Collecting High Quality Outcome Data, Part 2
Understanding Reliability, Validity, and Bias

Special Note

This skill building activity can be used to apply the concepts and principles covered in this module to real world situations.

Introduction

This exercise allows learners to review program scenarios to diagnose likely measurement problems or issues and to propose solutions to these problems.

The examples in the exercise are formulated to provide fairly clear cut illustrations of a single measurement problem—either reliability *or* validity *or* bias. However, this does not preclude learners from identifying additional measurement problems. The main purpose of the exercise is for learners to practice identifying measurement problems and to demonstrate an understanding of reliability, validity, and bias. In this sense, getting the "right answer" for any given scenario is secondary to correctly demonstrating the relevant knowledge/skills. As an example, the "right answer" for the waterway restoration project (scenario 4) is "validity." However, a learner would not necessarily be "wrong" to suggest that there is also a risk of bias due to poorly timed data collection, e.g., because a random event like a wildfire could undo the work of restoration crews. Indeed, this learner would be demonstrating a correct understanding and application of the concept of bias. The point of the exercise is to reinforce key concepts, not to memorize terms.

Key Points – Definitions

Reliability is the ability of a method or instrument to yield consistent results each time. Reliability is strengthened by using well-designed instruments and by providing data collectors and respondents with clear instructions on how to administer and complete instruments.

Validity is the ability of a method or instrument to accurately measure what it intends to or what it is supposed to measure. Measurement is valid when it produces results addressing the specific outcome you wish to measure. Valid measurement collects data on all relevant aspects or dimensions of an outcome. Validity is also supported when the results produced by an instrument are corroborated by information from other sources. For example, the validity of a math test is supported when students who score high (or low) on the test also perform well (or poorly) at solving math problems in class and on homework assignments.

Results are **biased** when they are systematically skewed or distorted. Results can be biased due to the over- or under-representation of particular groups in the dataset, and due to question wording that tends to encourage or discourage particular responses. The timing of data collection can also systematically bias results.

Sticking Points and Common Issues

Below are some issues that may come up as learners consider the material, along with notes on how to respond to these issues.

What do I do if I cannot get data collection in place before my program begins?
It is always best to make key decisions about methods and instruments before starting your program. However, we recognize this may not always happen perfectly. Development and improvement of methods and instruments is an ongoing process. Programs start from wherever they are, but should strive to develop and strengthen data collection systems as quickly as possible. For example, if a program needs to collect pre-and-post data, then, ideally, instruments need to be developed and tested before the program starts. Otherwise, the program will not be able to conduct a true pretest since the intervention will have already begun. In this situation, the program would still need to conduct the pretest as early as possible and note in the progress report that pretest data were collected late after the intervention started.

Under what circumstances may one modify an instrument?
In general, care should be taken when modifying instruments to avoid compromising the rigor, quality, and usefulness of data. Sample instruments provided in support of national performance measures may be modified as long as the instrument can still collect key data elements required by the performance measurement instructions. Instruments that come from other sources can be modified to fit your program context. However, modifying an instrument that has been validated will compromise the integrity of the instrument, so it is not advisable to revise these instruments or the instructions for their administration. When modifying an instrument always remain mindful of the instrument's original purpose and avoid modifications that deviate from this purpose or that will weaken the rigor, quality, or usefulness of the data.

On the one hand, I am advised to pilot test and revise instruments. On the other hand, I am advised not to revise standardized instruments. Does this mean that I should not pilot test standardized instruments, since I cannot revise them?
The purpose of pilot testing is to improve instruments. Piloting includes testing the instrument itself as well as the data collection process. If you plan to use a standardized instrument, pilot testing can help you understand how well (or poorly) the instrument works in your context. This is actionable information even if you are not revising the instrument or the procedures for administering it. Sometimes modest changes to how an instrument is administered can fix problems without compromising quality. If you encounter serious problems using a standardized instrument then you know in advance that it cannot be used and you will have to consider alternatives. Part of the value of pilot testing is simply learning about whether any problems exist – and gaining greater confidence in your instruments if you find that they are free from serious problems.

N

Exercise

Trainer Tip: It may be helpful to display the slide listing definitions for reliability, validity, and minimizing bias (slide 11, "Ensuring data quality: Reliability, validity, bias," in the module, "Collecting High Quality Outcome Data, Part 2") while learners complete this exercise.

Instructions: Read each scenario and use the checkbox list to identify the most likely measurement issue. Briefly explain your choice and propose a solution.

1. The **Elmwood Tutoring Program** provides one-on-one and small group literacy tutoring to students in grades 3-5 who read below grade level. Students at two local primary schools participate twice per week in 90-minute literacy tutoring sessions focusing on improving students' ability to read aloud (oral reading fluency). Outcome data are collected by school personnel who administer a standardized reading comprehension test to students at the beginning and end of the school year. Test results show trivial improvements in test scores from pretest to posttest. Assuming the problem is with measurement and not with the program design...

   What is the most likely measurement problem? ☐ Reliability ☐ Validity ☐ Bias

   Briefly explain your answer. _____

   _____

   Briefly, how might the program address this measurement problem? _____

   _____

   _____

2. The **Tupelo Housing Program** assists economically disadvantaged and homeless individuals to move into safe, healthy, and affordable housing. National service volunteers counsel individuals on their housing needs, help them apply for housing assistance, and follow up to provide continued assistance and to verify an individual's housing status up to 9 months after initial service. National service volunteers find that it is more difficult to track homeless men than other economically disadvantaged clients over the 9-month period to verify their housing status, so outcome data are missing for many homeless clients.

   What is the most likely measurement problem? ☐ Reliability ☐ Validity ☐ Bias

   Briefly explain your answer. _____

   _____

   Briefly, how might the program address this measurement problem? _____

   _____

N

3. The **Dawsonville Children's Health Initiative** partners with local schools to provide a twice-weekly afterschool physical fitness program for children ages 6-14. National service volunteers strive to increase children's exercise habits by teaching them about the benefits of regular physical exercise and by organizing a variety of age-appropriate sports activities for children to participate in. Changes in exercise habits are measured by a questionnaire that children complete at the end of the program. However, the program manager is concerned that the questionnaire is not producing high-quality data, particularly for questions that ask children about their exercise habits before participating in the program. Assuming the problem is with measurement and not with the program design...

   What is the most likely measurement problem? ☐ Reliability ☐ Validity ☐ Bias

   Briefly explain your answer. _____

   _____

   Briefly, how might the program address this measurement problem? _____

   _____

   _____

4. The **Easton Waterway Restoration Project** partners with the State Bureau of Conservation and Restoration to identify sections of the Victoria River for stream bank restoration. The goal of this work is to create stream bank conditions that can lead to eventual water quality improvements. Crews of national service volunteers implement remediation in accordance with the State Bureau's waterway management plan, including removal of trash and debris from stream banks, removal of invasive plants, reintroduction of native plants, and erosion abatement. Land managers from the State Bureau inspect project sites within two weeks of project completion. The assessment instrument used by land managers contains checkbox items to indicate whether various remediation actions were taken, but does not provide a way to assess the quality of these remediation actions with respect to environmental standards. Assuming the problem is with measurement and not with the program design...

   What is the most likely measurement problem? ☐ Reliability ☐ Validity ☐ Bias

   Briefly explain your answer. _____

   _____

   Briefly, how might the program address this measurement problem? _____

   _____

Answer Key and Points to Consider

1. Elmwood Tutoring Program

What is the most likely measurement problem? ☐ Reliability ☒ Validity ☐ Bias

Explanation: Measurement is focused on the wrong academic skill. The program tutors students to improve their oral reading fluency (i.e., their ability to read aloud at a fluid pace without mispronouncing words). However, the program is using a standardized test that measures students' reading comprehension (i.e., their ability to understand the meaning of what they read). **Validity** is the ability of a method or instrument to accurately measure what it intends to or what it is supposed to measure. By measuring the wrong thing the program is engaging in inaccurate measurement.

The measurement problem is **not reliability** because there is no indication that the measurement process is producing inconsistent results. If anything, the standardized test the program currently uses appears to be generating consistently poor results, and this makes sense since the children involved in this intervention are likely to perform poorly in other literacy skill areas, and the intervention is not designed to produce improvements in the outcome that the current instrument measures.

The measurement problem is **not bias** because there is no indication that a segment of the service population (students) is not getting measured. Nor is there any indication that the standardized test contains flawed questions that tend to produce biased answers (unbalanced scales, "leading" questions, double-barreled questions, etc.). There is also no indication that testing is poorly timed such that extraneous factors might influence the results.

How the program might address this measurement problem: The most obvious solution is for the program to use a standardized test that measures the right outcome, i.e., students' oral reading fluency.

2. Tupelo Housing Program

What is the most likely measurement problem? ☐ Reliability ☐ Validity ☒ Bias

Explanation: The absence of outcome data from a substantial subgroup of the service population (homeless men) means that results will underrepresent individuals in this subgroup. Since this subgroup is also less likely to achieve the intended outcome, results are likely to be **biased** towards overestimating the effectiveness of the program.

The measurement problem is **not reliability** because there is no indication that the measurement process produces inconsistent results. If anything, results may tend to appear more consistent as

the systematic (but unintentional) exclusion of a key subgroup from the dataset reduces variation in the results.

The measurement problem is **not validity** because there is no indication that the measurement process fails to accurately measure the intended outcome (participants move into safe, healthy, affordable housing) for those who the program is able to measure.

How the program might address this measurement problem: Any solutions to the measurement problem will involve finding ways to get outcome data from more homeless men so they are adequately represented in the dataset. Any reasonable sounding solution addressing the logistical problem of maintaining contact with homeless men during the 9-month follow-up period would be relevant to solving the measurement problem.

3. **Dawsonville Children's Health Initiative**

What is the most likely measurement problem? ☒ Reliability ☐ Validity ☐ Bias

Explanation: The program relies on children's ability to recall their exercise habits in some detail for a lengthy period of time. However, memory can be very **unreliable**, and will depend greatly on each child's ability to remember details about their exercise habits. Some children will be able to provide fairly detailed and accurate information, while others will only be able to provide sketchy and incomplete information.

While the possibility for inaccuracy also suggests a problem with **validity**, the main problem is with the lack of consistency in the responses. Information obtained from any two children will not be comparable because of difference in each child's capacity to remember exercise habits.

The measurement issue is **not a problem of bias** because there is no indication that some children are more or less likely to get included in the dataset. Nor does it appear that the measurement process tends to systematically over- or under-estimate exercise habits. While the limitations of children's memory could lead some children to underestimate exercise habits, it is also possible that some children will generalize specific memories (e.g., "I remember playing basketball 3 times last week, so that must be typical for me.") in ways that overestimates physical activity.

How the program might address this measurement problem: The most obvious solution is to switch from relatively infrequent or one-time measurement covering a long time period to more frequent measurement covering shorter time periods. This increases reliance on more recent, and therefore more reliable, memories. The program can ask children to describe their exercise habits when they enter the program and again upon leaving the program. The questionnaire can also be designed to facilitate the process of remembering, e.g., by providing prompts about specific types of physical activity. Yet another solution would be to switch to diaries or journals as a way to more regularly track children's exercise habits.

4. Easton Waterway Restoration Project

What is the most likely measurement problem? ☐ Reliability ☒ Validity ☐ Bias

Explanation: The program is seeking to measure whether completed stream bank restoration work is capable of supporting eventual improvements in water quality. The assumption is that water quality improvements can only happen if completed stream bank restoration work meets environmental standards. However, the assessment tool provided to land managers only allows them to indicate whether the work was done, but not if the work was done in accordance with environmental standards. This means that the tool **does not provide valid, accurate measurement** of the outcome, i.e., whether the completed work meets environmental standards. To measure the outcome accurately, the assessment tool also needs to allow land managers to rate the quality of the work.

The measurement problem is **not one of reliability** because there is no indication that the measurement process produces inconsistent results, or that land managers apply different criteria for assessing whether, for example, trash and debris was removed or whether native species were planted. Reliability could be compromised if the Easton Waterway Restoration Project worked with more than one environmental agency and if these agencies applied different standards to assess the rating items. However, the program is only working with one agency, the State Bureau of Conservation and Restoration, so this is not an issue.

The measurement problem is **not one of bias** because there is no indication that some project sites are excluded from measurement or that the instrument poses questions in a way that makes some responses more or less likely than others. The short timeframe for data collection (within two weeks of project completion) also minimizes the chance that a sudden change in environmental conditions (e.g., caused by a big storm or a wildfire) will bias results by undoing the efforts of restoration crews.

How the program might address this measurement problem: The most obvious solution is to revise the instrument to include a rating scale connected to the environmental standards that would have also informed the work done by project crews. This would allow land managers to provide a truer, more accurate assessment of the quality of completed projects, and strengthen the connection between completed projects and the long-term goal of promoting improved water quality.