What Works Clearinghouse™

IES INSTITUTE OF EDUCATION SCIENCES

WWC STANDARDS Brief

The What Works Clearinghouse (WWC) is an initiative of the U.S. Department of Education's Institute of Education Sciences. The WWC evaluates research studies that look at the effectiveness of education programs, policies, and practices, which the WWC calls "interventions." WWC Standards Briefs explain the rules the WWC uses to assess the quality of studies. For more information, visit the WWC's webpage at http://whatworks.ed.gov.

Baseline Equivalence

What is baseline equivalence?

The WWC uses the term "baseline equivalence" when determining if the intervention group (those that received the intervention of interest) and the comparison group (those that did not receive the intervention) had characteristics that were similar enough ("equivalent") at the start of the study (at "**baseline**").
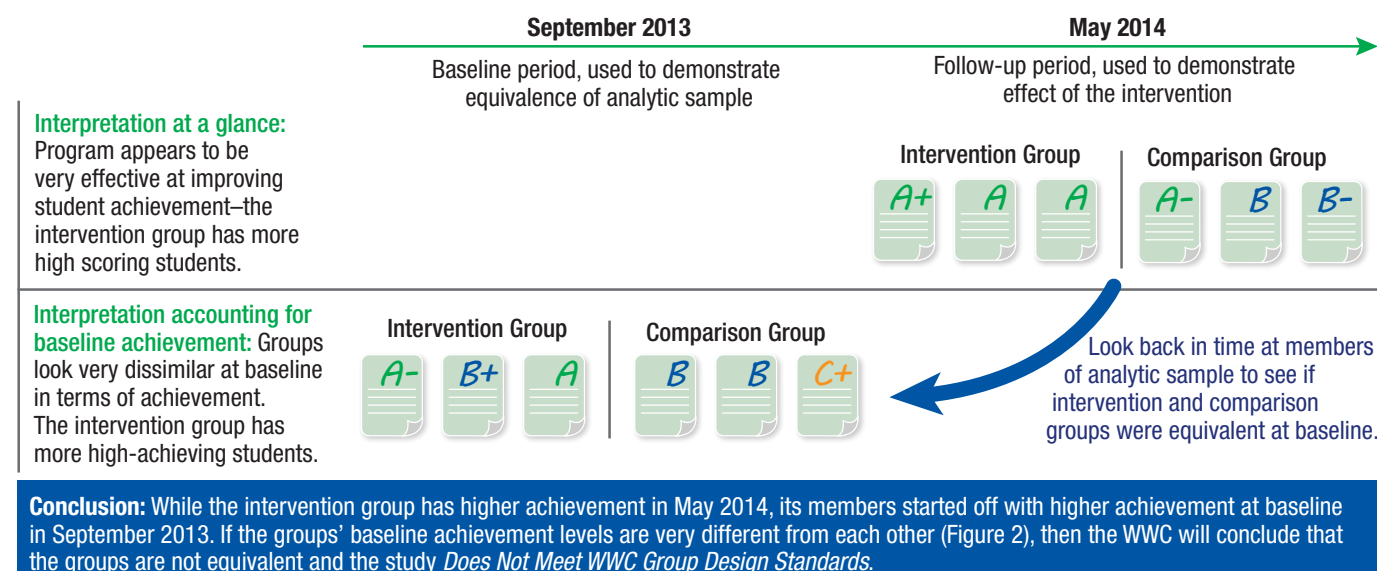
Why does baseline equivalence matter?

When two groups are similar at the start of a study (baseline) and, after that, the only difference between the groups is that one receives the intervention and the other does not, it is reasonable to conclude that any differences in the outcomes that are measured at the end of a study (**follow-up**) are caused by the intervention. However, if *the two groups are different* at baseline on key characteristics that could influence the outcomes, the effect found at the end of the study *might* be due to the differences that already existed at the beginning.

Demonstrating baseline equivalence is important in studies that did not assign participants randomly to the intervention and comparison groups. It is also important in random assignment studies with high attrition.[1]

In the example in Figure 1, students in the intervention group had higher academic achievement than those in the comparison group during the follow-up period, suggesting a positive program impact. However, we need to look back in time and examine the baseline achievement for the sample of individuals analyzed at follow-up (the **analytic sample**).

Figure 1: Why demonstrating baseline equivalence is important

September 2013 — Baseline period, used to demonstrate equivalence of analytic sample

May 2014 — Follow-up period, used to demonstrate effect of the intervention

Interpretation at a glance: Program appears to be very effective at improving student achievement–the intervention group has more high scoring students.

Intervention Group: A+ A A
Comparison Group: A- B B-

Interpretation accounting for baseline achievement: Groups look very dissimilar at baseline in terms of achievement. The intervention group has more high-achieving students.

Intervention Group: A- B+ A
Comparison Group: B B C+

Look back in time at members of analytic sample to see if intervention and comparison groups were equivalent at baseline.

Conclusion: While the intervention group has higher achievement in May 2014, its members started off with higher achievement at baseline in September 2013. If the groups' baseline achievement levels are very different from each other (Figure 2), then the WWC will conclude that the groups are not equivalent and the study *Does Not Meet WWC Group Design Standards*.

---

[1] Attrition refers to loss of sample, when individuals initially randomly assigned in a study are not included in the analysis. See the WWC Standards Brief for Attrition for more information on this topic.

The groups started out at very different achievement levels: the intervention group had more high-achieving individuals than the comparison group had at baseline. Therefore, the observed program impact may be biased: some of the differences in outcomes may result from having different types of students across intervention and comparison groups.

How does the WWC determine baseline equivalence?

Each WWC **review protocol** specifies the characteristics on which equivalence must be established at baseline. For academic outcomes, baseline equivalence is often established using a pre-intervention test. Some outcomes, such as high school graduation, do not have a pre-intervention measure. For these outcomes, the WWC often requires baseline equivalence for related demographic characteristics (e.g., age).

The WWC uses a standardized mean difference called an **effect size** to determine whether there is baseline equivalence between the intervention and comparison groups. The effect size is calculated as the difference between the intervention group mean and the comparison group mean, divided by the pooled standard deviation.[2] Figure 2 illustrates this calculation. In this example, the intervention group had higher average achievement at baseline than the comparison group. The WWC then determines whether the groups are similar by comparing the effect size for each baseline characteristic against the WWC standard for baseline equivalence in Figure 3.

The WWC considers intervention and comparison groups equivalent at baseline when the corresponding effect sizes are 0.05 or less in absolute value. When an effect size for a baseline characteristic is between 0.05 and 0.25 in absolute value, the WWC requires a **statistical adjustment**.[3] If the effect size for any baseline characteristic is greater than 0.25 in absolute value, the WWC concludes that the intervention and comparison groups are not equivalent at baseline.

Figure 2: Baseline equivalence calculation

Group	Baseline Mean	Baseline Standard Deviation
Intervention	90	10
Comparison	82	10

Difference in means ▶ 90 − 82 = 8
Pooled standard deviation ▶ 10

Baseline Equivalence Calculation

$$\text{Effect size} = \frac{\text{Difference in means}}{\text{Pooled standard deviation}} = \frac{8}{10} = 0.80$$

Decision

Does not satisfy baseline equivalence

Figure 3: WWC standard for baseline equivalence

Absolute value of effect size ≤ 0.05	0.05 < absolute value of effect size ≤ 0.25	Absolute value of effect size > 0.25
Satisfies baseline equivalence	Statistical adjustment required to satisfy baseline equivalence	Does not satisfy baseline equivalence

Baseline equivalence can only be established on observable characteristics—that is, only on characteristics that can be measured. While establishing baseline equivalence reduces the threat of observable baseline differences influencing the analysis, there may still be differences in unobservable characteristics that could influence the findings. That is why randomized controlled trials with high levels of attrition or other issues and quasi-experimental designs that meet the baseline equivalence standard can, at best, receive a rating of *Meets WWC Group Design Standards With Reservations*.

If baseline equivalence is not established, the study *Does Not Meet WWC Group Design Standards*.

Glossary

- The **absolute value** of a number is the size of the number without regard to its sign. For example, the absolute value of −0.10 is 0.10.
- The **analytic sample** is the sample on which the analysis is based.
- The **baseline** is the point in time before the intervention was implemented.
- The **effect size** is a standardized measure of the magnitude of a difference.
- The **follow-up** is the point in time during which the effect of the intervention is evaluated.
- A **review protocol** is a formal document developed to determine the scope of a review.
- A **standard deviation** indicates the variability of a measure across the observations in a sample.
- **Statistical adjustment** refers to including baseline measures in a statistical model to reduce bias in the estimate of the effect of the intervention.

For more information about the baseline equivalence standard and other WWC standards, please download a copy of the *WWC Procedures and Standards Handbook*.

[2] Pooled standard deviation can be conceptualized as an "average" standard deviation, calculated across the intervention and comparison groups. For more information, please see Section F of the *WWC Procedures and Standards Handbook*.
[3] A number of different techniques can satisfy the statistical adjustment requirement, including regression adjustment and ANCOVA.