

Measurement Issues in Applied Research

Notions of Reliability

Reliability is one way to conceive of the relationship between scores or data and the attribute we attempted to measure; the precision with which the score measures the attribute or the reliability of the data as an indicator of the attribute. We can also think of reliability as the degree of consistency between two measures of the same thing.

Reliability is not a property of a test or other measurement instrument

Reliability is a property of scores from which we make decisions and can be very different based on the object of measurement and universe of generalization. Interrater reliability may help provide information to improve the rating process, but it does not support the consistency of scores for making decisions about individuals.

There is no such thing as THE reliability of a set of scores. There are a number of estimates of reliability depending on the sources of error you are interested in capturing – and there is no right answer to the question “what sources of error should be captured?”

Common sources of error that can be captured in reliability estimates include error due to content sampling, trait instability, or subjective ratings. These estimates are often referred to as the coefficient of equivalence, coefficient of stability (test-retest), and interrater reliability. Without two or more observations or measures of the same thing, we can also estimate the internal consistency of a single instrument—as though each item was a measure of some attribute and contributes an internally consistent measure of the attribute.

In educational testing, the factors that influence reliability are known: test length, speed, group homogeneity, difficulty of items, objectivity.

Notions of Validity

Validity is another way to conceive of the relationship between scores or data and the attribute we expect to measure; the extent to which the score measures the attribute it was designed to measure or the validity of the inferences we derive from composite indicators of the attribute.

Validity is not a property of a test or other measurement instrument, or even scores.

Validity is a property of the interpretations and inferences we derive from scores. The three forms of quantitative score validity are conceptualized as construct, criterion, and content validity. The three categories are somewhat arbitrary and alone or together, they are incomplete.

There is no such thing as THE validity. For most applied research and program evaluation measurement instruments, it is hard to support interpretations or make inferences about results without evidence to support the content of the instruments from which the data are derived – are they comprised of the appropriate, relevant, and a representative sample of the domain of interest. If we cannot defend the content, little other evidence will make up for the deficiency.

The overriding principle is that the validity of our interpretations must be supported in some way. The evidence gathered must be directed at the specific interpretations made from the data.

Criterion-related validity is often relayed in the form of a correlation coefficient – the relationship between the data or scores from which we derive our interpretations and some other measure of the same thing (the criterion).

In educational testing, the factors that affect validity are known, including reliabilities of the predictor and criterion measures, sampling error, group homogeneity, the shape of the relationship (the actual relationship between the two measures, and restriction of range due to prescreening groups).

Notions of Reliability and Validity in Qualitative Approaches

Some have argued that reliability and validity as conceived in test theory are not applicable or are inappropriate for use in qualitative research.

In testing situations, subjectively scored assessment tasks (e.g., performance assessments and portfolios) are critiqued strongly because of their limited reliability and validity. Reliability is lowered because of the subjective nature of scoring and the few number of performance tasks that can be administered to any given individual. Validity is lowered primarily because of the limited coverage of the domain and also because of the greater level of measurement error in the total score variance (lower reliability).

Some educational researchers have suggested hermeneutic approaches to the evaluation of performance tasks or subjective ratings of performance and behavior—honoring the purpose that students bring to their work and the contextualized judgments of teachers. A common example is that of the dissertation defense . . . Also of growing concern is the role of consequences in validity theory. Some have suggested that there is a consequential basis for validity that must be determined regarding the appropriate uses of assessment results.

Understanding may be a more fundamental concept for qualitative research than validity. Maxwell (1992) and others have developed a typology of the kinds of understanding at which qualitative researchers aim for. This is a realist approach to validity as a form of understanding.

Again, there is no one truth to which we can compare some individual's account. We can have no direct knowledge of the objects we study and there is no independent entity to which we can compare our accounts of the objects we study. However, there exists ways of assessing our accounts that do not depend entirely on features of the accounts themselves, but relate to those things that the account claims to be about—gathering evidence about the relationship between the account and the object.

Descriptive Validity

Primary descriptive validity refers to the factual accuracy of what was seen or heard (physical objects, events, and behaviors). Secondary descriptive validity refers to the factual accuracy of what in principle could be observed but what was inferred from other data.

Descriptive validity is directly concerned with description and their meaning is generally not in dispute, only the accuracy. With data, disagreement can be resolved. The relevant consensus for the terms used in interpretation rests to a substantial extent in the research community.

Interpretive Validity

This pertains to what the objects, events, and behaviors mean to the people engaged in and with them. Included in this is intention, cognition, affect, belief, evaluation, and anything that is included in the participant's perspective. This addresses the evidence to support the comprehension of phenomena not from the researcher's perspective but from that of the participants of the study.

The relevant consensus for the terms used in interpretation rests to a substantial extent in the community studied. Unlike the case of descriptive validity (where access to data addresses the threat to validity), accounts of participants' meanings are always constructed by researchers on the basis of participants' accounts and other evidence.

Theoretical Validity

This pertains to the theoretical constructions that the researcher brings to or develops during the study – not the concrete description and interpretation of phenomena. Its purpose goes beyond describing these participants' perspectives, referring to an account's function as an explanation as well as a description or interpretation of the phenomena. This addresses the idea of construct validity and approaches causal validity. It includes the concepts that the theory employs and the relationships that are believed to exist among the concepts. It includes the researcher's proposed relationships, or how the objects as measured fit together in a model developed during the study.

Any challenge to the meaning of terms or the appropriateness of their application to a given phenomena moves us from descriptive or interpretive validity issues to theoretical validity.

Generalizability

We often hope to extend the account of a particular situation or population to other persons, times, or settings than those we directly studied. Often in qualitative research, there is no explicit sampling process developed to enable conclusions about specified populations through statistical inference. However, even in the most qualitative of research, some level of generalization is anticipated.

Internal Generalizability: The case where we wish to generalize within the community or institution studied to individuals, events, and settings that were not directly observed or interviewed.

External Generalizability: The case where we wish to generalize to other communities or institutions.

In the case of interviews, internal Generalizability may be most significantly at stake. The interview is a social situation and involves a relationship between the interviewer and the subject. Understanding the nature of that situation and relationship and how it may affect what occurs during the interview and even how the subject's actions and views could differ in other situations is critical to the validity of accounts from interviews.

Evaluative Validity

This refers to the application of an evaluative framework to the objects of the study. The evaluative statements made based on our investigations do not depend on the methods used to obtain the descriptions or data of the phenomena or the methods used to interpret or make theoretical sense of it. They do, however, depend on the particular description, interpretation, or theory the researcher constructs.

In most quantitative and experimental research, threats to validity can be eliminated in the design process, prior to the presentation of results. In much qualitative research, prior elimination of threats is not often possible – in part because of the focus on inductive reasoning. Qualitative researchers have to deal with threats to the validity of their accounts on phenomena – seeking evidence to rule out certain threats, generally after the accounts have been described.

Scriven refers to the process of addressing threats to validity after tentative accounts have been developed as the “modus operandi” approach.