Possible Futures for Assessing Educational Achievement

Richard J. Shavelson
Stanford University

Invited Talk
Learning Research and Development Center
University of Pittsburgh
January 14, 2000
Overview of Talk

• An Achievement Framework
  – What does it mean to “achieve” in a domain like science?
  – How might we measure achievement in science?
  – What evidence is there that we’re measuring different aspects of achievement?

• Items for the Design of Accountability Systems

• Concluding Comments
Toward An Achievement Framework: Knowledge Components

Characteristics That Vary According to Proficiency Level

<table>
<thead>
<tr>
<th>Extent (How much?)</th>
<th>Structure (How is it organized?)</th>
<th>Others (Precision? Efficiency? Automaticity?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declarative Knowledge (Knowing the “that”)</td>
<td>Procedural Knowledge (Knowing the “how”)</td>
<td>Strategic Knowledge (Knowing the “which,” “when,” and “why”)</td>
</tr>
</tbody>
</table>
| Domain-specific content:  
  - facts  
  - concepts  
  - principles | Production system--condition-action rules | Problem schemata/strategies/operation systems |

Cognitive Tools:
- Planning
- Monitoring

Low  
High
## 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><strong>Multiple-Choice</strong></td>
<td><strong>Performance Assessments</strong></td>
<td><strong>Models/Mental Maps</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fill-in</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conceptual Maps</strong></td>
<td><strong>Procedural Maps</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Procedural Maps</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extent of Declarative Knowledge

Air is made up of many gases. Which gas is found in the greatest amount?

A. Nitrogen
B. Oxygen
C. Carbon Dioxide
D. Hydrogen

Source: TIMSS Pop. 2
Structure of Declarative Knowledge

From White & Gunstone: “Probing Understanding” (1992, p. 16). Eleven-Year-Old’s Concept Map
Procedural Knowledge: Performance Assessments for FOSS Curriculum
Procedural Knowledge:
Performance Assessment: TIMSS Pop. 2

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?
Strategic Knowledge: Mental Models

• What causes day and night?
  A. The earth spins on its axis (.66)
  B. The earth moves around the sun (.26)
  C. Clouds block out the sun (.03)
  D. the sun goes round the earth (.04)

Strategic Knowledge: Mental Models

• (A) A rocket is moving along sideways in deep space, with its engine off, from point A to point B. It is not near any planets or other outside forces. Its engine is fired at point B and left on for 2 sec while the rocket travels from point B to point C. Draw in the shape of the path from B to C. (Show your best guess for this problem even if you are unsure of the answer.)

• (B) Show the path from C after the engine is turned off on the same drawing.

Some Empirical Evidence: 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
Some Other Empirical Evidence

- Correlations (covariances) tell you about rank ordering on different measures--that’s only part of the story.
- Evidence is needed for the cognitive claims that different methods measure somewhat different aspects of achievement:
  - Talk aloud
  - Focus group
  - Group work
Cognitive Validity Framework

Comparing assessment tasks on three dimensions

Intended Task Demands → Inferred Cognitive Activity → Scores Obtained

Congruence
Talk Aloud with Concept Maps: Method Variation

Correspondence Between “Directedness” and Inferred Activity

Correlation between "Directedness" and inferred activity in the context of methods.

- Explanation
- Monitoring
- Conceptual Errors
- No-Code

Verbal Codes

Proportion Scores

- Contract-A-Map
- Fill-In-Nodes
- Fill-In-Lines
Talk Aloud with Concept Maps: Performance Level

Construct-A-Map

Proportion Scores

<table>
<thead>
<tr>
<th>Explanation</th>
<th>Monitoring</th>
<th>Conceptual Error</th>
<th>No Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fill-In-The-Nodes

Proportion Scores

<table>
<thead>
<tr>
<th>Explanation</th>
<th>Monitoring</th>
<th>Conceptual Error</th>
<th>No Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Accountability’s Good and Bad News

• **Good News**: Demand for accountability is warranted and if done well, could improve teaching and learning

• **Bad News**: If current K-12 high-stakes accountability systems serve as models, the demand for accountability will harm not benefit education by significantly narrowing the curriculum and the diversity of teaching methods
Accountability Based On Faulty Logic

- Accountability must be inferred from observing outcomes in any system where all actions cannot be observed directly.
- To do this “inferencing,” the performance measure is an indicator of the desired behavior, not the behavior itself.
  - In business,
    - There is a clear performance measure--revenue or stock price--to guide business decisions and action.
    - You can’t manage a business if you can’t measure it’s performance.
  - In education,
    - Performances (goals) are many and debated.
    - The performance indicator--most often a multiple-choice achievement test--is but a distant proxy for the desired outcome.
    - When this indicator becomes an end in itself, and it does in education, well-intentioned accountability may very well distort the system it was intended to improve.
Linking Assessment To Learning Activities

<table>
<thead>
<tr>
<th>Type</th>
<th>Purpose</th>
<th>Agency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formative</td>
<td>Learning</td>
<td>student teacher teacher</td>
</tr>
<tr>
<td>Summative</td>
<td>Certification</td>
<td>external tests individual</td>
</tr>
<tr>
<td></td>
<td>Accountability</td>
<td>external tests individual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>external tests sample surveys</td>
</tr>
</tbody>
</table>

Paul Black 3/98
California Learning Assessment System: Data Sources

A. “On Demand” Matrix Sampling of Tasks & Tests

B. Standardized Curriculum-Embedded Assessments

Student’s Score

C. Portfolios
Conceptual Framework For CLAS

Aggregate Level of Performance

A. Matrix Sample Benchmark: Multiple-Choice & Performance-Based Assessment
   “Moderated” Score: Individual, School & District Score

Individual Level of Performance

B. Standardized Curriculum-Embedded Assessments

C. Portfolios
   Teacher Moderation
   Sample from Class for Aggregation
   Teacher Calibration & Professional Development
Governor Wilson’s Rationale For Discontinuing CLAS

SB 1273 [CLAS] takes a different approach . . . . Instead of mandating individual student scores first, with performance-based assessment incorporated into such scores as this method is proven valid and reliable, it mandates performance-based assessment now and treats the production of individual student scores as if it were the experimental technology--which it clearly is not. In short, SB 1273 stands the priority for individual scores on its head.
Concluding Comments on Assessment and Accountability

• A broader framework for achievement testing is feasible and is needed to improve teaching and assessment
• Teacher formative and summative evaluation practices, and accountability systems should include assessments linked to this broader framework
• Accountability systems should link formative and summative assessment of achievement to bring about education consistent with goals
  – Teachers’ formative classroom assessments are poorly developed but, if well developed, could raise standards significantly
  – Student self-assessment is poorly developed but, if well developed, would contribute most to self-directed, meaningful learning
  – Teachers’ summative assessments are currently problematic but if improved might be used for accountability purposes
    • more productive instructionally, and
    • more dependable for assessment purposes