WERA/OSPI State Assessment Conference
December 2009

Tricks of the Trade: Validity of Standards-Based Elementary Report Card Grades

Jack B. Monpas-Huber, Ph.D.
Director of Assessment & Student Information

SHORELINE PUBLIC SCHOOLS

jack.monpas.huber@shorelineschools.org
(206) 368-4774 Office
(206) 947-9926 Cell

Purposes of this Presentation

1. To offer validity as a lens for looking at standards-based report card grades

2. To offer a few examples of preliminary validity investigations of report card grades

3. To stimulate discussion about report card grades and/or validity work

What is Validity?

Validity

Question:
What does validity mean to you? How do you judge the validity of assessment information?

One definition from the assessment/measurement literature:

> *"An evaluative judgment of the extent to which theoretical rationales and empirical evidence support the adequacy and appropriateness of inferences and uses of test scores and other modes of assessment."*
>
> (Messick, 1989)

Such as report card grades?

Types of Validity and Validity Evidence

	Theoretical Rationale	*Empirical Evidence*
Type of validity	*Content validity*	*Criterion-related (predictive) validity* *Construct validity*
Primary concerns / questions	*Content of the item/test (in relation to domain of learning)* *Does the item/test measure more/less/something different than what we expect examinees to have learned?*	*Empirical integrity of the test* *Are the scores reliable?* *Do the scores predict some outcome of interest?*
Types of evidence	*Professional judgment of experts* *i.e., content experts, teachers* *Documentation* *Alignment documents* *Curriculum maps* *Technical documents*	*Statistical evidence* *Item analysis* *Inter-item correlations* *Inter-rater reliability* *Correlations w/ other scores* *Factor analysis*

Our Content Validity Argument

2008-09 Fourth Grade Math Report Card Template

GRADING KEY FOR ACADEMIC AREAS

4 **Area of excellence**
The student demonstrates **superior performance** and skills appropriate to content and grade level. The student consistently goes beyond requirements in subject areas.

3 **Area of competence**
The student demonstrates **solid performance** appropriate to content and grade level. The student applies skills in a variety of situations.

2 **Area of development**
The student shows **partial accomplishment** of grade level knowledge and skill in specific situations or with support. The student is showing progress over time.

1 **Area of concern**
The student demonstrates **little or no progress** or achievement at grade level. The student may be working on materials below grade level.

* Indicates not assessed at this time

Similar 4-point ordinal scale for reporting performance as state assessment (at other grades)

MATH	Support Services	F	W	S
Number sense				
Basic facts				
Multiplication concepts				
Fractions, Decimals and Mixed numbers				
Measurement: Area, perimeter, and time				
Data Analysis and Probability				
Algebraic sense				
Reasoning, problem solving, and communication				
Effort				

Some intention of aligning our own content area expectations to what expected statewide…at least at face value

Validity of Report Card Grades: Empirical Evidence

Three empirical investigations

Descriptive Results

What have we reported this year?

What was the distribution of grades?

How many students got 1s? 2s? 3s? 4s?

Validity Study 1: Reliability of Elementary Math Grades

Validity Study 2: Elementary Grades and WASL Performance

Descriptive Results: What We Reported Last Year

Fall 2007

Skill	N	Missing	1	2	3	4	% Competent
Number Sense	505	7.9	4.0	28.7	53.5	5.9	59.4
Basic Facts	505	17.6	7.1	29.5	30.9	14.9	45.8
Computation	505	7.5	5.1	25.0	57.2	5.1	62.3
Measurement	505	30.3	2.6	24.2	40.0	3.0	43.0
Geometric Sense	505	19.8	2.0	23.2	50.3	4.8	55.1
Probability and Statistics	505	81.6	4.2	6.3	7.3	0.6	7.9
Algebraic Sense	505	77.4	3.6	7.1	9.7	2.2	11.9
Reasoning/Problem Solving	505	12.5	4.6	33.7	44.8	4.6	49.4
Communicates Mathematically	505	46.3	3.4	18.6	28.9	2.8	31.7
Effort	505	8.1	0.2	12.7	63.2	15.8	79.0

Winter 2008

Skill	N	Missing	1	2	3	4	% Competent
Number Sense	505	8.5	2.6	27.3	52.9	8.7	61.6
Basic Facts	505	8.1	8.1	26.5	32.7	24.6	57.3
Computation	505	8.7	3.0	27.9	50.5	9.9	60.4
Measurement	505	13.5	2.4	27.9	49.7	6.5	56.2
Geometric Sense	505	42.6		13.9	36.6	6.9	6.9
Probability and Statistics	505	61.8	0.2	5.0	29.1	4.0	33.1
Algebraic Sense	505	35.6	2.4	21.6	32.3	8.1	40.4
Reasoning/Problem Solving	505	9.3	2.4	34.1	45.7	8.5	54.2
Communicates Mathematically	505	16.4	1.4	26.3	48.5	7.3	55.8
Effort	505	8.3	0.4	8.9	56.8	25.5	82.3

Spring 2008

Skill	N	Missing	1	2	3	4	% Competent
Number Sense	505	5.3	2.2	25.0	54.7	12.9	67.6
Basic Facts	505	4.0	4.2	22.8	37.6	31.5	69.1
Computation	505	5.5	2.0	28.3	49.7	14.5	64.2
Measurement	505	10.1	1.8	29.5	48.7	9.9	58.6
Geometric Sense	505	15.2	2.6	21.8	51.5	8.9	8.9
Probability and Statistics	505	14.5	1.4	21.0	55.8	7.3	63.1
Algebraic Sense	505	16.0	3.2	24.6	45.3	10.9	56.2
Reasoning/Problem Solving	505	11.9	1.8	25.9	50.7	9.7	60.4
Communicates Mathematically	505	27.9	2.2	24.6	38.8	6.5	45.3
Effort	505	5.1	0.6	9.7	54.9	29.7	84.6

What do you think?

Questions to answer in your group:

What do you see?

What did you expect?

What questions do these data answer for you?

What new questions does this raise?

Validity Study 1: Reliability of Report Card Grades

Reliability

What is reliability?
To what extent does this instrument (report card grade) capture consistent information across observations (students)?

***Example**s:*
John got a 4 in your class. Would he get a 4 in any other 4th grade class?

Seth got a 3 at Parkwood; Siobhan got a 3 at Brookside. How comparable are the achievements of these two students?

Why this matters
Reporting inconsistent information about student achievement creates confusion and inefficiency and is inequitable

Validity Study 1: Reliability of Report Card Grades

How do we assess reliability?

Test-retest reliability
Administer assessment, then administer again. A strong correlation is evidence that the assessment is measuring the same trait on both occasions.

Inter-rater agreement
At least two raters score the same piece of student work using a rubric. A strong correlation is evidence that the rubric (not the rater) is guiding the scoring of student work.

Internal consistency reliability
Administer a multi-item assessment once. Strong correlations among items are evidence that the assessment is measuring the same trait across observations.

Validity Study 1: Reliability of Report Card Grades

How do we assess reliability?

Test-retest reliability
Administer assessment, then administer again. A strong correlation is evidence that the assessment is measuring the same trait on both occasions.

Inter-rater agreement
At least two raters score the same piece of student work using a rubric. A strong correlation is evidence that the rubric (not the rater) is guiding the scoring of student work.

Internal consistency reliability
Administer a multi-item assessment once. Strong correlations among items are evidence that the assessment is measuring the same trait across observations.

Validity Study 1:
Reliability Evidence for Fall Math Grades

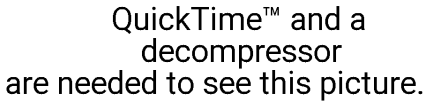
QuickTime™ and a
decompressor
are needed to see this picture.

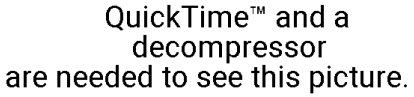
QuickTime™ and a
decompressor
are needed to see this picture.

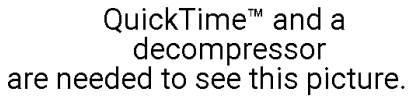
QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

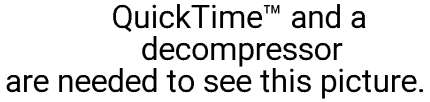
Validity Study 1:
Reliability Evidence for Winter Math Grades

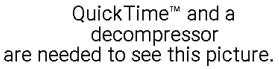


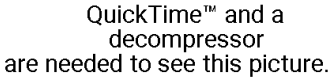




Validity Study 1:
Reliability Evidence for Spring Math Grades







Validity Study 1:
Reliability Evidence for Report Card Grades

Summary of the internal consistency data

- Students are receiving very similar grades across the traits
- No one skill stands out as particularly problematic

What remains unclear

Inter-"grader" reliability. What would be the inter-teacher agreement of the same students?

What do you think?

Questions to answer in your group:

What do you see?

What did you expect?

What questions do these data answer for you?

What new questions does this raise?

Validity Study 2: Grades and WASL Performance

How and why should grades relate <u>empirically</u> to WASL scores?

Why we *should* expect <u>some kind</u> of correlational type relationship:

1. WASL sorts students on the basis of achievement of skills assessed by the state standards.
2. Grades also distinguish which students have mastered skills from those who have not.
3. We should expect these two sorts to overlap considerably.

	Report Card Grades			
WASL Performance Level	1	2	3	4
1	Most	Some		
2	Some	Most	Some	
3		Some	Most	Some
4			Some	Most

Validity Study 2: Grades and WASL Performance

But why we should probably not expect a perfect relationship:

1. WASL samples from the population of standards BUT each trimester covers only a sample of the standards.
2. Different assessment formats.
3. Unreliability of either measure.
4. Some kids will always surprise us.

	Report Card Grades			
WASL Performance Level	1	2	3	4
1	Most	Some		
2	Some	Most	Some	
3		Some	Most	Some
4			Some	Most

Validity Study 2: Grades and WASL Performance

Fourth Grade -- 2008-09

	Modal December Trimester Grade				
WASL Level	1	2	3	4	Total
1	15 71.4%	36 24.2%	7 1.7%		58 9.5%
2	4 19.0%	37 24.8%	21 5.1%		62 10.2%
3	2 9.5%	59 39.6%	188 46.1%	4 13.3%	253 41.6%
4		17 11.4%	192 47.1%	26 86.7%	235 38.7%
Total	21 100.0%	149 100.0%	408 100.0%	30 100.0%	608 100.0%

Validity Study 2: Grades and WASL Performance

Fourth Grade -- 2008-09

	Modal March Trimester Grade				
WASL Level	1	2	3	4	Total
1	15 62.5%	35 25.7%	10 2.6%		60 9.9%
2	6 25.0%	35 25.7%	21 5.5%	1 1.5%	63 10.3%
3	3 12.5%	51 37.5%	190 49.6%	10 15.2%	254 41.7%
4		15 11.0%	162 42.3%	55 83.3%	232 38.1%
Total	24 100.0%	136 100.0%	383 100.0%	66 100.0%	609 100.0%

Validity Study 2: Grades and WASL Performance

Fourth Grade -- 2008-09

	Modal June Trimester Grade				
WASL Level	1	2	3	4	Total
1	14 56.0%	38 29.7%	11 2.9%		63 10.2%
2	8 32.0%	37 28.9%	18 4.7%	1 1.3%	64 10.4%
3	3 12.0%	42 32.8%	199 51.7%	13 16.3%	257 41.6%
4		11 8.6%	157 40.8%	66 82.5%	234 37.9%
Total	25 100.0%	128 100.0%	385 100.0%	80 100.0%	618 100.0%

Next Steps

We could always benefit from more common understanding of proficiency on the state assessment, via:

- Study of MSP Test & Item Specifications
- Participation in state professional development in assessment
- Rigorous review of prior year state assessment results

We need more shared operational definitions of:

- At trimester, what does a 3 or 4 in a skill area really mean?
- What does 2 or 3 mean in concrete terms?
- What do these look like in child- parent-friendly terms?

What arguments do we make about (or on the basis of) report card grades?

What is the evidence to support those arguments?