

Messick Award

Abstract

Validating Score Interpretations and Uses**Dr Michael Kane**

The score interpretations entailed by assessment uses can be complicated. In part, these interpretations depend on the kinds of tasks included in the assessment and the domains from which these tasks are sampled, the kinds of responses called for by the test and the rules used to score them, the procedures used to administer the test, and the contexts in which the test is administered. The interpretations are also shaped by the conceptual and social frameworks within which the scores are interpreted and by the theoretical assumptions implicit in these frameworks. And fundamentally, the assessments and the interpretations are developed (or adjusted) to meet the requirements inherent in proposed test-score uses; for example, high-stakes uses generally make stronger claims and require much stronger assumptions than low-stakes uses, and the mapping of test scores to proficiency level descriptors adds an extra layer of interpretation and complexity. As a result, test-score interpretations can be very complicated. To make things worse, many of the assumptions built into score interpretations and uses may be implicit.

Validity is currently defined in terms of the degree to which a proposed interpretation is justified by evidence and involves at least two stages. First, the proposed interpretations and uses have to be specified in some detail (e.g., in the form of an explicit *interpretive argument*, laying out the inferences and assumptions leading from the scores to the claims and decisions based on the scores). As noted above, the interpretations can be complex, and not completely spelled out, and therefore, specifying the proposed interpretation and use is not trivial. Second, the proposed interpretation/use are subjected to searching criticism, and where possible, empirical evaluation. The *validity argument* provides a systematic summary of this evidence.

The process is akin to the process of theory development and theory testing in applied contexts; the question is not whether the theory (interpretation/use) is true in general, but whether it works for this purpose in this context. Like scientific theory testing, validation is simple conceptually (say what you mean and justify your claims), but gets complicated as the interpretations and uses get more and more ambitious. Language assessments pose unique challenges, because they necessarily involve complex tasks embedded in social contexts and can serve ambitious goals, but they have the advantage of being associated with fairly clear performance domains. We have to pay attention to the challenges, and we should take advantage of the opportunities.

Michael T. Kane, Ph.D., has held the Samuel J. Messick Chair in Validity at the Educational Testing Service in Princeton, New Jersey since September of 2009. He was Director of Research for the National Conference of Bar Examiners from September 2001 to August 2009. From 1991 to 2001, he was a professor of kinesiology in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Before his appointment at Wisconsin, Kane was a senior research scientist at ACT, where he supervised large-scale validity studies of licensure examinations. Kane holds a B.A. in physics from Manhattan College, and an M.S. in statistics and a Ph.D. in education from Stanford University. His main research interests are in validity theory and practice, generalizability theory, and standard setting.