What Works Clearinghouse (WWC) Group Design Refresher Training

April 2024

Allison McKie
WWC-TOAST Training Lead
Mathematica

Elias Walsh
WWC-TOAST Deputy Project Director
Mathematica

IES Institute of Education Sciences

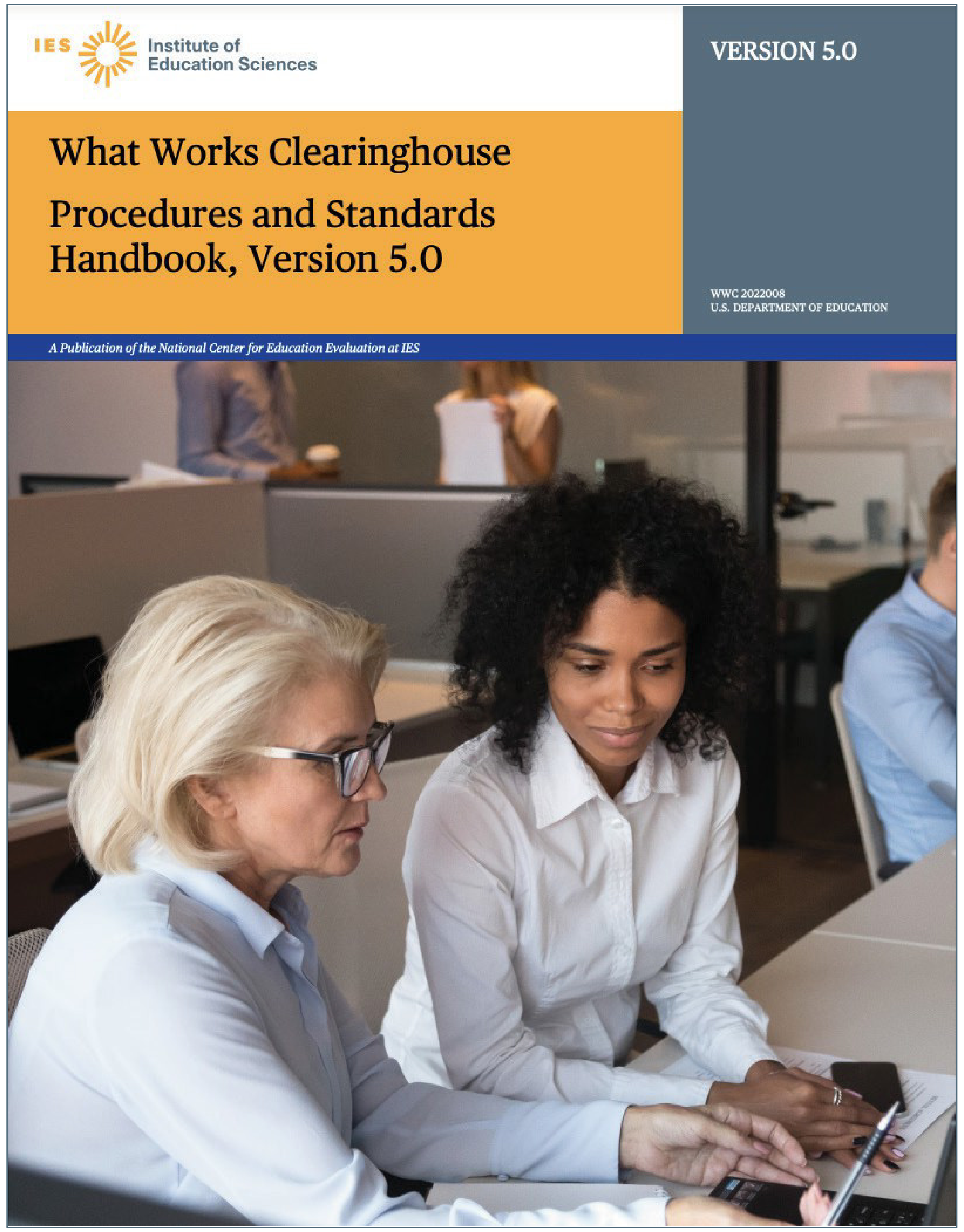
Who is this training for, and what are the learning objectives?

- Intended primarily for those who are already certified in WWC group design standards version 5.0
- This training aims to support you in:
  - Understanding areas of the WWC group design standards that changed between versions 4.1 and 5.0
  - Applying the baseline equivalence standard when authors use matching or weighting methods
  - Applying procedures and standards for repeated-measures analyses that estimate and report average impacts across multiple postintervention time points
  - Applying strategies to avoid or address common review issues

Major Changes Between Versions 4.1 and 5.0 of WWC Group Design Procedures and Standards

Under version 5.0, study-specific choices replace topic-specific choices

- For most standards, the WWC no longer allows topic-specific customization or application of the standards.
- However, some choices are now allowed to vary at the study (not topic) level, including:
  - Use of an optimistic or a cautious attrition boundary
  - Characterization of risk of bias due to joiners



IES Institute of Education Sciences

VERSION 5.0

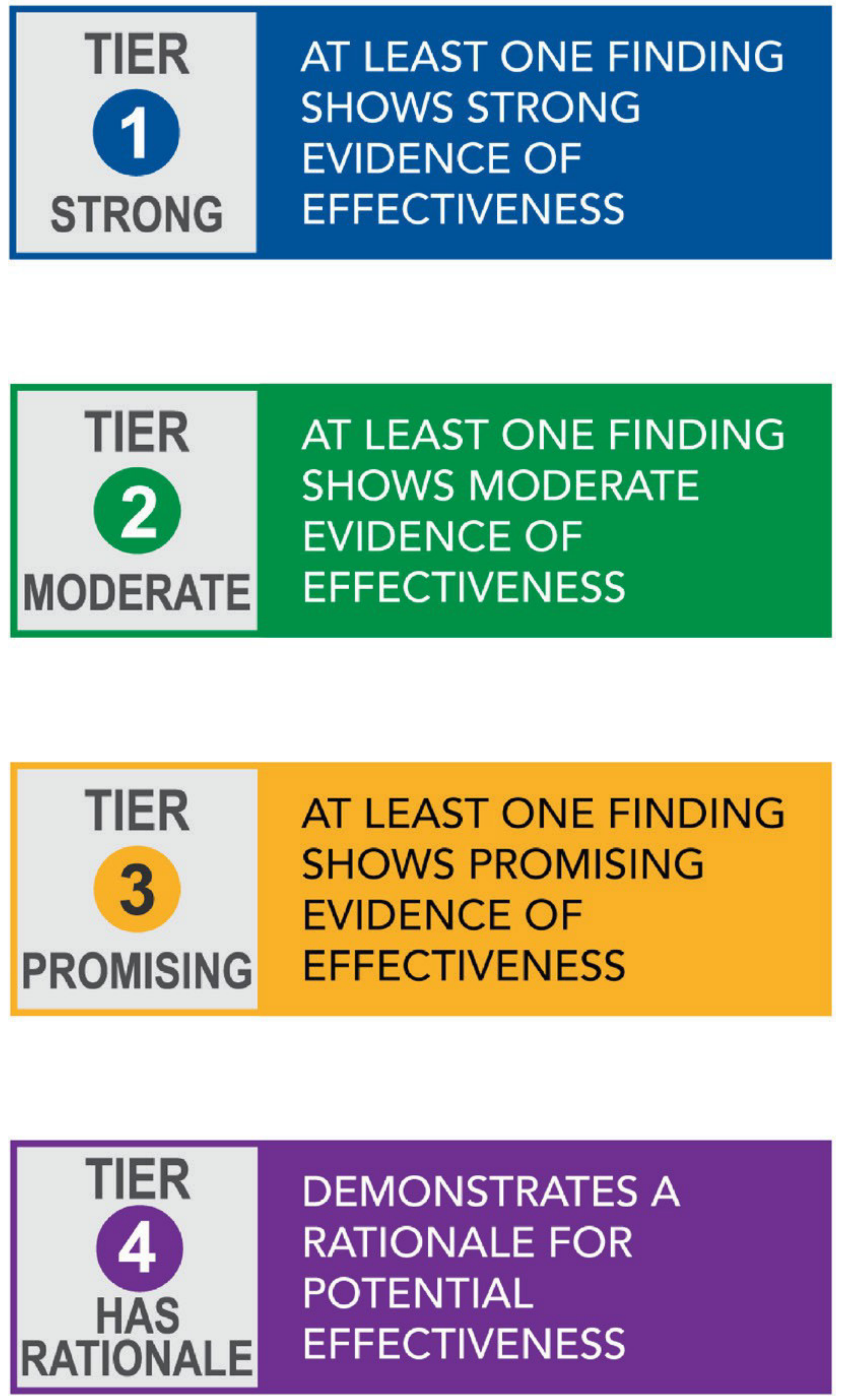
What Works Clearinghouse

Procedures and Standards Handbook, Version 5.0

WWC 2022008
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation at IES

Under version 5.0, effectiveness ratings now align with evidence definitions



TIER 1 STRONG — AT LEAST ONE FINDING SHOWS STRONG EVIDENCE OF EFFECTIVENESS

TIER 2 MODERATE — AT LEAST ONE FINDING SHOWS MODERATE EVIDENCE OF EFFECTIVENESS

TIER 3 PROMISING — AT LEAST ONE FINDING SHOWS PROMISING EVIDENCE OF EFFECTIVENESS

TIER 4 HAS RATIONALE — DEMONSTRATES A RATIONALE FOR POTENTIAL EFFECTIVENESS

- The WWC aligned its effectiveness ratings with the U.S. Department of Education’s evidence definitions for individual studies and synthesis products.
- This alignment means that the WWC characterizes evidence in a study only at the outcome domain level; evidence tiers are no longer assigned to individual findings.
- This change also eliminated the need for multiple comparison corrections.

Under version 5.0, consider an outcome measure’s independence for some domains

- The WWC considers a measure nonindependent if:
  - The measure was developed by the study authors and is not in broader use, or
  - The measure was developed by the intervention’s developers
- The independence of outcome measures is assessed for literacy- and math-related domains and domains related to broad student achievement, as identified in the Study Review Protocol.
- Studies that use nonindependent measures can still meet WWC standards.
- However, nonindependent measures will not contribute to effectiveness ratings (evidence tiers).

Under version 5.0, some studies with high attrition assessed using the optimistic boundary can satisfy the baseline equivalence standard via adjustment only

- To be rated *Meets WWC Standards With Reservations*, the following types of studies do *not* need to demonstrate that baseline differences between the analytic intervention and comparison groups are less than or equal to 0.25 standard deviations:
  - High-attrition randomized controlled trials (RCTs) without compromised random assignment when attrition bias is assessed using the optimistic boundary
  - Regression discontinuity design studies when attrition bias is assessed using the optimistic boundary
- Analyses in these studies do need to use an acceptable adjustment strategy.

For cluster RCTs under version 5.0, risk of bias from joiners has been simplified, and boundaries for assessing risk of bias due to leavers can differ across levels

- The WWC now classifies RCTs that assign clusters as having either a low or high risk of bias due to compositional change from joiners.
  - Reviewers will characterize the risk of bias due to joiners based on:
    - Unit of assignment
    - Unit of measurement
    - Potential for the intervention to affect joining
- The WWC now allows the attrition boundary for determining the risk of bias due to leavers to differ for the cluster and individual levels when assessing compositional change.

Matching and Weighting as Acceptable Methods for Baseline Adjustment

Assessing baseline equivalence using a baseline effect size

- For group design studies that must establish equivalence with a baseline effect size, the WWC calculates the difference between the analytic intervention and comparison groups at baseline.

- The magnitude of the effect size of that baseline difference will determine whether a study’s impact analysis requires a statistical adjustment for a baseline measure.

0.00 ≤ |Baseline effect size| ≤ 0.05	**0.05 < |Baseline effect size| ≤ 0.25**	**|Baseline effect size| > 0.25**
Satisfies the baseline equivalence standard	Requires statistical adjustment to satisfy the baseline equivalence standard	Does not satisfy the baseline equivalence standard

Acceptable baseline adjustment strategies for any baseline measure

- For studies that must adjust for baseline differences to satisfy the baseline equivalence standard, the WWC considers the adjustment strategies at right to be acceptable for any baseline measure.

Acceptable methods for any baseline measure

- Regression covariate adjustments in ordinary least squares models
- Regression covariate adjustments in hierarchical linear models
- Analysis of covariance
- Other approaches to regression covariate adjustments, including generalized linear models
- **Matching and weighting methods**
- Bounding techniques

Assessing baseline equivalence when matching or weighting methods are used for statistical adjustment

Study authors …

- Must demonstrate equivalence of the analytic intervention and comparisons groups *after matching or weighting* using the individual baseline characteristics described in the handbook.
  - If the impact analysis uses weights, the baseline means must also be calculated using the same weights.
- May use propensity score matching techniques to form groups.
- Cannot demonstrate equivalence using the propensity score.
- May use matching or weighting to create equivalent groups and simultaneously satisfy the statistical adjustment requirement.

Example of using matching to satisfy the statistical adjustment requirement

Researchers use nearest neighbor matching without replacement to form the analytic groups. The difference between the analytic intervention and comparison groups at baseline on the pretest measure of the outcome prior to matching is 0.30 standard deviations (SDs). The baseline effect size for the pretest in the matched analytic sample is 0.10 SDs. Having achieved what they believed to be reasonable balance between the analytic groups, the researchers chose not to include the pretest as a covariate in their impact model.

What is the highest research rating the study is eligible to receive?

The study is eligible to receive a research rating of *Meets WWC Standards With Reservations*.

- The baseline effect size *after matching* is in the range that requires statistical adjustment.
- The matching is considered an acceptable statistical adjustment. It is *not* necessary for the impact analysis using the matched data to also include the pretest as a covariate.

Example of using weighting to satisfy the statistical adjustment requirement

Study authors use inverse propensity score weighting to balance the analytic groups on important covariates. The difference between the analytic intervention and comparison groups at baseline on the pretest measure of the outcome prior to weighting is 0.18 SDs. The baseline effect size for the pretest in the weighted analytic sample is 0.06 SDs. The authors do not include the pretest as a covariate in their weighted impact model.

What is the highest research rating the study is eligible to receive?

The study is eligible to receive a research rating of *Meets WWC Standards With Reservations*.

- The baseline effect size *after weighting* is in the range that requires statistical adjustment.
- The weighting is considered an acceptable statistical adjustment. It is *not* necessary for the impact analysis using the weighted data to also include the pretest as a covariate.

Knowledge Check 1

A researcher uses inverse propensity score weighting to estimate the impact of a reading intervention on reading comprehension. The researcher demonstrates that the unweighted difference between the analytic intervention and comparison groups at baseline on the pretest measure of the outcome is 0.07 SDs. The impact analysis uses the inverse propensity score weights and includes the pretest measure of the outcome as a covariate.

Based on the available information, does the study satisfy the baseline equivalence standard?

□ A. Yes

□ B. No

Knowledge Check 1

A researcher uses inverse propensity score weighting to estimate the impact of a reading intervention on reading comprehension. The researcher demonstrates that the unweighted difference between the analytic intervention and comparison groups at baseline on the pretest measure of the outcome is 0.07 SDs. The impact analysis uses the inverse propensity score weights and includes the pretest measure of the outcome as a covariate.

Based on the available information, does the study satisfy the baseline equivalence standard?

□ A. Yes

□ B. No

Answer to Knowledge Check 1

□ **A is an incorrect answer.** The study does not satisfy the baseline equivalence standard because the researcher did not demonstrate equivalence of the analytic intervention and comparison groups after weighting.

☑ **B is the correct answer.** The WWC applies the baseline equivalence standard to the weighted data. Because the impact analysis uses weights, the baseline means must also be calculated using the same weights. The researcher has not demonstrated that the baseline effect size for the weighted analytic sample is less than 0.25 SDs.

Repeated-Measures Analyses that Estimate and Report Average Impacts Across Multiple Time Points

Assessing compositional change in RCTs with repeated-measures analyses that estimate an average impact across time points

- For RCTs, the WWC assesses compositional change separately at each postintervention time point included in the analysis.

- The amount of sample loss from the time of assignment can be different at each time point.

- The average of impacts across time periods can meet WWC standards without reservations if the risk of bias due to compositional change is low at each postintervention time point.

Assessing baseline equivalence in repeated-measures analyses that estimate an average impact across time points

For compromised RCTs, RCTs with high risk of bias due to compositional change, and quasi-experimental designs (QEDs)…

- The WWC applies the baseline equivalence standard to each postintervention time point included in the analysis.
- If the study authors observe more than one preintervention period, the WWC assesses baseline equivalence using the preintervention period closest to the start of the intervention.
- The average of impacts across time periods can meet WWC standards with reservations if the baseline equivalence standard is satisfied at each postintervention time point.

Some analyses with multiple postintervention time periods are ineligible for review

- A repeated-measures analysis with multiple postintervention time periods that must satisfy baseline equivalence is ineligible for review if the study manuscript:
  - Does not report point-in-time impact estimates, and
  - Does not report the baseline data needed to assess baseline equivalence for each postintervention time point
- The WWC will not send an author query to obtain the necessary baseline data for such a study.

Reporting findings for repeated-measures analyses that estimate an average impact across time points and meet standards

- The WWC can use *p*-values from impact estimates averaged over multiple postintervention time points to characterize study findings.
- Effect sizes may be calculated from impact estimates only if they are comparable to Hedges' *g* effect sizes.
  - Reviewers should briefly justify the decision to calculate or not calculate effect sizes in the notes of the Study Review Guide.
  - If unsure, submit a WWC Help Desk request.

Knowledge Check 2

The authors of a QED study describe their analysis as a comparative interrupted time series with five preintervention time points and three postintervention time points. They report a single impact estimate based on these data that measures the average impact across the three postintervention time points. They also report baseline effect sizes using data from the period closest to the start of the intervention for the analytic samples of students at each of the three postintervention time points.

Based on the available information, is this study eligible for review?

□ A. Yes

□ B. No

Answer to Knowledge Check 2

☑ **A is the correct answer.** The study is eligible for review because the manuscript reports the baseline data needed to assess baseline equivalence for each postintervention time point.

□ **B is an incorrect answer.** Although the study manuscript does not report point-in-time impact estimates, the average impact estimate is eligible for review because it reports the baseline data needed to assess baseline equivalence for each postintervention time point.

Common Review Issues and How to Address Them

Tip 1: Clearly document the (non)independence of an outcome measure when the determination is not obvious

- When assessing an outcome measure's independence, reviewers should consult the list of known independent measures appended to the Study Review Protocol.

- The WWC's list of known independent measures is **not exhaustive**; a measure might meet the definition of independence and not be on the list.

- If a study reports a measure that appears to meet the definition of an independent measure…
  - Check if the measure is currently included in the list of independent measures.
  - If not, submit a WWC Help Desk request for the WWC to consider whether the measure should be added to the list in the Study Review Protocol.

- Briefly justify the independence or nonindependence in the notes of the Study Review Guide if the determination is not obvious.

Tip 2: Clearly document the choice of outcome domain

- Consider carefully the outcome domain descriptions given in the Study Review Protocol.
- Consult content experts.
- Reach out to the WWC Help Desk for additional guidance if needed.
- Justify the outcome domain in the review notes if the appropriate domain is not obvious.

Tip 3: Always enter an intraclass correlation coefficient (ICC) when reviewing a cluster design study

- It is necessary to enter an ICC even when the authors’ analysis adjusted for clustering because the ICC can affect the WWC standard error calculation in some circumstances.

- Record the author-reported ICC when available.

- When study authors do not report an ICC, use the WWC defaults:
  - .20 for achievement outcomes
  - .10 for all other outcomes

Tip 4: Submit all study reviews that meet standards to peer review

- All study reviews that meet standards are peer reviewed, regardless of whether they are included in a product, such as a practice guide.

- The WWC Quality Control team aims to return materials submitted to peer review within 2 weeks.

Wrap-Up

Resources

- *What Works Clearinghouse Procedures and Standards Handbook*, Version 5.0
  - Summary of Changes in the *WWC Procedures and Standards Handbook, Version 5.0*
  - Questions and Answers about the *WWC Procedures and Standards Handbook, Version 5.0*
- Study Review Protocol, Version 5.0
- WWC Group Design Version 5.0 Online Training
- WWC Advanced Group Design Version 5.0 Online Training
- WWC Single-Case Design Version 5.0 Online Training

Thank you for your time, attention, and hard work!

- You may submit questions via the WWC Public Help Desk at https://ies.ed.gov/ncee/wwc/help or by emailing Contact.WWC@ed.gov.

- Reviewers currently working on WWC contracts should email the WWC Contractor Help Desk.