

Developing assessment tasks that lead to better instruction and learning

Mark Wilson

UC, Berkeley

Presented at the BEST PRACTICES IN STATE
ASSESSMENT Workshop 1, December 10-11, 2009
National Academy of Sciences, Washington DC

Outline

- Opening remarks
- Assessment Coherence
- Implementing this new logic for assessment:
The BEAR Assessment System
 - Principles I-IV
- Conclusion and Prospects



SYSTEMS FOR STATE SCIENCE ASSESSMENT

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES



What makes an assessment task “good”?

- I.e., What sort of assessment tasks lead to better instruction and learning?
- How does an assessment task affect instruction and learning?
 - By what it signifies
 - By the example of its content
 - By the information it generates

Relating summative assessments to formative assessments

- Assessment Coherence:
 - *the constructs being assessed at both levels need to be consistent,*
- although in practice, the constructs may be more differentiated at the classroom level and less differentiated at the large-scale level.
- More difficult to do this at the *classroom* level than at the large-scale level (!?!)

Threat Coherence

- where the large-scale assessment is used as a driving and constraining force, strait-jacketing the classroom instruction and curriculum to adhere to a specific state curriculum, and hence
 - “the tests used for accountability don’t have to be particularly good tests, they just have to serve their purpose—which is to ensure that teachers teach the standards as tested by the state!”

Threat Coherence

- Consequences
 - The classroom assessments are either
 - (a) parallel to the large scale assessments, or
 - (b) irrelevant for accountability
 - the large-scale test is used in a way that makes the classroom subject to the sanctions of an accountability system.

Threat Coherence

- Examples
 - NCLB tests in the USA
 - Key Stage Tests in the UK
 - US curricula: “a mile wide and an inch deep”

Threat Coherence

- Criticisms
 - (a) there are important aspects of school curricula that are not adequately assessable by multiple choice tests (Black and Wiliam, 2004; Thier, 2004);
 - (b) the sheer number of standards addressed in any given test (upwards of 100 in some states), ensure that the results cannot be used to gauge student accomplishment in a way that is useful in the classroom or the school for educational planning

Item coherence

- Most specific form of coherence
- For the assessments at the classroom and large-scale levels, the actual tests and items (or clones of the items) used at one level would also be used at the other.

item coherence

- Examples
 - “teach to the item” = take the items used in large-scale assessments and use them for practice in the classroom.
 - "benchmark testing" = slightly-altered versions of the large-scale items are used periodically to check up on students throughout the year.
- Criticism
 - Confuses what the item signifies with its surface content.

Systemic Coherence

- “the conceptual base or models of student learning underlying the various external and classroom assessments within a system should be compatible” (NRC, 2001, p. 255).
- Consider different *degrees of specificity* of coherence:
 - conceptual, information

Sub-types of *Systemic Coherence*

- *conceptual coherence*
- broadest sort of coherence
- the assessments at the classroom and large-scale levels share a common underlying framework

conceptual coherence

- Examples
 - "progress variables" designed in Australia by ACER, used in BEAR Assessment system
 - PISA "described variables"
 - the reduced set of prioritized standards as envisaged under the CISA requirements
[CISA=Commission on Instructionally Supportive Assessment (Popham, 2003)]

conceptual coherence

- Criticism:
 - coherence at the conceptual level or beyond requires a shared curriculum, and this is seen as being much more constraining than current conceptions of test alignment would indicate Shepard (2004)

Sub-types of *Systemic Coherence*

- *information coherence*
- the assessments at the classroom and large-scale levels
 - (a) shared a common framework, and
 - (b) shared information between the classroom and the large-scale levels, but
 - (c) did not necessarily use the same tests or results at the two levels

information coherence

- Examples
 - Statistically-moderated assessment systems
 - “work-sampling” systems
 - Wilson’s “*Community of Judgment*” based on the BEAR Assessment System...
- Advantage
 - Combines the effect of the content of the item with the effect of the information it generates.

*Towards Coherence
Between Classroom
Assessment
and
Accountability*

MARK WILSON

Editor

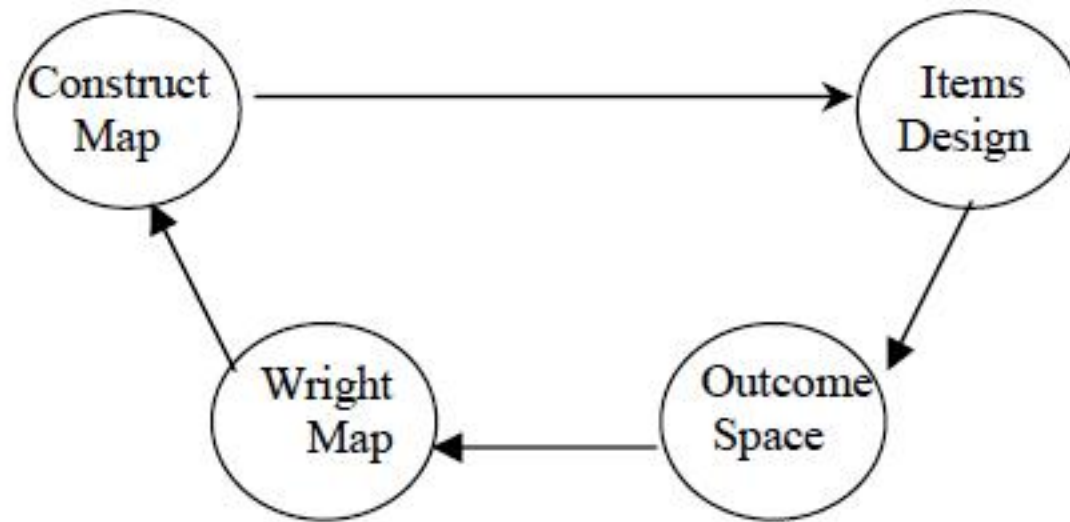
the 103rd Yearbook
of the National Society for the Study of Education
Part 2

The BEAR Assessment System

4 principles: 4 building blocks

Principle 1:
Developmental Perspective

Principle 2:
Match Between Instruction and
Assessment



Principle 4: Evidence of High
Quality

Principle 3:
Management by Teachers

Principle 1: Developmental Perspective

Building Block 1: Construct Map

- Developmental perspective
 - assessment system should be based on a developmental perspective of student learning
- Progress variable
 - Visual metaphor for
 - how the students develop and
 - how we think about how their item responses might change

Conceptions of Statistics (CoS)

Joint work with Rich Lehrer and Leona Schauble of Vanderbilt University

- *CoS4*. Discuss population parameters by referring to statistics and their sampling distributions.
- *CoS3*. Consider statistics as measure of qualities of a sample distribution.
- *CoS2*. Use statistics without relating them to qualities of the distribution.
- *CoS1*. Describe qualities of distribution informally.

Conceptions of Statistics (CoS)

CoS3. Consider statistics as measure of qualities of a sample distribution.	CoS3(f) Choose statistics by considering qualities of a particular sample.	- "It is better to calculate median because this data set has an extreme outlier. The outlier increases the mean a lot."
	CoS3(e) Attribute magnitude or location of a statistic to processes generating the sample.	- A student attributes a reduction in median deviation to a change in the tool used to measure an attribute.
	CoS3(d) Investigate the qualities of a statistic.	- "Nick's spreadness method is good because it increases when a data set is more spread-out."
	CoS3(c) Generalize the use of a statistic beyond its original context of application or invention.	- Students summarize different data sets by applying invented measures. - Students use average deviation from the median to explore the spreadness of the data.
	CoS3(b) Invent a sharable measurement process to quantify a quality of the sample	- "In order to find the best guess, I count from the lowest to the highest and from the highest to the lowest at the same time. If I have an odd total number of data, the point where the two counting method meet will be my best guess. If I have an even total numbers, the average of the two last numbers of my two counting method will be the best guess."
	CoS3(a) Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that other may not share.	- In order to find the best guess, I first looked at which number has more than others and I got 152 and 158 both repeated twice. I picked 158 because it looks more reasonable to me."

Principle 2: Match between curriculum and assessment

Building Block 2: Items design

- Instruction & assessment match
 - there must be a match between what is taught and what is assessed
- Items design
 - a set of principles that allows one to observe the students under a set of standard conditions that span the intended range of the item contexts

An Example CoS Item

Kayla's project

Kayla completes **four projects** for her social studies class. Each is worth **20 points**.

Kayla's Projects – Points Earned

Project 1	16 points
Project 2	18 points
Project 3	15 points
Project 4	???

The mean score Kayla received for **all four** projects was **17**.

Use this information to find the **number of points** Kayla received on **Project 4**. Show your work.

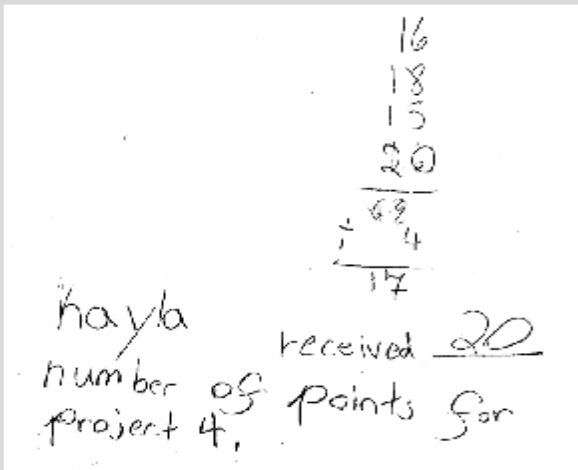
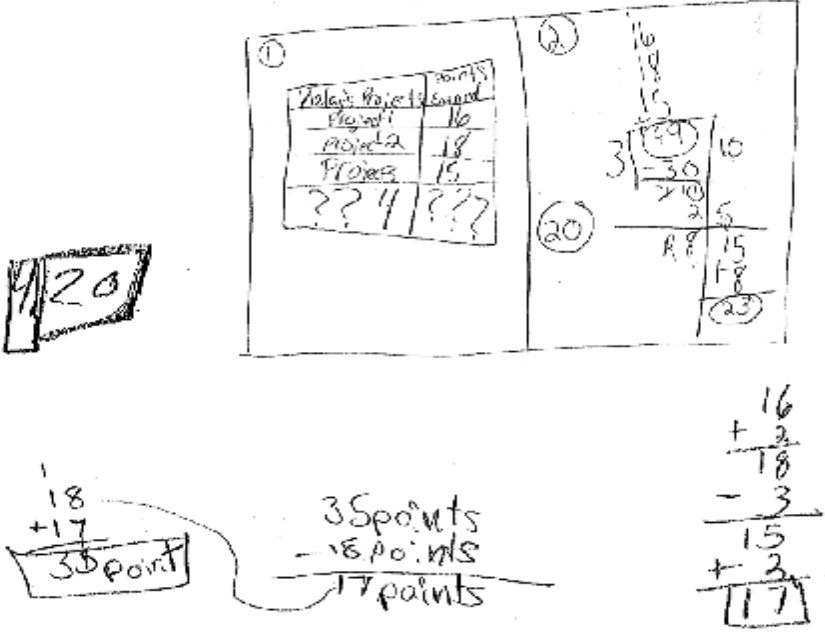
Principle 3: Interpretable by teachers

Building Block 3: Outcome space

- Management by teachers
 - that teachers must be the managers of the system, and hence must have the tools to use it efficiently and use the assessment data effectively and appropriately
- Outcome space
 - Categories of student responses must make sense to teachers

Kayla's Project Exemplar: Conceptions of Statistics (CoS)

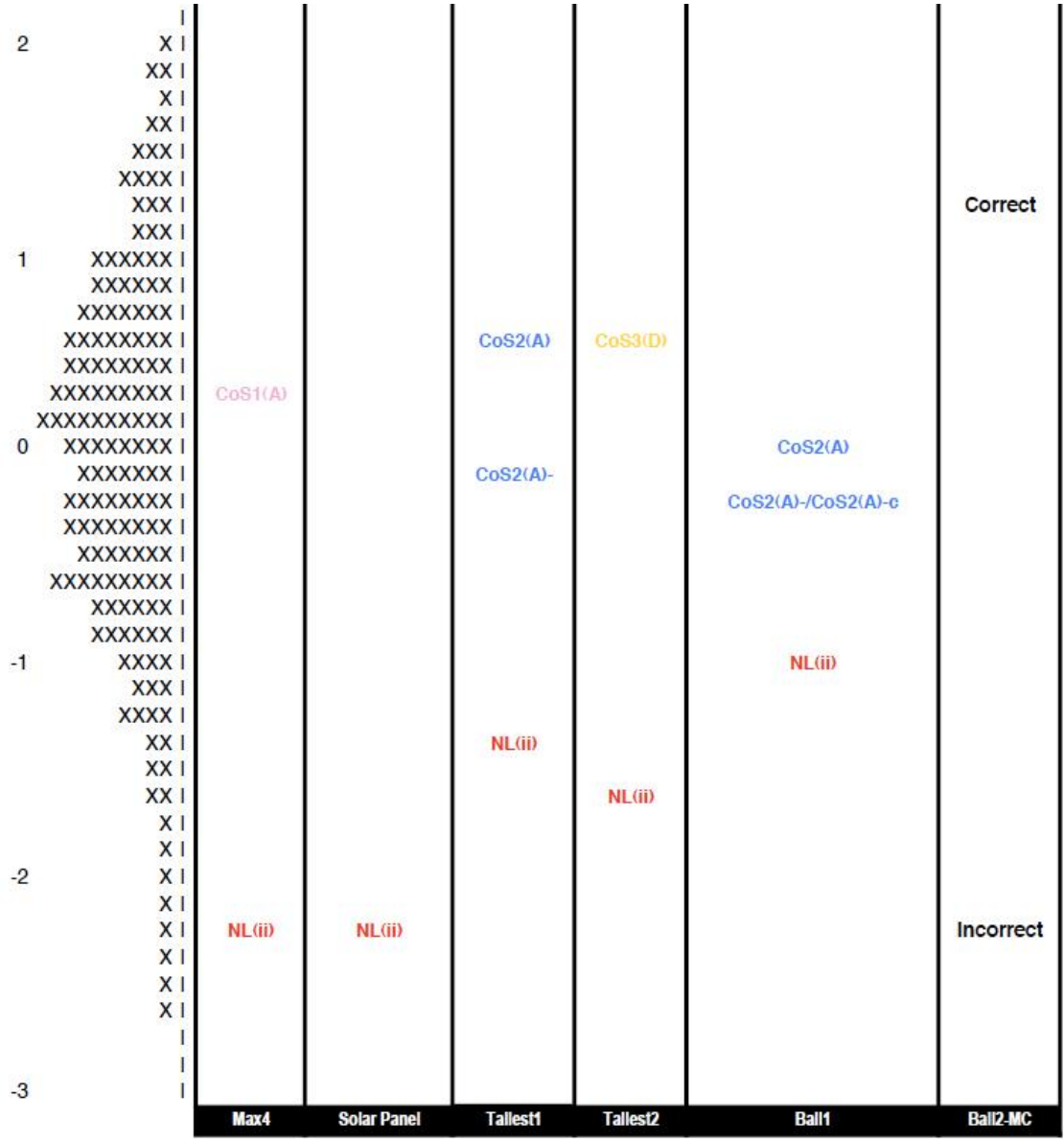
Levels	Response Exemplars	Example of Student Response
<p>CoS3 D</p>	<p>Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components.</p>	<p>"The differences between the mean and each score are -1, 1, -2, so the last difference must be 2 and the score must be 19."*</p>
<p>CoS2 A</p>	<p>Calculate statistics indicating central tendency correctly. Guess and check strategies are acceptable.</p>	<div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> $\begin{array}{r} 16 \\ 18 \\ +15 \\ \hline 49 \end{array}$ </div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> $\begin{array}{r} 24 \\ \times 1.7 \\ \hline 68 \end{array}$ </div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> $\begin{array}{r} 68 \\ -49 \\ \hline 19 \end{array}$ </div> </div> <div style="border: 1px solid black; padding: 10px;"> <p>She received 19 points.</p> $\begin{array}{r} 16 \\ +18 \\ 15 \\ +19 \\ \hline 68 \end{array}$ $4 \frac{12}{68}$ </div>

<p>CoS2 A -</p>	<p>Students write correct equation to calculate the missing measure.</p> <p>Students provide incorrect answer, but their work shows that they know the procedure.</p>	
<p>NL(ii)</p>	<p>Students do some idiosyncratic calculation with the measures. Answer is incorrect with no work shown but within a relevant range of answers.</p>	

Principle 4: Evidence of quality

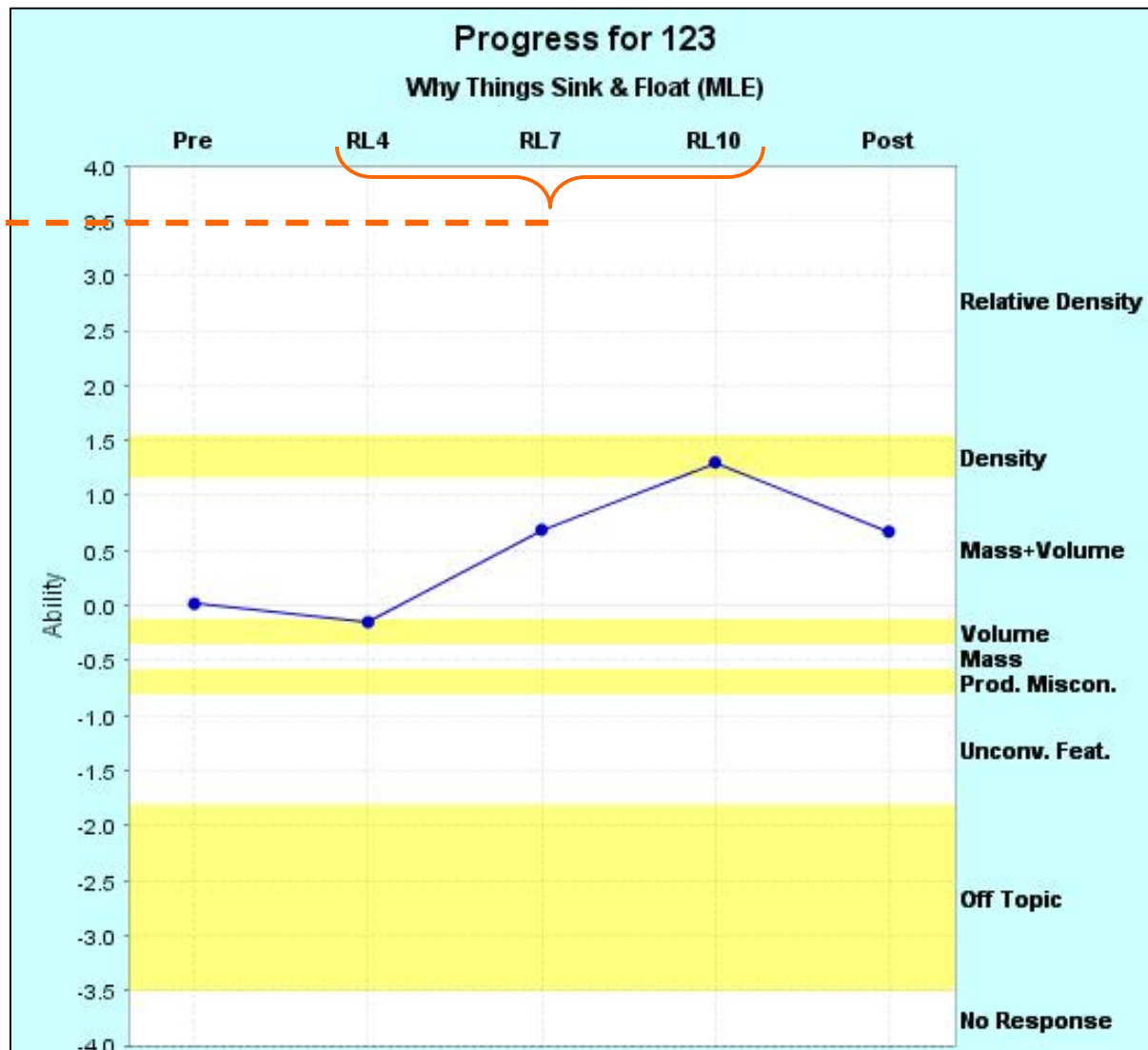
Building Block 4: Measurement model

- Evidence of quality
 - reliability and validity evidence, evidence for fairness
- Measurement model
 - multidimensional item response models, to provide links over time both longitudinally within cohorts and across cohorts



Evaluate a student's locations over time

Embedded Assessments



Set standards

Scale	Multiple Choice		WR 1		WR 2	
		P		P		P
620						
610						
600						
590						
580						
570	37	.30			2.3	.26
560	15	.34				
550						
540	28 39	.38				
530	27	.41				
520	19 38	.45				
510						
500	34 43 45 48	.50	1.3	.40		
490	17 18 20 40 50	.53				
480	4 31	.56				
470	11 32 33 44 47	.59				
460	5 9 12 46	.61				
450	3 6 7 10 16 29	.64				
440	36	.67				
430	8 14 22 23 26 35	.69				
420	13 24 25	.71				
410	41 42	.73				
400	1 21 30 49	.76				
390						
380					2.2	.56
370	2	.82				
360			1.2	.40		
350						

Second Example:
German Mathematical Literacy Test

Construct Map:

PISA Levels of Mathematical Literacy

Level	Description
VI	At Level VI students can conceptualize, generalize, and utilize information based on their investigations and modelling of complex problem situations. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply their insight and understandings along with a mastery of symbolic and formal mathematical operations and relationships to develop new approaches and strategies for attacking novel situations. Students at this level can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situations.
V	At Level V students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterizations, and insight pertaining to these situations. They can reflect on their actions and formulate and communicate their interpretations and reasoning.
IV	At Level IV students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilize well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments, and actions.
III	At Level III students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.
II	At Level II students can interpret and recognize situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions. They are capable of direct reasoning and making literal interpretations of the results.
I	At Level I students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

Items Design: German Mathematical Literacy Test

- Topic Areas: *Arithmetic, Algebra, and Geometry*
- Modelling Types: *Technical Processing*
 - requires students to carry out operations that have been rehearsed such as computing numerical results using standard procedures

Items Design: German Mathematical Literacy Test

- Modelling Types: *Numerical Modelling*
 - requires the students to construct solutions for problems with given numbers in one or more steps
- Modelling Types: *Abstract Modelling*
 - requires students to formulate rules in a more general way, for example by giving an equation or by describing a general solution in some way

An Abstract Modeling Item in Arithmetic

Difference

Put the digits 3 , 6 , 1 , 9 , 4 , 7 in the boxes so that the difference between the two three-digit numbers is maximized.

(Each digit may be used only once.)

1. number:

2. number:

Scoring Guide: The Abstract Modeling Item in Arithmetic

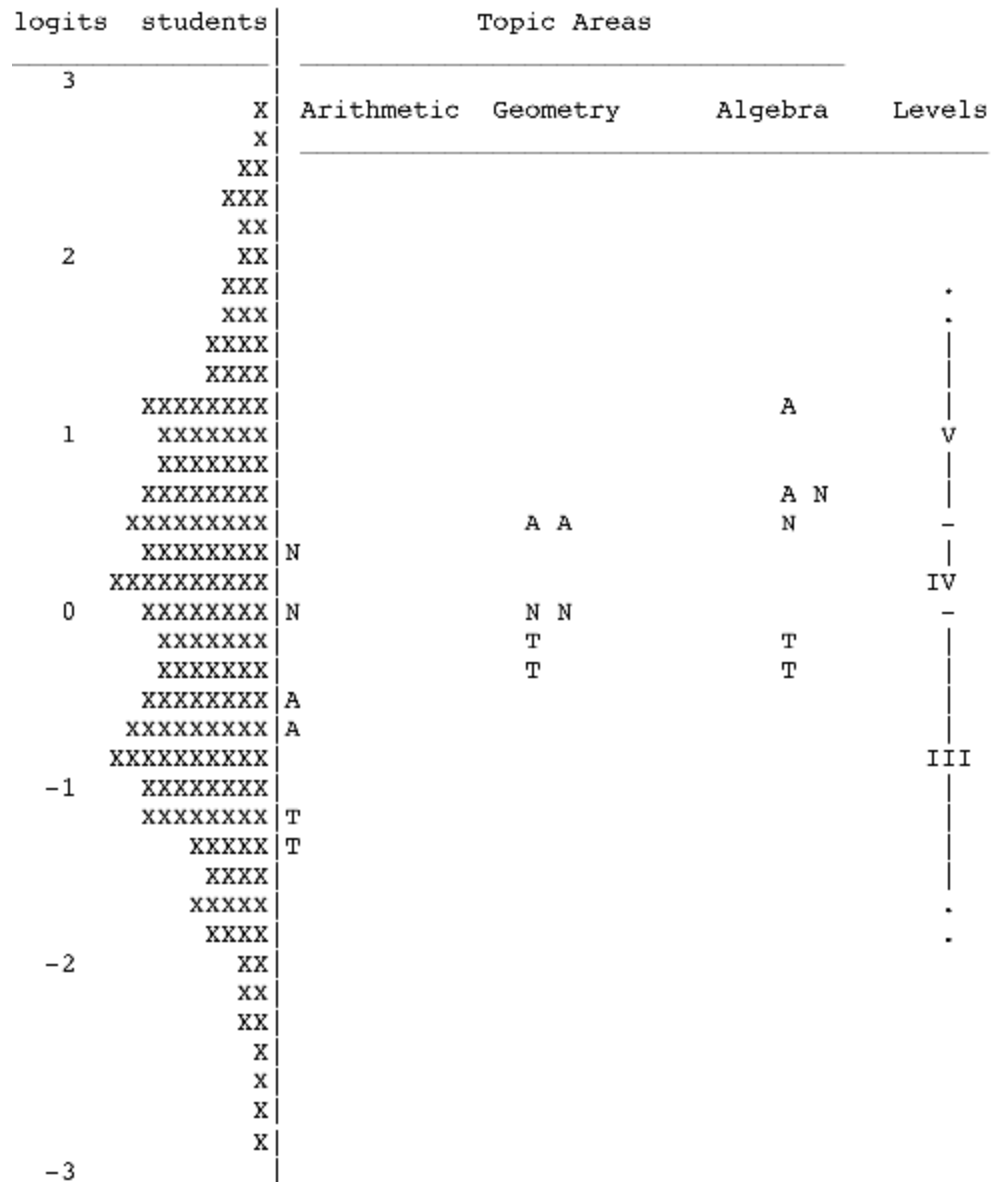
Scoring Guide:

Score 2: Number 1 = 976
 Number 2 = 134

Score 1: Uses all 6 digits and Number 1 > Number 2

Score 0: Anything Else

Wright Map: German Mathematical Literacy Test



Conclusion:

The wish of every large-scale testing program

- *To have the results of the large-scale tests be useful “diagnostically” to teachers in the classroom*
- Asked for by State Testing Directors, promised by testing companies

Suggested response

- both the large-scale and the classroom assessments must be constructed to be *coherent* in the information coherence sense.
- BEAR Assessment System is described as a process that can establish the coherence necessary to allow the micro and the macro to function coherently together.

Challenges

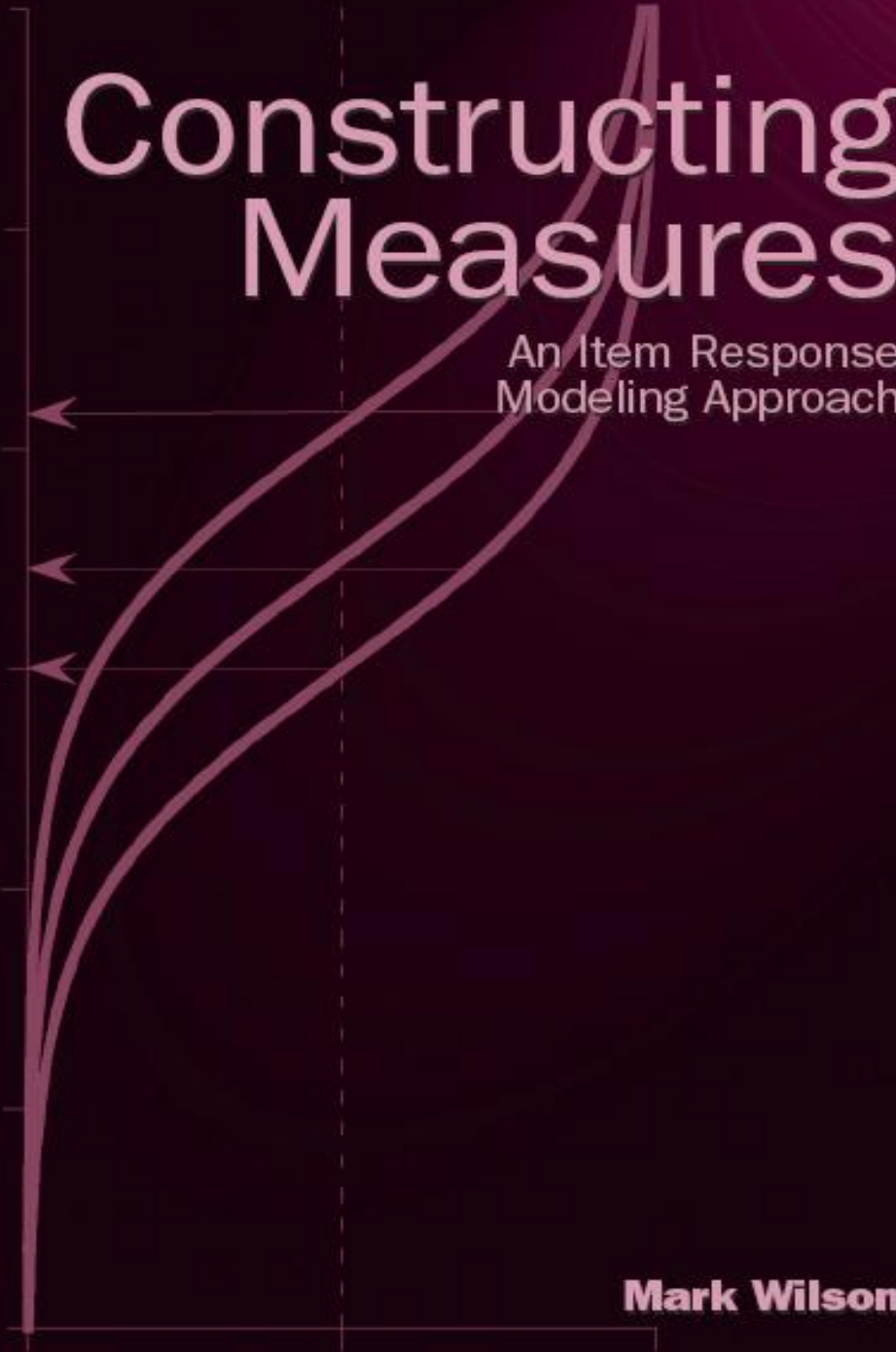
- Implementation of a BEAR Assessment System is not a minor matter
- requires a deeper analysis of the relationship between student learning and the curriculum and instructional practices than is commonly the case in assessment development
- requires a readiness to revise curriculum (i.e., "standards") based on empirical evidence available from the results of the assessments

Examples

- Implemented:
 - Living By Chemistry (high school)
 - SEPUP (middle school)
 - FOSS (elementary)
- Under development:
 - Carbon Cycle
 - Statistical Modeling
 - Reading Comprehension for “Striving Readers”
 - Evolution

Constructing Measures

An Item Response Modeling Approach



Mark Wilson