On the Large Scale Assessment of Academic Achievement:
The Role of Performance Assessment

Richard J. Shavelson
Stanford University

Invited Address
Congress of the German Society for Educational Research
Göttingen University
September 21, 2000
Overview

- What’s a performance assessment?
- What do they look like?
- What does it measure as part of a large-scale assessment?
- What do we know about its technical quality?
- How far along are we in building a technology for performance assessment?
What’s a *Science* Performance Assessment?

- One or more investigation tasks using concrete materials that *react to the actions taken* by the student
- A format in which students respond (e.g., drawing, table, graph, short-answer)
- A system of scoring involving professional judgment that considers both investigation *processes* and *accuracy* of findings
Comparative Tasks

• There are two or more categories (conditions) of an attribute or variable $A$
• There is a dependent variable $B$
• The problem consists of finding the effect of $A$ on $B$
• The problem solver has to conduct an experiment
• Correct solutions involve correct control, manipulation, and measurement of variables
Saturated Solutions Investigation

Students are asked: Find out which of three powders is the most and the least soluble in 20 ml. of water.
Component Identification Tasks

• There is a set of components which may be combined in a number of possible ways
• Each combination produces a specific reaction/result
• The problem consists of testing for the presence of each component
• Correct solutions involve using confirming and disconfirming evidence for the presence of the components in each combination
Mystery Powders Investigation

Students are asked to:

Part I: Examine four powders using five tests (sight, touch, water, vinegar and iodine).
Part II: Find the content in two mystery powders based on their observations.
Classification Tasks

• There is a set of specimens with similarities and differences
• The problem consists of sorting the specimens along two or more dimensions
• The problem solver has to use, construct, or formalize a classification with mutually-exclusive categories
• Correct solutions involve critical dimensions that allow finding relationships
Bottles Investigation

Students are asked: Find out what makes bottles [varying in mass and volume] float or sink.
Observation Tasks

• There is a set of phenomena that cannot be observed directly or in a short time
• The problem consist of finding facts
• The problem solver has to model phenomena and/or carry out systemic observations
• Correct solutions involve obtaining accurate data
• Correct solutions involve explaining conclusions satisfactorily
Daytime Astronomy Investigation

Students are asked to model the path of the sun from sunrise to sunset and use direction, length, and angles of shadows to solve location problems.
How Would You Classify This One from TIMSS?

PULSE

At this station you should have

A watch
A step on the floor to climb on

Read ALL directions carefully.

Your task:

Find out how your pulse changes when you climb up and down on a step for 5 minutes.

This is what you should do:

• Find your pulse and be sure you know how to count it. IF YOU CANNOT FIND YOUR PULSE ASK A TEACHER FOR HELP
• Decide how often you will take measurements starting from when you are at rest.
• Climb the step for about 5 minutes and measure your pulse at regular intervals.

1. Make a table and write down the times at which you measured your pulse and the measurements you made.
2. How did your pulse change during the exercise?
3. Why do you think your pulse changed in this way?
Response Formats

- Essay
- Graph
- Drawing
- Table

- Equation
- Short-Answer
- Record of Observations
- Other
Scoring Systems

• Analytic
  – Comparative task: Procedure based
  – Component task: Evidence based
  – Classification task: Dimension based
  – Observation task: Data-accuracy based

• Rubric
  – Likert-type rating scale
  – Likert scale usually collapses analytic dimensions
## Summary: Type of Tasks and Scoring Systems

<table>
<thead>
<tr>
<th>Scoring System</th>
<th>Type of Assessment Task</th>
<th>Comparative Investigation</th>
<th>Component Identification</th>
<th>Classification</th>
<th>Observation</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytic</td>
<td></td>
<td>Paper Towels</td>
<td>Bugs</td>
<td>Incline Planes</td>
<td>Friction</td>
<td>Bubbles</td>
</tr>
<tr>
<td>Procedure-Based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence-Based</td>
<td></td>
<td>Electric Mysteries</td>
<td>Mystery Powders</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimension-Based</td>
<td></td>
<td></td>
<td>Rocks and Charts</td>
<td>Sink and Float</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Accuracy-Based</td>
<td></td>
<td></td>
<td></td>
<td>Day-Time Astronomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holistic Rubric</td>
<td></td>
<td>Leaves (CAP Assessment)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


What Do PAs Measure As Part of a Large-scale Assessment?

Declarative Knowledge (Knowing the “that”)

Procedural Knowledge (Knowing the “how”)

Strategic Knowledge (Knowing the “which,” “when,” and “why”)

Cognitive Tools:
- Planning
- Monitoring

Domain-specific content:
- facts
- concepts
- principles

Domain-specific production systems

Problem schemata/strategies/operation systems

Proficiency
Low High

Extent
(How much?)

Structure
(How is it organized?)

Others
(Precision? Efficiency? Automaticity?)
## Linking Assessments to Achievement Components

<table>
<thead>
<tr>
<th>Extent</th>
<th>Declarative Knowledge</th>
<th>Procedural Knowledge</th>
<th>Strategic Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Multiple-Choice</td>
<td>Performance Assessments</td>
<td>• Performance Assessments</td>
</tr>
<tr>
<td></td>
<td>• Fill-in</td>
<td></td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• M-C Tests</td>
</tr>
<tr>
<td>Structure</td>
<td>Concept Maps</td>
<td>Procedure Maps</td>
<td>Models/Mental Maps</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some Empirical Evidence on Links between Knowledge and Measurement Methods

Correlations from Shultz’s Dissertation (N=109 6th Graders Studying Ecology):

- Reading and **Multiple-Choice**: 0.69
- Reading and **Concept Map**: 0.53
- M-C and CM: 0.60

- Reading and **Performance Assessment**: 0.25
- M-C and PA: 0.33
- CM and PA: 0.43
What Do We Know About the Technical Quality of Performance Assessments?

- Framework for evaluating reliability and some aspects of validity
- Summary of studies and findings
- Implications for large-scale assessment:
  - Are raters a significant source of sampling variability (error)?
  - Are task and occasion major sources of sampling variability (error)?
Sampling Framework

Standard
Science as Inquiry:
“Design and Conduct a Scientific Investigation”

Define

Domain
Force & Motion

Construct

Define

Observed Behavior on the Task/Response Sampled

Sample

Generalizable to Other Tasks in the Domain?

Task/Response
Friction

Task/Response

Task/Response

Task/Response

Task/Response

Task/Response

Task/Response

Task/Response

Task/Response
Sampling Framework

A score assigned to a student is but one possible sample from a large domain of possible scores that the student might have received if a different sample of assessment tasks were included, if different judges evaluated performance, and the like...

Is a score assigned generalizable, for example, across:

• Tasks?
• Occasions?
• Raters?
• Methods?
• Expertise?
Task or Occasion Sampling
Variability or Both?

- If task sampling variability, stratifying on tasks may reduce variability and number of tasks needed in large-scale assessment
- If occasion sampling, unlikely to increase the number of occasions
- If both, need for a large number of tasks (hint: both!)
Table 1
Variance Component Estimates for the Person x Rater x Task x Occasion
G Study Using the Science Data
(from Shavelson, Baxter & Gao, 1983)

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>n</th>
<th>Estimated Variance Component</th>
<th>Percent Total Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person (p)</td>
<td>26</td>
<td>.07</td>
<td>4</td>
</tr>
<tr>
<td>Rater</td>
<td>2</td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>Task (t)</td>
<td>2</td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>Occasion (o)</td>
<td>2</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>pr</td>
<td></td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>pt</td>
<td></td>
<td>0.63</td>
<td>32</td>
</tr>
<tr>
<td>po</td>
<td></td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>rt</td>
<td></td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>ro</td>
<td></td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>to</td>
<td></td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>prt</td>
<td></td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>pro</td>
<td></td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>pto</td>
<td></td>
<td>1.16</td>
<td>59</td>
</tr>
<tr>
<td>rto</td>
<td></td>
<td>0.00a</td>
<td>0</td>
</tr>
<tr>
<td>prto, e</td>
<td></td>
<td>0.08</td>
<td>4</td>
</tr>
</tbody>
</table>

Source: Shavelson, Ruiz-Primo & Wiley, 1999
Convergence of Hands-On and Computer Simulation PAs

\[ r_{H1H2} = 0.53 \]
\[ r_{C1C2} = ? \]
\[ r_{H1C1} = 0.52 \]
\[ r_{H2C2} = ? \]
\[ r_{H1C2} = ? \]
\[ r_{C1H2} = 0.45 \]
\[ r_{C1C2} = ? \]
<table>
<thead>
<tr>
<th>ISSUE</th>
<th>STUDY FOCUS</th>
<th>FINDINGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tasks</td>
<td>• Compare student performance across different tasks.</td>
<td>• Students performance is not consistent across tasks. Task sampling variability is large at both individual and school level.</td>
</tr>
<tr>
<td>Occasions</td>
<td>• Compare student performance across occasions</td>
<td>• Students performance is not consistent across occasions even though they receive about the same scores.</td>
</tr>
<tr>
<td></td>
<td>• Examine task and occasion interaction</td>
<td>• The interaction between person, task and occasion is the largest source of score variability.</td>
</tr>
<tr>
<td>Raters</td>
<td>• Examine consistency of scores across raters</td>
<td>• Inter-rater reliability coefficients are generally higher than .80. However, coefficients lower than .70 have been obtained in some assessments.</td>
</tr>
<tr>
<td></td>
<td>• Examine sequence of methods and domain expertise</td>
<td>• High reliability coefficients may mask important disagreements among raters.</td>
</tr>
<tr>
<td>Methods</td>
<td>• Examine the proximity of the assessments to the curriculum characteristics</td>
<td>• Exchangeability across methods is limited due to volatility in students performance across occasions.</td>
</tr>
<tr>
<td>Task Dimensions</td>
<td>• Examine the proximity of the assessments to the curriculum characteristics</td>
<td>• Close-curriculum assessments are more sensitive than proximal-curriculum assessments to changes in students’ performance.</td>
</tr>
</tbody>
</table>
How Far Along Are We in Building a PA Technology?

• A PA technology, analogous to a paper-and-pencil technology developed in the last century is within reach

• Dimensions of variation among PAs account for differences among PAs themselves and students’ thinking
  – Task x Scoring System Classification (Shavelson, Ruiz-Primo, Baxter)
  – Content x Process Characterization (Baxter & Glaser)
  – Basic, Quantitative and Spatial Reasoning (Ayala, Shavelson, & Ruiz-Primo)

• Item shells (design specifications) have been developed that guide, only very generally, PA development (Solano-Flores & Shavelson)

• Computer simulation is the next step in building the technology to address cost, logistics, and time issues but research must examine the exchangeability of PA simulation with hands-on equivalents