MODULE 3

Baseline Equivalence

whatworks.ed.gov

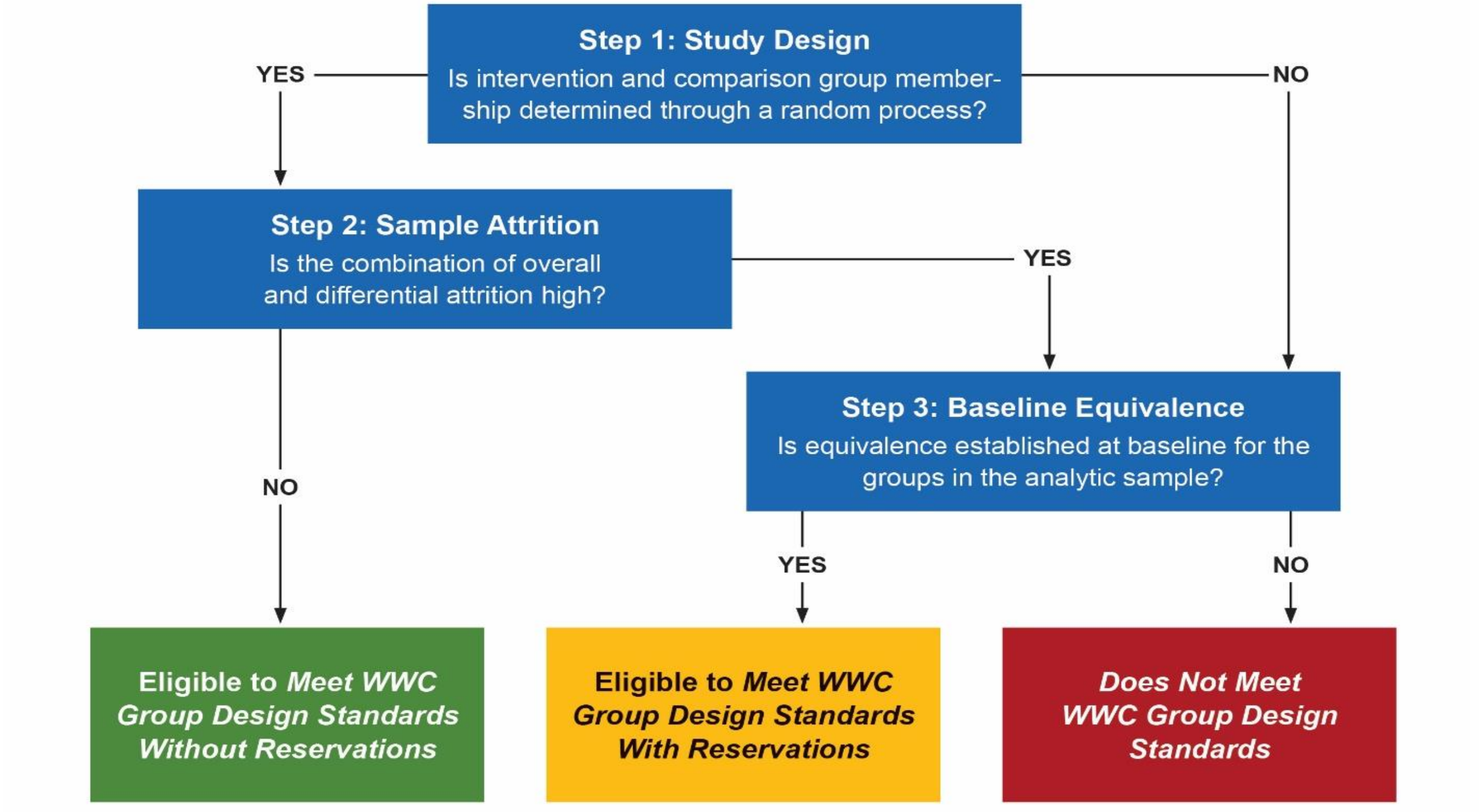
MODULE 3 Baseline Equivalence

This module covers the WWC baseline equivalence standard, which the WWC applies to studies that use randomized controlled trials (RCTs) with high attrition or compromised random assignment and those that use quasi-experimental designs (QEDs).

After completing this module, you will be able to:

- Describe baseline **equivalence**
- Identify when the WWC would assess baseline equivalence
- Demonstrate how to apply the WWC's equivalence standard
- Describe how the WWC determines the characteristics that it uses to assess baseline equivalence

Group Design Standards Framework



Step 1: Study Design
Is intervention and comparison group membership determined through a random process?

YES

NO

Step 2: Sample Attrition
Is the combination of overall and differential attrition high?

YES

NO

Step 3: Baseline Equivalence
Is equivalence established at baseline for the groups in the analytic sample?

YES

NO

Eligible to *Meet WWC Group Design Standards Without Reservations*

Eligible to *Meet WWC Group Design Standards With Reservations*

Does Not Meet WWC Group Design Standards

What Is Baseline Equivalence and Why Does It Matter?

The WWC uses the term "baseline equivalence" when determining whether the intervention and comparison groups had key observed characteristics that were similar enough ("equivalent") before the start of the intervention (at "baseline").

Differences between the two groups at the start of the intervention could bias the estimated impact of the intervention.

When Does the WWC Assess Baseline Equivalence?

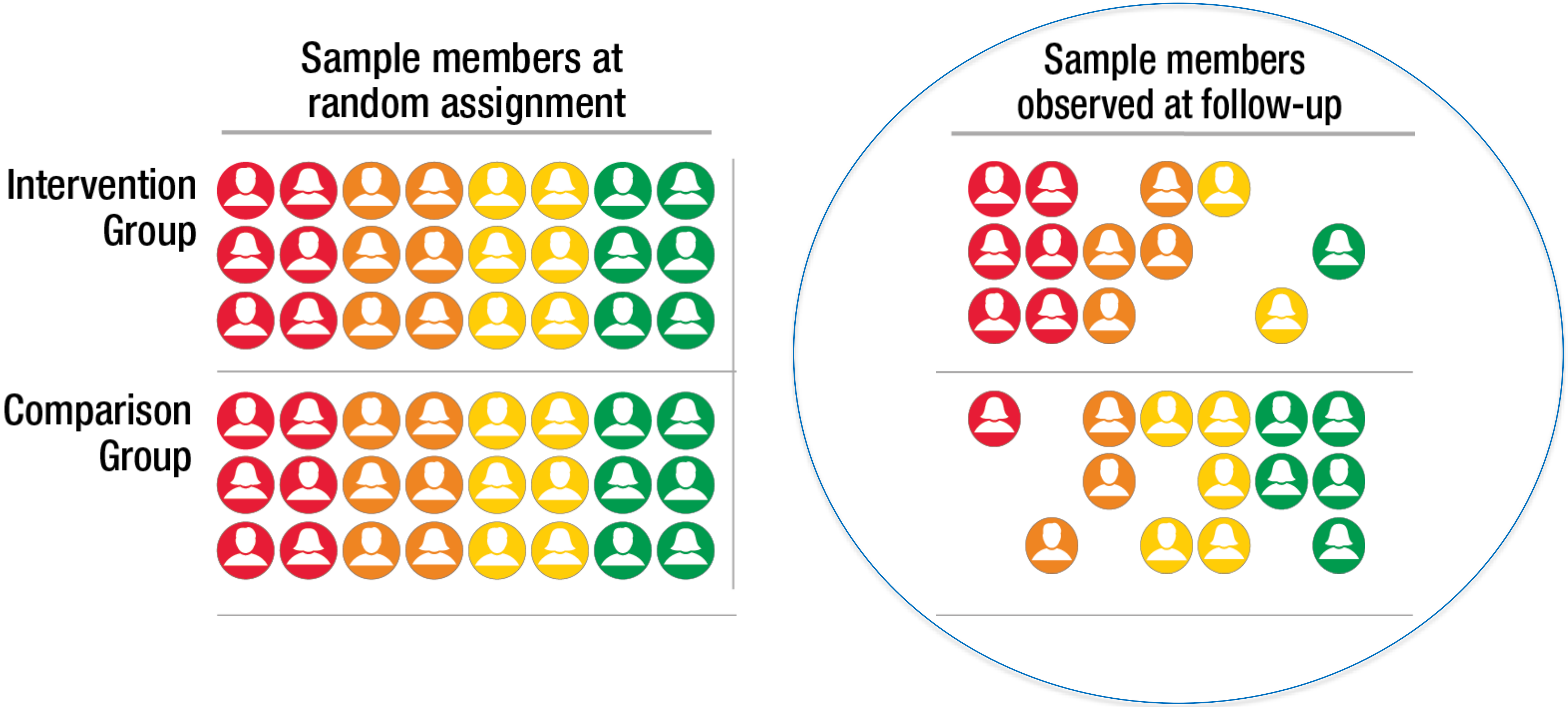
The WWC assesses equivalence when there is reason to be concerned that the units in the analytic intervention and comparison groups may be dissimilar.

- In RCT studies that have high attrition or compromised random assignment, the analytic intervention and comparison groups may not be equivalent.
- In QED studies, the groups may differ due to nonrandom formation.

Because of these concerns, the WWC requires that QEDs and RCTs with high attrition or compromised random assignment demonstrate equivalence of the analytic intervention and comparison groups to *Meet WWC Group Design Standards With Reservations*.

The WWC Assesses Equivalence on the Analytic Sample Using Baseline Data

The WWC assesses equivalence on the analytic sample, not on the initially assigned sample.


Sample members at random assignment
Sample members observed at follow-up
Intervention Group
Comparison Group

WWC Standard for Equivalence

The WWC assesses baseline equivalence on each outcome measure to determine whether baseline differences are:

- Small, i.e., the groups are equivalent,
- Moderate, i.e., the analysis requires a statistical adjustment to meet WWC Group Design Standards, or
- Large, i.e. the differences at baseline are too large to meet WWC Group Design Standards, even with statistical adjustment.

WWC Standard for Equivalence

This assessment compares baseline differences measured in standardized effect size (ES) units against the WWC standard for baseline equivalence.

0.00 ≤ |Baseline ES| ≤ 0.05	0.05 < |Baseline ES| ≤ 0.25	|Baseline ES| > 0.25
Satisfies baseline equivalence	Requires **statistical adjustment** to satisfy baseline equivalence	Does not satisfy baseline equivalence

Calculations to Determine Equivalence

- **The WWC calculates effect sizes in different ways based on the type of variables the study uses**
  - For continuous variables (many possible values), the WWC calculates **Hedges' *g***.
  - For dichotomous variables (only two possible values), the WWC calculates **Cox's Index**.

- **The WWC has tools to assist with the calculation of Hedges' *g* and Cox's Index**

Calculating Equivalence: Continuous Variables

- **The WWC calculates Hedges' *g*, a common effect size index.**

- **It is the difference between the average characteristic for the intervention group and the average characteristic for the comparison group, divided by the pooled standard deviation (SD) of the characteristic.**

Hedges' *g*

$$g = \frac{\omega\left(y_i - y_c\right)}{\sqrt{\dfrac{\left(n_i - 1\right)s_i^2 + \left(n_c - 1\right)s_c^2}{n_i + n_c - 2}}}$$

- y_i is the adjusted (or unadjusted) mean for the intervention group
- y_c is the adjusted (or unadjusted) mean for the comparison group
- n_i is the sample size of the intervention group
- n_c is the sample size of the comparison group
- s_i is the unadjusted standard deviation for the intervention group
- s_c is the unadjusted standard deviation for the comparison group
- Omega (ω) is the small sample size correction
  - It is equal to $1 - \frac{3}{4N - 9}$ where N is the total sample size (or $n_i + n_c$)

Example: Calculating Hedges' *g*

Information:

- Intervention group sample size (n_i) = 100; mean (y_i) = 25; SD (s_i) = 10
- Comparison group sample size (n_c) = 50; mean (y_c) = 20; SD (s_c) = 10

$$g = \frac{\left(1 - \frac{3}{4(150-9)}\right)(25-20)}{\sqrt{\frac{(100-1)\,10^2 + (50-1)\,10^2}{100 + 50 - 2}}}$$

Answer:
g = 0.50

|Baseline ES| > 0.25
Does not demonstrate equivalence

Calculating Equivalence: Dichotomous Variables

- **The WWC calculates Cox's Index for dichotomous variables.**
- **The formula for Cox's Index is more complex than Hedges' *g*, but is designed to produce a comparable effect size.**

Cox's Index

- p_i is the prevalence rate of an outcome for a student observed in the intervention group
- p_c is the prevalence rate of an outcome for a student observed in the comparison group
- Omega (ω) is the small sample size correction
  - It is equal to $1-\frac{3}{4N-9}$ where N is the total sample size (or $n_i + n_c$)
- See the *WWC Procedures Handbook for more details*

$$d_{Cox} = \omega\left[ln\left(\frac{p_i}{1-p_i}\right) - ln\left(\frac{p_c}{1-p_c}\right)\right]/1.65$$

Example: Calculating Cox's Index

Information:

- Intervention group sample size (n_i) = 100; p_i = .45
- Comparison group sample size (n_c) = 50; p_c = .47

$$d_{Cox} = \left(1 - \frac{3}{4(150) - 9}\right)\left[ln\left(\frac{.45}{1 - .45}\right) - ln\left(\frac{.47}{1 - .47}\right)\right]/1.65$$

Answer:
d = -0.049

0.00 ≤ |Baseline ES| ≤ 0.05
Demonstrates equivalence

Protocols Identify Characteristics for Equivalence

- Each WWC protocol lists characteristics for which studies must demonstrate equivalence.
- These are characteristics that have a strong association with **outcomes**.
- For example, the *WWC Beginning Reading review protocol* states: Baseline equivalence must be demonstrated for the intervention and comparison groups in the analytic sample on one of the following pre-intervention (or baseline) characteristics:
  - A pre-intervention measure of the outcome used in the analysis; or
  - If a pre-intervention measure of the outcome used in the analysis is not available, a pre-intervention measure of an outcome from any of the four **outcome domains** detailed in the protocol.

Demonstrating Equivalence for Outcomes With No Pretest

- If a study examines an outcome for which a pretest is not possible, the protocol specifies a set of related characteristics to use in demonstrating equivalence.

- In the *Dropout Prevention review protocol,* there is no baseline measure of dropping out. Instead, groups must demonstrate equivalence on:
  - Race/ethnicity,
  - Gender,
  - One measure of degree of disadvantage, and
  - One measure of school performance.

Other Relevant Characteristics

- Protocols sometimes require or consider equivalence on demographic characteristics or other measures that could affect outcomes.

- In the *Primary Math review protocol*, studies must demonstrate equivalence for groups on a math pretest, but the review team will also look at measures the authors provide, including:

  Grade
  Gender
  Race/ethnicity
  At-risk status
  Tracking level
  Special education
  School location
  Average class size
  English-learner students

- If groups differ on these measures, the review team leadership has discretion to indicate that the groups lack equivalence, even if pretest measures are equivalent.

Default: Assess Equivalence for Each Domain

- Most protocols assess equivalence within outcome domains.
- If there is a baseline difference between the groups greater than 0.25 SD for any required characteristic, then all outcome measures in the domain that use the same analytic sample receive the *Does Not Meet WWC Group Design Standards* rating.

Domain	Measure (Baseline ES)	Rating for domain
Operations	Addition Test 1 (0.05 SD) Addition Test 2 (0.24 SD) Subtraction Test 1 (0.20 SD)	*Meets WWC Group Design Standards With Reservations*, provided the analysis includes acceptable statistical adjustments.
Patterns and Classification	Patterns Test 1 (0.05 SD) **Classification Test 1 (0.26 SD)** Patterns Test 2 (0.20 SD)	*Does Not Meet WWC Group Design Standards*. Analyses in other domains could still meet standards.

- If the analysis of Classification Test 1 used a different analytic sample, other analyses in the domain could *Meet WWC Group Design Standards With Reservations*.

Equivalence in One Domain That Affects Other Domains

The protocol will state whether equivalence in one domain affects other domains using the same analytic sample or the full study.

- In the *Beginning Reading review protocol*, if any outcome measure in any domain has a difference of 0.25 SD or greater, then no outcome measures in any domain can meet standards.

Domain	Measure	Rating for Domain
Reading fluency	Fluency 1 (0.05 SD) Fluency 2 (0.24 SD) Fluency 3 (0.20 SD)	Would be acceptable if the study had only this domain. However, because Comprehension 2 has a difference of 0.30 SD, the whole study receives the *Does Not Meet WWC Group Design Standards* rating.
Comprehension	Comprehension 1 (0.05 SD) Comprehension 2 (0.30 SD) Comprehension 3 (0.20 SD)	*Does Not Meet WWC Group Design Standards*.

Equivalence on One Outcome Measure That Does Not Affect Other Outcome Measures in the Domain

The protocol will state whether equivalence on one outcome measure does not affect other outcome measures in the domain.

- In the *Secondary Math review protocol*, baseline measures must be pretests of outcomes.
- Non-equivalence on one measure has no bearing on the equivalence of other outcomes in the same domain.

Domain	Measure	Rating
Algebra	Writing Equations (0.51 SD) Graphing (0.04 SD) Polynomials (0.16 SD)	Writing Equations *Does Not Meet WWC Group Design Standards*. Graphing can *Meet WWC Group Design Standards With Reservations*. Polynomials can *Meet WWC Group Design Standards With Reservations* if the study makes an acceptable statistical adjustment.

Issue: Equivalence and Propensity Score Matching

- Study authors may use propensity score matching techniques to form groups.
- They cannot demonstrate equivalence on the propensity score.
- They must demonstrate equivalence using the characteristics identified in the protocol.

Issue: Equivalence and Imputed Baseline Data

- Studies must base imputed baseline data on an approach the WWC considers acceptable.

- The WWC Standards Handbook describes standards for assessing baseline differences in studies with missing or imputed data.

- Imputed baseline data not used to demonstrate equivalence do not affect the study's rating, as long as the study used an acceptable approach to impute the data.

Summary of Requirements for Statistical Adjustments

If the effect size for the analytic sample is	Then the impact analysis	And the highest possible rating is
Less than or equal to |0.05| SD	Does not require statistical covariate adjustment	*Meets WWC Group Design Standards With Reservations*
Greater than |0.05| SD and less than or equal to |0.25| SD	Requires statistical covariate adjustment for at least that outcome measure	*Meets WWC Group Design Standards With Reservations,* provided the analysis appropriately controls for differences; otherwise, the highest possible rating is *Does Not Meet WWC Group Design Standards*
Greater than |0.25| SD	Cannot control for the differences	*Does Not Meet WWC Group Design Standards*

Methods to Adjust for Pretest Differences

Acceptable methods of statistical adjustment include:

- Regression covariate adjustments (including covariates in a hierarchical linear model [HLM])
- Analysis of covariance (ANCOVA)

Statistical adjustments acceptable only in certain circumstances include:

- Gain scores
- Difference-in-differences adjustments
- Fixed effects for individuals

- Fixed effects for individuals are acceptable statistical adjustments only when (1) pretest and posttest are measured in same units and (2) correlation between pretest and posttest is at least 0.6.
  - The WWC considers a pretest to be measured in the same units as the posttest only when both the same test and the same scoring procedures were used, so that the two measures can be directly compared.
  - The WWC will perform its own difference-in-differences adjustment (see the Reporting Module) to satisfy the statistical adjustment requirement if the author performed no adjustment, but these conditions hold.
- Methods that do not provide acceptable statistical adjustments can meet standards if the WWC does not require an adjustment.
  - Examples include low-attrition RCTs and studies with baseline differences less than |0.05| standard deviations.

Knowledge Check 1

For what types of designs does the WWC require studies to demonstrate baseline equivalence?

□ A. Low-attrition RCTs

□ B. High-attrition RCTs

□ C. QEDs

□ D. Options B and C, or

□ E. Options A, B, and C

Answer to Knowledge Check 1

□ A and E are incorrect answers. The WWC does not require RCTs with low attrition to demonstrate equivalence.

□ B is a partially correct answer. RCTs with high attrition must demonstrate equivalence on the analytic sample.

□ C is a partially correct answer. All QED studies must demonstrate equivalence on the analytic sample.

■ D is the correct answer. RCTs with high attrition and QED studies must demonstrate equivalence on the analytic sample.

Knowledge Check 2

Does the WWC assess equivalence on baseline characteristics for the original randomly assigned sample or for the analytic sample?

□ A. The original randomly assigned sample

□ B. The analytic sample

□ C. Both the randomly assigned sample and the analytic sample

Answer to Knowledge Check 2

□ A and C are incorrect answers. The WWC does not require studies to demonstrate equivalence for the original randomly assigned sample.

■ B is the correct answer. The WWC requires studies to demonstrate equivalence for the analytic sample, to show that the two groups in the analysis were similar before the intervention occurred.

Knowledge Check 3

Is it possible to need to assess equivalence for some, but not all, outcome measures in a study?

□ A. Yes
□ B. No

Answer to Knowledge Check 3

■ A is the correct answer. In an RCT, the samples for some outcome measures can have high attrition, while others have low attrition. The study must demonstrate equivalence for the high attrition analyses only.

□ B is an incorrect answer. It is possible to need to assess equivalence for some, but not all, outcome measures in a study because the study must demonstrate equivalence for the high-attrition analyses only.

Knowledge Check 4

A study identifies the 20 lowest-scoring 1st-grade students in a school as eligible for a supplemental reading intervention. The study randomly assigns 10 students to receive the program and 10 not to receive it. The researchers give pre- and posttests to all students. At pretest, the students in the intervention group had a mean score of 23 (SD = 5), and students in the comparison group had a mean score of 35 (SD = 5).

Does the reviewer need to assess equivalence?

- □ A. Yes
- □ B. No

Answer to Knowledge Check 4

□ A is an incorrect answer. This is an RCT with low (no) attrition: “The researchers give pre- and posttests to all students.” Thus, it does not need to demonstrate equivalence.

■ B is the correct answer. This is an RCT with low (no) attrition, so it does not need to demonstrate equivalence.

Knowledge Check 5

In a QED study with 50 students participating at follow-up (25 in the intervention group, 25 in the comparison group), these 50 students took a pretest that resulted in the following scores: intervention mean = 46, intervention SD = 20; comparison mean = 50, comparison SD = 20.

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

- □ A. *Meets WWC Group Design Standards Without Reservations*
- □ B. *Meets WWC Group Design Standards With Reservations*, if the analysis includes an acceptable statistical adjustment
- □ C. *Meets WWC Group Design Standards With Reservations*, regardless of whether the analysis includes an acceptable statistical adjustment
- □ D. *Does Not Meet WWC Group Design Standards*

Answer to Knowledge Check 5

- □ A is an incorrect answer. This study uses a QED; therefore, it cannot receive the *Meets WWC Group Design Standards Without Reservations* rating.
- ■ B is the correct answer. The study uses a QED, so it must demonstrate equivalence. The baseline difference is 0.20 SD (ignoring the small sample size correction), which falls in the “adjustment zone”; the study must therefore include an acceptable statistical adjustment to account for this difference to *Meet WWC Group Design Standards With Reservations*.
- □ C is an incorrect answer. The baseline difference is in the “adjustment zone,” so the study must include an acceptable statistical adjustment.
- □ D is an incorrect answer. Because baseline differences fall within the adjustment zone, it is possible for the study to *Meet WWC Group Design Standards With Reservations* if the analysis appropriately controls for baseline differences.

Knowledge Check 6

A study examining a science intervention collected outcome data at three points in time (immediate posttest, one-year follow-up, and two-year follow-up). Researchers used propensity score matching to identify a comparison group at the beginning of the study. The propensity score included demographic characteristics, as well as the pretest (science achievement). From the final report, which focuses on the two-year follow-up, we can calculate the mean difference on the propensity score for the initial sample (200 youth, 100 in each group), which is 0.03 SD reported as Hedges' *g*. We can also calculate the mean difference on the propensity score of the two-year analytic sample (150 youth, 80 intervention and 70 comparison), which is 0.10 SD when reported as Hedges' *g*.

Has the study demonstrated equivalence?

☐ A. Yes, the study has demonstrated equivalence
☐ B. No, the study has not demonstrated equivalence

Answer to Knowledge Check 6

□ A is an incorrect answer. A study cannot use a propensity score to demonstrate equivalence. It must demonstrate equivalence on the individual measures required by the review protocol.

■ B is the correct answer. A study cannot use a propensity score to demonstrate equivalence. It must demonstrate equivalence on the individual measures required by the review protocol.

MODULE 3

Baseline Equivalence

Conclusion

MODULE 3 Baseline Equivalence

- You can access all the resources mentioned in this module through the WWC website, **whatworks.ed.gov**.
- The full slide deck for this module, including detailed responses to the knowledge check questions, is available on the WWC website.
- To receive a certificate of completion for viewing these training modules, you must view the videos on the WWC website.

Thank you!