Validity and Reliability – Assessment Basic Principles

Validity versus reliability

Șișianu A.
Technical University of Moldova,
Faculty of Computers,Informatics and Microelectronics
Chișinău, Moldova
ala_at@mail.ru

***Abstract* - For every dimension of interest and specific question or set of questions, there are a vast number of ways to make questions. Although the guiding principle should be the specific purposes of the research, there are better and worse questions for any particular operationalization. How to evaluate the measures?**
Two of the primary criteria of evaluation in any measurement or observation are:
Whether we are measuring what we intend to measure.
Whether the same measurement process yields the same results.
These two concepts are validity and reliability.

***Key words*: assessment, concept, test, tool, validity, reliability.**

I. INTRODUCTION

Nowadays when life enforces us a lot of tests, these must be not only effective and easy to solve. The most important in a test is its quality. To make a test qualitative it should possess some basic principles. The present paper proposes as its objective the study of the main features that make the very quality of a test.

The principles of validity and reliability are fundamental cornerstones of the scientific method/assessment. In order for assessments to be sound, they must be free of bias and distortion. Reliability and validity are two concepts that are important for defining and measuring bias and distortion.

Reliability refers to the extent to which assessments are consistent. Just as we enjoy having reliable cars (cars that start every time we need them), we strive to have reliable, consistent instruments to measure student achievement. Another way to think of reliability is to imagine a kitchen scale. If you weigh five pounds of potatoes in the morning, and the scale is reliable, the same scale should register five pounds for the potatoes an hour later (unless, of course, you peeled and cooked them). Likewise, instruments such as classroom tests and national standardized exams should be reliable – it should not make any difference whether a student takes the assessment in the morning or afternoon; one day or the next.

Validity refers to the accuracy of an assessment - whether or not it measures what it is supposed to measure. Even if a test is reliable, it may not provide a valid measure. Let us imagine a bathroom scale that consistently tells you that you weigh 59 kg. The reliability (consistency) of this scale is very good, but it is not accurate (valid) because you actually weigh 66 kg (perhaps you re-set the scale in a weak moment)! Since teachers, parents, and school districts make decisions about students based on assessments (such as grades, promotions, and graduation), the validity inferred from the assessments is essential - even more crucial than the reliability. In addition, if a test is valid, it is almost always reliable.

II. RELIABILITY AND ITS TYPES

Reliability is the degree to which an assessment tool produces stable and consistent results.

Types of Reliability

Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time.

Example: A test designed to assess student learning in psychology could be given to a group of students twice, with the second administration perhaps coming a week after the first. The obtained correlation coefficient would indicate the stability of the scores.

Parallel forms reliability is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions.

Example: If you wanted to evaluate the reliability of a critical thinking assessment, you might create a large set of items that all pertain to critical thinking and then randomly split the questions up into two sets, which would represent the parallel forms.

Inter-rater reliability is a measure of reliability used to assess the degree to which different judges or raters agree in their assessment decisions. Inter-rater reliability is useful because human observers will not necessarily interpret answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed.

Example: Inter-rater reliability might be employed when different judges are evaluating the degree to which art portfolios meet certain standards. Inter-rater reliability is especially useful when judgments can be considered relatively subjective. Thus, the use of this type of reliability would probably be more likely when evaluating artwork as opposed to math problems.

Internal consistency reliability is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results.

Average inter-item correlation is a subtype of internal consistency reliability. It is obtained by taking all of the items on a test that probe the same construct (e.g., reading comprehension), determining the correlation coefficient for each *pair* of items, and finally taking the average of all of these correlation coefficients. This final step yields the average inter-item correlation.

Split-half reliability is another subtype of internal consistency reliability. The process of obtaining split-half reliability is begun by "splitting in half" all items of a test that are intended to probe the same area of knowledge (e.g., World War II) in order to form two "sets" of items. The *entire* test is administered to a group of individuals, the total score for each "set" is computed, and finally the split-half reliability is obtained by determining the correlation between the two total "set" scores.

III. Validity and its Types

Validity refers to how well a test measures what it is purported to measure.

Why is it necessary?

While reliability is necessary, it alone is not sufficient. For a test to be reliable, it also needs to be valid. For example, if your scale is off by 5 lbs, it reads your weight every day with an excess of 5lbs. The scale is reliable because it consistently reports the same weight every day, but it is not valid because it adds 5lbs to your true weight. It is not a valid measure of your weight.

Types of Validity

1. Face Validity ascertains that the measure appears to be assessing the intended construct under study. The stakeholders can easily assess face validity. Although this is not a very "scientific" type of validity, it may be an essential component in enlisting motivation of stakeholders. If the stakeholders do not believe the measure is an accurate assessment of the ability, they may become disengaged with the task.

Example: If a measure of art appreciation is created, all of the items should be related to the different components and types of art. If the questions are regarding historical time periods, with no reference to any artistic movement, stakeholders may not be motivated to give their best effort or invest in this measure because they do not believe it is a true assessment of art appreciation.

2**. Construct Validity** is used to ensure that the measure is actually measure what it is intended to measure (i.e. the construct), and not other variables. Using a panel of "experts" familiar with the construct is a way in which this type of validity can be assessed. The experts can examine the items and decide what that specific item is intended to measure. Students can be involved in this process to obtain their feedback.

Example: A women's studies program may design a cumulative assessment of learning throughout the major. The questions are written with complicated wording and phrasing. This can cause the test inadvertently becoming a test of reading comprehension, rather than a test of women's studies. It is important that the measure is actually assessing the intended construct, rather than an extraneous factor.

3. Criterion-Related Validity is used to predict future or current performance - it correlates test results with another criterion of interest.

Example: If a physics program designed a measure to assess cumulative student learning throughout the major. The new measure could be correlated with a standardized measure of ability in this discipline, such as an ETS field test or the GRE subject test. The higher the correlation between the established measure and new measure, the more faith stakeholders can have in the new assessment tool.

4. Formative Validity when applied to outcomes assessment it is used to assess how well a measure is able to provide information to help improve the program under study.

Example: When designing a rubric for history one could assess student's knowledge across the discipline. If the measure can provide information that students are lacking knowledge in a certain area, for instance the Civil Rights Movement, then that assessment tool is providing meaningful information that can be used to improve the course or program requirements.

5. Sampling Validity (similar to content validity) ensures that the measure covers the broad range of areas within the concept under study. Not everything can be covered, so items need to be sampled from all of the domains. This may need to be completed using a panel of "experts" to ensure that the content area is adequately sampled. Additionally, a panel can help limit "expert" bias (i.e. a test reflecting what an individual personally feels are the most important or relevant areas).

Example: When designing an assessment of learning in the theatre department, it would not be sufficient to only cover issues related to acting. Other areas of theatre such as lighting, sound, functions of stage managers should all be included. The assessment should reflect the content area in its entirety.

What are some ways to improve validity?

Make sure your goals and objectives are clearly defined and operationalized. Expectations of students should be written down.

Match your assessment measure to your goals and objectives. Additionally, have the test reviewed by faculty at other schools to obtain feedback from an outside party who is less invested in the instrument.

Get students involved; have the students look over the assessment for troublesome wording, or other difficulties.

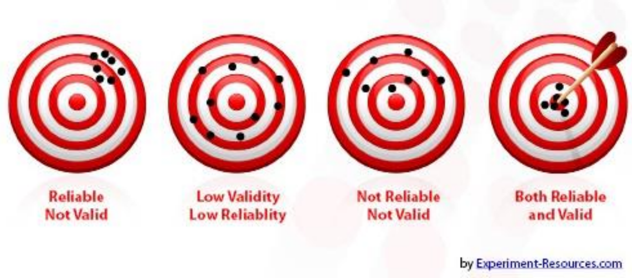
If possible, compare your measure with other measures, or data that may be available.

IV. Conclusion

If you have constructed your experiment to contain validity and reliability then the scientific community is more likely to accept your findings.

Eliminating other potential causal relationships, by using controls and duplicate samples, is the best way to ensure that your results stand up to rigorous questioning. So what is the relationship between validity and reliability? The two do not necessarily go hand-in-hand. At best, we have a measure that has both high validity and high reliability. It yields consistent results in repeated application and it accurately reflects what we hope to represent. It is possible to have a measure that has high reliability but low validity - one that is consistent in getting bad information or consistent in missing the mark. It is also possible to have one that has low reliability and low validity - inconsistent and not on target.

Finally, it is not possible to have a measure that has low reliability and high validity - you cannot really get at what you want or what you are interested in if your measure fluctuates wildly.



by Experiment-Resources.com

Tips for boosting measurement validity and reliability:

General:

- Always consider pilot testing your instrument with the target population of your research.
- Have experts in the area of your research check or provide guidance on your data collection tools.
- It is imperative that the test you select will collect data on the types of skills your research is targeting. For example, if you are teaching math to children with cognitive impairments, you will need a test that will be sensitive enough to detect growth in their learning within your timeframe. This is a question best determined in consultation with a professional researcher in your area of study.

Surveys, interviews and focus groups:

- You need to check your questions to determine if they are prompting the types of responses you expect. Run a pilot test with a small set of people from your target population. Note that these must be people who will not otherwise be involved in the study.

Observations:

- The observation protocol or record keeping sheet is critical to getting credible data. Spend time to pilot the protocol with your observers. Try it out in a shared observation (a videotaped classroom would be effective for this) and then discuss ratings. Did all the raters mark events in the same manner? If not, why not? It is critical to work this out in advance.

References

[1] Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational*

[2] Measurement (2nd ed.). Washington, D. C.: American Council on Education.

[3] Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation, 7(10). [Available online: http://pareonline.net/getvn.asp?v=7&n=10].

[4] The Center for the Enhancement of Teaching. How to improve test reliability and validity: Implications for grading. [Availableonline:http://oct.sfsu.edu/assessment/evaluating/htmls/improve_rel_val.html].