Washington Educational Research Association (WERA) State Assessment Conference, December 2012

The Course Challenge Test: Catalyst for Change?

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

Marcia Garrett, NBCT
District K-12 Science Specialist

SHORELINE PUBLIC SCHOOLS

Organization of this session

<u>Key Questions</u>
What happens when we let students challenge courses?
What evidence suggests the challenge will be successful?
Can this policy improve instruction and increase student achievement?

Background
Conceptualizing the test
Content review
Form assembly
Piloting
Results
Standard setting
Outcomes

Background

2008 - Present
Shoreline increasing committed to professional learning communities (PLC), Response to Intervention (RtI), and systemic **alignment**
Program Alignment and Coherence Team (PACT) organized to bring alignment and coherence to course sequences, instructional programs

2010-11
PACT reviewed secondary science courses
Decided to enable more upper-level science courses to able students

2011-12
Charged with development of Integrated Physical Science (IPS) Course
Challenge Test to be administered to outgoing 8th grade students

Purposes of the IPS Course Challenge Test

- To allow 8th grade students an opportunity to challenge IPS and enroll in Biology in 9th grade

- To lay foundation for a common district end-of-course IPS assessment for 2012-13

The concept of a course challenge test

In your group (2-3 people)

How do you go about designing the test?

What content and scientific practices should be on the test? How should lab skills be assessed?

How do you avail this test to students?

What counts as "passing"?

What counts as evidence of validity?

Validity of the course challenge

What counts as evidence of validity for a test? For a successful course challenge?

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests."
AERA, APA, & NCME (1999)

Validity evidence for course challenge test

Content validity

Science teachers and Specialist declare, based on content expertise, that the Challenge items / test:

- are scientifically sound
- measure appropriate science content (IPS Power Standards, course content)
- validly measure the expectations of the IPS course

Empirical "data" validity

- Challenge items/test predict successful performance in IPS course. Challenge items/test are correlated with successful performance in IPS based on current IPS students.
- Challenge items/test are of good quality (moderate difficulty and discrimination, evidence of reliability)

Project plan for the year

Obtain or develop item pool
Content review and form assembly
Pilot fall content
Revise fall items based on pilot results
Pilot spring content
Revise spring items based on pilot results
Administer 8th grade test in June
Standard setting

Shared in October with teachers, administer via Google Docs
Open for input

Developing the item pool

Where would you get items?

We searched for a model
Science TOSA network and online, found one
Reviewed the model for match to the standards

Critical decision: content come from course syllabi, or state standards for the units covered in the course?

Collaborative content review

1- Marcia examined items on the test, compared to state standards, found gaps

2- Got teacher volunteers (1 per school)
Task: Teachers reviewed the items on the test against their course objectives (power standards) and they found gaps.

3- Identified different gaps due to district power standards not covering all state standards intended for this grade band in physical, earth & space science

4- Got more time for teachers to work on this -- editing, proofing, wrote more items to fill gaps (or used items from the curriculum materials)

Assembling the test forms

How large to make the test?
Simply based on standards and practicality.

How many items to adequately cover the standards?
We went through state standards and noted proportionality of each and attempted to reflect that in the design of the test.

What should guide construction of the test?

Missing question: Level of Cognitive Demand

Test map of 1st semester pilot test

IPS Content Topics	Number of items/points	% of items/points
Additional Forces	4	8%
Atomic Structure & Periodic Table	5	10%
Chemical Reactions (physical change, chem. change, bonding)	9	18%
Energy	6	12%
Kinematics	9	18%
Laws of Motion	8	16%
Scientific Method/Inquiry	7	14%
Solutions	2	4%
Total	50	100%

Piloting the items

Rationale for piloting

- Need empirical data as well as content review
- Plan to have current 9th grade students take portions of test: Can current students pass this test?
- Fall content
- Spring content
- Use to refine items for operational administration
- Provide data for IPS teachers

Critical decision: Scheduling and coordinating data collection... different 1st & 2nd Semesters

Piloting the items

Evidence of item quality
(Allen & Yen, 1979)

Difficulty
p-value: percent of students getting item correct
Should cluster toward "easy" range (above 70-100%) if measure content taught

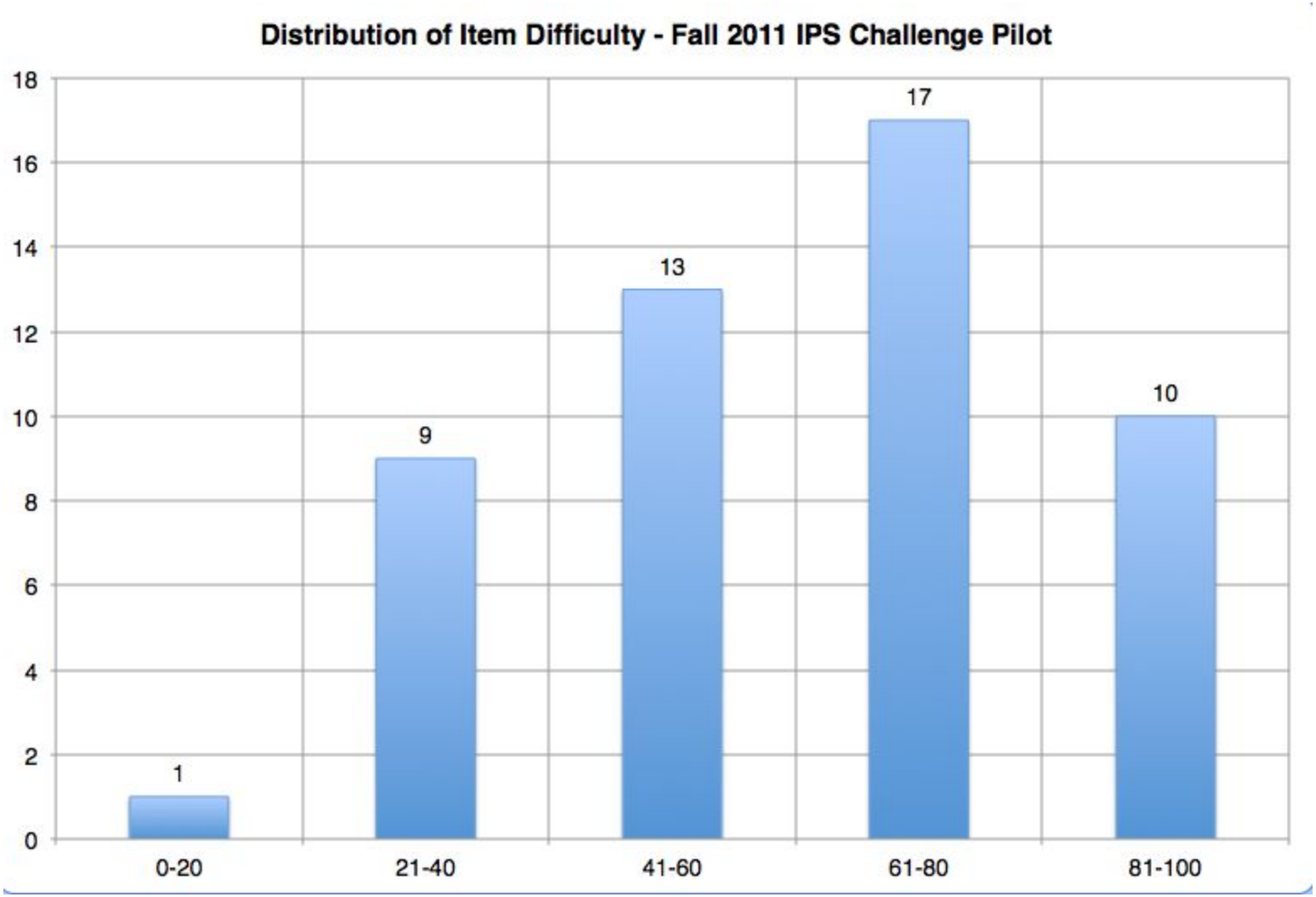
Discrimination
item-total correlation
item-course grade correlation
Typically want item-total correlations at least .3

Results of 1st semester pilot: Item analysis

A Google spreadsheet shared with teachers

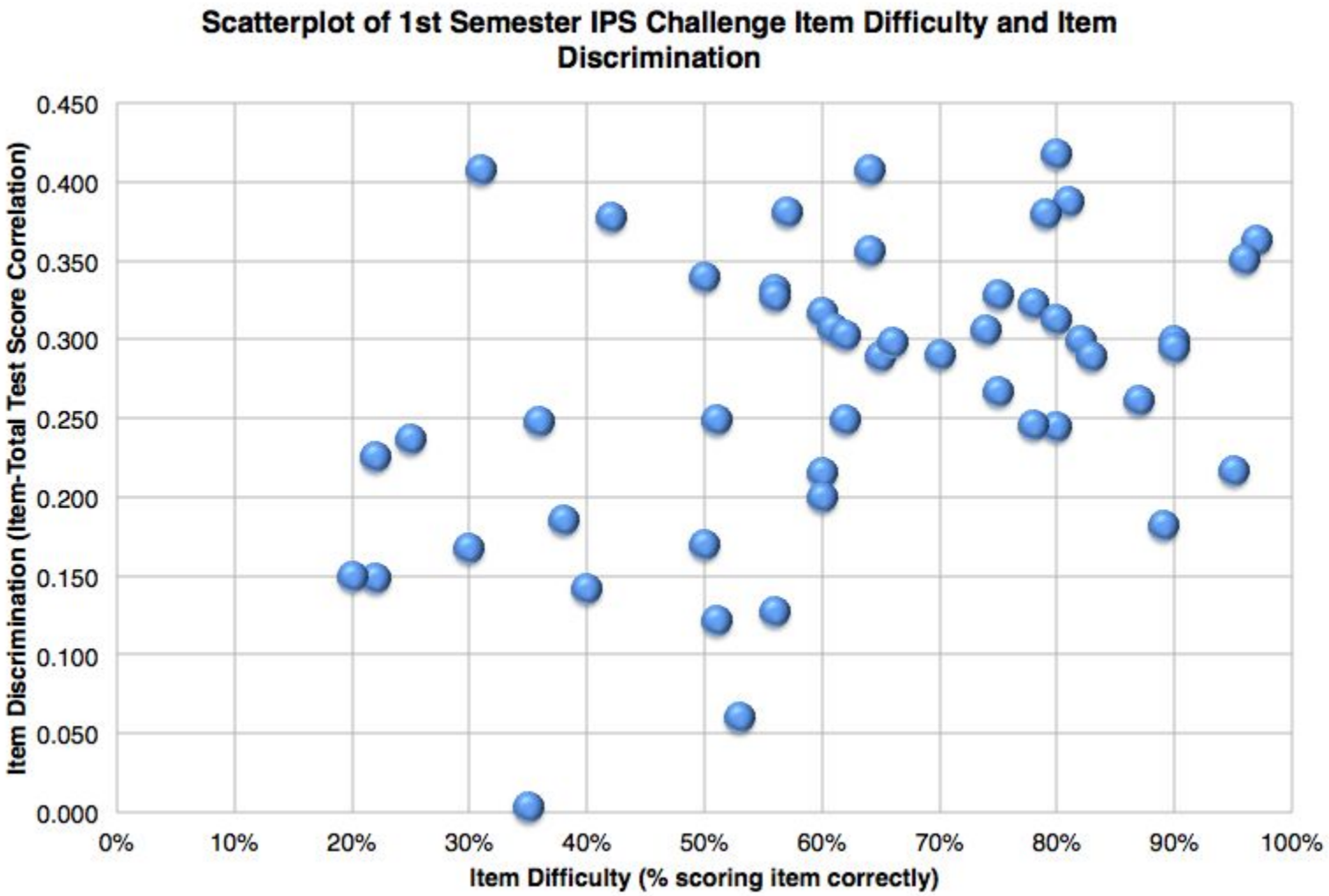
	A	B	C	D	E	F
1	**Comments**	**Item Number**	**Item Content**	**Item Difficulty (% scoring item correctly)**	**Item Discrimination (Item-Total Test Score Correlation)**	**Item Discrimination (Item-Course Grade Correlation)**
2		1	Scientific Method/Inquiry	75%	0.267	0.181
3		2	Scientific Method/Inquiry	82%	0.299	0.188
4		3	Scientific Method/Inquiry	81%	0.387	0.182
5		4	Scientific Method/Inquiry	79%	0.380	0.137
6		5	Scientific Method/Inquiry	56%	0.332	0.162
7		6	Scientific Method/Inquiry	65%	0.289	0.175
8	Prefer to dismiss. Basic skill students should know from lower grade levels	7	Scientific Method/Inquiry	80%	0.244	0.290
9		8	Kinematics	78%	0.323	0.306
10	Confusing; not properly taught because teachers disagreed on correct answer.	9	Kinematics	35%	0.003	0.015
11	Not part of standards	10	Kinematics	56%	0.127	-0.022
12		11	Kinematics	83%	0.289	0.060
13		12	Kinematics	57%	0.381	0.203
14		13	Kinematics	90%	0.299	0.117
15		14	Kinematics	87%	0.261	0.089
16		15	Kinematics	42%	0.377	0.201
17	Correct answer is not there. None of the answers have "direction"	16	Kinematics	22%	0.149	0.003
18		17	Laws of Motion	64%	0.356	0.033
19		18	Laws of Motion	78%	0.246	0.044
20		19	Laws of Motion	64%	0.407	0.186

Results of 1st semester pilot: Item difficulty

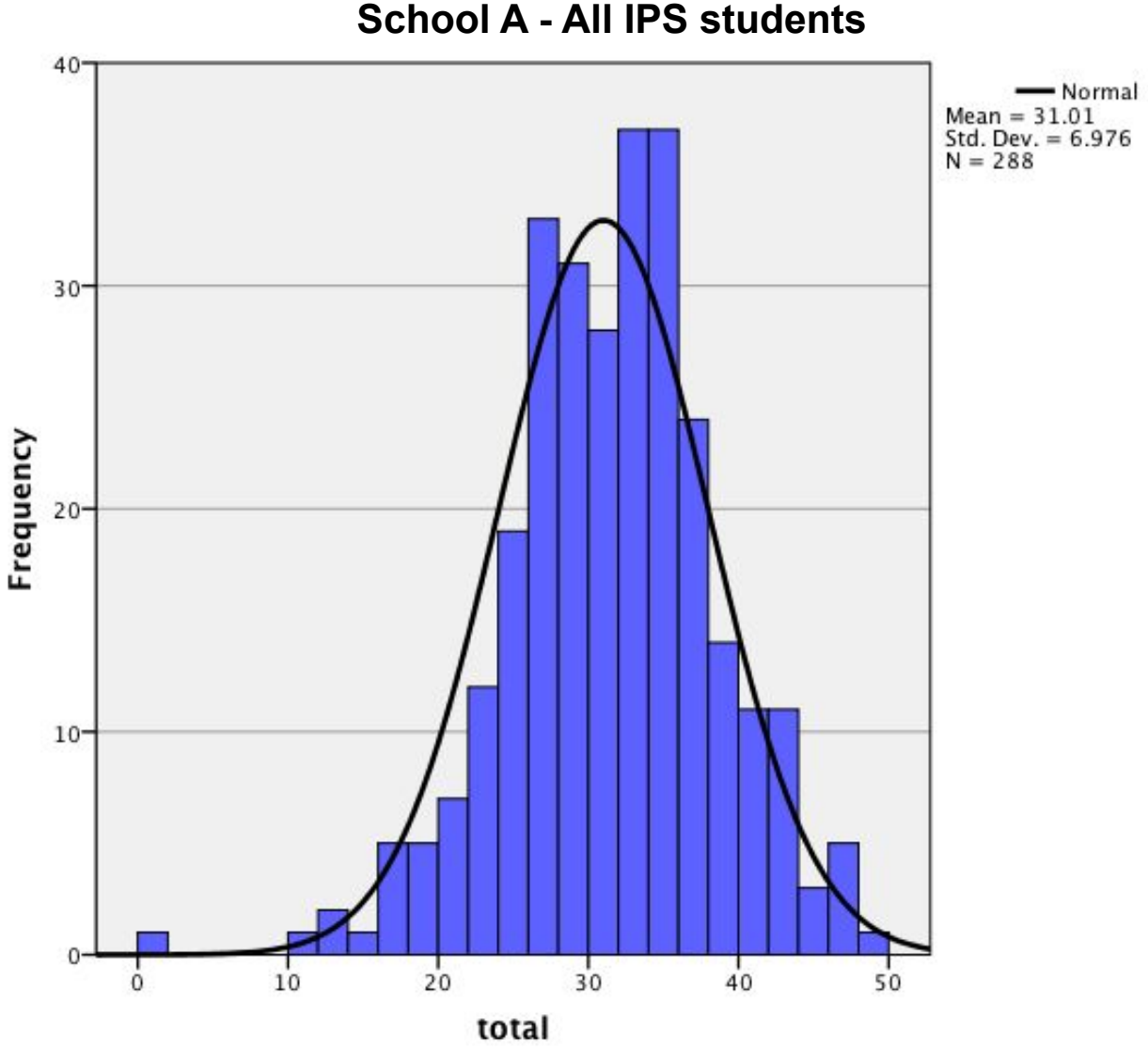


Percent of students scoring item correctly

Results of 1st semester pilot: Item difficulty and discrimination



Results of 1st semester pilot: Student scores

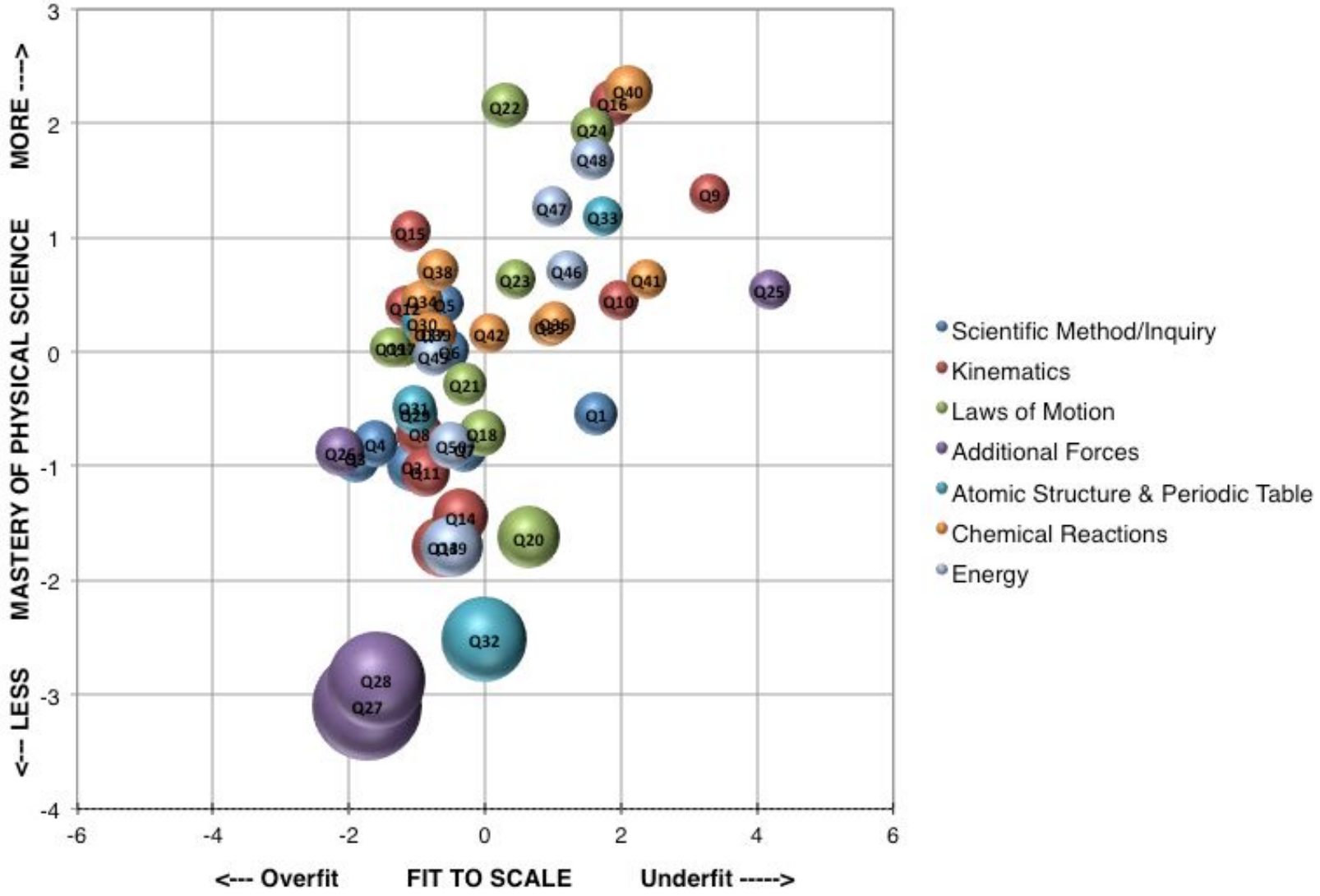


Results of pilot data: Summary statistics and reliability evidence

Instrument	Examinees	N students	N items	Mean	SD	Cronbachs' α	SEM
Fall pilot	School A All	288	50	31.01	6.98	0.82	2.97
Fall pilot	School B Honors	64	50	38.47	4.12	0.61	2.56
Fall pilot (embedded in final)	School B Regular Form A	60	36	25.10	5.52	0.82	2.37
Fall pilot (embedded in final)	School B Regular Form B	54	36	23.61	6.53	0.86	2.43
Spring pilot	All	499	30	17.02	4.35	0.66	2.55
Combined 9th grade anchor	All	433	80	48.20	10.40	0.87	3.75
Final 8th grade form	All	193	73	41.39	8.49	0.81	3.73

Looking forward: A pathway diagram of IPS test

Applying the Rasch Model, Bond & Fox, 2007



Teacher item review (1st semester)

Took item analysis data back to teachers

Revised 1st semester Challenge items based on those data

Began to write items for 2nd semester based on:

- Standards
- Model test we started with
- Test bank items from the new curriculum materials

Teacher item review (2nd semester)

IPS Challenge Test – Spring 2012 Pilot
IPS TEACHER ITEM REVIEW

Teacher Name:		Period:	

Dear IPS Teacher,
Board Policy now allows 8th grade students the ability to challenge IPS through examination. This exam will measure the Washington state standards for physical science, earth science, space science, etc. The items for this exam were thoroughly reviewed by IPS teacher representatives from each high school for accuracy, scientific soundness, and match to Washington State science standards. The purpose of this pilot testing is to gather data on item quality that will inform the operational form of the IPS Challenge Test and future development of an IPS end-of-course exam. As your students take this exam, please take a few moments to review the items and test *in reference to this class period* and provide your feedback in the spaces below. Thanks!

Item	Standard	How difficult will this be for your students? What percent will get it correct?	How much time did you spend on this concept/skill in this class? (0=none, 1=a little, 3=some, 4=thorough)	Comments about this item?
1	9-11 PS # D			
2	9-11 PS # D + E			

Standard setting

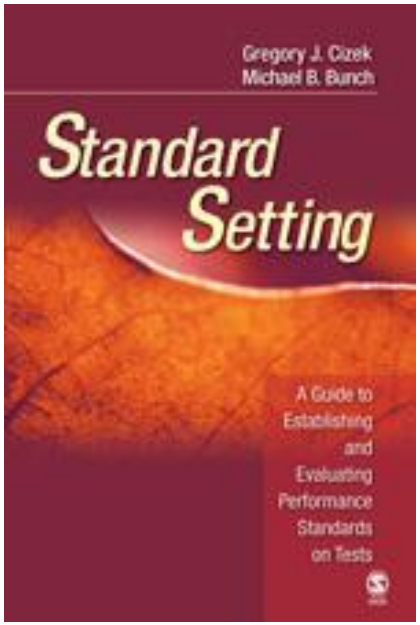
What counts as passing? How to set the cut score?

Could use the traditional 80%

Wanted to consult the literature (Cizek & Bunch, 2007)

Wanted to incorporate the data from current 9th graders

Could use an Angoff or Bookmark Method



Standard setting

<u>The Beuk Method</u> (Beuk, 1984)

Surveyed teachers and administrators:

1. What should be the passing *score* for this exam?-- that is, what *percent of the exam* should a student score in order to pass and successfully challenge IPS?
(Example: A score of at least 80% *on the exam* is needed to pass and thereby mount a credible challenge to IPS.)

2. What should be the pass *rate* for this exam?
(Example: Only 10% *of the examinees* should be able to pass it.)

Standard setting

The Beuk Method Results

1. What should be the passing *score* for this exam?-- that is, what *percent of the exam* should a student score in order to pass and successfully challenge IPS?

A handful of teachers responded. Average response was 80%

2. What should be the pass *rate* for this exam?

Nobody responded to this question.

Standard setting

<u>Final standard setting procedure</u>

Teachers want the cut score of 80%

How would 9th graders have performed on this test?
Combined fall and spring pilot data sets, performed Rasch calibration to get theta scores for 9th grade students. Found the theta score for students who got an A in IPS (Regular or Honors). That theta score landed on a raw score of approximately 70% on the test.

Decided on **75%**

Assessment of lab skills

How to assess lab skills?

Ideal would be lab practical assessment

Alternative: affidavit from 8th grade science teacher

Assessment of lab skills

8th grade Science:
Affidavit for Performance Skills, Safety & Responsibility for Science Investigations

Student Name:_____

For each item rate the student level of mastery on the scale given with
0 = not demonstrated to 4 = fully mastered.

Process Skills	Circle appropriate rating:				
Follows instructions and procedures	0	1	2	3	4
Uses tools to make accurate measurements	0	1	2	3	4
Carries out specific techniques with care	0	1	2	3	4
Uses lab materials safely	0	1	2	3	4
Collects and records data accurately	0	1	2	3	4
Skilled data organization(graphs, charts, etc.)	0	1	2	3	4

Work Habits	Circle appropriate rating:				
Self-motivated	0	1	2	3	4
Responsibly uses time & materials	0	1	2	3	4
Works collaboratively as part of a team	0	1	2	3	4

Teacher signature: _____

Outcomes: Where are they now?

Grade	Shorecrest	Shorewood	Total	Percent
A	4	3	7	58%
B	1		1	8%
C	2	1	3	25%
C-		1	1	8%
Total	7	5	12	100%

Outcomes

Discovered Systems & Applications strands were missing from IPS Power Standards; discovered strength of the Inquiry strand (Implications of adoption of NGSS)

Full-day professional development on Systems strand October 5

IPS teachers time to work on standards, have some baseline data, foundation for EOC assessment

Recent outcomes

Cross-department agreements about use of Challenge Test...

- Review test for content
- Look at data for instructional improvement
- Report strand scores to 8th grade students/parents in letter

...and common course assessment

- Look for Semester 2 units taught in common, assessments used, and design unit common assessment for PLC work..eventually Semester EOC

Questions raised

Should "skippers" be held to a different performance standard than students who take the course?

What is (and should be) the difference among:

- the course finals
- the Challenge test
- the common course assessment (EOC)

Is a written test enough?
Should there be a lab practical to assess skills needed for success in a Biology Lab course?

"We look at the present through a rear-view mirror. We march backwards into the future."

Marshall McLuhan

References

Allen, M. H., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

American Educational Research Association, American Psychological Assocation, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing.

Beuk, C. M. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21, 147-152.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human science*. Second Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards*. Thousand Oaks, CA: Sage Publications.

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Second edition. Thousand Oaks, CA: Sage Publications.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. Third Edition. McGraw-Hill.

Thank you!