWC) Standards

For experiments, the WWC has specific criteria for the conditions that must be met when researchers conduct random assignment. However, for quasi-experiments, the WWC does not scrutinize the appropriateness of the matching approach. Any of the approaches reviewed by Rosenbaum or used in the literature are acceptable.

The WWC *does* scrutinize two aspects of the comparison group:

- Whether the characteristics on which treatment and comparison groups were matched align with characteristics specified in a WWC topic area's protocol[3]
- How "close" the groups are on those characteristics.

  Example 1. In the topic area of *beginning reading*, the treatment and comparison groups must be equivalent on the pre-test score of the reading outcome and on other social and demographic characteristics.

  Example 2. For studies of college access or success interventions, the *WWC postsecondary education topic area protocol* requires that the groups be equivalent on a pre-intervention measure of the outcome or a close proxy (or, if pre-intervention measures are not available, baseline measures of socio-economic status and pre-intervention academic achievement measures such as SAT scores or high school grades).

The WWC applies other standards in addition to equivalence. Some apply to both experiments and quasi-experiments, and others apply to one design or the other.[4] For QED studies to meet WWC group design standards with reservations, they must:

- **Compare two distinct groups**, not the same group before and after a treatment.
- **Use appropriate outcomes**[5] that are:
  - **Valid**: that is, the study measures what it says it measures
  - **Reliable:** that is, the outcome is measured consistently and accurately
  - **Measured in the same way** (using the same instrument and at the same time) for the treatment and comparison groups.
  - **Not too closely aligned with the treatment.** "Overalignment"—such as when a reading test includes questions that relate only to the reading program being studied—

Because a strong QED study cannot rule out potential bias in the impact estimates, the highest rating it can receive from the WWC is "meets WWC group design standards with reservations."

[3] The WWC has identified a number of priority topic areas for review covering a range of K–12 and postsecondary issues. Each WWC review protocol within a topic area specifies the population and types of interventions that can be reviewed, criteria for study relevance (for example, time frame and study design), acceptable outcome measures, and statistical and analytic requirements, including the required list of pre-intervention characteristics on which groups must demonstrate equivalence.

[4] For example, for experiments but not for quasi-experiments, the WWC applies an attrition standard which measures sample loss from random assignment until follow-up data collection. Experiments with low sample attrition are the only studies eligible to obtain a rating of "meets WWC group design standards without reservations."

[5] If a study does not have any acceptable outcomes, then it could not meet WWC group design standards. If a subset of outcomes is acceptable, a QED study is eligible to meet WWC group design standards with reservations.

leads to inaccurate measurement of program effects because one group will naturally score differently than the other.

- **Demonstrate *baseline (pre-intervention) equivalence* of the treatment and comparison groups** on characteristics specified in the relevant WWC review protocol. Studies must demonstrate pre-intervention equivalence for their **analysis sample** (in other words, the sample used to measure program impacts). Equivalence of characteristics before participation in the intervention is required to be within 0.25 standard deviations. And if differences are between 0.05 and 0.25 standard deviations, impact estimates must statistically adjust for these differences by using methods such as regression analyses or analysis of covariance. If all differences are smaller than 0.05 standard deviations, then effects can be measured as simple differences of means.

- **Be free of *confounding factors***. A confounding factor is one that affects outcomes of one group but not the other and is not part of the treatment being studied. When confounds arise, it is impossible to know whether measured effects are from the treatment or from the confounding factor, or some combination. Examples of common confounds include time (treatment and comparison groups come from different school years) and when either the treatment or comparison group come from a single school or classroom, but there are others.[6]

Additional guidelines for sound research

Other considerations fall within the larger context of conducting sound research. The WWC does not have standards for these aspects of studies, but they are useful as guidelines for researchers. In sound research:

- Studies should specify clear research questions up front.

- Studies should determine sample design and data collection approaches to answer the research questions. The sample design should specify clear eligibility criteria, methods for forming the research sample, and sample sizes necessary to detect meaningful impacts of the intervention on key outcomes. The data collection plan should identify valid, reliable outcome measures needed to answer the research questions.

- Plans for analysis should reflect the research design and sample selection procedures.

---

[6] When outcomes for the treatment group are from one school year and outcomes for the comparison group come from a different school year, the measured effect is confounded with differences that may arise between the different school years because of, for example, year-to-year changes in leadership and staffing, other programs that were implemented or taken away from one year to the next, or external issues that may have affected outcomes in one year but not the other (for example, major weather-related interruptions). A second common example is a case in which a treatment is implemented in only one classroom. In this situation, the measured effect is confounded with the effects of the classroom teacher. A third example is a study in which all treatment schools are in one district and all comparison schools are in another school district. In this case, the measured effect confounds the treatment effects and differences between the two districts. A fourth example is a case in which a treatment is always delivered in combination with another treatment—for example, when a first-year college transition course is offered at a student support center where there is easy access to tutors and mentors. In this case, the study is measuring the combined effect of both the transition course and the enhanced access to supports. If only one of the treatments fits the WWC topic area, the confound means that the study does not meet WWC group design standards.

These plans and the intervention ideally will be well implemented for a study to produce the strongest evidence of a treatment's effectiveness. Researchers who conduct effectiveness research in field settings such as classrooms and schools often encounter challenges and hurdles to implementing the intervention or analyzing its effects. But starting with a clear plan and being flexible will yield stronger research than starting with a vague plan and hoping for the best. In field settings, researchers should expect that unplanned or unforeseen events will hamper study designs rather than strengthen them.

Appendix A presents (1) a checklist of issues to consider when designing strong QEDs and (2) a supporting table that provides a comprehensive discussion of the checklist. The table defines each key design issue, explains the extent to which the WWC considers the issue in determining a study rating, and documents general good practice considerations. Readers may want to review the table before delving into the next two sections of this document ("Frequently asked questions" and "Common pitfalls").

Frequently Asked Questions about Meeting WWC Group Design Standards

The following list presents frequently asked questions *most relevant to QED studies*. General frequently asked questions about the WWC are available here. Also, we encourage you to browse the resources section of the WWC website, which provides useful documents and links to WWC webinars and databases. Notably, the *What Works Clearinghouse Procedures and Standards Handbook, Version 3.0* provides the most detailed discussion of the WWC evidence standards.

Q1: What kinds of outcomes are reviewed by the WWC?

> **Answer:** Evaluations designed to meet WWC standards should always include at least one outcome that falls within the acceptable list of outcomes as defined by specific WWC review protocols. In situations where a relevant WWC protocol is not available, then researchers should focus on outcomes related to achievement, progression through school, completion of education programs, degree attainment, and student behaviors. Outcomes such as attitudes or perceptions, while often important to measure as mediators, are not reviewed by the WWC and thus should not be the sole focus of a research study that hopes to meet WWC standards.

Q2: What kinds of comparison group designs are eligible to meet "WWC group design standards *without* reservations"?

> **Answer:** The only studies that are eligible to obtain a rating of "meets WWC group design standards without reservations" are well-implemented randomized controlled trials (RCTs). A WWC webinar on July 21, 2014 provides extensive advice on implementing a strong RCT study. Quasi-experimental designs are not eligible for this rating.

Q3: What kinds of *nonrandomized* comparison group designs are eligible to "meet WWC group design standards *with* reservations"?

> **Answer:** The WWC does not have any requirements about how treatment and comparison groups are constructed. All QED studies in which there are distinct treatment and comparison groups will be eligible for review under WWC group comparison designs, and the highest rating that those studies can achieve is "meets WWC group design standards with reservations."

Q4: Can a study that measures the same sample before and after a treatment meet WWC group design standards?

> **Answer**: No. To be eligible to be reviewed under WWC group design standards, there must be two distinct groups: a *treatment group* that receives a treatment and a *comparison group* that does not. Studies of groups that serve as their own controls can only be reviewed by the WWC under very specific circumstances and only under WWC pilot single-case design standards. For more information on these pilot standards, consult Appendix E of the *What Works Clearinghouse Procedures and Standards Handbook, Version 3.0*.

Q5: Can a QED study that compares an earlier cohort to a different later cohort meet WWC group design standards with reservations?

> **Answer**: No. Comparing two different cohorts is a type of confound that makes it impossible to determine whether the program being tested is responsible for differences in outcomes between a treatment and comparison group or whether some other competing explanation

accounts for differences in outcomes. In most schools, classrooms, and programs, a variety of things change from year to year other than the intervention being evaluated. When a historical cohort is used as a comparison group, it is not possible to assess whether other activities have affected outcomes. For example, suppose that a new program is implemented in a set of schools. Outcomes for students at the end of that year are then compared to outcomes of students enrolled in the same schools in the prior year. It is not possible to separate differences in outcomes between the earlier and later cohorts that are due to the treatment and differences that changed between the two time periods. A new district-wide or institution-wide policy or some external force may have affected outcomes.

Q6: Can a QED study that compares two different doses—for example, 1 year versus 2 years of exposure—of the same intervention be eligible to meet WWC group design standards with reservations?

Answer: No. WWC reviews focus on measuring full program effects. A study measuring differences in dosage would not be eligible for review by the WWC because it would not be possible to determine whether the intervention was having an effect.

Q7: Can a QED study that uses pre-existing data to identify a comparison group be eligible to meet WWC group design standards with reservations?

Answer: Yes. Any retrospective or prospective analyses of two distinct groups would be eligible to be reviewed under WWC group design standards. For example, researchers could use existing administrative longitudinal datasets to compare (1) outcomes for students attending schools that had implemented a particular program model with (2) outcomes for students attending schools that did not implement that model during that same time period. The WWC would review these comparisons under group design standards, regardless of whether the administrative data had already been collected in prior years (retrospective) or whether the administrative data is in the process of being collected (prospective).

Q8: Can a QED study in which the treatment and comparison groups are clustered—for example, within classrooms or schools—meet WWC group design standards with reservations?

Answer: Yes. Nonrandomized clustered QED studies are eligible to meet WWC group design standards with reservations and must demonstrate equivalence at the cluster level if the analysis is measuring cluster-level effects. If a treatment is clustered at, for example, the school or classroom level, but the analysis makes inferences at the student level, then the study must establish equivalence at the student level to meet WWC group design standards with reservations.

Q9: Can a study that compares only one treatment school with one or more comparison schools meet WWC group design standards with reservations?

Answer: No. The WWC would consider this design to be *confounded* because there would be only one unit (school) assigned to at least one of the treatment or comparison conditions. If there is only one unit, then some other characteristic (for example, teacher quality or alternative curricula available at that school) could explain differences in outcomes. For this reason, the WWC requires that both the treatment and control conditions contain at least two units to be eligible to meet WWC group design standards with reservations. This is also true when the intervention is clustered at the district, teacher, classroom, or any other level.

Q10: When a treatment is a curriculum and the teacher must implement the curriculum, does the teacher need to teach both the treatment and comparison groups so that there are no concerns about the particular effects of that teacher?

Answer: The WWC would consider a QED study with one teacher who teaches multiple classrooms in each of the treatment and comparison conditions as eligible to meet WWC group design standards. However, teachers need not teach both the treatment and comparison groups as long as there are multiple teachers or classrooms in the study in both the treatment and comparison groups. In this latter case, there must be at least two teachers in the treatment group classrooms and two different teachers in the comparison group classrooms to be eligible to meet WWC group design standards with reservations.

Q11: On what characteristics must QED studies demonstrate treatment and comparison group equivalence to meet WWC group design standards with reservations?

Answer: In general, most WWC topic area protocols require that studies demonstrate equivalence on a baseline measure of the outcome. Many also require or consider equivalence on other baseline characteristics, such as race-ethnicity and socioeconomic status. The WWC posts all topic area protocols that describe requirements for baseline equivalence on its website. If a relevant WWC topic area protocol is not available, researchers should consider using the key baseline characteristics that are highly correlated with outcomes when demonstrating equivalence of the treatment and comparison samples.

Q12: How does the WWC determine whether groups are equivalent?

Answer: The WWC examines equivalence by calculating the effect size difference between the treatment and comparison groups for each of the required baseline characteristics. An effect size is calculated as the difference in means between the treatment and comparison groups, divided by the pooled (treatment and comparison group) standard deviation. If an effect size difference is greater than 0.25, then the comparison does not meet WWC group design standards. If an effect size difference falls between 0.05 and 0.25, then the study must statistically control for this baseline measure in its impact analysis in order for the result to meet WWC group design standards with reservations. If the effect size difference is less than 0.05, then the study result is eligible to meet WWC group design standards with reservations, regardless of whether this measure is controlled for in the analysis. The WWC *does not* consider whether differences in baseline measures are statistically significant when assessing whether groups are equivalent. Requirements for which characteristics must demonstrate equivalence varies by WWC topic area (see Q10 for more information about which characteristics must be equivalent at baseline).

Q13: May survey data be used to demonstrate equivalence?

Answer: Yes. The WWC considers data from any source eligible for determining baseline equivalence, including surveys, test scores, reliable and valid observations, or administrative records.

Q14: Is establishing equivalence using a pretest that is different from the outcome measure acceptable?

Answer: In general, the WWC will allow demonstration of baseline equivalence on a similar outcome measure, provided that this measure has appropriate validity and reliability characteristics and falls within what the WWC calls the same outcome "domain." WWC

topic area protocols specify these outcome domains. For example, in reviews of post-secondary interventions, the "credit accumulation" domain includes outcomes such as number of credits earned, ratio of credits earned to credits attempted, or persistence measures such as number of continuous semesters enrolled. Researchers should review specific WWC topic area protocols to learn more about how domains are defined and also topic- and domain-specific equivalence requirements. For example, in the science topic area, math pretest scores are acceptable for demonstrating equivalence if a science pretest is not available.

Q15: In cases where propensity score matching is used to construct a comparison group, will the WWC accept equivalence of the propensity scores as evidence of equivalent groups?

Answer: No. Although researchers may choose to use baseline outcomes and other demographic characteristics to calculate the propensity scores, it is not acceptable to present equivalence of only the propensity score. Baseline equivalence for the analysis sample (the sample used to measure program impacts) must be demonstrated for each of the required baseline characteristics.

Q16: What if I have only test scores from a prior school year as a baseline measure? Will the WWC accept this as a measure of baseline equivalence?

Answer: The WWC allows the use of test scores from prior school years to demonstrate equivalence. Because achievement can change over time (for example, achievement levels may increase or decrease over the summer depending a youths' summer experiences), measuring achievement immediately prior to the start of the treatment will provide the most accurate depiction of each study participant's starting point and is the preferred approach. However, the WWC accepts the approach of demonstrating equivalence on measures from a previous year when immediate measures are not available.

Q17: I am able to collect pretest information only after groups are formed and the intervention has begun. Will the WWC accept this as a measure of baseline equivalence?

Answer: The WWC allows the use of pretest scores obtained after groups are formed and intervention has begun to demonstrate baseline equivalence. However, authors should be cautious in collecting these data too long after the intervention has begun because, assuming that the intervention will have an effect on test scores, it is possible that scores may already start to diverge by the time students take the pretest. If the analysis sample includes groups that are not equivalent when pretest scores are obtained, then the study would be rated as "does not meet group design standards," even if careful matching procedures had been implemented before the intervention began. Also, the difference between the pretest and posttest will no longer measure the full impact of the program, which may reduce the chance that the study will detect significant differences in outcomes.

Q18: Should I also present equivalence information for subgroup analyses?

Answer: Yes. The WWC reports subgroup results as supplemental evidence of program effectiveness for the subgroups identified in the specific WWC review protocol. The WWC reviews equivalence information for these subgroups in order to determine whether these analyses meet WWC group design standards.

Q19: I am using gain scores from pretest to posttest as my outcome. Do I also need to adjust for differences using covariates?

Answer: If pretest differences in a QED fall between 0.05 and 0.25 standard deviations, the study must include the pretest as a covariate in a statistical model (such as ANCOVA or regression analysis), regardless of whether the study uses gain scores.

Q20: Do I need to worry about attrition (sample loss) in my QED?

Answer: Not for the WWC study rating. The WWC assesses attrition for RCTs but not QEDs.[7] Although the WWC does not factor attrition into the rating of QEDs, in studies where the treatment and comparison groups are carefully matched at baseline and followed over time, sample loss could lead to groups no longer looking the same when outcomes are measured. Also, substantial sample loss will reduce the analysis sample size, making it harder to detect statistically significant differences. For these reasons, researchers should always design studies that employ procedures that will maximize data collection efforts and minimize sample loss.

Q21: Is statistical power a factor in meeting WWC standards with reservations?

Answer: No. Statistical power is not a factor in whether a study meets standards. However, power may affect how the WWC characterizes results. A study finding that has an effect size of smaller than 0.25 that is not statistically significant will be considered by the WWC as an "indeterminate" effect. When designing their studies, researchers should consider what a reasonable expected effect size is and the associated sample size requirements to ensure that their study is adequately powered to detect effect sizes of that magnitude as statistically significant.

Q22: Can sample sizes be reduced as part of the matching process and still meet WWC group design standards with reservations?

Answer: Yes. A subsample of participants can be matched and included in the analysis to ensure that groups are equivalent on observed pre-intervention characteristics. Researchers should be consistent and transparent about their methods of sample selection and matching. If WWC reviewers have concerns that the method for achieving equivalence in the sample compromises the integrity of the study (for example, if there is evidence that participants were inconsistently excluded from the research sample across the treatment and comparison conditions on the basis of specific characteristics of the participants), it may lead to a rating of "does not meet WWC group design standards."

Q23: Can I impute missing outcome values in a QED and meet WWC group design standards with reservations?

Answer: No. Only the results from the unimputed analysis can even be considered as meeting the standards. When researchers impute outcomes, they substitute missing values with a best-guess estimate of what the value would have been if the information had been available. Researchers use a variety of methods to impute missing values. However, for QEDs (and RCTs with substantial sample loss), the WWC does not permit any outcome imputation, because of a concern that not enough information is known about the missing

[7] RCTs that have high attrition run the risk of no longer having equivalent treatment and control groups. For this reason, high attrition RCTs cannot meet WWC standards without reservations. Like QEDs, they must demonstrate that their analytic samples are equivalent at baseline to be eligible for a rating of "meets WWC standards with reservations."

sample members to be able to estimate what an outcome would have been if it had not been missing. For example, in a study of a postsecondary mentoring program where youths volunteered to be part of the program and are compared to those who chose not to participate, treatment group members who do not complete a follow-up survey may be those who did not engage with the program; whereas missing comparison group members may be a more diverse set of students who weren't available for follow-up and feel less connected to the study. In this case, it would be difficult to make assumptions about what the outcomes would have been if the researchers had been able to obtain the outcomes for everyone, and inaccurate assumptions made during imputation could inappropriately change the estimated magnitude of the program effect (or, in other words, the estimated effects could be "biased.")

Q24: Can a QED meet WWC standards with reservations if the comparison group receives an alternate treatment? For example, could I examine how one new curriculum compares to another new curriculum, or how one strategy for delivering a developmental math course compares to a specific different strategy for developmental math?

Answer: Yes, but, under some circumstances, the study may be excluded from a WWC review of overall intervention effectiveness. In the education research field, comparison groups often receive an alternate educational experience, whether it be "business as usual" (meaning that a school continues its status quo curriculum or set of supports) or a new, alternative curriculum or set of supports. When a study compares a new treatment to either a "no treatment" or a "business as usual" condition, it tries to answer the question, "what is the overall effect of this treatment in comparison to what students would have had if they not had the opportunity to receive the new treatment?" When a study compares two treatments, then it tries to answer the question, "How much better (or worse) do students fare when they participate in one treatment versus the other?"

For example, if a study is comparing the effects of a new developmental math course to no additional supports, then it will be able to measure how students who were given additional supports fare compared to those who did not have that opportunity. The WWC would report the results of this kind of study (and other similar studies) in an intervention report to help provide researchers, practitioners and policy makers with evidence of overall program effectiveness. Conversely, if two developmental math course curricula are compared "head to head" in a study, then the study results will help provide information to practitioners and policy makers who are trying to choose which one to implement. While this is useful information for those who are selecting between two developmental math programs, the results will never provide any evidence to support or refute the fact that a developmental math course will improve outcomes. This kind of "head to head" comparison can meet WWC standards, and may be included in a WWC product such as a single study review, but very likely would not be considered for inclusion in an assessment of overall intervention effectiveness in an intervention report.

Q25: Does the WWC require the research sample to represent the population of interest?

Answer: The WWC does not factor sample size or sample specification criteria into a study's rating. It is important to note, however, that some U.S. Department of Education grants require evaluations to be designed not only to meet WWC standards (with or without reservations) but also to focus on a specific population of interest to that grant program.

Q26: Are there cases where authors report a statistically significant effect but the WWC will report the effect as nonsignificant?

Answer: Yes. This occurs when a study does not take certain design features into account in the analysis.[8] When this happens, the WWC will either request information from the study authors in order to re-analyze the data or will re-analyze the data using defaults. Researchers planning a study should consider these issues at the design stage to ensure consistency of interpretation. Specific design shortcomings that can lead to WWC adjustment or reinterpretation of findings include

- **When a study has a clustered design but does not account for clustering in the analysis**. For example, if the treatment occurs at the classroom or school level, the study must statistically adjust for the clustered structure of the data. Researchers should design and analyze studies that appropriately adjust for clustering to obtain more precise estimates than the WWC would be able to calculate based on simple defaults.
- **A study does not adjust for testing multiple comparisons within the same outcome domain**. The more outcomes included in an analysis that measure a similar concept, the higher the probability that researchers may find at least one statistically significant effect by chance. The WWC uses the Benjamini-Hochberg procedure when multiple comparisons arise. Researchers should conduct multiple comparison adjustments when they are presenting impacts for multiple similar outcomes.[9] This issue also has bearing on study design decisions and sample size requirements. Researchers who are sparing in the number of outcomes measured within a domain, focusing on the strongest and most relevant measures, will have a greater ability to detect statistically significant program effects than if they include a wide range of outcomes. The more outcomes included, the larger the sample that will be necessary to detect statistically significant effects. Study designers should factor in the number of outcomes within a domain (in addition to other factors necessary to do an appropriate statistical power analysis) when assessing the sample size requirements for a study.

[8] Adjustments to statistical significance do not affect whether a study meets WWC group design standards but will affect how the WWC characterizes the findings.

[9] The What Works Clearinghouse Procedures and Standards Handbook, Version 3.0 provides detailed guidance on acceptable methods for adjusting for multiple comparisons.

Common Reasons Why QED Findings Are Rated *Does Not Meet WWC Group Design Standards*[a]

- **An inability to attribute measures of effectiveness solely to the intervention**, or, in other words, the study has a "confounding" factor. As described in the "WWC Standards" section and in Appendix A, a confounding factor occurs when a characteristic is completely aligned with either the treatment or comparison condition. Most commonly, the WWC has rated studies as having confounding factors because
    - A study compares outcomes for a cohort of participants in one year to a cohort in an earlier year.
    - A treatment or comparison condition is clustered within one classroom, school, or district. For example, all students assigned to the treatment have the same teacher.
    - The study uses different methods to collect data for the treatment and comparison groups. For example, researchers collect survey data from the treatment group and administrative records for the comparison group.
- **The outcomes do not meet WWC requirements**. Most commonly, this is because
    - The outcome is directly aligned with the content of the intervention (for example, a reading assessment that includes comprehension passages to which the treatment group students had already been exposed).
    - Either reliability statistics for outcomes (for example, inter-rater reliability of internal consistency) fall below the WWC's topic area protocol's standards or reliability information is not provided.
- **Equivalence of the analysis sample on pretest or other WWC protocol-specified demographic measures is not demonstrated**. This commonly occurs when
    - The authors do not collect baseline data to determine equivalence.
    - The study does not present equivalence information for the analysis sample. This often occurs when researchers either exclude equivalence information or provide equivalence information only for the original baseline sample and not the sample that is being used in the analysis. The WWC submits an author query to obtain these data and will rate a study as not meeting WWC group design standards if they receive no author response.
    - The analysis sample is not equivalent on the required measures according to WWC standards for demonstrating equivalence (see Appendix A for more details). This could occur even if groups are equivalent at the beginning of a study, because sample makeup can change over time. (Note that a study could meet WWC group design standards with reservations in some outcome domains and not meet standards in others if the topic area protocol specifies that equivalence needs to be demonstrated only within an outcome domain (such as reading, math, behavior, etc.).
- **The study uses analysis methods that do not meet WWC standards**. This often happens when
    - The study needs to adjust for baseline differences but does not include appropriate covariates in the analysis.
    - The QED study uses imputation methods to fill in missing values for outcome measures.

[a] A webinar presented by the WWC on March 3, 2015, also discusses common pitfalls that lead a QED study not to meet WWC group design standards. Issues most relevant to RCT studies are discussed in a July 24, 2014, webinar conducted by the WWC.

Appendix A. Checklist for Quasi-Experimental Designs; Study Design Characteristics to Consider

The checklist and accompanying table in this appendix highlight key issues that researchers should consider when designing strong QED studies. The checklist is broken into two sections. The first section focuses on design issues that may influence a study's WWC evidence rating. The second section covers other general design issues that researchers should factor in at the planning stage.

Checklist for QEDS during the Study Design Phase

Is my QED study designed to meet WWC group design standards with reservations?

A. The study will compare two distinct groups—a treatment group and a comparison group.
B. The comparison group will be drawn from a population similar to that of the treatment group, and groups will be equivalent on observable pre-intervention characteristics.
C. The contrast between the treatment and comparison groups will measure the impact of the treatment that I am interested in.
D. There will be no known confounding factors.
E. The study will collect pre-intervention measures of the primary outcomes of interest as well as background characteristics at baseline.
F. The study will collect valid and reliable outcome data that are most relevant to assess intervention effects.
G. The data collection process will be the same—same instruments, same time/year—for the treatment and comparison groups.

Is my study designed with additional qualities of a strong QED?

H. The study has pre-specified and clear primary and secondary research questions.
I. The study results will generalize to a policy or program-relevant population.
J. The study has established clear criteria for research sample eligibility and matching methods.
K. The study will have an analysis sample size large enough to detect meaningful and statistically significant differences between the treatment and comparison groups.
L. The study is designed to detect meaningful and statistically significant effects for specific subgroups of interest if this is a high priority for my study.
M. The planned analysis methods will appropriately reflect the research design and sample selection procedures.
N. The study includes a clear plan to document the implementation experiences of the treatment and comparison conditions.

Table A.1. Study Design Characteristics to Consider When Planning a Strong QED

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
A. The study will compare two distinct groups—a treatment group and a comparison group.	To measure the effect of a program or practice, a treatment group that receives the intervention must be compared to a separate comparison group that has not received this intervention. When these groups are not distinct (for example, the same group of students before and after a treatment), then it is impossible to isolate the effect of the intervention (for example, regular maturation could explain changes in outcomes over time).	To be eligible for review under group design standards, a study must have at least two distinct groups that are compared (sometimes there are more than two groups if multiple interventions are being tested or if there are multiple comparison groups). The WWC has no specific criteria for how researchers form these groups in QEDs. Retrospective data based on extant data and prospective nonrandomized design studies that rely on new data are both eligible for review as QEDs. Despite the fact that the WWC has no restrictions on how groups can be formed, choosing the right groups can have major implications for what will be tested and for WWC evidence ratings (see items B through G in the "Study design characteristic" column of this table).	At the study design stage, researchers should confirm that groups are distinct. Researchers must weigh issues related to cost, convenience, and timing when determining which groups will be included in the study. For more issues to consider when determining how to form treatment and comparison groups, see items B through G in the "Study design characteristic" column of this table.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
B. The comparison group will be drawn from a population similar to that of the treatment group, and groups will be equivalent on observable pre-intervention characteristics.	A comparison group that is drawn from a similar population has a stronger chance of serving as a proxy for what a treatment group would have experienced if it had not been exposed to the intervention. When the comparison group is drawn from a different population or setting, differences in outcomes between the treatment and comparison groups may be related to the characteristics of different settings rather than to the effect of the intervention. For example, it may not be possible to attribute differences in outcomes to a treatment for high-needs students if all of the treatment group students attend schools that serve predominantly urban, high-needs students and comparison group students attend schools that serve a more diverse set of suburban students.	If the treatment and comparison groups are drawn from different populations or settings, the WWC may conclude that the populations or settings are too dissimilar to provide an adequate comparison condition and may assign a rating of "does not meet WWC group design standards."	In a sound QED study, the comparison group should serve as a "mirror" to the treatment group. Researchers should analyze data at the study design stage to assess whether potential groups may be drawn from different populations. If so, then they should either (1) determine whether a different population can serve as a comparison group or (2) use careful matching techniques, such as propensity score matching or direct matching, on key characteristics that are highly related to desired outcomes. These efforts will help ensure that groups will be equivalent on observable characteristics. While it is not possible to match on unobserved characteristics, it is possible to use observable characteristics to match groups.
C. The contrast between the treatment and comparison groups will measure the impact of the treatment that I am interested in.	The contrast between the experiences of the treatment and comparison groups influences the interpretation of the program impacts in group design studies. The strongest contrast occurs when a fully implemented intervention experience is compared to either no alternate intervention or a "status quo" educational experience (like the established curriculum). In group design studies, the contrast can be minimized if the comparison group receives a new or existing alternative treatment that is similar to the intervention or if there are low rates of program participation.	In general, the nature of the contrast between the treatment and comparison groups would not exclude an eligible QED from receiving a rating of "meets WWC group design standards with reservations" and perhaps be included in a single study review. However, the WWC may deem a study ineligible for inclusion in a WWC Intervention Report if the treatment group receives only a slight incremental change above the comparison group's experience or if there is a "head to head" comparison of two new interventions.	Researchers should think carefully about what the contrast between the treatment and comparison groups will likely be. This contrast has implications for sample selection (that is, choosing a comparison group that is not participating in a similar intervention). This also highlights the importance of planning to measure participation rates and program implementation to learn more about the experiences of both the treatment and comparison groups.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
D. There will be no known confounding factors.	Researchers can more confidently attribute effects to an intervention when the only explanation for differences is that the treatment group received a program and the comparison group didn't. When another characteristic (a "confounding factor") that is unrelated to the intervention is present in either the treatment or comparison condition but not both, it is no longer possible to say with confidence that differences are due to the treatment. Differences could be due to that other characteristic. This can occur when there is only one "unit" in one or both conditions. For example, if there is one treatment teacher and two comparison teachers, and the treatment teacher is highly motivated and engaging, then which is having an effect— the treatment or the attributes of the treatment teacher? A similar situation can occur, for example, when a study compares students from one academic year to a prior academic year. What accounts for the differences—the program or other things that occurred during these academic years (like changes in leadership, staffing, or alternate curricular offerings?	Any study that has a confounding factor in which there is a known characteristic that is completely aligned with the treatment or comparison condition will be rated as "does not meet WWC group design standards." One exception is if a treatment is bundled with another intervention. A QED study of this type could meet WWC standards with reservations but may be excluded from specific WWC products, such as an intervention report if only one piece of the intervention (unbundled) is relevant to the WWC review.	Although it is not always possible to plan in advance for all contingencies that may arise during a study, researchers should carefully consider potential confounds during the sample selection process. Any potential confounds that arise during the course of a study should be documented carefully to help inform interpretation of study findings. For example, a QED study may have initially been designed to compare a new science supplemental program to no supplement. However, during the course of the study, all of the school principals in the treatment group schools decided jointly to also use a new science curriculum while the comparison group continued with its existing science curriculum. In this example, the study would no longer be able to isolate the effects of the science supplement alone, and researchers would need to be clear that they are now testing the effects of a combination of a new science curriculum and supplement.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
E. The study will collect pre-intervention measures of the primary outcomes of interest as well as background characteristics at baseline.	QED studies that collect pre-intervention measures can help provide evidence that the treatment and comparison groups were similar before program implementation, thus making evidence of program effectiveness more plausible. Because participants are not selected at random in QEDs, the treatment and comparison groups may differ in ways that we can observe as well as ways that we cannot. These initial differences, if they are related to program outcomes, could bias estimates of program effects.	The WWC requires that all QED studies demonstrate baseline equivalence for their analysis sample. Without collecting pre-intervention measures, a study would not meet WWC group design standards.	Researchers can use pre-intervention data to help formulate well-matched groups, to assess whether groups are matched, and to analyze and statistically control for pre-intervention differences in outcomes and other background characteristics.
F. The study will collect valid and reliable outcome data that are most relevant to assess intervention effects.	Studies with strong outcomes will provide the most useful evidence of program effectiveness. The most useful outcomes are not overly aligned with the intervention being tested, are general enough to be policy relevant, are replicable in other studies, and are specific enough so that researchers would expect that the intervention would affect them.	In general, QED study findings for outcomes that lack validity (i.e., don't measure what they are supposed to measure), reliability (i.e., aren't measured consistently), or are overaligned (i.e., measure content that is covered explicitly in the intervention but not comparison condition) will not meet WWC group design standards. If all outcomes in a study do not meet standards, then the entire study would be rated as "does not meet WWC group design standards."	Researchers should carefully select outcomes that have strong psychometric properties and are most relevant to measuring program effectiveness. Whenever possible, researchers should try to use strong pre-existing measures. Researchers can access many resources to see the wide array of outcomes currently available. When it is not possible to use existing measures, then researchers should carefully design outcomes that are not overly aligned with the intervention, and they should document the development process of the outcomes and psychometric properties.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
G. The data collection process will be the same—same instruments, same time/year—for the treatment and comparison groups.	When data are collected in a similar fashion, researchers can be more confident that differences in outcomes between the treatment and comparison groups are not due to the method of data collection. Differences in data collection can occur when, for example, data for the treatment group come directly from a student or teacher survey, but data for the comparison group come from administrative records. The timing of the data collected also needs to be the same for both groups.	The WWC considers differences in data collection, if completely aligned with the treatment and comparison conditions, as a "confound" (see "C" in the "Study design characteristic" column), and WWC would rate the results as "not meeting WWC group design standards."	During the design phase, researchers should plan data collection procedures to ensure that no confound is related to data collection. In addition, careful preparation and a clear data collection process can help to improve the quality of data collected and reduce sample attrition. In particular, if the study is a prospective study, researchers should make every effort to reduce both overall sample loss and differential sample loss between the treatment and comparison conditions (which could lead to the analysis sample no longer being equivalent, even if careful matching had occurred at the beginning of the study).
H. The study has pre-specified and clear primary and secondary research questions.	A carefully planned study that has specified primary and secondary research questions is more credible to its audience because it shows that researchers were not going on a "fishing expedition" to find significant results. It also helps to focus analyses on the most critical and relevant outcomes.	The WWC does not factor research questions in its reviews. WWC reviews focus on all outcomes specified in its topic area protocols, regardless of whether authors report these outcomes as primary or secondary.	Researchers should take the time at the beginning of designing a study to consider the most critical research questions and should use these questions to frame other design issues, such as sampling, matching techniques, outcome selection, and analysis and reporting plans.
I. The study results will generalize to a policy or program-relevant population.	Even if a sound study is designed in which similar groups are being compared and the contrast is clear, if the design is not representative of a relevant population, the results of the study will be of limited use to policymakers and practitioners.	The WWC does not consider sample size or composition of study populations in study ratings. However, WWC topic area protocols do have study participant requirements that determine whether a study will be eligible for review. Also, the WWC has created an "extent of evidence" rating that captures both the number of studies reviewed and the sample sizes of studies.	Researchers should choose a study population that is most relevant to answering questions of program effectiveness for a policy- and practitioner-relevant population and, when applicable, whether the population is relevant to particular grant program to which they are applying. If a convenience sample is the only sample available, researchers should think carefully about whether results from the study will be useful and relevant.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
J. The study has established clear criteria for research sample eligibility and matching methods.	Having clear eligibility criteria enables researchers to focus recruitment and sample selection on the most relevant population. It also helps researchers formulate matching strategies to ensure the equivalence of treatment and comparison groups, which reduces that concern that there may be an alternative explanation if differences between the groups are detected.	The WWC does not have any specific requirements regarding sample eligibility or matching techniques that would affect a study's evidence rating. However, when the WWC determines whether a study is eligible for inclusion in a specific review (for example, for inclusion in a WWC intervention report), the study must include sample characteristics that are aligned with a particular topic area protocol. (For example, studies in the Early Childhood Education topic area must include studies of interventions for children between the ages of 3 and 6 who are not yet in kindergarten and are attending school- or center-based programs.)	Researchers who conduct well-implemented QED studies consider issues related to sample recruitment and participant eligibility early in the study design process. Appropriate matching procedures should be determined on the basis of (1) each study's particular situation and (2) factoring in key issues related to availability of baseline data, sample size availability, and key concerns about baseline characteristics that are most related to the outcomes of interest.
K. The study will have an analysis sample size that is large enough to detect meaningful and statistically significant differences between the treatment and comparison groups.	A study with adequate statistical power will have a large enough sample size to detect expected statistically significant effects. This will prevent the danger of making an incorrect assessment that a program doesn't affect outcomes when it actually does.	The WWC does not consider statistical power in evidence ratings. However, the WWC would report underpowered results as having "no discernable effect" if there are no statistically significant differences and the effect size difference is less than 0.25. (For example, see Question Q20, "Does the WWC consider statistical power in WWC evidence ratings?" in the FAQ section.)	Statistical power should be carefully analyzed in the study design phase to ensure that there is an adequate sample to detect the expected differences between the treatment and comparison contrasts. Researchers should carefully consider what the expected effect may be and how well other covariates may help reduce variation in outcomes. Also, researchers should analyze the power ramifications due to the clustered nature of results if the intervention will be provided at the cluster (e.g., classroom, school) level.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
L. The study is designed to detect meaningful and statistically significant effects for specific subgroups of interest if this is a high priority for my study.	If researchers are specifically interested in knowing whether a program works for specific subgroups, then a study should be designed with a large enough sample within these subgroups to detect expected differences for these subgroups.	The WWC does not consider statistical power in evidence ratings. The WWC reports subgroup findings as supplemental evidence if they are specified as subgroups of interest in WWC topic area protocols. If a study is not powered to detect significant effects for subgroups, then researchers run the risk of the WWC reporting that the intervention had no discernible effects for the subgroups, even if the results are in the expected direction and of the expected magnitude.	Researchers should determine, in advance, whether there are specific high-priority subgroups of interest and should design a study with enough of a sample to be able to detect expected effects. They may also review relevant WWC topic area protocols to see whether the WWC would report these findings as supplemental evidence of intervention effectiveness.
M. The planned analysis methods will appropriately reflect the research design and sample selection procedures.	Well-designed analysis plans improve a researcher's ability to report credible estimates of program effects. Analyses that are not well designed and implemented run the risk of yielding imprecise estimates of program effects that researchers and policymakers may not consider useful.	Certain analysis methods affect a WWC's rating, and others could affect how the WWC portrays results. QED studies that require statistical adjustment for baseline differences (see the equivalence discussion on page 4 and Study design characteristics "B" in this table) will not meet WWC group design standards if appropriate covariates are not included in the analyses by using methods such as regression or ANCOVA (gain score, ANOVA, or difference-in-difference analyses would not be acceptable). QED studies that impute baseline or outcome data also will be rated as not meeting standards by the WWC. Other design issues can affect whether the WWC will deem results to be statistically significant, including (1) using a model that appropriately adjusts for clustering in the research design and (2) applying statistical adjustments for multiple comparisons within specified outcome domains.	Researchers should carefully plan their analyses in advance, including determining the best statistical model that fits the research design and sampling methods, determining the primary and secondary outcomes and planned adjustments, and planning appropriate sensitivity analyses to see how results vary, depending on assumptions made.

Study design characteristic	This is important because . . .	How does the WWC consider this issue in its ratings?	General considerations
N. The study includes a clear plan to document the implementation experiences of the treatment and comparison conditions.	Careful documentation of program and comparison experiences provides invaluable evidence to help understand why a program did or did not find significant results. It helps to document the contrast between the treatment and comparison conditions, whether the program was implemented as intended, and the degree to which research subjects participated in the program.	The WWC does not consider implementation issues in the study rating. The WWC does narratively describe program implementation in its reviews for studies that meet WWC group design standards.	Researchers who are planning a prospective study should develop a strong implementation analysis plan that measures and documents adherence to the intervention, the contrast between the treatment and comparison conditions, and contextual issues specific to the study (such as changes in the environment or adaptations that were made over time). Researchers should also consider including in their study a careful assessment of program quality, although it might be costly.