Invalid measurement validity

Editorial

Methodological concerns are a major reason behind having to turn down papers offered to *Developmental Medicine and Child Neurology*. One important cause is inadequate validation of measures used in the work. The validity of a test is one of its most important attributes, whether it is a clinical assessment, laboratory assay, electrophysiological or neuroimaging investigation, or a measurement of function, such as motor or IQ scales. With such a plethora of tests used in both everyday clinical work and research, and with the difficulties involved, it is not surprising that some are incompletely validated.

Measurement validity concerns whether a test measures what it is supposed to measure. The several processes that are required for validation have been defined but do not always seem so widely recognized. They appear complex, time consuming, and carry confusing jargon which is both off-putting and obscures what really matters. The different aspects that need to be addressed have been defined as 'the 3 Cs': content, construct, and criterion. Some of these domains overlap, which could question the validity of separating them. Clinical tests, such as the plantar response in childhood, are often the least well validated. Many tests seem to have been part validated in one or more of these domains, even if not in all. Thus, with the impossibility of making perfect measurements, validity remains a matter of degree.

In the papers we see, tests usually have some degree of content validity which covers whether the items measured are properly representative of whatever is being assessed. The simplest form of this, termed face validity, is the opinion of a group of experts that the test concerned is likely to measure what is wanted, and looks rather subjective. Showing that a scale has full content validity is a much more difficult task that overlaps with construct validity, a closely related concept showing the logic behind the development of the test and its function. Asking if an IQ assessment really measures intelligence, or just some aspects of mental function that we happen to value in present day society, is to examine its construct validity. In this sense, IQ tests could be criticized for functions that they do not measure. However, the common practice of using only some of the subscores in assessing children may be invalid in terms of content validity. Similarly, if the application of a test is changed, the validity may be lost. Scales developed to assess spasticity may not be valid if used for a different purpose, such as measuring the effect of intrathecal baclofen. A measure, or scale, measuring normal development of an ability in a child may not be applicable to degenerative conditions that cause loss of that ability where a different process is being assessed. If modified versions are developed, they in turn must be validated.

The subdivisions of criterion-related validity, concurrence or predictive, validity are the domains least frequently assessed. Concurrence looks at how a measure relates to a gold standard. If, as frequently occurs, there isn't one, the next best is the correlation with other established measures. This assumes that the latter are themselves valid, which may not be true, and risks stifling innovation, a source of concern if a paper is criticized on this ground. More straightforwardly, concurrence also means that a test should be consistent with previous approaches to the function measured and should be internally consistent. This might be relevant to current attempts to redefine cerebral palsy. Predictive validity concerns a different aspect: correlating the scale with a future assessment, such as the value of the admission Glasgow Coma Score in predicting the outcome from traumatic brain injury. As it is not particularly well correlated, other measurements are needed if intervention studies are being planned.

Accuracy and reliability are associated concepts. A test with high concurrent validity on correlation measures may still be inaccurate. If different categories are being defined, are they unambiguous and clearly separated from each other? If the results are part of a continuum, do they reflect that? The difference between centiles and standard deviations affects widely used instruments, such as growth charts. The mean is the 50th centile, 1 standard deviation below it approximates to the 33rd centile, 2 to the 2nd centile, etc. It is not immediately obvious that in statistical terms the difference between the 50th and the 33rd centile is the same as between the 33rd and the 2nd. Reliability, as in intra- and interobserver variation or test–retest reproducibility, is essential for validity, but a reliable test may not be valid if it does not measure what is intended. The final aspect, which relates to the study as a whole rather than measurement itself, concerns internal and external validity. Internal validity means that all confounding factors have been controlled for. External validity indicates how well a finding can be generalized to other groups of people (population validity) or situations (ecological validity).

Essentially, measurement validity is an assessment of how good the test is at doing what it is intended to do. Many of the tests we use in clinical life and in research are not ideal in these terms, particularly some of those hallowed by time. If a researcher is developing a new test, such as a new quality of life instrument, all these issues need to be considered. It is upsetting to have to send back a piece of work that has clearly taken many hours in terms of preparation, data collection, analysis, and writing up, all because of a problem in the fundamental validity of the tests used.

Peter Baxter

DOI: 10.1017/S0012162205000551