Is psychometrics science?

Geoff Norman¹

Published online: 8 August 2016

© Springer Science+Business Media Dordrecht 2016

If you have been counting, this is my third editorial (some might say “rant”) in the past 4 issues with “validity” in the title. It’s not exactly Chapter 3 of a Gothic romance; the issues raised in each editorial are overlapping but distinct. Certainly there was no grand design like a Len Deighton or Ken Follett trilogy. More like individual chapters around a common theme.

The idea for this one occurred only a few days ago. Here’s what happened. I had become involved in a study on campus of the relationship between teacher self-perception and student perceptions about some aspects of teaching skills, using an introductory undergraduate course. (I am being deliberately vague in order to protect the guilty).

We completed the study; basically a comparison of student, teacher and peer ratings, and had the opportunity to present a preliminary analysis to the author of the instrument. Well, things did not go well, to put it mildly. We appeared to disagree at every turn. Even the mildest starting point turned into a heated exchange. Several times I came very close to simply walking out.

However, to the subject at hand, in the course of a rancorous debate, he challenged me with the question “How would you validate it?” I knew what he wanted me to say. Every paper he wrote on his scale involved a factor analysis of the responses, and typically came out with 2 factors (but sometimes one). I pointed out that, while we had about 80 responses, they were derived from student ratings of 6 or 7 teachers, so our sample size was 6, which was too small to factor analyze a 22 item questionnaire.

AN ASIDE:

This appears to be the most widely unrecognized error in psychometrics. When you administer a learning environment questionnaire to 100 students, a “satisfaction with the simulation lab” survey to 500 users, a “how do you like the food” card to 1000 patrons, the person filling it out is the **rater** of the environment, the lab or the cafeteria. The sample size is not 100, 500 or 1000—it’s **ONE**. One learning environment (maybe more, but we don’t

✉ Geoff Norman
norman@mcmaster.ca

¹ McMaster University, Hamilton, ON, Canada

know how many), one simulation lab, one cafeteria. So all the variance between items we use to do the factor analysis or the internal consistency or the correlations is rater variance, error variance, not true variance. The results are meaningless. I have turned down 3 papers this year for precisely this reason. A sad waste of time and energy.

Meanwhile back at the meeting.....

I tried to describe a holistic approach to validation, à la Messick, including looking at inter-rater reliability, predictive validity, concurrent validity, and so on. That too led to a long and unprofitable back and forth, including a comment from the celebrity that there was no point in looking at inter-rater reliability because we already know the tool is unreliable. (Well if you knew the tool was unreliable, why are you still using it? You can't have validity without reliability).

And yet, one part of me was sympathetic to his position. I was very conscious that I was advocating a bunch of analyses that somehow were supposed to sum up to a decision about whether it was valid or not. But as I reflected on it, there was no obvious way to aggregate all these studies into a final decision.

Fortunately, after about an hour and a half, he had to break off to prepare for a public talk and I went home, relieved that the ordeal was over. But the whole episode got me thinking? What I *should* have said in response to his challenge, "How are you going to validate it?" was "How would you INvalidate it?" But that begs the question of what kind of evidence constitutes evidence of invalidity.

How, precisely, do we go about deciding whether an assessment is valid or not? Or even more critically, can we make that decision? That is, if the investigation of an assessment instrument is to be viewed as a scientific enterprise, it must pass a critical test. If we set out to validate a test, this must include some strategy or criterion that makes it possible to INvalidate it. Yet I could not think of a single instance when the community collectively decided that a tool was not valid and should be abandoned.

Well, maybe sort of. Past editorials have challenged measures of learning style, emotional intelligence, personality. Certainly, there have been reviews strongly rejecting these constructs. Yet all of these instruments continue to attract practitioners and researchers. Even The Meyers-Briggs Type Inventory, originally developed in the 1940s and repeatedly challenged by study after study, remains a mainstay in the toolbox of the HR department.

How can we approach the question? There are threads in various domains that provide a direction for further thinking. Within the psychometric domain itself, the notion of a refutable theory maps nicely onto Messick's formulation in which all empirical research can be viewed as a form of construct validity, in which the research seeks to test a hypothetical construct—a theory—concurrent with the measure itself. If the results are positive, the theory is supported and the measure is validated. On the other hand, a negative result may mean that the theory is wrong, the measure is poor, or both. So at least we have transformed validity into a hypothesis-testing framework. But can the Messick framework lead to a decision of invalidity? It is not clear how this would come about. And yet, such a dichotomous decision is part of the bedrock of science.

Time for some philosophy of science. There is no accepted definition of science, but one characteristic of science, championed by Popper, is that scientific hypotheses must be falsifiable. In his philosophy, a single refutation of a theory spells the demise of the theory, whereas no amount of confirmations can verify the theory. All you need is one black swan to reject the hypothesis that all swans are white.

Popper's philosophy was not quite right; while logically, one falsification does doom a theory, in the social situation of a scientific discipline, things are not quite that simple. In our field, and in medical research generally, just about any question can be associated with

confirmations and falsifications. The fact that systematic reviews exist is explicit recognition that many important questions are accompanied by contradictory evidence. On the other hand, the systematic review is explicitly directed at verification and falsification. It generally addresses a question of the form “It works—it doesn’t work”. While the answer is messier than in Popper’s clean logic, nevertheless, the conclusion is still possible.

In fact, this is generally true of science. Rarely does a single experiment lead to rejection of a theory; instead the process of theory development is much more gradual and social. That is the essence of Kuhn’s “scientific revolutions”.

But even this is far from the whole story. Neither Kuhn nor Popper have anything to say about scientific progress. For both, progress amounts to starting all over again after you’ve thrown out the last theory. But it’s not like that. Practitioners know that every theory is an approximation relying on a mountain of simplifications. And its ultimate fate is to be replaced with a better one. As one source (that I cannot track down) said, “All theories are wrong. But some are more useful than others.”

So is psychometrics any different from any other kind of science. I think so. Generally speaking, in experimental research, we advance hypotheses on which the study stands or falls. Either the teaching intervention was successful or it wasn’t ($p < .05$). Either senior students performed better than juniors or they didn’t ($p < .05$). None of these study findings will destroy a theory. Social science is too vague for that. But they do have a hypothesis which is accepted or rejected.

Somehow psychometrics isn’t like that. If we look at reliability, certainly there are accepted approaches to reliability—ICCs or G coefficients, or Kappa. A study can have poor reliability, which usually arises in two ways: either the internal consistency is too high (something few researchers seem to recognize) suggesting that the multiple items are really measuring the same thing, or other kinds of reliability are too low. For example, multiple studies of assessments of CanMEDS roles have shown that the internal consistency of ratings of different competencies typically exceeds 0.85 (Ginsburg et al. 2013; Park et al. 2014) that shows that these ostensibly different competencies cannot be distinguished by observers. Conversely, reliability of end-of-rotation evaluations is very low in the same studies —0.10 to 0.11 for one rating. Similarly, in one very large study (Wright et al. 2012) of multi-source feedback involving over 1000 physicians, inter-rater reliability of peer ratings was 0.11, and of patient ratings was 0.08. It would take 15 peers and 34 patients to achieve a reliability of 0.7.

Did this kind of dismal performance constitute sufficient evidence to lead to abandoning the approach? Not at all. We just need more observations. Occasionally poor reliability has led to abandonment, as happened with Patient Management Problems many years ago and essays a bit more recently. But the decision typically arises from practical grounds related to realistic testing times, not from reliability concerns per se. The fact is that a reliability study does not yield any test of a hypothesis which can be determined to be supported or refuted.

Construct validity is even more problematic. With reliability, at least we can examine how good the measure is, knowing that the criterion is a reliability of 1.0. On the surface it may look a bit better. Constructs ARE hypotheses or mini theories. They may be accepted or rejected. But the problem is that we do not distinguish between weak hypotheses (women are different than men) and strong (performance on this test is related to later clinical performance in practice). Indeed, far too often we see validity analyses done on things like age and gender simply because they are there, with no prediction as to where a difference should arise.

Which brings us back to our original situation. Factor analysis is frequently used to assess content validity, demonstrating consistency (or not) between the hypothesized factors or subscales and the evidence. But FA is not really a statistical procedure—there is no universal criterion for accepting a factor. And inevitably an unrotated FA yields one “g” factor, but if you want more factors, do a varimax or oblique rotation. Inevitably there will be more than one factor. In short, it is again almost impossible to reject validity based on FA.

It seems to me that, while we can proceed to design complex generalizability studies with variance components all over the place that look more scientific than just about anything else in medical education, in the end none of this constitutes critical evidence for theory-testing and decision-making. We are guilty of institutionalizing confirmation bias—every study we conduct, regardless of the findings, can be interpreted either as supporting the validity of the measure, or non-contributory. As a result, we don’t accumulate sufficient evidence to “falsify the hypothesis” and to claim that the instrument is not valid and should be abandoned.

Central to this problem is that some evidence of validity is strong, and can add weight to a falsifiability judgment. For example, if an admissions test does not predict any aspect of performance in medical school, such as is pretty well the case for personality tests (Eva 2005). Similarly, if studies of a learning environment repeatedly show that matching learning style yields no instructional benefit (Pashler et al. 2008) that should be sufficient evidence for falsification and indeed history has pretty well borne that out (although we still see some studies of learning styles and personality tests). But too many validity tests (Are senior students scoring higher than junior students?) constitute at best, very weak evidence of validity.

To return to philosophy of science. Popper once said that good science amounts to affirming bold hypotheses and rejecting cautious hypotheses. A similar approach to test validation may go far toward putting psychometrics on a scientifically defensible footing.

References

- Eva, K. W. (2005). Dangerous personalities. *Advances in Health Sciences Education: Theory and Practice*, 10(4), 275.
- Ginsburg, S., Eva, K., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Academic Medicine*, 88(10), 1539–1544.
- Park, Y. S., Riddle, J., & Tekian, A. (2014). Validity evidence of resident competency ratings and the identification of problem residents. *Medical Education*, 48(6), 614–622.
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles concepts and evidence. *Psychological Science in the Public Interest*, 9(3), 105–119.
- Wright, C., Richards, S. H., Hill, J. J., Roberts, M. J., Norman, G. R., Greco, M., et al. (2012). Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires. *Academic Medicine*, 87(12), 1668–1678.