# An Approach To Testing And Modeling Competence

Richard J. Shavelson

Stanford University& & SK Partners, Inc.

Invited Address

Conference on Modeling and Measurement of Competencies in Higher Education

Berlin, Germany
February 24, 2011

richs@stanford.edu

# Overview

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

(Hartig, Klieme and Leutner, 2008, p. v).

- Definition of competence that delineates the domain of tasks, responses and scoring
- Observation of responses on tasks within a competence domain
- Interpretation involving inference from behavior on the assessment to competence
- Model of competence measurement

# The Construct, Competence, Defined: Six Facets

1. Complexity
2. Performance
3. Standardization
4. Fidelity
5. Level
6. Improvement

# A Competence Measure Should:

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

# Competence Measurement Illustrated: Job Performance

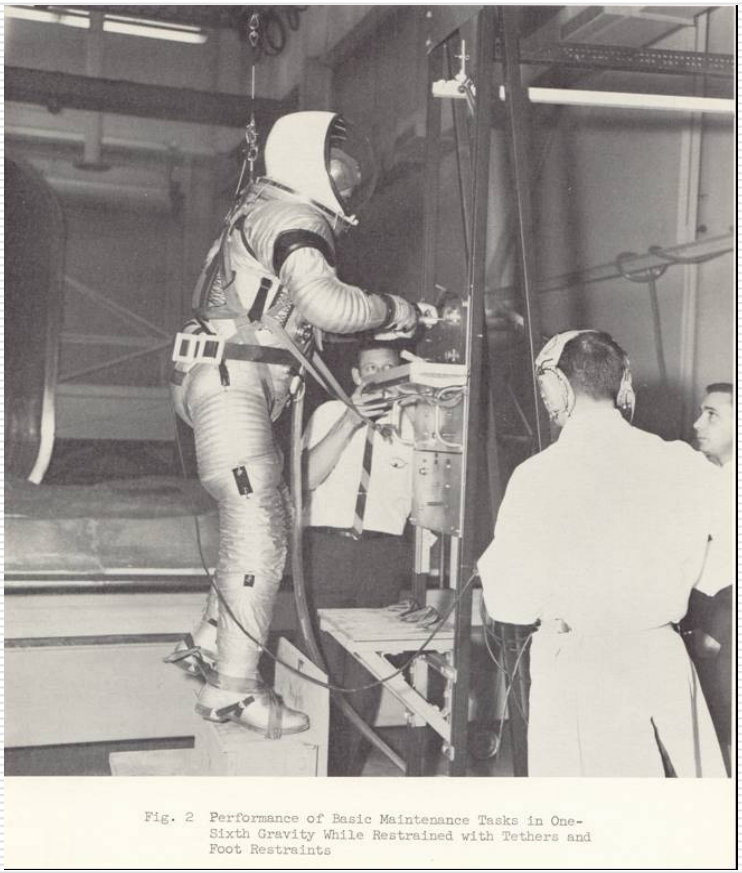
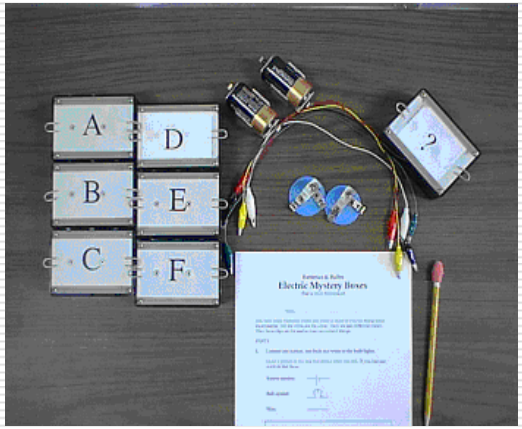


Fig. 2 Performance of Basic Maintenance Tasks in One-Sixth Gravity While Restrained with Tethers and Foot Restraints

Source: Shavelson (1967, p. 36)

# Competence Measurement Illustrated: Science Education PA




Electric Mysteries


Paper Towels

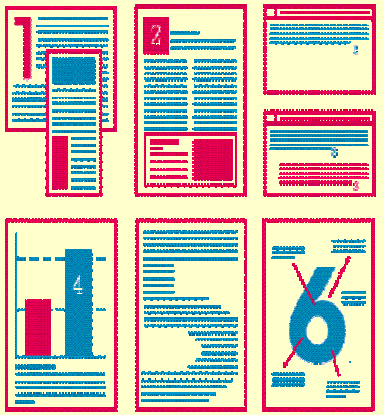
Source: Shavelson, Baxter & Pine, 1991

# Competence Measurement Illustrated: Collegiate Learning Assessment

## DynaTech Performance Task

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.

Source: Shavelson (2010) and Council for Aid to Education (www.cae.org)

# Observation of Performance: Job and Education

## Astronaut Performance

- Identify and enumerate the performance domain—tasks and corresponding response
- Enumerate "generic occupational tasks"--tasks that were to be carried out across a number of mission activities
- Purposively sample tasks and their corresponding responses from the universe of generic tasks.
- Observe performance on all tasks in all gravity conditions and clothing conditions
- Measure accuracy and time
- Draw inferences to performance in the domain of generic tasks

## Science Education Performance

- Identify a domain of science investigations ("domain specific skills").
- Sample tasks/ responses from that domain
- Create a performance assessment from the tasks/responses that fits within space/safety restrictions in classrooms
- Score performance using trained raters
- 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

# Observation: Sampling Tasks From Military Jobs

- Identified the universe of job tasks as specified in military "doctrine" for a job such as Navy machinist mate ("domain-specific skills")
- Sample tasks/responses from that domain. At issue was whether the sample should be drawn purposively or randomly
- Put the sample of tasks/responses on a job-performance test
- Scored performance either using objective evidence such as an infantryman's rifle accuracy in a simulated combat field using laser technology.
- Interpret enlistees' performance scores as representative of their job performance.

# Observation Of Performance: Task Sampling

| Task Feature | Domain/Sample Center: Infinite Simple Random | Finite Simple Random | Finite 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 is between means calculated from the purposive sample and the alternative random sample models in standard error units. PCTSUP—percent supervised; PCTPERF—performance performed; IMPORT—importance; ERROR—frequency perform incorrectly; FREQ—frequency of performance; COMP—complexity of task.

Source: Wigdor & Green (1991, p. 137).

# Observation: Task Sampling From Fuzzy Universes—The CLA

- Everyday situations
- Document library
- Constraint of comprehensibility
- Document utility
  - Reliability of information
  - Validity of information
  - Susceptibility of information to judgmental error
- Product of deliberation

# Interpretation Of Performance

- *Interpretation* refers to the inferences drawn from an individual's behavior on a sample of tasks to what his or her behavior would be, on average, if he or she performed on all of the tasks in the large universe of possible tasks.
- In other words, can one reliably and validly *interpret* (infer) from a person's performance on *a small sample of tasks* the presence or absence of competence, or the level of competence in the full domain?

# Interpretation: Evidence

- Quantitative
  - Reliability
  - Validity
    - Predictive
    - Concurrent
    - Convergent
    - Etc!
- Qualitative
  - Content validity
  - Cognitive validity

# Interpretation: G Theory

- A competence assessment may be conceived of:
  - As a sample of a person's behavior
  - On a sample of tasks/responses
  - On a sample of occasions
  - Delivered by several methods
  - As scored by a sample of raters

# Interpretation: G Theory Continued

- Inter-rater reliability → rater sampling variability
- Internal consistency reliability → task ("item") sampling variability
- Retest reliability → occasion sampling variability
- Convergent validity → method sampling variability

# Interpretation: Generalizability Of Machinist Mates' Scores

| 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 x E | 0.00 | 0.00 |
| P x T | 25.85 | **60.00** |
| E x T | 0.03 | 0.00 |
| P x E x T, error | 1.46 | 3.37 |

*Over 100 percent due to rounding.

Source: Webb, Shavelson, Kim & Chan (1991, p. 137).

# Interpretation: Generalizability Of CLA Performance-Task Scores

| Source Variability | Variance Component | Estimate | %Total |
|---|---|---|---|
| School ($s$) | $\sigma_s^2$ | 817.47 | 20.90 |
| Task ($t$) | $\sigma_t^2$ | 0[a] | 0 |
| Judge ($j$) | $\sigma_j^2$ | 62.56 | |
| 1.60 | | | |
| $s \times t$ | $\sigma_{st}^2$ | 671.42 | 17.10 |
| $s \times j$ | $\sigma_{sj}^2$ | 62.18 | |
| 1.60 | | | |
| $t \times j$ | $\sigma_{tj}^2$ | 0[a] | 0 |
| $s \times t \times j, e$ | $\sigma_{stj,e}^2$ | 2305.77 | 58.80 |

[a]Negative variance component set equal to zero.

Webb, Shavelson & Steedle (2009).

# Interpretation: Standard Setting

- Judgmental procedures
- Empirical procedures

"College Readiness Benchmark Scores for the ACT represent median test scores that are predictive of student success [B or better] in relevant college courses" (ACT, 2011, p. 3). http://www.act.org/standard/

**College Readiness Benchmark Scores**

| Subject Test | EXPLORE Test Score | | PLAN Test Score | ACT Test Score |
|---|---|---|---|---|
| | Grade 8 | Grade 9 | | |
| English | 13 | 14 | 15 | 18 |
| Mathematics | 17 | 18 | 19 | 22 |
| Reading | 15 | 16 | 17 | 21 |
| Science | 20 | 20 | 21 | 24 |

# Summary: Model Of Competence Measurement

- Definition of competence with six facets circumscribes the domain of tasks, responses and scoring to which inferences about competent performance are to be drawn
- Tasks, responses, scorers, methods, and the like are sampled from this domain to create a competence measurement
- Standards are set for what constitutes competent performance
- Assuming random sampling from the domain, statistical models such as G theory can be used to model and evaluate interpretations of the measurement
- Qualitative methods including content validity and cognitive validity can also be used to evaluate proposed interpretations

# Limitations Of The Model

- Need for multiple performance tasks to get reliable estimate of performance (due to task-sampling variability)
- Construct under-representation
- Time, cost and logistical challenges
- Paucity of human resources capable of building competence measures
- Cost of human scorers (but possibility of machine scoring)

# Thank You!