FUNDAMENTAL ISSUES OF EDUCATIONAL MEASUREMENT AND ASSESSMENT

DR. NAVEED YOUSUF

PhD Medical Education, University Ambrosaina Italy;
Dip- HPE AKU, MBA , Department for Educational Development Faculty of Health Sciences
Aga Khan University.

It is appropriate to begin by defining the term 'assessment'. It is important to note that the dictionary definition of the term 'assessment' is similar to 'measurement'. In educational science, assessment is a systematic process to measure, evaluate, and document students' academic achievement against expected outcomes or competencies. Hence, the assessment process in education is akin to measuring – that is, assessors are attempting to quantify, in grades or scores, the knowledge, skills or attitudes that students demonstrate against the intended learning objectives defined in the curriculum. Compare this to the measurement of tangible constructs such as weight, height, length, volume, etc.

On the other hand, student assessment primarily deals with intangible constructs such as intelligence, aptitude, cognitive ability, and attitude. This raises the question if the measured construct is the one which was intended and if it is being measured as precisely as possible. The former part of the question pertains to validity, and the latter to reliability issues. The other challenge for institutions involved in educational assessments is to investigate for evidence of reliability and validity of decisions and record the evidence for the defensibility of outcomes and accreditation of programmes.

The present article will first discuss the types of reliability to ensure precise measurement and validity evidence later to facilitate accurate assessment decisions.

Reliability and Errors of Measurement

Reliability is a necessary but not a sufficient condition to ensure validity. Reliability is the degree to which the scores are consistent when an assessment is repeated on a population or a group.

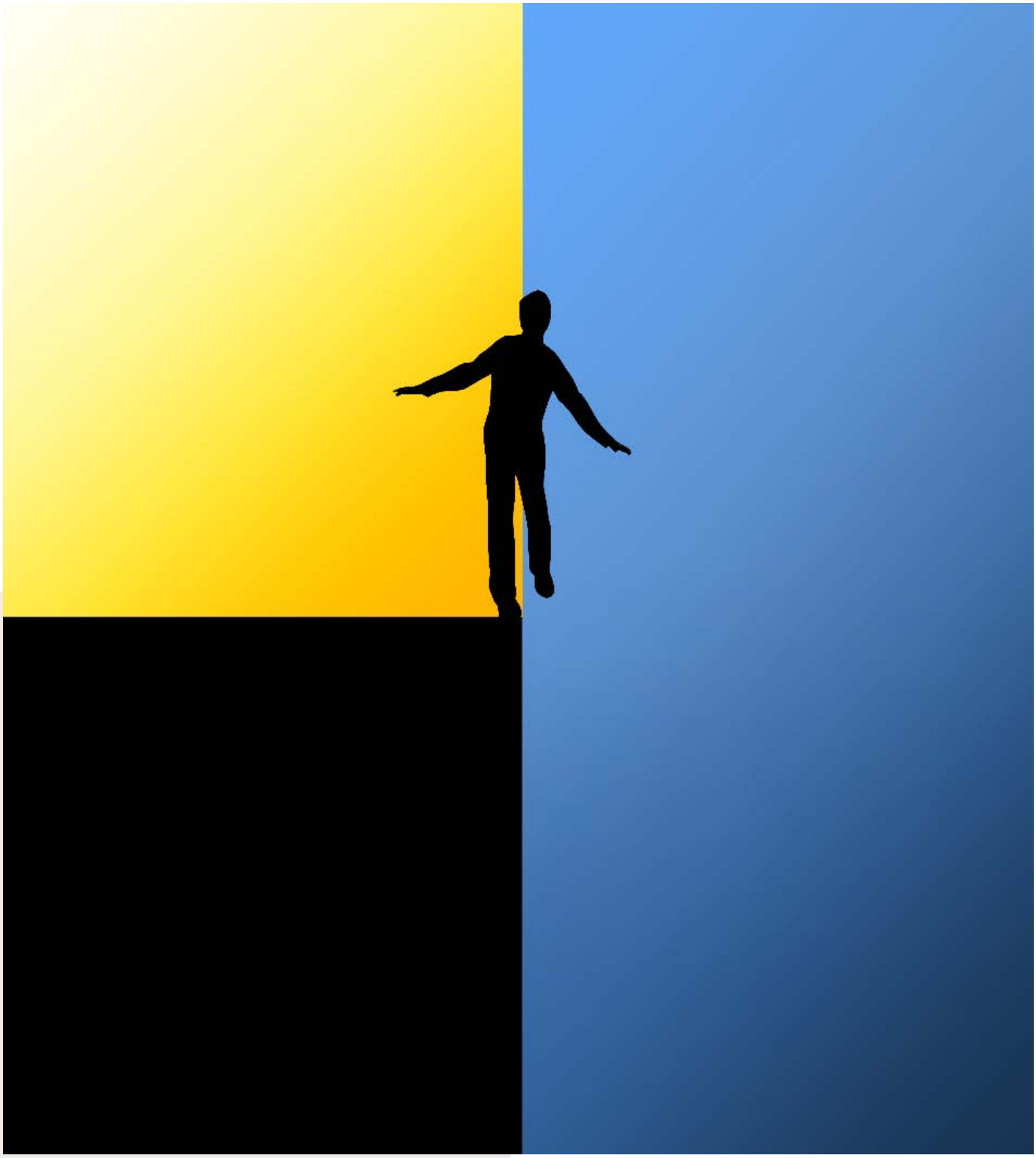
There are four main types of reliability.

1. Test-retest reliability:

The same test should produce the same or similar results when given to the same group of students at different times. This type of reliability helps measure the stability of the test over time.

2. Parallel forms (equivalent forms) reliability

This type of reliability focuses on the consistency of the measurement of different forms of the same test. Two different forms of the same test are given to the same group of students. The two forms can be administered simultaneously or at different times. If administered at different times, the stability



of measurement over time is also taken into account.

3. Internal consistency

Internal consistency is concerned with the consistency of the measurement of different items within the test/instrument. It is estimated using the split half method, KR20 or Cronbach's alpha (α). The test is administered once for these methods. For the split-half method, the test is divided into odd and even items, and the scores are correlated. For KR20 or Cronbach's α, the reliability indices are calculated using their formulae. *(Also see Evidence Based on Internal Structure under Validity)*.

4. Generalizability

The analysis of variance is applied to determine the reliability of the measurement using generalizability. All other types of reliability provide a reliability coefficient, an indicator of the consistency of measurement. When Generalizability is used, in addition to estimating reliability, it also provides measurement errors introduced by different sources. This helps identify the extent of error introduced by each factor and minimise these if possible. Generalizability is useful for assessments like OSCEs and admissions tests.

Validity

Validity is the essential criterion to be met by any measurement and assessment, including educational assessment. It refers to the degree to which the assessment measures what it is supposed to measure; that is, the scores and their interpretations truly reflect the candidates' ability or achievement in the domain to be assessed. In short, validation is a process that involves gathering evidence to ensure correct interpretations of the assessment scores.

Once thought to consist of many sub-types, validity is now considered a unitary concept, all being construct validity. The construct is the purpose or intent of the assessment and interpretation of the scores it produces. Hence, the validation process usually begins with a clear statement of the proposed intention of the assessment and the purpose of score interpretation. This is followed by gathering evidence to ensure the interpretation of scores as intended in the construct. The validation process includes one or more of the following types of evidence-gathering steps aligned to the stated purpose:

1. Evidence Based on Test Content

One of the preliminary types of evidence for validity usually involves ensuring alignment of the assessment content against the construct. This often includes using a table of specifications which evolves from the construct and explicitly describes the domain, depth and breadth of the content to be assessed, along with the type and number of assessment items defined for the content. Content experts or specialists are involved in the confirmation of the assessment content with the construct. The term 'assessment content' includes, but is not limited to, themes or concepts to be assessed, wording and format of the questions or items, and guidelines for administering and scoring the tests.

Let's consider the following examples:

Example 1: If the purpose of the assessment is to evaluate the application of basic health sciences to patient problems, assessing recall and understanding of facts or principles of basic health sciences would compromise the validity of results.

Reason: Assessment would not be aligned to the cognitive level of the learning outcome, that is, the application of scientific knowledge versus recall and understanding.

Example 2: If the purpose of the assessment is to assess the skills to use the laryngoscope effectively and appropriately, assessing such an objective using a paper and pencil test would be invalid.

Reason: Assessment would not be aligned to the learning outcome, psychomotor skill versus knowledge.

2. Evidence Based on Response Processes

There are two primary sources of evidence for the response process: the examinees and the examiners. Evidence needs to be sought for the fit between the engagement process of the candidates while performing the test and the true construct of the examination.

For example, if the students were required to reason out their responses in a reasoning test. Also, for assessments depending upon the examiners to observe and score candidates' performance, the process followed by the assessors and the extent of adherence to the pre-decided scoring criteria may provide significant evidence for validity.

Example: During marking, if the marker is not trained or the marking scheme is not clearly spelt, it will compromise the response process.

Reason: During scoring by an untrained marker or in the absence of a (clearly developed) marking scheme, the marker may award scores based on subjectivity/ individual biases, compromising the scores' validity.

3. Evidence Based on Internal Structure

Evidence of internal structure works best for assessments measuring a single construct, that is, unidimensional tests. The high correlations between test items or between different components of an examination form the basis for evidence regarding internal structure.

For assessments focusing on more than one construct, there should be high correlations between test items within a sub-construct and not with items across different sub-constructs.

4. Evidence Based on Relations to Other Variables

Valid assessment scores should demonstrate high correlations with scores from other external assessments measuring similar constructs and low or negligible correlations with scores from assessments measuring different constructs. These types of evidence are known as

convergent and discriminant evidence, respectively. If the scores from external assessments are available for comparison simultaneously, this is known as concurrent validity. This is, however, not always possible, especially if the comparison must be with some future performance, known as predictive validity.

5. Evidence Based on Consequences of Testing

Consequential validity requires gathering evidence for the intended and unintended consequences of assessment scores and decisions as compared to the construct.

6. Integrating the Validity Evidence

A well-built validity study would integrate all the validity evidence to support or disavow the proposed interpretation of assessment scores. Such a study would include, but would not be limited to, evidence for appropriate test construction, administration, scoring, and standard setting.

7. Threats to Validity

There are two main threats to the validity, construct underrepresentation and construct irrelevant variance. Construct underrepresentation means that the assessment could not capture the breadth or depth of the assessment content or construct. On the other hand, construct irrelevant variance is the influence on the scores by factors external to the intended construct. Examples of construct irrelevant variance may include a requirement of good vocabulary knowledge to perform on a science test, anxiety during the examination, and more than the intended difficulty of a test. Attention to the issues of validity and reliability of measurement and assessment ensures accurate, precise, and fair outcomes. Gathering evidence for validity and reliability and sharing the evidence with stakeholders makes assessment decisions more transparent and defensible.in student assessment. Medical Teacher, 41(4): 457-464.

References

- Bloxam, S., and Boyd, S. (2007). Developing effective assessment in higher education. Berkshire: McGraw-Hill Education.
- Downing, Steven. M. (2003). Validity: on the meaningful interpretation of assessment data. Medical Education, 37:830-837.
- Downing, Steven. M. (2004). Reliability: on the reproducibility of assessment data. Medical Education, 38:1006-1012.
- Standards for Educational and Psychological Testing. (1999). Washington, United States of America: American Educational Research Association.
- Yousuf N (2018). Principles of Assessment: Leaders' Perspective. The Reformer, Issue 3 (May): 77-84.
- Yousuf N., Mohammad M, Nisar R., Jeeva S (2019). Assessment and Quality Assurance Cycles: An Interwoven Thread. The Reformer, January: 21-31.
- Zuberi RW., Klamen DL., Hallam J, Yousuf N, Beason AM., Neumeister EL., Lane R & Ward J (2018): The journeys of three ASPIRE winning medical schools toward excellence