AN APPROACH TO TESTING & MODELING COMPETENCE

This paper presents an approach to measuring competence, and to statistically modeling the reliability and validity of the scores produced. To be sure, there are many possible approaches. By presenting this model, I hope to stimulate debate and data. My goal is to illustrate how the field of competency testing might develop the best possible working model of competence measurement through improving the model and measurements over time.

In my approach, competence is defined by a set of six facets. These facets carve out the domain in which measures of competence – their tasks, response formats and scoring – might be developed. Assuming an indefinitely large number of possible forms of a competence measure, a particular competence test may be viewed as a sample of tasks and responses from this large domain. Under certain reasonable assumptions, the assessment tasks/responses and the raters who score test-takers’ performance can be considered as randomly sampled. In such cases, a statistical theory for modeling the reliability and validity of competence scores, generalizability (G) theory, can be used to evaluate the quality of the competency measurement.

In sketching the model, I follow the now well-known assessment triangle (National Research Council, 2001): cognition or, more generally, the construct to be measured; observation of behavior; and the interpretation of observed behavior with inference back to cognition. First, then, I attend to the definition of the construct, competence. Then I turn to observation. This entails sampling tasks from the domain bounded by the construct definition. In this way, a test of competence is built. The intent is to produce an observable performance from which to infer competence. Finally, consideration is given to the interpretation of performance scores. To what extent are competence test scores reliable? To what extent are interpretations of competence test scores supported by logical and empirical evidence? These questions call for a combination of quantitative and qualitative studies. Throughout this paper, I provide concrete examples drawn from the fields of business, the military and education. I conclude with a summary of the model.

THE CONSTRUCT: COMPETENCE

The term construct refers to an attribute of a person that is to be measured. In this case, the construct is competence. Competence, therefore, is an idea, a construction created by societies; it is not directly observable. Instead, it is inferred from...
observable performance on a set of tasks sampled from a domain of interest, such as a job or an educational discipline.

In broad terms, competence is a “… complex ability… that … [is] closely related to performance in real-life situations” (Hartig, Klieme, & Leutner, 2008, p. v; see also McClelland, 1973 and Weinert, 2001). More specifically, I (Shavelson, 2010a) identified six facets of competence from the literature: (1) complexity – a complex physical and/or intellectual ability or skill; (2) performance – a capacity not just to “know” but also to be able to do or perform; (3) standardization – tasks, responses, scoring-rubric and testing conditions (etc.) are the same for all individuals; (4) fidelity – tasks provide a high fidelity representation of situations in which competence needs to be demonstrated in the real world; (5) level – the performance must be at a “good enough” level to show competence; and (6) improvement – the abilities and skills measured can be improved over time by education, training and deliberative practice (see Shavelson, 2010a for details).

Tasks and responses that are included in competence measurement, therefore, should meet the following criteria:

1. Tap into complex physical and/or intellectual skills and…
2. Produce an observable performance using a common…
3. Standardized set of tasks with…
4. High fidelity to the performances observed in “real-world” “criterion” situations from which inferences of competence can be drawn, with scores reflecting…
5. The level of performance (mastery or continuous) on tasks in which…
6. Improvement can be made through deliberate practice.

Ideally, competence assessments would satisfy all six criteria. Practically, competence assessments will most likely tap into a subset of these criteria. Criterion 2, combined with the other criteria, emphasizes constructed responses, for example, an extended written response, a physical performance or a product. However, this criterion does not preclude the possibility that some portion of the assessment may include selected responses such as multiple-choice questions that will probably focus on the declarative knowledge that underpins competence. Criterion 4 is an ideal and the level of fidelity (low to high) may vary due to cost, time and logistical constraints. It seems that criteria 1, 3, 5 and 6 should be satisfied on any competence assessment.

In this chapter, I focus on assessments that meet, as closely as possible, all six criteria to a greater or lesser extent. Examples are drawn from several fields. As a first example of how the construct definition circumscribes an assessment (tasks, responses and scoring system), consider a measure of job performance, albeit an unusual one – a measurement of an astronaut’s performance on general maintenance tasks on Earth and in lunar and zero gravity. Qualified astronaut-like participants performed tasks in three clothing conditions – shirtsleeves, deflated space suit and inflated space suit (Shavelson & Seminara, 1968). This study, the first of its kind, found a considerable performance decrement, measured by error rate and time, as participants went from Earth’s 1 gravity to the moon’s 1/6 gravity.