pedocs
Open Access Erziehungswissenschaften

DIPF

Shavelson, Richard J.

On the measurement of competency

Empirical research in vocational education and training 2 (2010) 1, S. 41-63

urn:nbn:de:0111-opus-52350

in Kooperation mit / in cooperation with:

SKBF | CSRE

http://www.skbf-csre.ch

Nutzungsbedingungen / conditions of use

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.
By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

pedocs
Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Mitglied der Leibniz-Gemeinschaft
Informationszentrum (IZ) Bildung
Schloßstr. 29, D-60486 Frankfurt am Main
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Empirical Research in Vocational Education and Training Vol. 2(1), 2010, 41-63

On the measurement of competency

Richard J. Shavelson[a,*]

[a]*Stanford University*

Abstract

Across multiple societal sectors, demand is growing to measure individual and group *competencies*. This paper unpacks Hartig et al.'s (2008) competency definition as a complex ability construct closely related to real-life-situation performance to make it amenable to measurement. Unpacked following the assessment triangle (construct, observation, inference), competency measurement is exemplified by research from business, military and education sectors. Generalizability theory, a statistical theory for modeling and evaluating the dependability of competency scores, is applied to several of these examples. The paper then pulls together the threads into a general competency measurement model.

Keywords: Assessment, competency, generalizability theory, measurement, performance assessment

1. Introduction

Today we see increasing demand from employers, educators, policy makers and the public to prepare citizens for 21st Century Skills and to measure achievement of, or performance on them. These skills include critical thinking, analytic reasoning, problem-solving, and communicating, both in vocational and educational domains (polymechanics skills) and generic skills such as critical thinking that transfer across specific domains. The demand goes beyond simply *knowing* and includes *applying knowledge* to everyday problems and tasks. That is, the demand is for both knowing and being able to use that knowledge at work, in higher education, and in the context of individual and civil engagement with everyday activities and decisions.

While there is some agreement on the need to prepare citizens for 21st Century Skills, there is little understanding of how these skills might be measured. Images of 40-item multiple-choice tests or a series of small essays fall short of what these stake holders have in mind. But, as the report of the U.S. Commission on the Future of Higher Education[1] demonstrated, stake holders know what they want in a competency measure when they see it, endorsing the performance-based Collegiate Learning Assessment as the prototype for measuring undergraduate's learning. While a

[*]Corresponding author: 308 Cubberley–School of Education, 485 Lasuen Mall, Stanford University, Stanford, CA 94305-3096, Tel: 650-723-4040, richs@stanford.edu

[1]*A Test of Leadership: Charting the Future of U.S. Higher Education*. Retrieved from www.ed.gov/about/bdscomm/list/hiedfuture/ reports/final -report.pdf on August 10, 2009. See also Shavelson (2010).

© *2010 Sense Publishers. All rights reserved.*

prototype for generic skills measurement, its approach to competency measurement can be and as we will see has been applied to domain-specific skills.

In this paper I provide one possible vision, or more formally, model of what the measurement of competencies - be they in education, work, or everyday life - might look like and how such a measurement might be developed and evaluated. It addresses, in part, the challenge posed by Hartig et al. (2008, p. v):

> The theoretical modeling of competencies, their assessment, and the usage of assessment results in practice present new challenges for psychological and educational research.

The paper is organized along the lines of the assessment triangle (Figure 1; National Research Council, 2001): Construct - Observation - Interpretation. What is measured, indirectly, is typically called a *construct* - in our case, the construct is *competence or competency*. Competence is a "... complex ability ... [construct] that ... [is] closely related to performance in real-life situations" (Hartig et al., 2008, p. v). Note that the construct, competency, is an idea, a construction created by Western societies. It is hypothetical and cannot be observed directly. It can only be inferred from a person's behavior.

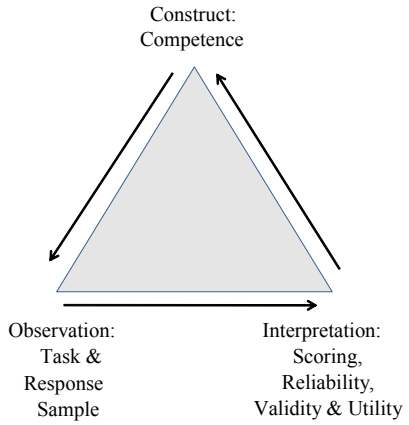


Figure 1: The Assessment Triangle

Competencies may be simple (e.g., fastening a bolt) or complex (e.g., troubleshooting a malfunctioning engine). Underlying performance and competence are a complex set of abilities. These abilities are cobbled together when a person attempts to meet task and response demands. The ability complex changes over the duration of task performance as sub-goals are met and new goals are set. These complexes are inextricably intertwined and while psychologists untangle them, analyze them separately, and add them up to provide a measure of the whole, this is not the

case of competency measurement. Rather, competency measurement focuses on the real-world tasks and responses to them recognizing a multitude of abilities are involved in performance and pulling them apart distorts the performance of greatest interest (e.g., McClelland, 1973; Shavelson et al., 2002).

From this construct, a task, or problem or stimulus can be identified that is thought to evoke the construct. By engaging in the task, a person's behavior - her response - can be *observed*. Either the presence or absence of the construct can be observed, or the person's level of *performance* can be *observed*. The universe of possible tasks and responses for observing competency performance, then, logically follow from the definition of the construct. For the purpose of building an assessment, a sample of the tasks/responses is drawn from this universe.

Having created an assessment and observed behavior on the sample of tasks and responses, the question remains: Do the scores actually measure-reliably and validly-the construct? That is, can one reliably and validly *interpret* (infer) from a person's behavior on the assessment the presence or absence of the construct, or the level of performance on the construct?

In what follows, one possible definition of the construct, competence, is set forth in a preliminary way. From that definition, concrete examples of how that definition might be put into practice are provided. These examples are drawn from business, military and education. From this characterization, a statistical theory for modeling performance on competence measurements is briefly described. It is then applied to several of the example assessments and some general findings of competency measurement are summarized. The entire package is put together in a final concrete example of an operational assessment of college students' learning. The paper concludes with one possible synoptic model for developing measures of competency.

As will be seen, the model is a work in progress. The examples of competency assessment are incomplete given the definition developed below.

2. The construct: competence

All measurement is driven by an idea, definition or mini-theory of what is to be measured-in our case, *competence*. The theory may be explicit but often it is implicit. Implicit theories become somewhat explicit when an assessment is built. The assessment contains some kinds of tasks but not others. For example, a mathematics achievement test does not contain items that probe a person's mathematics self-concept. That is, the construct definition guides the selection or *sampling* of tasks and responses that go into an assessment. The definition also is the object of inference and *interpretation*. From a sample of observed behavior, an inference is drawn about the degree of competence a person possesses in some domain.

While many of us have some notion of what it is to be competent, say in physics or psychology, we don't necessarily all share the same notion. One way to begin the search for a definition of competence is to look in dictionaries or on the web. To this end, I searched the web for definitions of competence (selectively as there are many

notions of competence). A sampling of converging definitions is provided in Table 1.

Synthesizing the gist of these definitions, I found that competence involves: (1) a physical or intellectual ability, skill or both; (2) a performance capacity to do as well as to know; (3) standardization of the conditions under which performance is observed; (4) some level or standard of performance as ”adequate,” ”sufficient,” ”proper,” ”suitable” or ”qualified”; and (5) improvement. Definition 2 (Table 1) comes close to encompassing the other definitions - proper performance combining knowledge, skills and behavior. Yet it adds two important ingredients: standardization and improvement.

Table 1: Sample of Definitions of Competence Found on the Web

1.	Quality of being adequately or well qualified physically and intellectually wordnetweb.princeton.edu/perl/webwn
2.	Standardized requirement for an individual to properly perform a specific job. It encompasses a combination of knowledge, skills and behavior utilized to improve performance. en.wikipedia.org/wiki/ Competence_(human_resources)
3.	Quality or state of being competent, i.e. able or suitable for a general role en.wiktionary.org/wiki/competence
4.	Competent - properly or sufficiently qualified or capable or efficient; ”a competent typist wordnetweb.princeton.edu/perl/webwn
5.	Competence is the acquisition of knowledge skills and abilities at a level of expertise sufficient to be able to perform in an appropriate work setting (within or outside academia) http://www.qualityresearchinternational.Com /glossary/competence.htm

Competencies have recently been defined for educational contexts by Hartig and colleagues as, ”. . . complex ability constructs that are closely related to performance in real-life situations” (Hartig et al., 2008, p. v). This definition is consistent with those in Table 1 but adds the notions of complexity and ”real-life” situations. The definitions in Table 1 are either mute as to the nature of the task to be performed, or assume a particular job task (e.g., typist). So, we can add (6) complex ability and (7) ”real-life” situations to round out the facets of our competence definition:

> Competence (1) is a physical or intellectual ability, skill or both; (2) is a performance capacity to do as well as to know; (3) is carried out under standardized conditions; (4) is judged by some level or standard of performance as ”adequate,” ”sufficient,” ”proper,” ”suitable” or ”qualified”; (5) can be improved; (6) draws upon an underlying complex ability; and (7) needs to be observed in real-life situations.

2.1 Complex physical and/or intellectual ability or skill

Let's first unpack ability and skill. Carroll (1993) defined ability as some kind of task performance or potential for performance under suitable conditions. He uses the example of weight lifting-an ability to lift a certain amount of weight or the potential for doing so. He goes on to say that a cognitive ability is any ability that concerns some class of cognitive tasks where: A cognitive task is "*...any task in which* ***correct or appropriate processing*** *of mental information is critical to successful performance*" (Carroll, 1993, p. 10, italics in original; bold author's own). It seems, then, that a complex ability would be one that places high cognitive demand on the performer. That is, the ability required to successfully perform a complex task would be multifaceted, and perhaps comprised of a set of simpler abilities and skills cobbled together to perform the task under certain situational supports and constraints (Shavelson et al., 2002).

2.2 Performance

The notion of not just knowing but being able to do is integral to Carroll's definition of ability. He captures, then, the performance aspect of ability and skills, be they physical, mental or both.

2.3 Standardization

Standardization means that the tasks/responses used to elicit performance should be identical in what is demanded of the individual and the conditions under which it is demanded; they should not vary from one person to another. This is the same as the notion of a standardized test. The tasks/responses are the same and the administration, time of day (if relevant) and the like are held constant for all test takers. Standardization rules out, for example, the use of portfolios to assess competence because typically tasks-responses, administration, and conditions vary from one portfolio to another and there is sometimes little control over actually knowing to whom the contents of a portfolio belong (Shavelson et al., 2009). This said, portfolios could be justified if they are used in a criterion situation (e.g. assessing artistic competence).

2.4 Real-life situations

The notion of a "real-life" situation suggests the nature of the tasks that a person would perform to demonstrate competency. It too needs a bit of unpacking. In his seminal work on measuring competence, McClelland (1973, p. 7) formalized the notion of "real-life" situation in reference to *criterion sampling*: "Testers have got to get out of their offices where they play endless word and paper-and-pencil games and into the field where they actually analyze performance into its component parts." For example, "If you want to know how well a person can drive a car (the criterion),

sample his ability to do so by giving him a driver's test." Or "If you want to test who will be a good policeman, go find out what a policeman does."

2.5 Level or standards

The notion of competence seems to entail some notion of "enough." That is, what level of performance is needed to be considered as performing adequately, properly, or sufficiently? This suggests that some criterion or performance level needs to be set. Those whose performance falls at about a certain level are declared "competent" and those who fall below are not.

2.6 Improvement

Competence involves the notion of improvement. That is, a person's underlying level of ability or skill is not fixed but malleable. Competence can be improved through deliberate practice, education, or some other environmental intervention. Tasks/responses on a competency assessment, then, should be amenable to improvement. (This is in contrast to abstract, figural tasks such as the Ravens Matrices which are intended to measure a stable aptitude.)

One implication of this improvement notion for the assessment of competence, as McClelland (1973) pointed out, is that it should be okay to "cheat" on a competence measurement by teaching and practicing the abilities and skills underlying performance of a real-life task. If a person cheats and learns to do real-world, complex tasks, that is what is wanted; the person is learning to perform competently on a set of tasks she will likely encounter in some form in life.

3. Observation of performance

The definition of the construct, competence, provides a very rough blueprint for building an assessment. The definition should identify a domain or universe of tasks and responses that might be sampled for an assessment. To be sure, the universe might be broader than that which can be accommodated in a time, cost and logistics-bound assessment. But it should rule in certain tasks and certain types of responses and rule out others. For *competence* as defined here, the tasks and responses should: (a) be real-life in nature, (b) tap complex abilities and skills, (c) be amenable to practice and improvement, and (d) be amenable to standard setting.

McClelland's (1973) criterion-sampling approach to assessment provides the clues for us. His approach contrasts starkly with the traditional approach to creating assessment tasks and responses. Traditionally, complex tasks are analyzed into their component parts and psychological traits underlying them are identified. Tests are then built to measure each trait, and the sum of the scores on each test is supposed to put the pieces back together again to represent the whole of performance. For McClelland (and me), something is lost in the traditional approach; the whole is greater

than the sum of the parts. A person might form a response to a situation by calling on a complex set of skills; those skills might change as the task changes over the course of the assessment (Shavelson et al., 2002). The closer the task reflects real life situations, the more likely the person's responses on the task reflect responses she makes in life.

For McClelland (1973) the best sampling of tasks for an assessment is criterion sampling. In his case this is job-performance measurement; the best predictors of job performance are an applicant's performance on samples of the job itself. Related to criterion-sampling is his recommendation that "tests should involve *operant* as well as respondent behavior" (p. 11). "Life outside of tests seldom presents the individual with such clearly defined alternatives as: 'Which dog is most likely to bite?' or 'Complete the following number series: 1 3 6 10 15 _'" (p. 11). That is, by criterion sampling, both the tasks and the responses on an assessment map onto real-life situations. Except in formal education, life rarely presents itself as a question with a set of possible answers, with only one being correct. Moreover, tasks are typically coupled to their allowable responses. Therefore, tasks and responses are not independently sampled but often are linked.

McClelland goes on to argue that "tests should assess competencies involved in clusters of life outcomes" (p. 9). That is, there is more to performance than just physical and intellectual abilities. Competent performance cannot be separated from "everyday skills" and "people skills"; in addition to cognitive abilities, competence involves conation and emotion. Our definition of competence is devoid of such skills and we may want to return to it and expand it to include personal and social responsibility skills (e.g. Shavelson, 2010).

Moreover, "tests should be designed to reflect changes in what the individual has learned" (p. 8), and while criterion-sampling permits this; multiple-choice *aptitude* tests do not. And information should be made available to those taking the test as to how to improve their performance. For example, teaching a person to drive a car is not considered cheating. Generally, teaching a person to perform well on criterion-sampled tasks is teaching them the physical and intellectual abilities and skills needed to perform at a level of competence in the domain.

In sum, to address the observation vertex of the Assessment Triangle, I recommend a criterion-sampling approach-one in which tasks and responses are sampled from criterion (real-life) situations. To this end, a sampling frame is needed:

- The domain of behavior in which competence is to be assessed should be specified.
- Next, the domain should be analyzed into its task and response make-up. This should be done by enumerating all potential tasks and responses, or at least by describing them broadly so as to include certain tasks and responses and exclude others.
- Then a sample of tasks and responses should be chosen. At issue - to be addressed later - is whether these task/responses should be sampled randomly

or purposively and how the sample might be evaluated for representativeness. This sample of tasks and responses comprises the observational part of the assessment.

The next step in characterizing the assessment of competence is to turn to the interpretation vertex of the assessment triangle. At this vertex, we specify the nature of scoring and methods for evaluating interpretations (inferences) from the sample of behavior on the assessment to our construct of interest - competence.

However, before proceeding, it seems appropriate to provide some concrete examples of assessments of performance and competence. To do so, I draw somewhat chronologically on my own research with colleagues over the past 30-plus years. The first example comes from research published in 1968 on astronauts' performance on generic maintenance tasks in lunar and zero gravities (Shavelson & Seminara, 1968). This was the first study of its kind and found a considerable performance decrement, measured by error rate and time, as the astronaut-qualified participants went from earth's 1 gravity to the moon's 1/6 gravity to zero gravity of space (Figure 2).

The performance assessment was built as follows: (a) the performance domain - tasks and corresponding responses - was identified from a lunar mission set by the U.S. National Aeronautics and Space Agency (NASA). (b) "Generic occupational tasks" were then enumerated. These were tasks that were found across a number of mission activities. (c) Tasks and their corresponding responses were purposively sampled from the universe of generic tasks. (d) Performance was observed on all tasks in all gravity conditions (and various space suit and shirt-sleeve conditions - inflated space suit shown in Figure 2). (e) Accuracy and time were measured. And (f) inferences were drawn to performance in the domain of generic tasks.

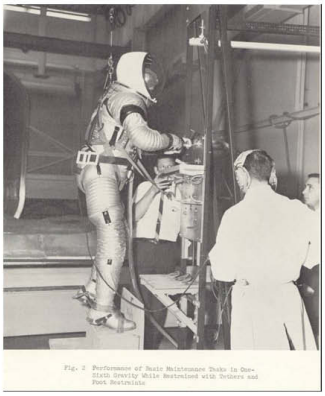


Figure 2: Assessment of Performance on Basic Maintenance Tasks (Shavelson, 1967, p. 36).

The second example is taken from the measurement of military job performance

carried out in the 1980s (Wigdor & Green, 1991). In 1980, the United States moved from a conscription military force with a draft of men 17 years and older to an all volunteer force. Of great concern to the military, and to the U.S. Congress (but for different reasons), was the ability to ”man the force” so that a sufficient number of talented and highly talented people would enlist and defer entering college or the labor force. To do so, recruiting budgets were expanded and various educational and other incentives were provided for recruiting. This increased military budget created Congressional concern. Moreover, the rationale for increasing the recruiting budget was based on a strong positive correlation between enlistment aptitude test scores and the criterion - grades in training. Congress rightly questioned whether training grades were the right criterion; shouldn’t actual job performance be the criterion? This led to the study of enlistees’ job performance in the four branches of military service in the U.S (Wigdor & Green, 1991).

To build a job performance assessment, developers: (a) Identified the universe of job tasks as specified in military ”doctrine” (”domain-specific skills”). (b) Sampled tasks-responses from that domain. At issue was whether the sample should be drawn purposively or randomly. (c) Put the sample of tasks-responses on a job-performance test. (d) Scored performance across tasks, occasions and raters. And (e) interpreted enlistees’ performance scores as representative of their job performance.

The task/response-sampling issue turns out to be important. The goal is interpretation - to draw inferences from a person’s performance on a sample of tasks to what the person would do on another sample or over the entire universe of tasks comprising the job. Scientifically, sampling theory is the preferred method as the nature of representativeness and the margin of sampling error can be specified. However, typically performance measurements employ a small sample of tasks, and leaving the composition of the assessment to chance may, as many argue, often produce an unrepresentative test. The alternative is purposive sampling. With purposive sampling, although complete control is exercised over task-response selection, and using the judgment of experts produces the tasks that ”look” representative of the job, the issue remains as to how representative the purposive sample is and how we would know it.

I (Green & Shavelson, 1987) had an opportunity to develop a method for evaluating the representativeness of a purposive sample against various forms of random sampling. For each task in a Navy Radioman job, job-incumbents (experts) rated the task as to: (1) whether they had performed it (PCTPERF), (2) how frequently they performed it (FREQ), and (3) how complicated it was to perform the task (COMP). Also, for each task, supervisors indicated (4) whether they had supervised it (PCTSUP) and rated both its (5) importance for mission success (IMPORT) and how often it was (6) performed incorrectly (ERROR). From these data, the ”universe” mean (μ) and standard deviation (σ) over all 124 job tasks could be calculated for each of the six ratings.

Then, job experts drew a purposive sample of 22 tasks from the radioman job. For these tasks, the mean and standard deviation was calculated and compared to the

universe parameters. Moreover, three random sampling schemes were identified for drawing 22 tasks: simple random sampling from an infinite universe, simple random sampling from a finite universe of 124 tasks, and stratified random sampling from a finite universe. Using the central limit theorem and sampling ratios, for each rating (e.g., ERROR) I calculated the distance (in σ units) between the purposive sample mean based on the selected 22 tasks to what would be expected from each of the random sampling methods. The results are presented in Table 2.

Table 2: Evaluation of Purposive Sampling (Wigdor & Green, 1991, p. 137)

	Domain/Sample Center		
Task	Infinite	Finite	Finite
Feature	Simple Random	Simple Random	Stratified Random
PCTSUP	5.06	6.29	3.11
PCTPERF	5.06	6.25	3.25
IMPORT	1.00	1.20	1.06
ERROR	1.00	1.20	0.29
FREQ	3.12	4.08	1.50
COMP	-1.50	-1.80	-0.89

Note: Distance between purposive and random samples in standard deviation units. (PCTSUP - percent supervised; PCTPERF - performance performed; IMPORT - importance; ERROR - frequency perform incorrectly; FREQ - frequency of performance; COMP - complexity of task)

It turned out that the purposive sampled tended to include tasks that ”looked like” the job (PCTSUP), and were performed frequently (PCTPERF). That is, the purposive sample disproportionately included tasks frequently performed on the job. The purposive sample also included tasks that incumbents rated as less complicated to perform than the average task (COMP). For this and other scientific reasons, the National Academy of Sciences urged the use of some form of random sampling for selecting tasks/responses for job-performance measurement. These sampling methods included stratified random sampling where the most important tasks could be sampled with probability 1.00.

So far I have drawn from work dealing with jobs; this area of everyday life seems an obvious candidate for producing observations from which competence can be inferred. What about education? Drawing from the military performance work, Jerry Pine, Gail Baxter, and I along with other colleagues in the Stanford Education Assessment Laboratory (SEAL) began applying the technology to science inquiry assessment in the mid 1980s. A science performance assessments is shown in Figure 3 (see Shavelson et al., 1993; Shavelson et al., 1999). The same steps used in job performance task sampling were applied to these ”hands-on” science assessments: (a) Identify a domain of science investigations (”domain specific skills”). This was done by examining hands-on science materials, textbooks, teacher and student work books and the like. (b) Sample tasks/responses from that domain. We drew purposive

samples that were highly representative of the kinds of activities students carried out in inquiry science. (c) Create a performance assessment from the tasks/responses that fits within space/safety restrictions in classrooms. (d) Score performance using trained raters. And (e) interpret a student's score over tasks, raters, occasions and methods (e.g., hands-on, paper-and-pencil) as a reflection of their capacity to carry out science investigations.

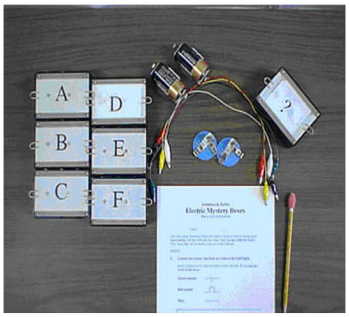


Figure 3: Science performance assessments

One important finding from this work is that paper-and-pencil measures (e.g., multiple-choice, short answer) are *not* exchangeable for hands-on performance measures (Baxter & Shavelson, 1994). It is one thing to carry out an investigation with materials that react to the actions you take and quite another to carry out an abstract, hypothetical investigation as prompted with pencil-and-paper. For example, a student with a perfect hands-on score on an electricity task received a score of 0 on the same short-answer version of the task. When asked on the pencil-and-paper task how she would know what is inside the mystery box, she said she would take a screw driver and open the box, or weigh the box or shake the box! These actions did not occur to her nor could they be accomplished to any advantage when doing the hands-on task.

Towards the end of the paper, I will describe in some detail another educational performance-task assessment of competence. The example will be drawn from higher education and the assessment of learning 21st Century Skills, in contrast to the domain-specific skills shown so far. Let's now turn to the final vertex of the assessment triangle - interpretation.

4. Interpretation: from observation to competence

The interpretation vertex of the assessment triangle focuses on the methods and tools used to score performance and draw inferences from a sample of (fallible) observations to an interpretation that reflects a person's competency level. This is an enormously broad area. My focus here will be primarily on scorning and the dependabil-

ity or reliability of drawing inferences *from* individuals' scores based on a sample of tasks/responses *to* their characteristic behavior in the full universe of tasks/responses.

4.1 Scoring

The requirement that a competency measure include observations of operant as well as respondent behavior means that most likely, a person will *construct* her responses to the tasks on the assessment. In those cases of competence measurement where responses are *respondent* and *selected* - such as a multiple-choice test of knowledge of driving laws as part of a driving examination - concern about scoring is minimal. Typically one and only one correct answer exists among the selection set and a machine can score the responses without error (or with negligible error).

Operant responses are expected to be the norm on competency measures. Consequently, the degree of accuracy or adequacy of a constructed response must be judged. And typically that involves human judgment in the form of trained raters scoring performance. Two general approaches have been taken to accomplish this (e.g., Ruiz-Primo & Shavelson, 1996). The first approach is *holistic*. It uses the typical 6-point rating scale in which performance is judged from "absent" or "extremely poor" to "outstanding." Often the scoring rubrics provide gradations from 1 to 6 such that a score of 2 has minimally one piece of relevant information, a score of 3 and 4 has several pieces and a score of 5 almost all. The problem is that the scoring rubric is so generic that it fits everything but is specific to none so is limited in its diagnosticity.

The alternative is *analytic* scoring where each task/response is scored with a rubric specifically designed for its contents. While this scoring may be highly diagnostic, it is limited in that it cannot be applied to multiple tasks/responses and therefore tends to produce a hodgepodge of scores across tasks[2].

An alternative, a kind of hybrid, is now emerging in scoring constructed responses. It is a combination of the holistic and analytic and seeks to find common features across a universe of tasks/responses that comprise the observation component of a competence assessment. The astronaut tasks/responses described above provide one obvious example: time and accuracy scores can be used across diverse tasks-responses.

However, especially on educational tasks, such measures may not be consistent with real-life behavior. For example, the Collegiate Learning Assessment employs a diverse set of tasks/responses (Shavelson, 2010). Yet all of them include a number of underlying facets including whether students recognize that the information provided may be reliable or unreliable, whether the information is valid or invalid for the particular problem at hand, and whether one or another judgmental heuristic has erroneously been used (e.g., correlation is not causality, consider baseline when

[2] A major concern, historically, has been as to whether raters are able to judge complex performance especially when observed in real time with either holistic or analytic scoring. This topic is addressed later in the paper.

making comparisons). Also clarity of expression is considered. Ratings are made on each of the dimensions; the dimensions are constant across tasks/responses while the specifics vary.

What will be shown below is that human raters can be trained to score performance consistently and accurately; high ”inter-rater” reliability is typically found. However, four problems persist. The first is rater drift. Raters need to be constantly monitored and retrained as they drift away from consistency over time. The second is cost. The use of human raters makes competency measurement expensive. The third is logistics. Bringing raters together or orchestrating them from afar presents challenges. And the fourth is time. There is a considerable delay between the person's observed performance and when a score becomes available.

New developments in natural language processing software have made it possible to score complex performance about as reliably with a computer as with human raters when operant responses come in the form of writing (e.g., Klein, 2007; Shavelson, 2010). Computers are trained to score performance on a benchmark set of papers. They then can score thousands of papers quickly and without rater drift. The logistics are easy. The assessment, delivered on an internet platform, produces electronic file responses that are scored by the computer in real time (Shavelson, 2010).

4.2 Generalizability of competency scores

The criterion-sampling approach espoused for constructing a competence assessment with real-world tasks and responses can now be formalized statistically. Suppose a competency assessment contains a random sample of tasks/responses. Performance on the tasks/responses is observed on several occasions and scored by a set of randomly-selected, well-trained raters on each of a set of task-response at each occasion. With this formulation, we are in a position to evaluate the dependability or reliability of the competence measurement statistically.

More specifically, a competence assessment can be viewed as a sample of a person's performance drawn from a complex universe defined by a combination of all possible tasks/responses, occasions and raters. The task/response sample is representative of the task/response universe. The occasion sample is viewed as representative of all possible occasions on which a decision maker would be equally willing to accept a score on the competence assessment. And the rater facet includes a representative sample of all possible individuals who could be trained to score performance reliably[3]. These three facets are, traditionally, thought of as sources of unreliability in a measurement.

In addition, it might be necessary to include different methods for observing performance on a competence assessment. For example, in evaluating the performance of jet-engine mechanics in the military job-performance project (Wigdor & Green, 1991) some tasks were carried out exactly as done on the job. When it came to work-

[3]The issue arises as to whether the conditions of the facets of the competence measurement are random. See Shavelson and Webb (1981) on the exchangeability theorem for a discussion.

ing specifically on a very expensive jet engine, a mistake would be quite expensive. So a ”walk-through” method was used and enlistees explained how they would carry out the task instead of doing the task.

By incorporating a method facet into the definition of the complex universe for observing performance, this formulation moves beyond reliability into a sampling theory of validity (see Kane, 1982). Specifically, the method facet represents all possible methods (e.g., short answer, computer simulation, hands-on, walk-through, multiple-choice, video) that a decision maker would be equally willing to interpret as bearing on a person's competence.

Once conceived as a sample of performance from a complex universe, the statistical framework of generalizability (G) theory can be brought to bear on the technical quality of competency-assessment scores (Cronbach et al., 1972; see also Brennan, 2001; Cardinet et al., 2009; Shavelson & Web, 1991). From the G theory perspective, an assessment score is but one of many possible samples from a large domain of assessments defined by the particular combination of sampled tasks, occasions, raters, measurement methods, etc.

The theory focuses on the magnitude of sampling variability due to tasks, raters and so forth, and their combinations. It provides estimates of the magnitude of measurement error in the form of variance components. In addition, it provides a summary coefficient reflecting the ”reliability” of generalizing from a sample score to the much larger universe of interest. This coefficient is called a *generalizability* coefficient in G theory, recognizing that generalization may be different across facets, depending on how a performance assessment is used. The theory also can be used to estimate the magnitude of variability among scores due to method sampling, thereby providing an index of the degree to which alternative measurement methods converge (cf. Kane, 1982).

From a generalizability perspective, sampling variability due to raters, for example, speaks to a traditional concern about the variability of performance assessments-namely, inter-rater reliability (cf. Fitzpatrick & Morrison, 1971). Sampling variability due to tasks reflects the variation in task difficulty in the task domain. Traditionally, task sampling has been thought of as related to internal consistency reliability. One goal of test developers has been to make ”items” homogeneous to increase reliability. (This does not necessarily work with performance assessment; see Shavelson et al., 1999.) Within the sampling framework, task-sampling variability is dealt with *not* by homogenizing the tasks but by increasing sample size - that is, increasing the number of tasks drawn from the universe of interest (cf. Shavelson et al., 1993). Sampling variability due to occasions corresponds to the classical notion of retest reliability. From a sampling perspective, it reminds us that decision makers are willing to generalize a person's performance on one particular occasion to many possible occasions. Finally, sampling variability due to measurement method bears on convergent validity (cf. Kane, 1982). Large method sampling variability indicates that measurement methods do not converge, as has commonly been assumed when arguing for the cost efficiency of multiple-choice testing.

To be concrete, several applications of G theory will be presented. In a study of hands-on performance of Navy Machinist Mates, we (Webb et al., 1989) examined the consistency of expert raters' real-time judgments of incumbents' performance on the assessment. In this case, machinist mates (person) were observed by expert examiners (examiner) performing 11 job tasks (task). Each machinist mate's performance was scored by two examiners on each of the 11 tasks. The total variability among these scores could be partitioned into person - the variance the measurement was designed to measure - rater, task and their combinations. Statistically, a random-effects model of the analysis of variance was used to partition and estimate variance components statistically.

The partitioning of the total variability among scores can be found in the "Source of Variance" column in Table 3. The magnitude of the variability in scores contributed by each source in the assessment is shown in the estimated variance column. And the proportion of the total variability among scores contributed by each source of variability is shown in the last column. This column provides a quick glance at where the major sources of variability are - wanted or expected variability among persons and error variability among the facets of the measurement and in interaction with person.

Table 3: Generalizability Analysis of Scores from the Navy Machinist Mates Hands-on Assessment (data from Webb, Shavelson, Kim & Chan, 1989)

Source of Variance	Estimated Variance Component (x 1000)	Percent of Total Variation Due to Each Source*
Person (P)	6.26	**14.45**
Examiner (E)	0.00	0.00
Task (T)	9.70	**22.40**
$P \times E$	0.00	0.00
$P \times T$	25.85	**60.00**
$E \times T$	0.03	0.00
$P \times E \times T$, error	1.46	3.37

*Over 100 percent due to rounding.

The variability due to person, 14.45 percent of the total variability, was predicted. Machinist mates vary in the levels of their performance. Some are more competent performers than others. The variability due to examiner and the interaction of examiner with person was zero, contrary to expectation at the time. Raters did not introduce error into the measurement. But the variability due to task was large, 22.40 percent, indicating that the sample of tasks on the assessment differed in difficulty. Most importantly, the Person × Task interaction accounted for a whopping 60 percent of total score variability, also contrary to expectation at the time. The reliability (generalizability) of the scores using 1 examiner and 11 tasks was 0.72 (in a range from 0 or chance to 1.00, perfect reliability). Adding another examiner had

no influence on reliability as examiners scored performance consistently. However, by adding another 6 tasks, reliability could be raised to 0.80.

The results of this study exemplify what has been found in job performance measurement and other domains (e.g., education) generally (e.g., Shavelson et al., 1993). At the time, these results and others on military performance measurement were surprising. Contrary to expectation, examiners were able to rate Navy machinist mates' performance reliably; they closely agreed in their scoring of complex performance in real time. Heretofore, examiner disagreement was expected to be a major source of measurement error. Moreover, contrary to expectation, very large task sampling variability was observed. That is, an incumbent's performance varied in level from one job task to the next and some tasks that were easier for certain machinist mates were more difficult for others. Generalized job expertise, then, might be more in the beholder than in observable performance. Additionally, task sampling variability, not examiner sampling, was a major cost, time, and logistics concern and continues to be.

Turning to education, we (e.g., Ruiz-Primo et al., 1993; Shavelson et al., 1993; Shavelson et al., 1999) examined the generalizability of science performance-assessment scores on tasks like those shown before (Figure 3). In one study (Shavelson et al., 1999, p. 64), each student received a score on each of three tasks (Paper Towels, Electric Mysteries and Bugs), from two raters, on two occasions with two methods of observation (notebook-recorded and direct-observation of responses). Consequently, the total variability among all of these scores can be partitioned into components for: (a) student, (b) task, (c) occasion, (d) method, and (e) all their combinations.

The partitioning of the total variability among scores can be found in the "Source of Variance" column in Table 4. The magnitude of the variability in scores contributed by each source in the assessment is shown in the estimated variance column. And the proportion of the total variability among scores contributed by each source of variability is shown in the last column. Again, this column provides a quick glance at where the major sources of variability are - wanted or expected variability among persons and error variability among the facets of the measurement.

Several findings stand out. First, the largest sources of variability in students' scores are due to task *and* occasion sampling (Person × Task, Person × Task × Occasion, and Person × Task × Occasion × Method, error). This finding, especially the task sampling finding, has been replicated across many performance assessments. Moreover, these findings suggested that not only task sampling, but occasion sampling influenced performance-assessment measurement error. While often performance is measured on only one occasion, this finding reminds us that the occasion facet is a "hidden facet" influencing the magnitude of the task-sampling variance. And finally, in this particular case, variation in measurement method introduced an unappreciable amount of variance in scores. However, in other studies, it has contributed to greater inconsistency (e.g., Shavelson et al., 1999).

Table 4: Variance Components Contributing to Reliability (Person) and Unreliability of a Science Performance Assessment (data from Ruiz-Primo et al., 1993)

Source of Variance	Estimated Variance Component (x 1000)	Percent of Total Variation Due to Each Source
Person (P)	330.00	**13.11**
Task (T)	0.00	0.00
Occasion (O)	100.00	4.14
Method (M)	0.00	0.00
$P \times T$	650.00	**25.83**
$P \times O$	0.00*	0.00
$P \times M$	10.00	0.50
$T \times O$	30.00	1.03
$T \times M$	120.00	4.67
$O \times M$	0.00	0.00
$P \times T \times O$	790.00	**31.35**
$P \times T \times M$	20.00	0.63
$P \times O \times M$	0.00	0.00
$T \times O \times M$	0.00	0.00
$P \times T \times O \times M$, error	470.00	**18.70**

*A small negative variance component was set to zero.

We summarized some of these findings on science performance assessment as follows. Task sampling is consistently a major source of measurement error; rater sampling is not. Occasion sampling, typically a hidden facet, is also a major source of measurement error, especially in combination with task sampling. We conclude from these findings that multiple tasks/responses will need to be included on a competency assessment to attain reliable scores for individuals.

5. Putting it all together: the Collegiate Learning Assessment (CLA)

The CLA was developed to measure undergraduates' learning of "21st Century Skills" - in particular their ability to think critically, reason analytically, solve problems, and communicate clearly. The assessment focuses on the institution or on programs within an institution, not on individual students. Institution or program-level scores are reported, both as observed performance and as value added beyond what would be expected from entering students' admission (e.g., SAT) scores. The CLA also provides students their scores on a confidential basis so they can gauge their own performance.

The assessment consists of two major components: a set of performance tasks and a set of two different kinds of analytic writing prompts (see Figure 4). The performance tasks component presents students with problems and related information and asks them either to solve them or recommend a course of action based on the evidence provided. The analytic writing prompts ask students either to take a position

on a topic or to critique an argument.

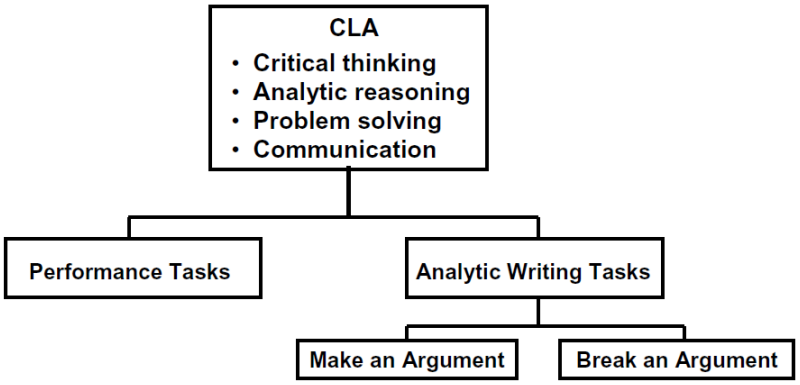


Figure 4: Collegiate Learning Assessment structure

5.1 The CLA performance tasks

The CLA's conceptual underpinnings are embodied in McClelland's (1973) criterion sampling approach to measurement. As noted above, this approach assumes that the whole is greater than the sum of the parts and that complex tasks require the integration of abilities that cannot be captured when divided into and measured as individual components.

The criterion-sampling notion is straightforward: If you want to understand what a person knows and can do, you must select a sample of tasks from the domain in which that person is to act, observe her performance on these tasks, and infer competence and learning. For example, if you want to understand whether a person knows not only the laws that govern driving a car but also whether she can actually drive a car, a multiple-choice test will not suffice. You also need to administer a driving test with a sample of tasks from the general driving domain (e.g., starting a car, pulling into traffic, turning right and left in traffic, backing up, parking). Based on this sample of performance, it is possible to draw valid inferences about her driving performance more generally.

The CLA follows the criterion-sampling approach by defining a domain of real-world tasks that are holistic and drawn from life situations. It samples tasks and collects students' *operant responses*. There are no multiple choice items in the assessment. Finally, the CLA provides CLA-like tasks to college instructors so they can "teach to the test." With the criterion-sampling approach, "cheating" by teaching to the test is not a bad thing. If a person "cheats" by learning and practicing to solve complex, holistic, real-world problems, she has demonstrated the knowledge and skills educators seek to develop in students. That is, she has learned to think

critically, reason analytically, solve problems and communicate clearly.

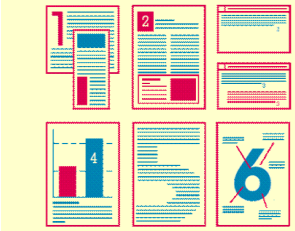
5.1.1 The Collegiate Learning Assessment's criterion-sampling approach

Recall that the CLA is composed of performance tasks and analytic writing tasks. DynaTech is an example of a performance task (see Figure 6). DynaTech is a company that makes instruments for aircraft. The company's president was about to approve the acquisition of a SwiftAir 235 for the sales force when the aircraft was involved in an accident. As the president's assistant you (the student) have been asked to evaluate the contention that the SwiftAir is accident prone. Students are provided an "in-basket" of information that might be useful in advising the president. They must weigh the evidence - some relevant, some not; some reliable, some not - and use this evidence to support a recommendation to the president. (Incidentally it might be that the SwiftAir uses Dynatech's altimeter!) DynaTech exemplifies the type of performance tasks found on the CLA and their complex, real-world nature.

Figure 6

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

1: Newspaper articles about the accident
2: Federal Accident Report on in-flight breakups in single engine planes
3: Pat's e-mail to you & Sally's e-mail to Pat
4: Charts on SwiftAir's performance characteristics
5: Amateur Pilot article comparing SwiftAir 235 to similar planes
6: Pictures and description of SwiftAir Models 180 and 235



Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.

Figure 5: CLA's DynaTech Performance Task

5.1.2 CLA analytic writing tasks

The CLA contains two types of analytic writing tasks. The first type of task asks students to build and defend an argument. For example, students might be asked to agree or disagree with the following premise, justify their position with evidence and show weaknesses in the other side of the argument: "College students waste a lot of time and money taking a required broad range of courses. A college education should instead prepare students for a career."

The second type of task is one in which a student is asked to critique an argument such as the following:

> A well-respected professional journal with a readership that includes elementary school principals recently published the results of a two-year study on childhood obesity. (Obese individuals are usually considered to be those who are 20 percent above their recommended weight for height and age.) This study sampled 50 school children, ages 5-11, from Smith Elementary School. A fast food restaurant opened near the school just before the study began. After two years, students who remained in the sample group were more likely to be overweight relative to the national average. Based on this study, the principal of Jones Elementary School decided to confront her school's obesity problem by opposing any fast food restaurant openings near her school.

In this case, the student must evaluate the claims made in the argument and either agree or disagree, wholly or in part and provide evidence for the position taken.

5.2 CLA technology

Many of the ideas underlying the CLA are not new. The history of learning assessment shows that assessments similar to the CLA have been being built for decades. In the late 1970s, John Warren at the Educational Testing Service (ETS) was experimenting with constructed-response tasks; American College Testing (ACT) created the College Outcomes Measurement Project (COMP); and the state of New Jersey created Tasks in Critical Thinking to assess undergraduates' learning. Although these assessments had marvelous performance tasks, the attempts to build these assessments failed. They were costly, logistically challenging and time consuming to score.

What makes the CLA different is that it tackles these problems of time, cost and scoring by capitalizing on internet, computer and statistical-sampling technologies. The advent of these technologies has made it possible to follow in the tradition of the criterion-sampling approach. Students' complex performance is still scored by human judges but computers with natural language processing software are increasingly doing this; computers score performance on the analytic writing prompts. In neither case is reliability or validity compromised (e.g., Shavelson, 2010). Moreover, the CLA uses matrix sampling so that not all students answer all questions, which reduces testing time. (Nevertheless, even with this technology, it takes a fair amount of time - 90 minutes - to answer subsets of questions.) Finally, reports can be produced rather quickly because of the technology used[4].

[4]Technical quality information including reliability and validity can be found in Chapter 4 of Shavelson (2010).

6. One possible model for measuring competency

Following the Assessment Triangle (construct - observation - interpretation), I have presented a working definition of competence, shown how this definition permits the sampling of certain tasks and responses and rules out others, and presented methods for scoring and generalizing from scores on an assessment to the universe of performance interpreted as competence. The working definition had seven or eight features, depending on how you count, only some of which were dealt with in detail here (1, 2 and 4 below). These features are:

1. Complex physical and/or intellectual ability or skill (combines features 1 and 6 enumerated above) required
2. Overt ”operant” performance involved
3. Standardization across individuals
4. Real-life situations based on sampling ”criterion situations”
5. Level or standard of performance indicating competence or levels of competence
6. Improvement possible on tasks-responses in competence domain
7. Conative and emotional as well as cognitive involved in competent engagement with tasks-responses

Tasks/responses, consistent with the definition, are sampled from the universe of tasks/responses that is either explicitly or implicitly specified in the definition. This sampling of criterion tasks - typically operant-response tasks found in real-life situations - is to be done randomly; the random sampling scheme can be quite sophisticated.

Because most tasks/responses will be operant or constructed, scoring will have to be done initially by humans and then perhaps subsequently by computers. This calls for the development of scoring rubrics to capture performance; these rubrics should also create a common framework for scoring performance across tasks/responses in the universe. The recommendation is to go beyond typical holistic and analytic scoring rubrics to either generalizable scores such as accuracy or time, or to include hybrid rubrics in which a common set of dimensions, based on the competence domain measured, can be used across the universe of tasks/responses.

The sampling framework underlying this model of competency measurement leads to the statistical evaluation of the quality of the measures - their generalizability and interpretability - within the framework of generalizability theory. The theory statistically evaluates the dependability of scores and can be used to determine the size of the samples of tasks/responses (or number of human judges, or number of occasions) needed in an operational assessment to attain a reliable measurement of competence.

My hope, then, is that this (or some other model) is adopted across research and development groups involved in measuring competency. In this way, a center of gravity will be created and new advances in one domain will most likely inform measurement in another competency domain. The goal, in the end, is to create continuous improvement in both our measurement methods and our theories of competency.

References

Baxter, G.P. & Shavelson, R.J. (1994.) Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, 21, 279-298.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cardinet, J., Johnson, S. & Pini, G. (2009). *Applying generalizability theory using EduG*. New York: Routledge/Psychology Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, MA: Cambridge University Press.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Fitzpatrick, R. & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.) *Educational measurement*. Washington, DC: American Council on Education.

Green, B. & Shavelson, R.J. (1987). Distinguished panel discussion on issues in the joint-service JPM program. In H.G. Baker and G.J. Laabs (Eds.) *Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies*. , Washington, D.C.: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Hartig, J., Klieme, E. & Leutner, D. (Eds.) (2008). *Assessment of Competencies in Educational Contexts: State of the Art and Future Prospects*. Göttingen: Hogrefe & Huber.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6, 125-160.

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In D. Nolan & T. Speed (Eds) *Probability and Statistics: Essays in Honor of David A. Freedman. IMS Collections, Volume 2*. Beachwood, OH: Institute for Mathematical Statistics.

McClelland, D.C. (1973). Testing for competence rather than testing for "intelligence". *American Psychologist*, 28(1), 1-14.

National Research Council (2001). *Knowing what students know: The science and design of Educational Assessment*. Washington, DC: National Academy Press.

Ruiz-Primo, M.A. & Shavelson, R.J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, 33(10), 1045-1063.

Shavelson, R.J. (1967). *Man's performance of basic maintenance tasks in lunar gravity*. Unpublished master's thesis, San Jose State College.

Shavelson, R.J. (1991). Generalizability of military performance measurements: I. Individual performance. In A.K. Wigdor & B.F. Green, Jr., (Eds.) *Performance assessment for the workplace (Vol. II): Technical issues*. Washington, D.C.: National Academy Press.

Shavelson, R.J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.

Shavelson, R.J., Baxter, G.P. & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215-232.

Shavelson, R.J., Klein, S. & Benjamin, R. (2009). The limitations of portfolios. *Inside Higher Education*. Retrieved August 10, 2009 from on http://www.insidehighered.com/views/2009/10/16/shavelson.

Shavelson, R.J., Roeser, R.W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Quihuis, G. & Gallagher, L. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment*, 8(2), 77-100.

Shavelson, R.J., Ruiz-Primo, M.A. & Wiley, E.W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36 (1), 61-71.

Shavelson, R.J. & Seminara, J. (1968). Effect of lunar gravity on man's performance of basic maintenance tasks. *Journal of Applied Psychology*, 52(3), 177-183.

Shavelson, R.J. & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.

Webb, N.M., Shavelson, R.J., Kim, K-S & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinists mates. *Military Psychology*, 1(2), 91-110.

Wigdor, A.K. & Green, B.F. Jr., (Eds.) (1991). *Performance assessment for the workplace (Vol. I)*. Washington, D.C.: National Academy Press.