SAGE reference

The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Construct Irrelevance

Construct irrelevance, as the name might suggest, refers to measuring phenomena that are not included in the definition of the construct. This is generally considered to be one of the two biggest threats to the validity of an assessment, along with construct underrepresentation. Although construct underrepresentation involves an assessment not adequately measuring all aspects of the construct of interest, construct irrelevance (sometimes referred to as construct irrelevant variance) occurs when an assessment is measuring more than just the construct of interest. This entry examines how construct irrelevance occurs and its implications and then provides illustrative examples of construct irrelevance.

Construct Irrelevance in Practice

The underlying assumption of all assessments is that the score that is produced reflects a test taker's ability on the construct of interest. In an educational setting, this construct is often an educational content area. For example, if a student takes a math assessment, we assume that the score that student receives represents, and is an accurate measure of, his math ability.

As noted previously, construct irrelevance occurs when the score produced by the assessment is dependent on more than just the construct of interest. Continuing with the example of the math assessment, suppose that some of the questions were word problems. These word problems will require some level of reading and comprehension ability in order to understand the question and respond appropriately. This is construct irrelevance. We are not interested in measuring a student's reading ability with this assessment. That is, reading ability is irrelevant to math ability, as we have defined the construct.

The implication of including these items on the assessment is that a student's score may not accurately reflect that student's ability on the construct of interest. Suppose a student with poor reading ability but high math ability took the assessment previously described. This student might be fully capable of performing the math necessary to correctly answer the word problems, but because of her poor reading ability, she doesn't understand the question and gets the item wrong.

As this type of construct irrelevant error variance accumulates over many word problems, the student's math ability score is going to be biased down. The student will get many questions wrong that someone with a high level of math ability should get correct, due simply to her reading ability. Thus, the score she receives is not an accurate representation of her true math ability. This can in turn influence which performance levels students are placed in, which then has implications for many of the decisions in educational settings that are based on student test scores.

This example clearly demonstrates the importance of construct irrelevance to the overall validity argument of an assessment. If an assessment contains variance that is irrelevant to the construct, the scores will be biased, and thus it becomes extremely difficult to justify using those scores to make decisions. Therefore, it is important to ensure that sources of construct irrelevance are minimized as much as possible to safeguard the validity of the scores and their intended uses. Common sources of construct irrelevance and methods used to detect some of these sources are discussed in the following section.

Examples of Construct Irrelevance

In operational psychometrics, construct irrelevance is most commonly associated with differential item functioning (DIF) and differential test functioning. This is because at their core, both DIF and differential test functioning represent construct irrelevance. Thus, possible sources of DIF are also possible sources of construct irrelevance. If performance on an item can be predicted by group membership after accounting for ability level on the construct on interest, then construct irrelevance is present. Most commonly, the groups of interest in these analyses are gender, ethnic, and socioeconomic groups.

When a reading assessment is administered, there is no intention for that assessment to also measure a student's gender or race. Thus, if one group (e.g., females) is favored over another (e.g., males) after accounting for ability level, construct irrelevant variance is introduced into the scores. Similarly, if an item or the assessment overall favors a particular racial or socioeconomic group after accounting for ability, then the assessment is not only measuring the construct of interest, but also the racial or socioeconomic group membership.

Construct irrelevance is not limited to group membership, however. For example, individuals' performance on an assessment may be influenced by their motivation or their preconceived notions about how they will perform (e.g., stereotype threat). Thus, the assessment would be measuring not only the construct of interest but also the individual's motivation.

This can also be true of tests that are overly long. If individuals get tired at the end of a long assessment, then the assessment now measures exhaustion in addition to the construct of interest. When providing scores, only the construct of interest should be included. Therefore, it is of the utmost importance construct irrelevance is considered throughout the assessment development from item writing (to lower the chance of construct irrelevance through DIF) to the blueprint of assessment (to avoid test designs that may invite additional sources of error).

W. Jake Thompson
http://dx.doi.org/10.4135/9781506326139.n143
10.4135/9781506326139.n143

Further Readings

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education, 7, 235–241.

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1), 17–27.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.