Assessment

Fall Quarter 2007



High-Stakes Testing



- All Tests Have Shortcomings
- Tests Tend to Concentrate on Discrete Thinking
 - Knowledge/skill in bits and pieces
 - Emphasis on recognition rather than production of knowledge

When Testing Has A Lesser

- Role

 Discrete Test Items
 Sample From A Larger
 Valued Set of
 Proficiencies
- But These Test Items Do Not Embody All that We Mean by A High-Quality Education.
- Still, Inferences to the Domain are Valid



When Testing has a Larger

Rolline Stakes are Raised, Then the Content of Testing Begins to Define Educational Goals

- Instruction is Shaped Accordingly
- Test Scores Are No Longer Predictive of a Broader and More Significant Educational Achievement
 - Test scores lose their meanings
 - They no longer "sample" from the larger domain"

How to Respond?

- Possibly, Reduce the Extent of Testing
 - But there is growing accountability
- Possibly, Make Tests More Worthy of Their New Role
 - So that teaching to the test does good rather than harm

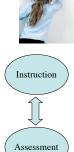


Features of New Assessments

- Alternative, Authentic, and Performance Assessment All Involve Tasks That Are Complex, Extended in Time, and Intrinsically Worthwhile
- The Emphasize Student Production
- It's Implicit That Students Are Thinking at a High Level
 - Critical thinking and problem solving rather than recognition and recall

Assessment in the Service of Instruction

- · Blending of Assessment and Instruction
 - Classroom or formal assessment
- · Teaching (Like Testing) Should Involve Diagnosis and Response
- Assessment (Like Teaching) Should Always Have Value for Learning





here Do Assessments Come From?

- · Goals and Objectives for Education Are the Prerogative of Multiple Stakeholders
- Standards Frameworks
 - -National and state frameworks
 - Illustrative tasks
 - -General Rubric
 - E.g., characteristics of a good essay

Creating the Assessment Creating the Assessment Framework

- (Specifications)
- · Knowledge objectives by process dimensions
 - Knowledge: Planets, molecules, vertebrates
 - · Process: Reasoning, representational forms, types of items (e.g., constructed response, multiple-choice).
- · Fairness Entails Opportunity to Learn



Ser Fress					
Tupos	Fo yigon or posters hes been unded	A poster. Sarbera Marc, but is set	A partice harbon tokened convolut convolut	A solf portion has been taken with strong expension	
Oquater	The letter is 1907 to reposited and yeaps social	orpeand, for days set have, steag forg.	The letter is copesited.	The letter is very unit opposed and ar has sufficing story.	
Helissan	posty softward reduced sorted springed posterial	The bone or facty well earlier, but contains spread of printing and printing and pr	The latter or mall within, lect contains stoom operators presented costs.	The latter or very well writter. There are to spelling or grammatical extens.	
Former	The letter is not written. is the consect format	The letter has acless comment compressate	The letter is written in saleuret Ou- coment Output	The letter is written in the correct flower.	

Scoring Rubrics

- · Scoring Must Be Reliable and Fair
- Scoring Rubrics Can Focus Efforts of Students and Teachers
 - This is what constitutes good work
- · Clear, Unambiguous, Meaningful
- · Separated Into Performance Levels
 - Descriptions of each level (anchoring)

Types of Scores

- · Holistic or Analytic?
 - Holistic is overall impression
 - Analytic involves aggregation of several dimensions
- Norm-Referenced or Criterion-Referenced?
 - Norm-referenced is comparative (e.g., standard scores)
 - Criterion-referenced cites performance against a standard
 - Can be both, and the two can cross-map



Reliability



- · Reliability is the Replicability of the Test Score
 - What happens when a tape measure is made of rubber?
- One Positive Feature of Traditional Testing
 - Large number of questions (items) tends to produce high reliability
- Performance testing often has lower reliability
- Whenever Possible, Increase the Number of Observations
- On Performance Grading, Ensure that Scorers are Well-Trained
- · Reliable Measurement Does Not Promise Validity



Ensuring Reliable Scoring

- · Clear, Well-Defined, and Defensible Criteria
- · Well-Trained Scorers
- Evidence of Inter-Rater Agreement
 - Document this
- Procedure of Resolving Score Divergence
 - A third scorer?
 - Beware of Rater Fatigue; Rater Differences
- Is Training of Scorers Professional Development?

	Validity	
 Valid Purpo Cons A o Other Co Pro Relia 	dity is Concerned with the Meaning of What dity is an Argument That a Test Really Mean worts to Measure struct Validity is the Most General Form construct is the psychological dimension measured or Forms of Validity ontent: Established partly by test specifications redictive: Does the test relate to future success? ability Places an Upper Bound on Validity you can't measure it accurately, it can't be	sures What it

-	
-	
_	
-	
-	
-	
-	
-	
-	
_	
-	
-	
-	