MODULE 5

Outcome Measures

whatworks.ed.gov
This module covers the WWC standards related to outcome measures. By the end of this module, you will be able to:

❖ Define an outcome measure
❖ Explain why and how the WWC evaluates outcome measures used to demonstrate impacts in evaluations
❖ Describe the WWC outcome measure standard
❖ Illustrate how to apply the WWC outcome measure standard
Definitions

- An outcome is the knowledge, skills, attitudes, behaviors, or other measurable characteristic that researchers measure to learn about the impact of an **intervention**. Studies that compare measured outcomes from a group that receives the intervention to outcomes for a group that does not receive the intervention are **eligible** for review.

- An **outcome measure** is an instrument, device, or method that provides data on the desired outcome.

- An **outcome domain** is a group of closely related outcome measures that provide information on the same underlying skill or ability.
Outcomes, Measures, and Domains: Example from Beginning Reading Review Protocol

**Outcome Domain**
Alphabets

**Outcomes**
- Phonemic awareness
- Phonological awareness
- Letter identification
- Print awareness
- Phonics

**Outcome Measures**
*Examples*
- Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Letter Naming subtest
- Test of Word Reading Efficiency (TOWRE)
- Phonological Awareness Test (PAT)
The WWC Outcome Measures Standard

❖ To meet the requirements for a review, a measure must:

• Demonstrate **face validity**
  – Does it measure what it claims to be measuring?

• Demonstrate sufficient **reliability**
  – Does the measure produce the same types of scores if it is administered over and over to different people, in different settings, etc.?

• Avoid **overalignment**
  – Is the outcome measure tailored to the intervention?

• Use **consistent data collection procedures**
  – Did the researchers collect data for the measure in the same way for the intervention and comparison groups?
Definition of Face Validity

An outcome measure must capture what it claims to measure to appropriately label the intervention’s effect on an outcome of interest.

❖ To demonstrate face validity, a measure must:

• Have a clear definition

• Assess what it purports to assess
Assessing Face Validity: Does the measure capture what it claims to measure?

❖ According to the WWC, the description of an outcome measure must align with its content.

• For example, if a measure is called a “math achievement test,” it must measure students’ math knowledge and skills.

❖ If an outcome measure does not have face validity, any impact estimated using that outcome will receive the *Does Not Meet WWC Group Design Standards* rating.
Definition of Reliability

The reliability of a measure captures whether it would yield similar scores from different administrations and whether the observed scores are free of measurement error.
Assessing Reliability: Does the measure produce the same types of scores over different administrations?

1. Do the outcome measure’s items provide consistent information about individual respondents?
   - The WWC requires an internal consistency statistic (for example, Cronbach’s alpha) of 0.50 or higher.

2. Do individuals get similar scores in repeated assessments?
   - The WWC requires a test-retest reliability statistic of 0.40 or better.

3. Do raters score individuals similarly?
   - The WWC requires an inter-rater reliability statistic (for example, percent agreement or Kappa) of 0.50 or higher.

If authors do not demonstrate that an outcome measure is reliable, any contrast using that outcome measure will receive the *Does Not Meet WWC Group Design Standards* rating.
Review Team-Specific Requirements for Reliability

- The **review protocol** will document any additional or different reliability requirements for a specific product or review team.

  - Review teams may require higher reliability depending on the nature of the measures common to the review.

**Example:** The *Teaching Elementary School Students to Be Effective Writers Practice Guide* focuses on a holistic writing score, which is usually determined by a rubric. The review team decided the measure must demonstrate inter-rater reliability of 0.80 or higher, and documented this in the *Elementary School Writing Practice Guide Review Protocol.*
Outcome Measures Without Reliability Statistics

- Outcomes that are administratively measured behaviors automatically pass the reliability requirement.
  - For example: graduation, enrollment in school, or grade retention.

- A review team’s content expert or lead methodologist can determine that an outcome measure without reliability statistics is reliable if a single rater could score the responses with low error.
  - For example: number of words spelled correctly in a passage, or counting to 10 correctly.

- The review protocol may stipulate how to deal with other outcomes that are unlikely to have reliability information.
Standardized Tests

❖ The *WWC Standards Handbook* defines a standardized test as an assessment given to all test takers using the same set of procedures.

❖ In most cases, standardized tests have established administration and scoring procedures.

❖ Generally, standardized tests report scores in relation to a normed sample.
Standardized tests do not need to demonstrate face validity and reliability. This is because:

- Developers design standardized tests to have a specified set of concepts. Therefore, by definition, the measure has face validity.
- Developers of standardized tests establish administration and scoring procedures that support reliable scale scores.
Identifying Standardized Tests

The WWC considers a measure standardized if any of the following are true:

❖ It has a trademarked name.
❖ A reputable testing organization developed it.
❖ It has a technical manual.
❖ It is well known in the field.
❖ It has established scale scores.
❖ It is a state or district achievement test.

The WWC does not typically consider researcher-developed measures standardized tests.
Potential Concerns About Standardized Tests

Did the study administer a subset of items from a standardized test?

- Scores must come from established scales or **subscales**, not a selection of items.
- For example, the Woodcock-Johnson III (WJ-III) includes the following subscales:
  - WJ-III Calculations
  - WJ-III Letter-Word Identification
  - WJ-III Passage Comprehension

Did the study deviate from administration or scoring procedures?

- Deviating from established procedures could affect validity or reliability.
- If there is evidence that a study did not follow established procedures, the WWC does not consider the measure standardized.
Examples of Issues with Standardized Tests

In the following examples, the WWC might not consider the measures standardized:

❖ To measure outcomes for low-achieving 1st graders, researchers selected a measure designed for kindergartners.

❖ The developer designed a measure for one-on-one administration, but the study administered it in small groups.

❖ The developer requires valid responses for 95% of items to generate a score, but the research team calculated scores when only 90% of items had valid responses.
Outcome Measures and Overalignment: Does the outcome measure provide an unfair advantage to one group over another?

Overalignment between a measure and a condition occurs when a study closely tailors a measure to a condition (or the measure repeats some aspect of a condition).

❖ Example: Teaching test content to students in the intervention group but not to those in the comparison group.

Analyses of an overaligned outcome measure receive the *Does Not Meet WWC Group Design Standards* rating.

Exception: sometimes material covered by an outcome measure must be explicitly taught.

❖ Example: Reciting the alphabet requires being taught the alphabet.
Similarity of Data Collection Across Groups: Did the study collect data the same way for the intervention and comparison groups?

- If the mode of data collection differs between groups, differences in the outcome could be due to the intervention or to the data collection method.

- Analyses of studies that collected outcomes measures differently by condition receive the *Does Not Meet WWC Group Design Standards* rating.
Other Differences in Measures to Consider

- The WWC considers several issues to determine whether data collection differed across groups:
  - Differences in the mode of data collection
  - Differences in the timing of data collection
  - Differences in personnel involved in test administration
  - Differences in the construction of scores

- Many studies do not discuss these aspects of data collection directly. If there is nothing to suggest a difference, then the WWC assumes the processes were the same across groups.
Baseline Measures

- Pre-intervention measures used to demonstrate equivalence must satisfy the same reliability criteria as outcome measures.

- If the pre-intervention measure is different and the study does not meet the reliability requirement, the study cannot use the measure to demonstrate equivalence.

- If a pre-intervention measure is a pretest of the outcome measure, the study automatically satisfies this reliability requirement if it did so for the outcome measure.
A study of a math intervention reports on an analysis using an outcome measure called the “Math Problem Comprehension Test,” which the study authors developed. The test contained 15 problems: 5 asked respondents to match an equation to a sentence from a word problem, 5 asked them to identify the numbers needed to solve a word problem, and 5 asked them to write the arithmetic equations needed to solve the word problem.

1.1. What type of measure is this?
- A. This is a not a standardized test.
- B. This is a standardized test.

1.2. Does it have face validity based on the information provided?
- A. Yes
- B. No

1.3. Is it reliable based on the information provided?
- A. Yes
- B. No
Answer to Knowledge Check 1.1

- **A is the correct answer.** This is a not a standardized test. The item notes that the study authors developed the test. It does not describe the measure as a state or district assessment, does not include a citation to a technical manual, and does not say that the test meets any other criteria for a standardized test.

- **B is an incorrect answer.** The measure does not meet any of the WWC’s criteria for a standardized test.
A is the correct answer. The test is called “Math Problem Comprehension Test,” and the items are concrete math tasks based on word problems. Based on this description, the measure meets the WWC requirement for demonstrating face validity.

B is an incorrect answer. Measures have face validity if they capture what they claim to capture. Based on the description, the measure meets the WWC requirement for demonstrating face validity.
A is an incorrect answer. The item does not mention reliability. The WWC requires a measure of internal consistency (like Cronbach’s alpha) of 0.50 or higher, a test-retest statistic of 0.40 or higher, or inter-rater reliability (like a Kappa coefficient) of 0.50 or higher.

B is the correct answer. The authors do not provide any evidence of reliability.
Knowledge Check 2

From the total population of 4-year-olds in day care centers, a study randomly selected 30 for a group that heard a different story at each of six sessions, and 30 for a group that heard each of three different books twice during the six sessions. The purpose of the repeated readings in the second group was to assess the development of children’s comments and questions in response to familiar stories. The study administered a reading comprehension outcome measure to both groups. The researchers based the questions in the outcome measure on a book used in the second (repeated-book) group, but not in the first group. The researchers scored the responses based on transcripts of the readings and coding of the children’s responses to the story.

Does the outcome measure in this study meet review requirements?

☐ A. Yes
☐ B. No
Answer to Knowledge Check 2

☐ A is an incorrect answer. The outcome does not meet review requirements because the study did not establish reliability and the outcome may be overaligned.

■ B is the correct answer. The study did not use a standardized outcome measure, so the WWC would require reliability information. The study also did not establish the reliability of the outcomes. Additionally, there are concerns about overalignment. Members of the repeated-book group had practice with the book used in the assessment; therefore, they had an advantage over members of the other group on the outcome. If the study used an overaligned outcome measure, the WWC would rate analyses of the outcome measure *Does Not Meet WWC Group Design Standards.*
MODULE 5

Outcome Measures

Conclusion
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!