

A New Challenge: Making Test Score Reports More Understandable and Useful

Ronald K. Hambleton
University of Massachusetts,
At Amherst

CCSSO Annual Meeting, Nashville,
June 18, 2007

Important Time in the Testing Field

- More testing today than ever, and results have become very important to many persons.
- Testing, generally, is a central component of educational reform in the US and other countries—the information from tests needs to be used, and correctly!

Introduction

1. Considerable investment of time and money has been made to address technical problems in testing programs--IRT modeling of data, test score equating, reliability estimation, DIF analyses, standard-setting, and validity studies.

Introduction

2. Surprisingly, given importance, test score reporting attracts very little research attention!

--Name one research study?

--Without clear and meaningful reporting of information, the other steps are of little value!

--On this topic, more than other technical problems, people think they are experts! ⁴

AERA, APA, NCME Test Standards: What do they say about score scales and reporting?

5.10. When test score information is released...those responsible should provide appropriate interpretations.

--information is needed about content coverage, meaning of scores, precision of scores, common misinterpretations, and address use.

AERA, APA, NCME Test Standards: What do they say about score scales and reporting?

13.14 ...Score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Major Problems in Score Reporting!

- Reporting scales are confusing to many persons (e.g., percent vs. percentile; IQ; SAT vs. ACT; NAEP, MCAS scores, etc.)
- Quantitative literacy among adults is not high in this country (three kinds of persons; half of adult pop can't read bus schedules; what's 20 million dollars for testing--1/3 of 1%)

Major Problems in Score Reporting!

- Body of evidence highlighting score reporting problems
 - Reporting scores without error bands
 - Too much meaningless score information on some reports
 - Not providing meaningful diagnostic information...

Goals of the Presentation

1. Describe some of the recent research on score reporting.
2. Identify several promising directions for score scales and score reporting

Example: Review of Student Reports from One State

- Lots of good information—background information (name, grade, form, etc.), performance classifications, scores, graphical presentations, percentile, instructional information.
- Basically, lots of information to work with in a revised and improved form.

Suggestions for the State to Consider in Revising the Reports

- No stated purpose, no advanced organizer, no cue about where to start reading an array of numbers.
- The four performance categories are **not** defined, even briefly.
- No error bands on any of the reported scores, or even a hint that errors of measurement are present.

Suggestions to Consider in Revising the Displays, cont.

- Font is too small in presenting the instructional needs information.
- Some instructional needs information, but not user-friendly—e.g., You need help in “extending meaning by drawing conclusions and using critical thinking to connect and synthesize information within and across text, ideas, and concepts.”

Suggestions to Consider in Revising the Displays, cont.

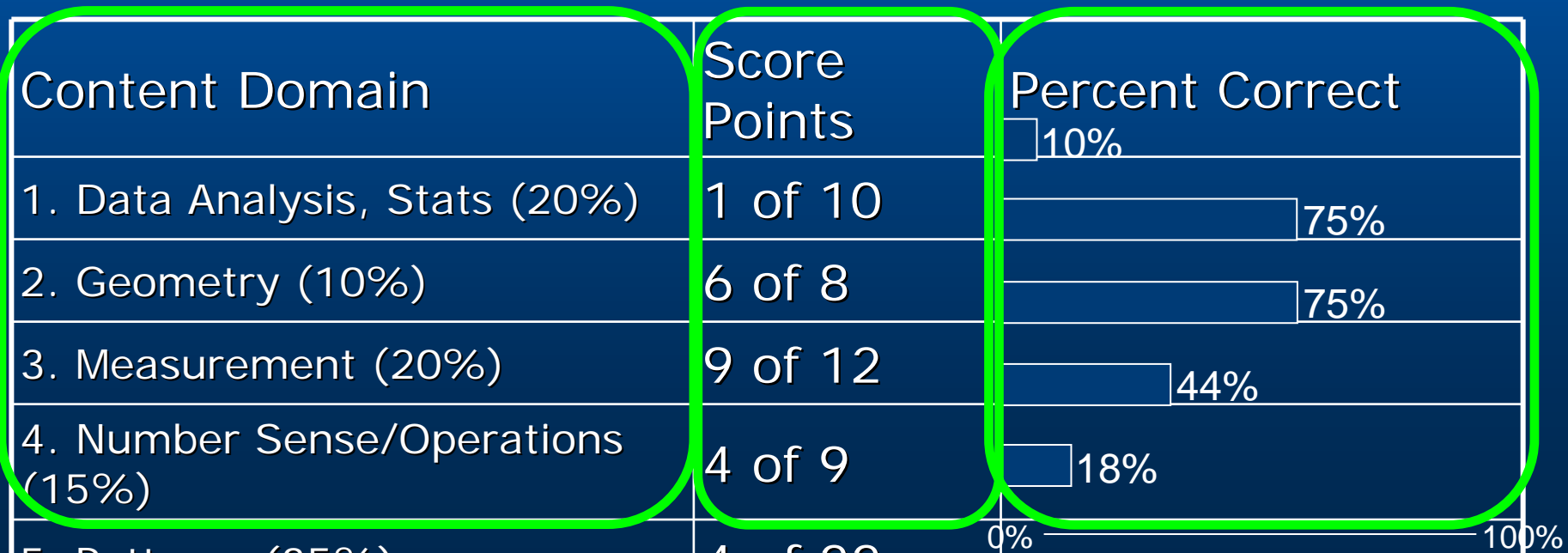
- Several undefined terms on the display: percentile, prompt, stand-alone prompt.
- Basically, the reports were crowded, making reading even more difficult.

Hambleton & Goodman Studies (2003, 2004, in progress)

- Review of score reports in 12 states, two provinces in Canada, and the three major test publishers in Canada
- Time for one example of diagnostic score reporting and recommendations

One Popular Way of Reporting Diagnostic Information

- Student results by content domain:
The mathematics assessment measures skills in five content domains. The graph below shows how many points you received in each content domain.



Highly Problematic Report!!

- No indication of measurement error
- No guarantee that the items are a representative sample (needed for a % score interpretation)
- No meaningful basis for score interpretation

Mathematics

Your Performance Compared to Passing Students

Content Domain	Your Performance	Passing Student Performance	Weaker	Comparable	Stronger
1. Data Analysis, Stats-20%	10%	20%		✓	
2. Geometry - 10%	75%	60%			✓
3. Measurement -20%	75%	90%	✓		
4. Numbers/ Operations -15%	44%	60%	✓		
5. Patterns -35%	18%	65%	✓		

Overall Performance	Weaker	Comparable	Stronger
Multiple Choice (70%)	✓		
Constructed Response (30%)		✓	

Goal 1: Conclusions

- Appear to be lots of problems in the score reports and score scales I've seen (and Goodman).
- Ultimately, though, these reports and scales will get better if research for how they might be improved is carried out, and APA-AERA-NCME guidelines are followed.

G & H Recommendations

- Report results in multiple ways (e.g., using numbers, graphics, and narrative texts).
- Highlight main findings from the test.

Recommendations, cont'd

- In student score reports, include all information essential to interpretation.
 - Purpose
 - Explanation of how results should be used
 - Description of scores
 - Example illustrating use of confidence bands

Recommendations, cont'd

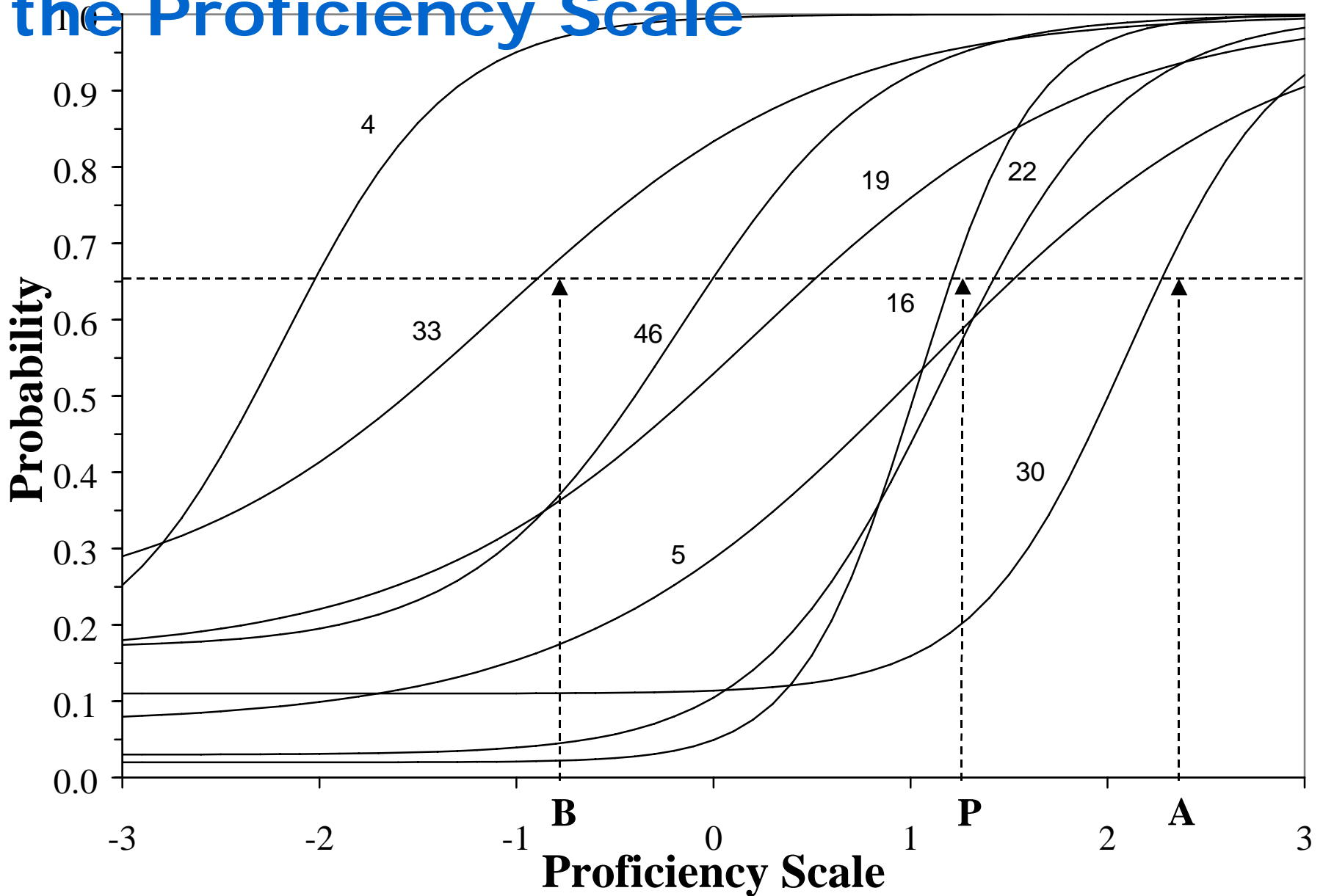
- Include more detailed information in a separate interpretive guide.
 - Detailed description of test content
 - Sample test questions that reflect relevant performance levels

Goal 2: Promising Approaches

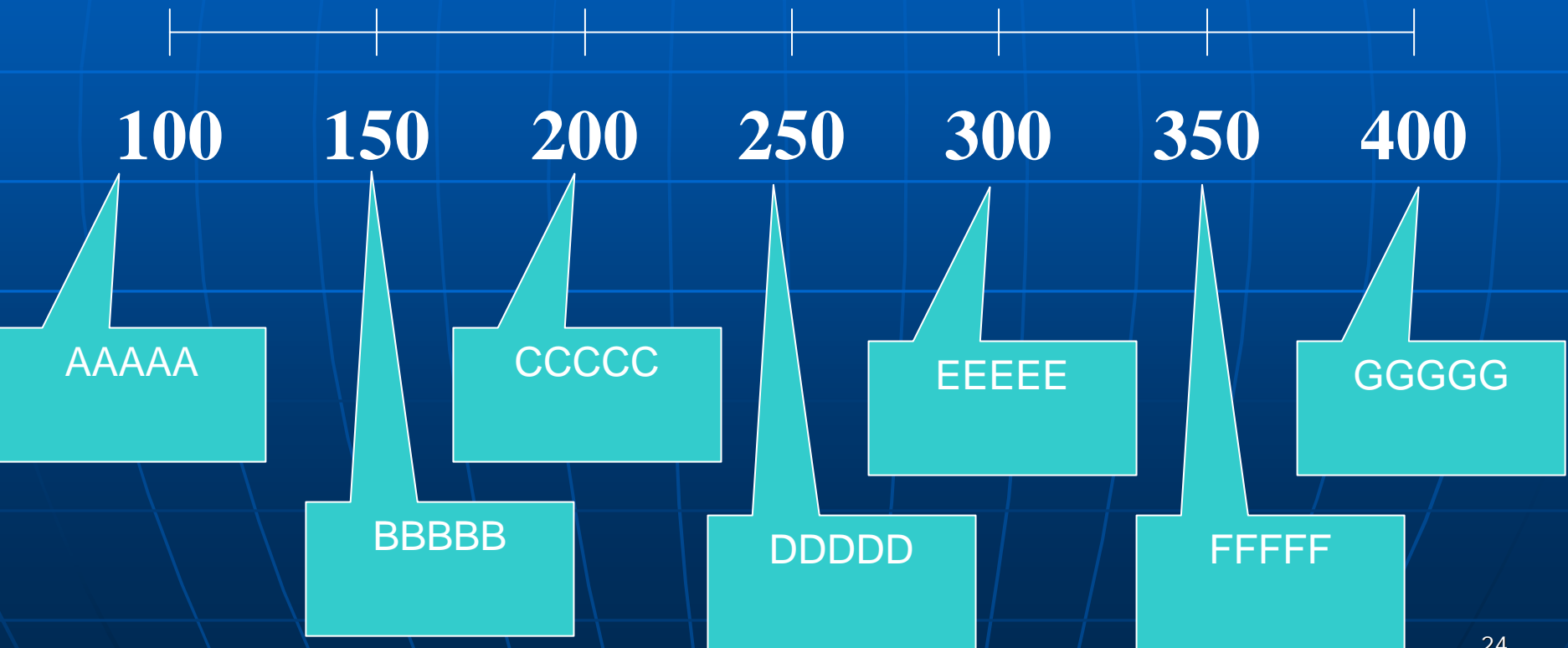
1. Item mapping and benchmarks on the proficiency (reporting) scale.

--Aim: Make scores on a reporting scale more meaningful. (NAEP, College Board, several states, publishers)

Item Mapping and Benchmarking to the Proficiency Scale



Benchmarking the Score Reporting Scale



Goal 2: Cont.

2. Use of exemplary items to describe performance categories (e.g., NAEP, MCAS)
(Bringing together items, scores, actual work, results.)
 - select items that highlight change (e.g., failing and passing candidates)



Mathematics, Grade 10

36. Which of the following functions will yield the largest value for $x = 50$?

A. $f(x) = 5 + x$

B. $f(x) = 5x$

C. $f(x) = x^2$

✓ D. $f(x) = 5^x$

Hit **Enter** key for answer

Reporting Category/Substrand for item **36**:
**Patterns, Relations, and
Functions/Functions (p. 172)**

Percentage of Students in Each
Performance Level Category Who
Answered this Question Correctly

Failing	32%
Needs Improvement	65%
Proficient	86%
Advanced	96%
Overall	54%

Goal 2: Cont.

3. Research To guide score report design

- Consider Wainer's four guidelines with **Individual Reports**—(1) What's my score? (2) How do I compare? (3) How stable is my score? (4) What does the score mean?
- Basically, apply these questions when designing/evaluating a report.

Goal 2: Cont.

4. **Seven step model**

(Hambleton, Allalouf, & Slater, in progress)--e.g., focus on purpose, intended audience, field-testing, trying alternative displays.

5. **Think-aloud experiments—**

listen to persons working through the score reports and intended interpretations.

Goal 3: Cont.

6. **Experimentation/Field-Testing/Focus Groups** (e.g., Wainer, et al., 1999) (see work being done by AICPA, for example, with focus groups)

Final Remarks

- Important advances in score reporting (NAEP, College Board, Test Publishers, States) (Goodman & Hambleton, 2004)
- More research needed on matching score reports to intended audiences, and field testing all score reports.

Future Research

- More research to build a base of knowledge to support score reporting.
- Diagnostic score reporting is especially important—new ideas needed (e.g., MIRT, Bayesian).
- We need outstanding examples of score reporting and research to guide practice.
- So, get busy, no research in testing is more important!!

Please contact Ron Hambleton
if you would like a copy of the
slides:

rkhambleton@educ.umass.edu