

# **Using Common Standards to Enable Cross-National Comparisons**

**Ronald K. Hambleton**

**Center for Educational Assessment**

**University of Massachusetts Amherst**

**Best Practices in State Assessment**

**BOTA Meeting, Washington, Dec. 11, 2009**

# Introduction

- An implicit purpose of common core state standards is to facilitate state to state comparisons of educational achievement.
- Many proposed features of suitable state tests in the coming years: computer-based, use of innovative item types, and designed to address multiple uses.
- All of these proposals for assessment changes will create challenges for reporting scores over time.

# **Purposes of the Summative Assessments for States/Consortia**

- To inform teaching and learning
- To determine school effectiveness
- To determine teacher and principal effectiveness
- To determine student readiness for college and careers
- To determine if a student is on track for college and career readiness

# **Purposes of the Summative Assessment for States/Consortia**

- To measure student growth or change in achievement
- To determine high school graduation
- To determine college course placements
- To inform college admissions

- Simply NOT possible for any test vendor to meet all of these competing purposes, regardless of the amount of funding available, with any reasonable test length!
- Reminds me of the 1970s, and demand for tests that could provide both norm-referenced and criterion-referenced information. A **single** test was NEVER optimal for providing both types of information. NR and CR inferences require different kinds of tests.

# Goals of the Presentation

- Address several of the statistical and psychometric complications of equating scores so that valid state to state comparisons can be made [I'll focus on summative tests].
- Provide a couple of examples of analyses for checking equating feasibility.
- Add a few personal remarks about test score reporting.

# Test Score Equating Concerns

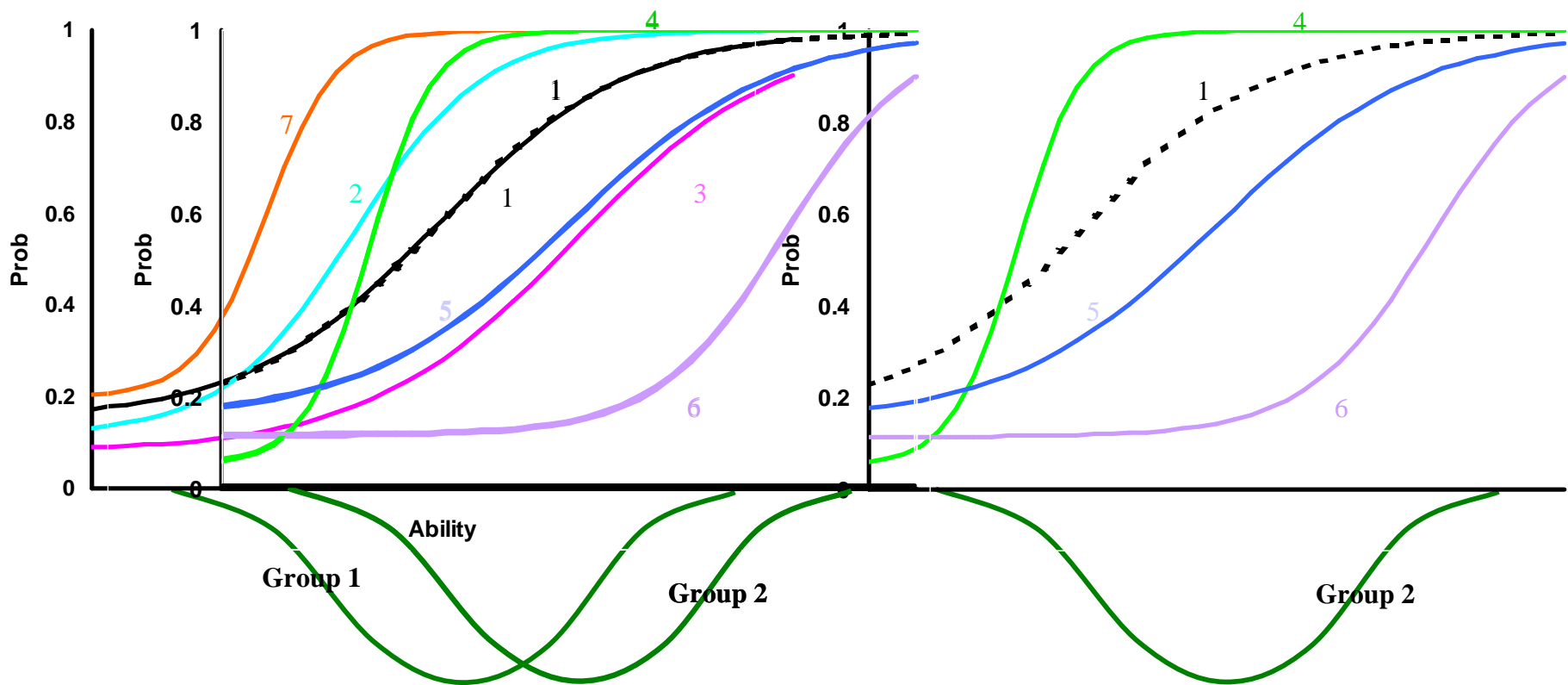
- It is near impossible from year to year to produce strictly equivalent tests to permit scores from one test to the next to be compared.
- **Solution:** Equate the test scores and this is done via complex statistical modeling of the data—and having available either common persons, or common items. Scores on one test are adjusted (up or down) for the fact that this test is harder or easier than the other form.

- Al Beaton once said that if you want to measure growth change the measure.
- But, in this instance, we must! Otherwise, the lack of test security would undermine the validity of any use of the test scores on the second administration. And, the problem would worsen over time.
- The consequence is that test score equating is essential.

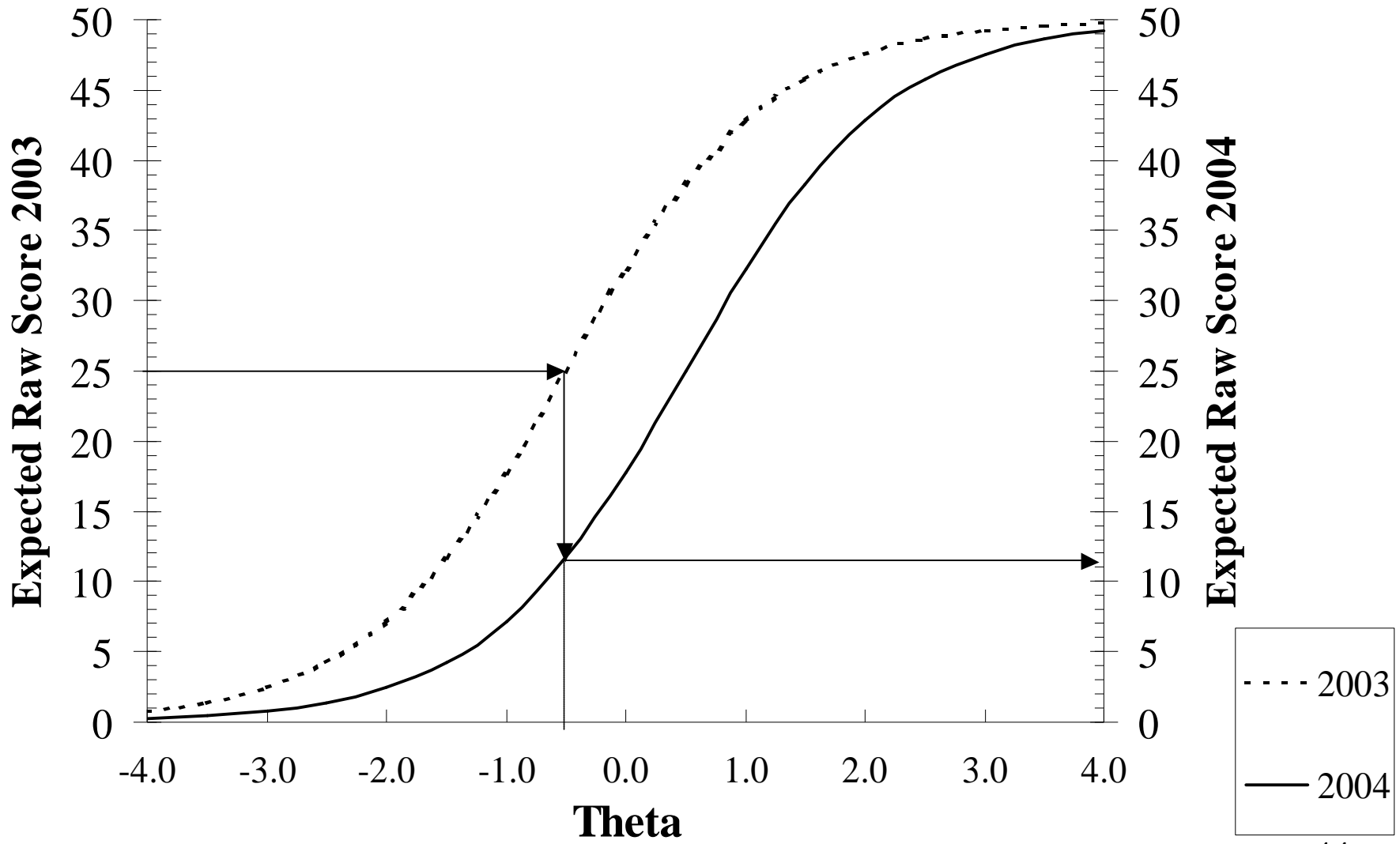
# Purpose of Test Score Equating

- By means of a statistical adjustment of test scores, the purpose of equating is to adjust for differences in difficulty among forms of a test so that the scores on the forms can be used interchangeably.
- After successful equating, examinees can be expected to earn the same score regardless of the test form administered. **This is needed for state to state comparisons.**

# Using Common Items in Equating 2 Groups (unequal) , 4 items (1 common) in each test ability scores from two tests not comparable; item statistics not comparable



## 2003 to 2004 Test Characteristic Curve Equating



- Valid equating requires model fit, lots of common items, test unidimensionality, and more.
- Books written on the topic, years of research, experience, and practice, and my two colleagues (Laurie Wise, and Rebecca Zwick) are two of the most experienced persons around.
- But equating is a complex process, and there are many complications that arise that can undermine the methodology.

- Even adjusting scores from year to year within a single state to account for non-equivalent forms is complicated. (And done just about every year in every state.)
- Moving to a cross-state comparison of results will be even more complicated.
- You have all of the usual problems associated with equating and then more.

# Possible State Test Designs

- **Model 1.** (NE) Multiple states, common content standards, the same proficiency standards, and the participating states administer the **same** tests. Different tests are administered each year and equated.
- **Model 2.** Same as Model 1, except that states use their own tests.
- **Model 3.** Multiple consortia, similar content standards, similar proficient standards, and similar tests.

# Model 1

- This is the New England Common Assessment Program (NECAP)—Laurie Wish has discussed-- and it appears to be working well (our group at UMass actually does the redundancy analysis work on equating).
- But NECAP doesn't have CBT/P and P mix, doesn't have different tests across states, and has no vertical scales.

# Other Models

- Now the situation gets more complicated.
  - tests may be different across states,
  - more performance tasks are used
  - may be a mix of CBT and P and P administrations (even within states)
  - curricula may be different, or at least teaching methods may differ.
- Let me move next to consider some of the complications of equating.

# Factors Not Considered

- Choice of classical or modern equating methods or specific methods (e.g., Stocking-Lord, FCIP)
- Data collection designs
- Pre-equating (necessary with CAT) vs. post-equating.

# Challenge 1

- There is the “push” to have these summative tests on a computer (perhaps even “CAT” administrations).
  - Both modes of administration will be needed for some time, and research evidence suggests that computer and paper and pencil administrations may function differently by grade level, subject matter, ethnic group, and student experience with computers. Equating of forms across modes is now a new task.<sup>18</sup>

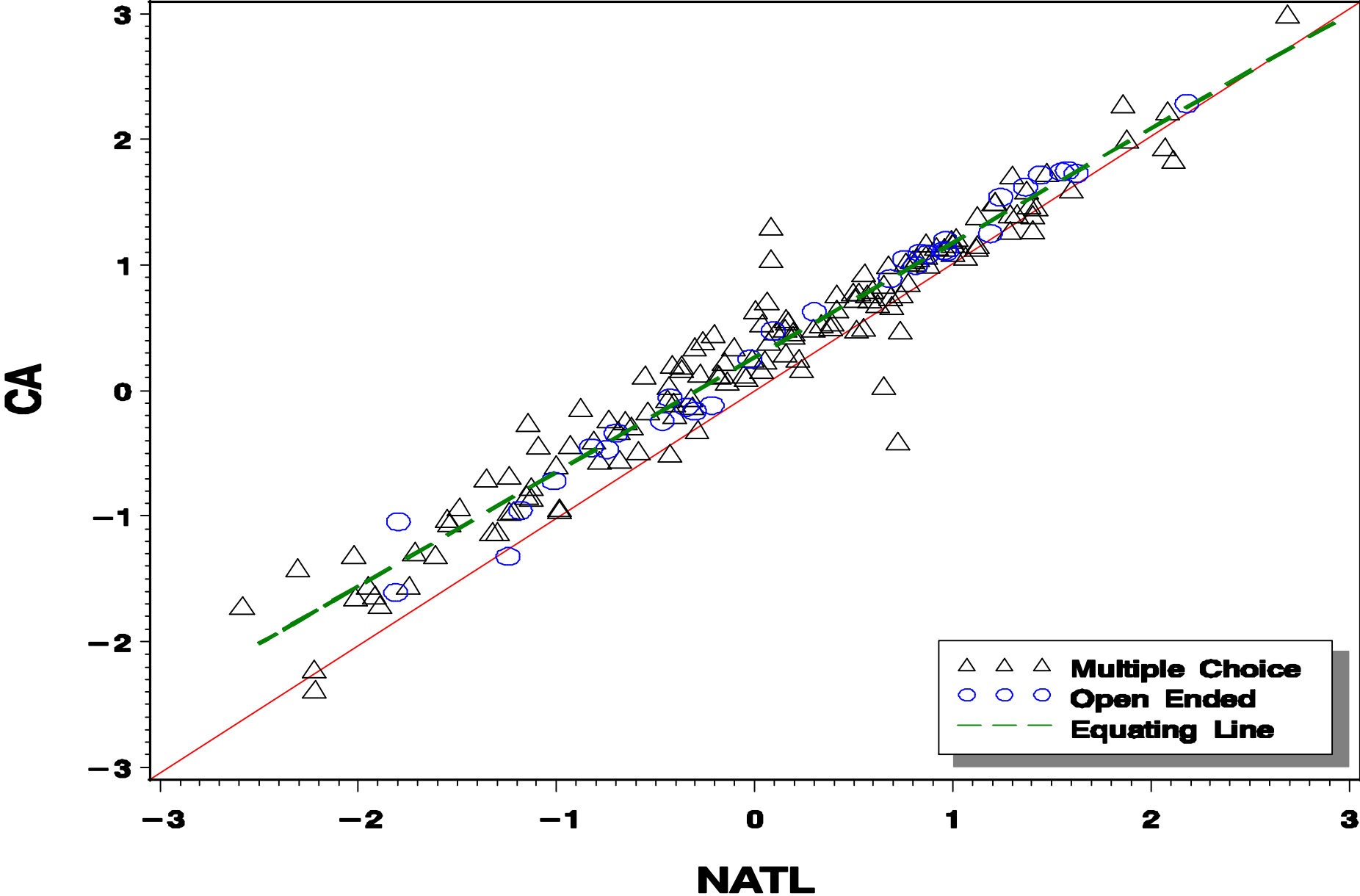
# Challenge 2

- If there are insufficient numbers of computers for all students to take the test administration at the same time, then security problems arise.
  - can be reduced with a big item bank, or multiple forms, but this raises the cost of test development.
  - we have states in which the test window is as much as a month.

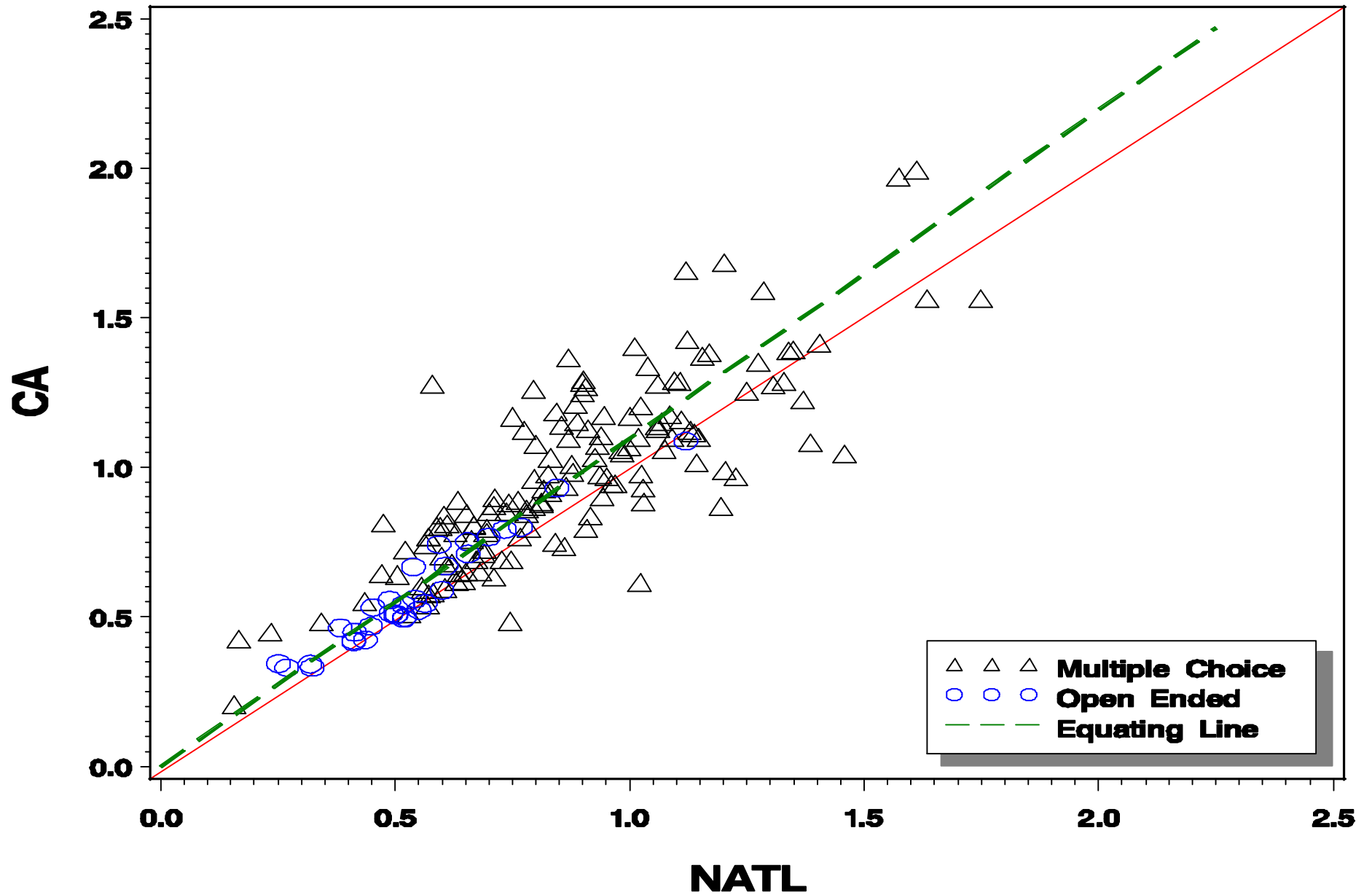
# Challenge 3

- Extension of the testing program to **multiple states** (with same content standards, performance standards and test).
  - Will the test items function in the same way? (Recall, that teaching methods will vary even if content standards and test are the same.)

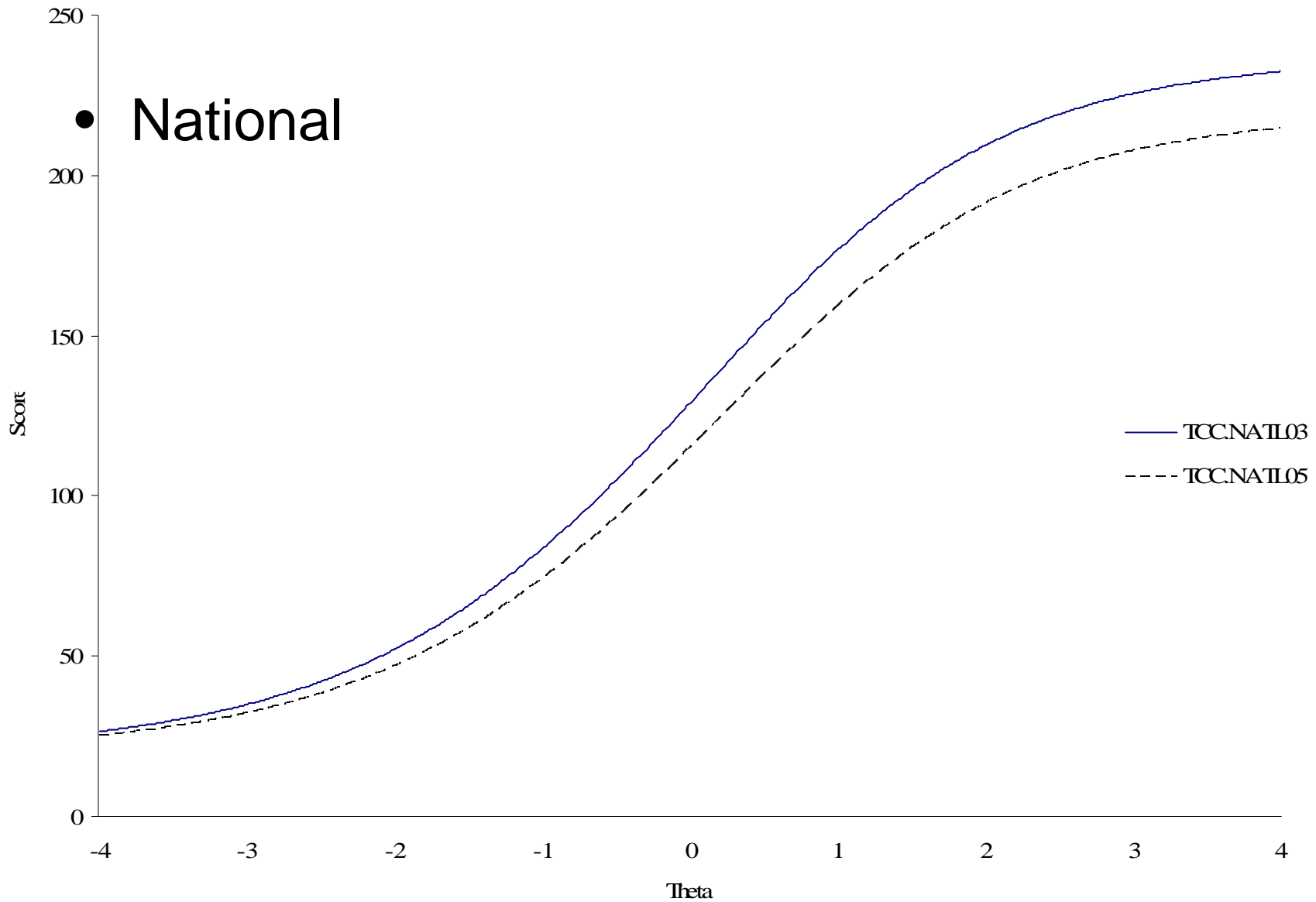
# 2005 NAEP Math Gr 8 b – plot: CA vs NATL



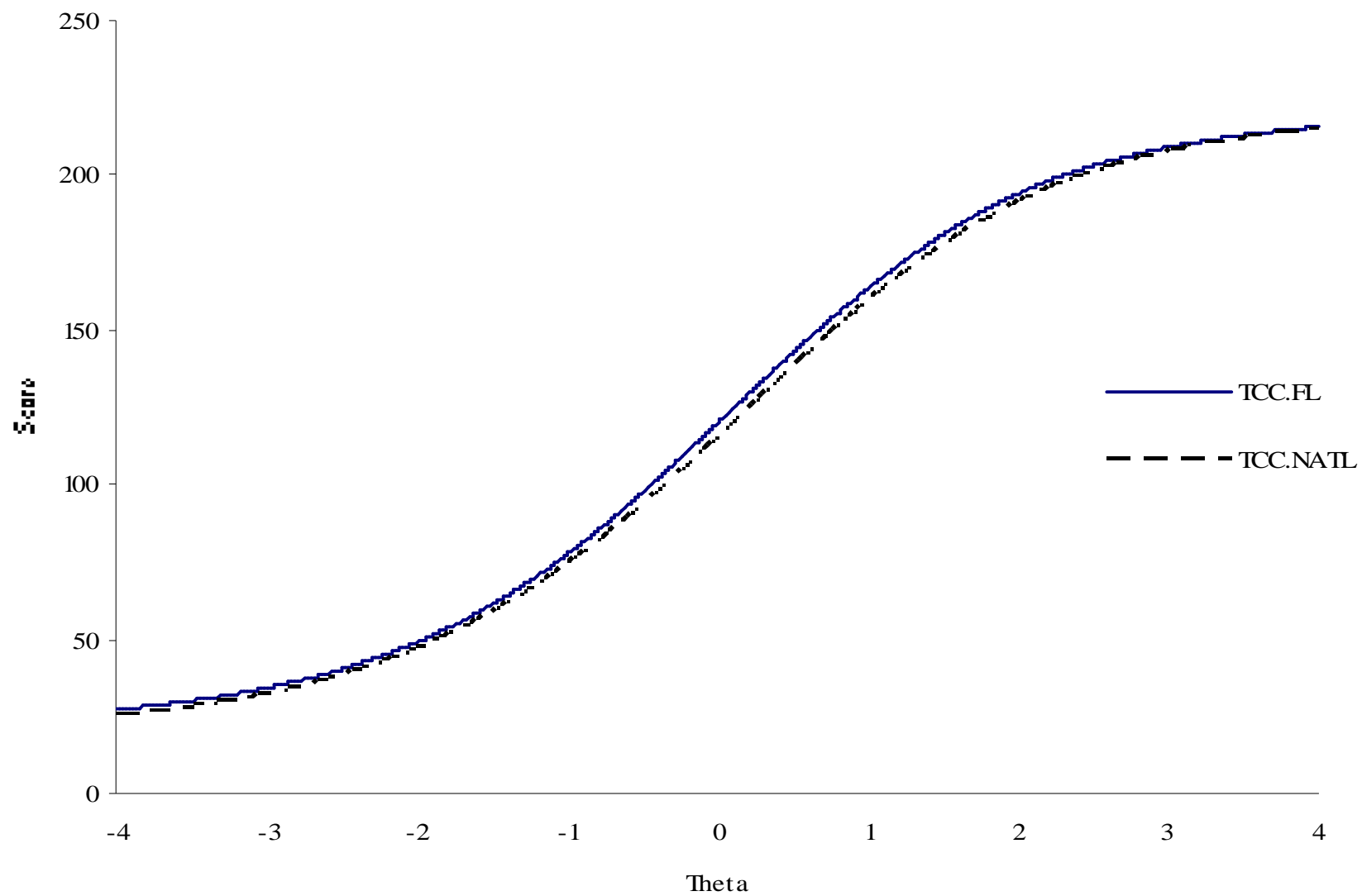
# 2005 NAEP Math Gr 8 a – plot: CA vs NATL



# Comparison of TCCs



# Comparison of TCCs



# Achievement Level Results: Math Assessment using the National Item Statistics

State	Below Basic	Basic	Proficient	Advanced
FL	34.7	39.5	20.9	4.9
MA	19.4	35.9	32.3	12.4
CA	43.0	34.3	17.3	5.3
NC	27.0	40.5	25.1	7.4
OK	34.7	45.4	17.7	2.2
Mean	31.76	39.12	22.66	6.44 <sup>25</sup>

# Achievement Level Results: Math Assessment using the State Item Statistics

State	Below Basic	Basic	Proficient	Advanced
FL	34.9	39.7	20.2	5.2
MA	19.4	36.7	31.8	12.1
CA	42.8	34.7	17.3	5.3
NC	27.1	40.8	24.6	7.5
OK	34.9	45.0	18.1	2.0
Mean	31.82	39.38	22.4	6.42 <sup>26</sup>

# Achievement Level Results (Math): National versus State (less than 1% difference in an comparison)

State	Below Basic	Basic	Proficient	Advanced
FL	-0.20	-0.20	0.70	-0.30
MA	0.00	-0.80	0.50	0.30
CA	0.20	-0.40	0.00	0.00
NC	-0.10	-0.30	0.50	-0.10
OK	-0.20	0.40	-0.40	0.20
Mean	-0.06	-0.26	0.26	0.02 <sup>27</sup>

## Summary on Challenge 3

- **Even with the same items**, but variations in the curriculum and teaching methods across states, you must check that the test items are functioning in the same way, or at least determine that there are no consequences at the test score level.

## Challenge 4

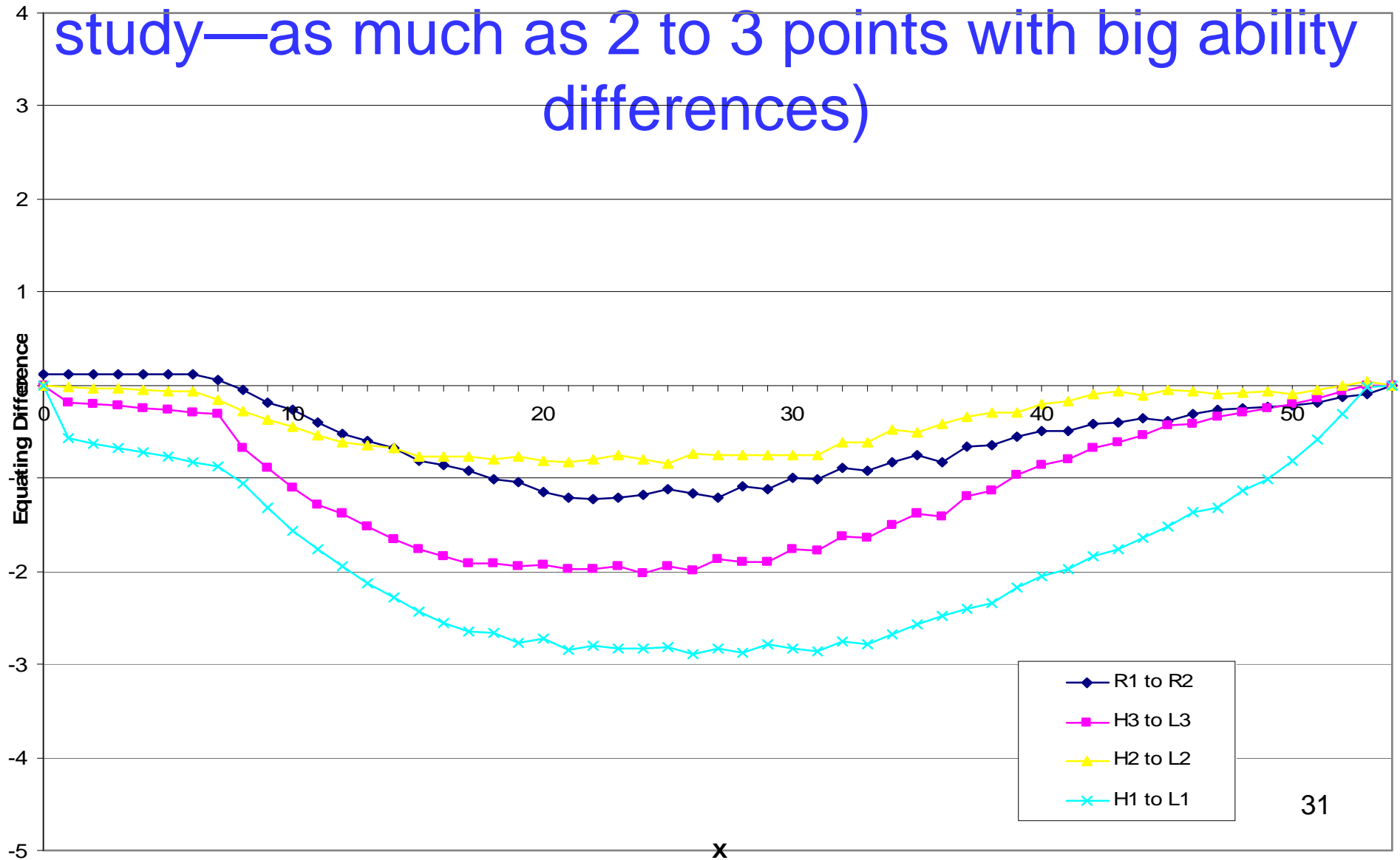
- “Push” to introduce more constructed response items to assess higher level thinking skills.
  - As desirable as this change might be instructionally, it substantially complicates the problem of test development from year to year, and equating of scores from year to year.
  - but CR items needed to assess higher level skills.

# Challenge 5

- How about creating a vertical scale in each consortium linking tests at the various grade levels? (not required to measure growth, but many policy-makers would like this feature)
  - This is controversial too, and more so with the use of more CRs and complicating the common test unidimensionality assumption underlying many applications today.

# Vertical Equating/Stocking-Lord

(shows impact of ability group differences in our study—as much as 2 to 3 points with big ability differences)



# Challenge 6

- Positioning of items in a test can have an influence on item statistics (see Zwick, and more recently Hill). This becomes especially important when “pre-equating” of forms.
  - Challenge with fixed forms is to place items near where they were field-tested, or if they are common items, then they need to be in a similar test position.

# Challenge 7

- New item types may impact on the dimensions measured by a test. New item forms—but dimensionality too, and this impacts on scoring model, and reporting, and not well developed. Unidimensional scoring could undervalue the new item types in the scoring model (3p/grm).

## Challenge 8

- Even if states in a consortium are required to adopt 85% of the common standards, would it be the same 85%? If not, in principle, theoretically possible that only 70% of the curriculum would be common.
- And, if states are using their own assessments, even with 85% match, equating will be necessary.

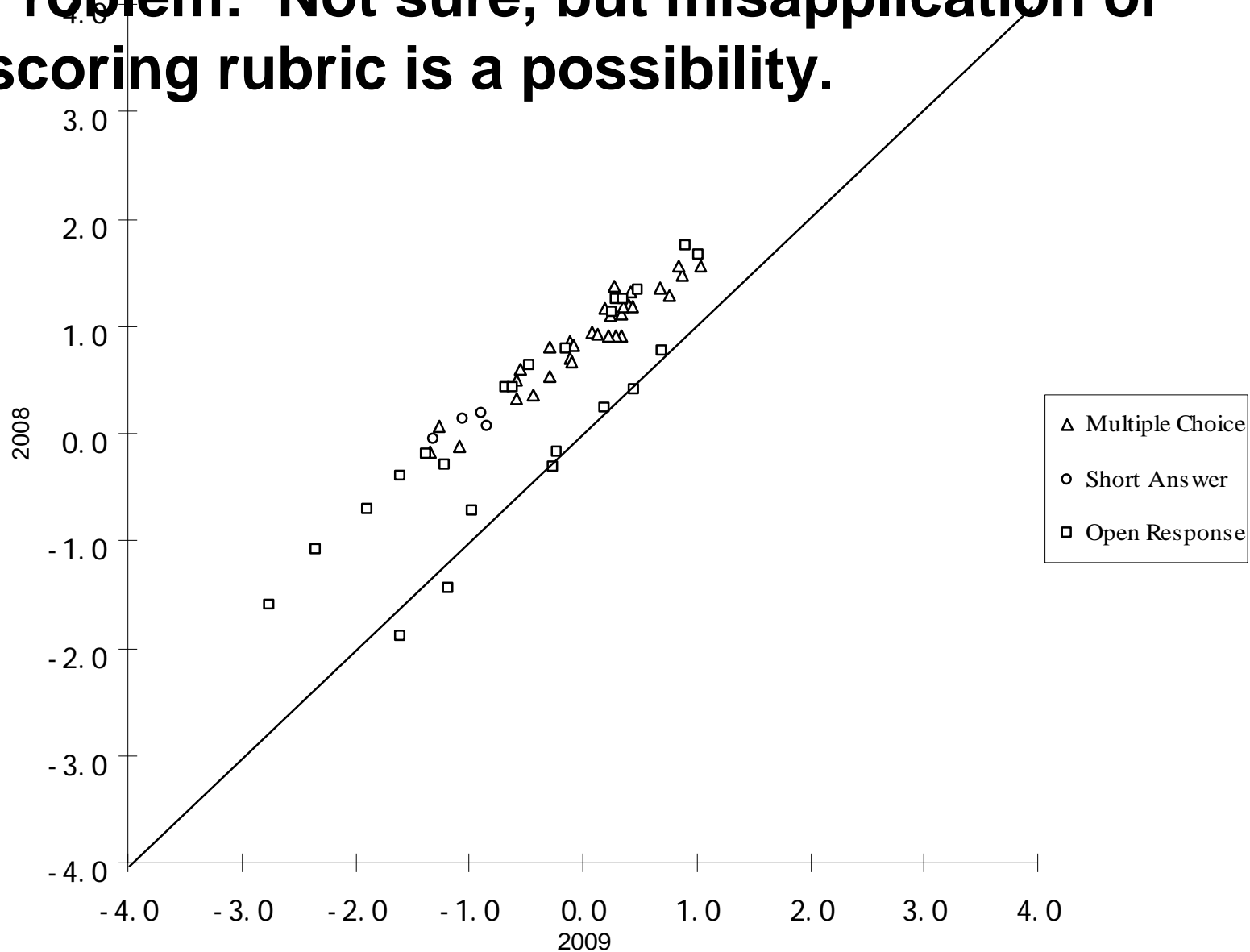
# Challenge 9

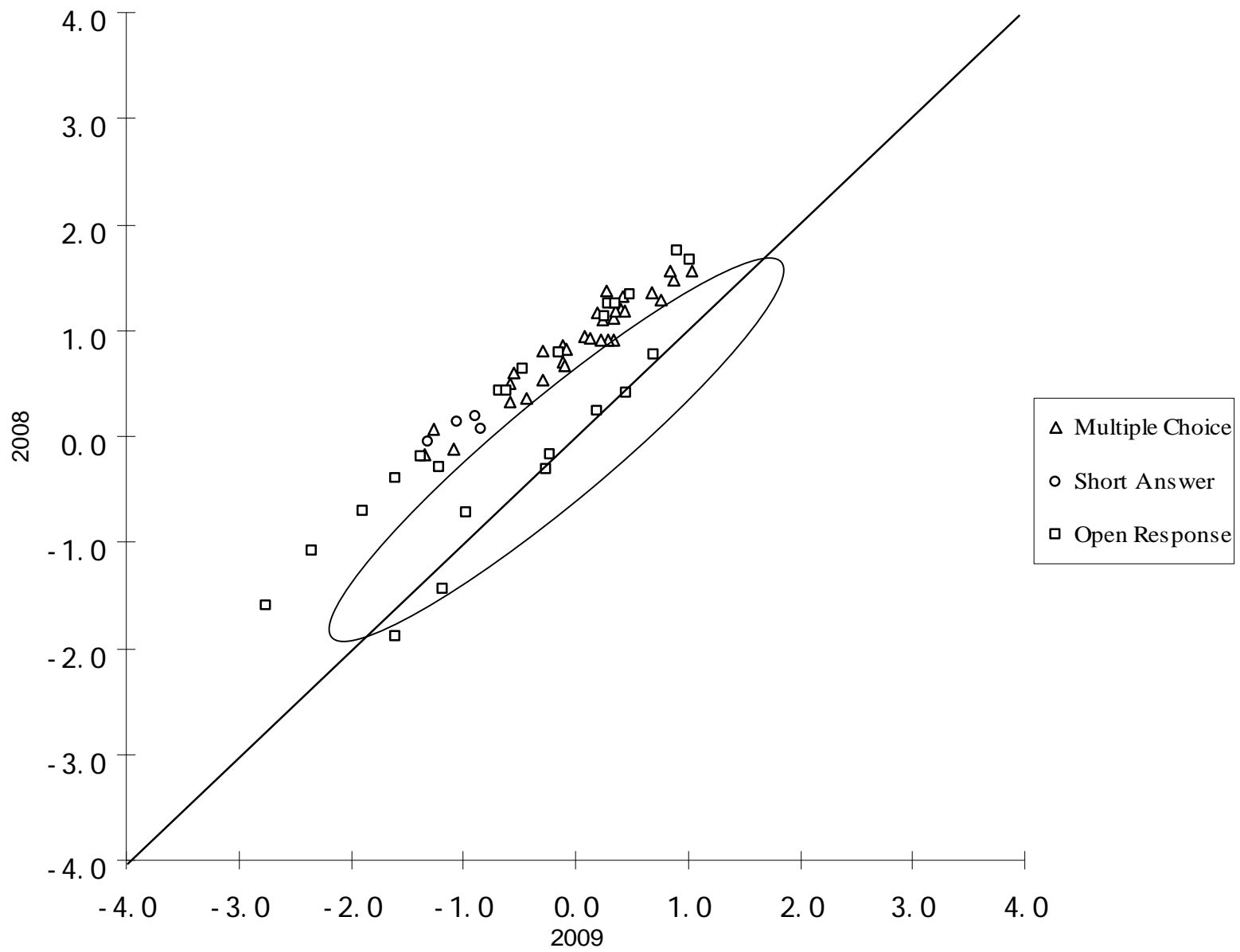
- “Common items” for linking purposes within a state across time is a frequent equating design.
  - Would be critical to use for state to state comparisons.
  - But it is not just “common items”.  
There are some strong requirements for these common items.

# Challenge 10 (continued)

- But representative items matching the contents of the two tests being linked will be essential.
  - without a “common item” link that is representative of the content of the two tests being linked, the possibility exists for over or under-representing achievement differences.
  - this means that CR items are needed in the link, and scoring must be consistent across states.

**Situation below could arise in one state over time, or the same year over two states.  
Problem: Not sure, but misapplication of scoring rubric is a possibility.**





# Challenge 11

- Near identical testing conditions are important:
  - test directions need to be consistent.
  - if time limits, they need to be the same across states. (deal with school preferences, and school starting dates)
  - test security plans
  - stakes would need to be consistent (i.e., to maximize motivation to perform well)
  - unique test content should not impact on test items in the main assessment.

# Challenge 11

- Near identical testing conditions are important:
  - minimize complications of having different vendors in different states, and minor variations in directions, answer sheets, booklet layouts, etc.
  - especially important that the “common items” look similar to students in the two or more states.

# Challenge 12

- Suppose now too you want to “link” scores in one consortium to other consortia to scores across consortia can be compared.
  - curricula may be at least a bit different
  - tests almost certainly will be different
  - performance standards may be different
  - instructional methods will vary
  - scoring of CR items will not be the same
  - testing times of the year may vary

# Challenge 12 (continued)

- This may be a design, where test score equating would be highly problematic. More common elements are needed to make equating feasible.

# With NCLB, and the Common Standards Project, no psychometricians will be left behind!



**Actually, the field needs many more psychometricians to meet the needs!**



# Excellent References

- Both publications below came from BOTA, and the National Academy Press:
  - Uncommon Measures: Equivalence and Linkages Among Educational Tests (1999).**
  - Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests (1999)**

# Score Reporting

- BOTA will take up this problem soon, but just wanted to say a few words now.
- If test score equating is done well, then improving the score reports to communicate the test information better has considerable value.  
[Rearranging deck chairs on the Titanic otherwise.]

# Important Time in the Testing Field

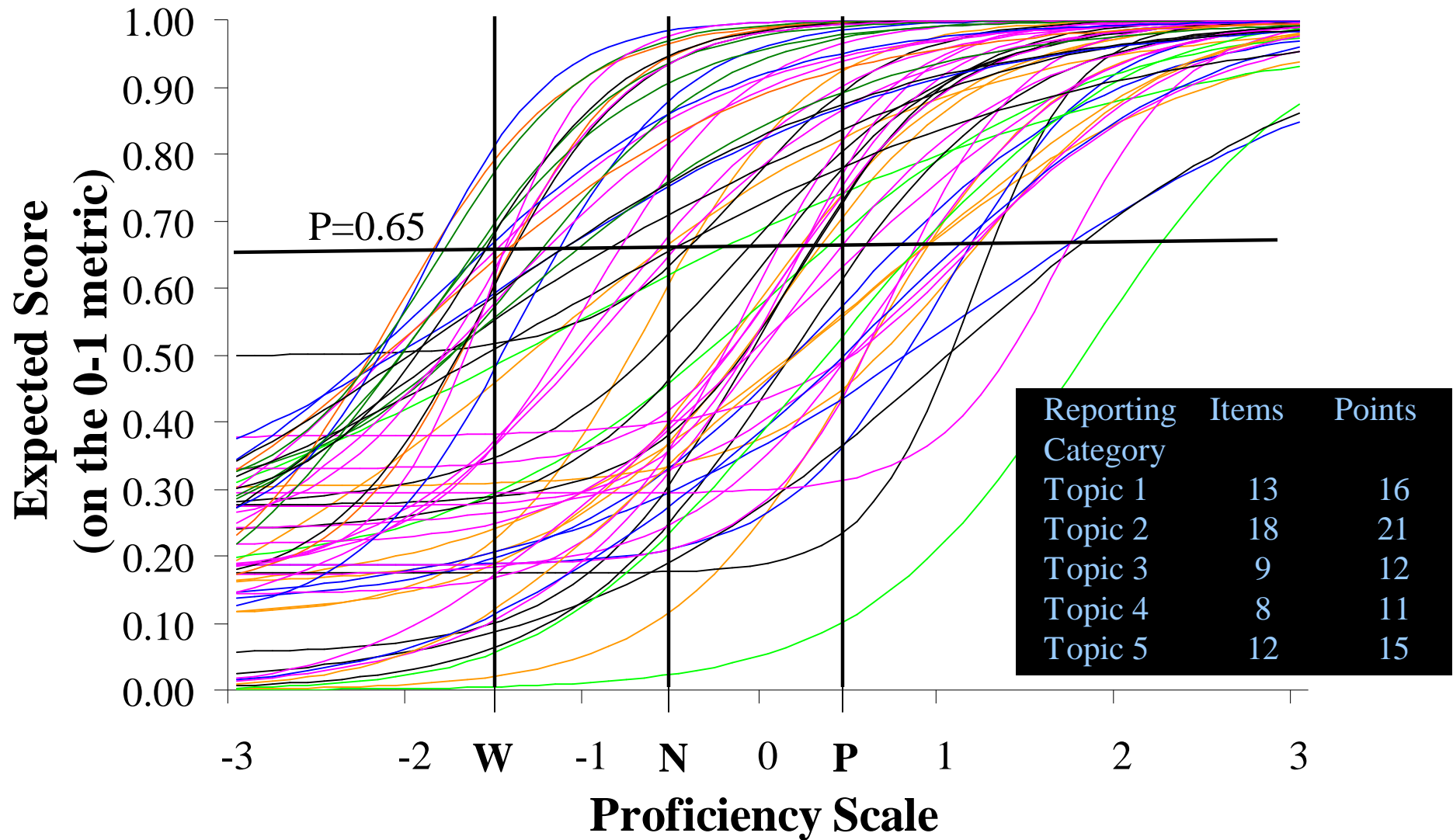
- Many new tests have been/will be introduced, with NCLB, AYP, Reach for the Top, etc. and the large number we have, have become increasingly important.
- The public, educators, policy-makers, parents, and examinees want to understand scores and score reports.
- **Good equating is a prerequisite**—need the right score adjustments for valid measurement over time and over states.

1. Considerable investment of time and money has been made to address technical problems in testing practices:  
(1) IRT modeling, (2) scoring of performance data, (3) test score equating, (4) reliability estimation, (5) computer technology, (6) DIF analyses, (7) standard-setting, and (8) validity studies.

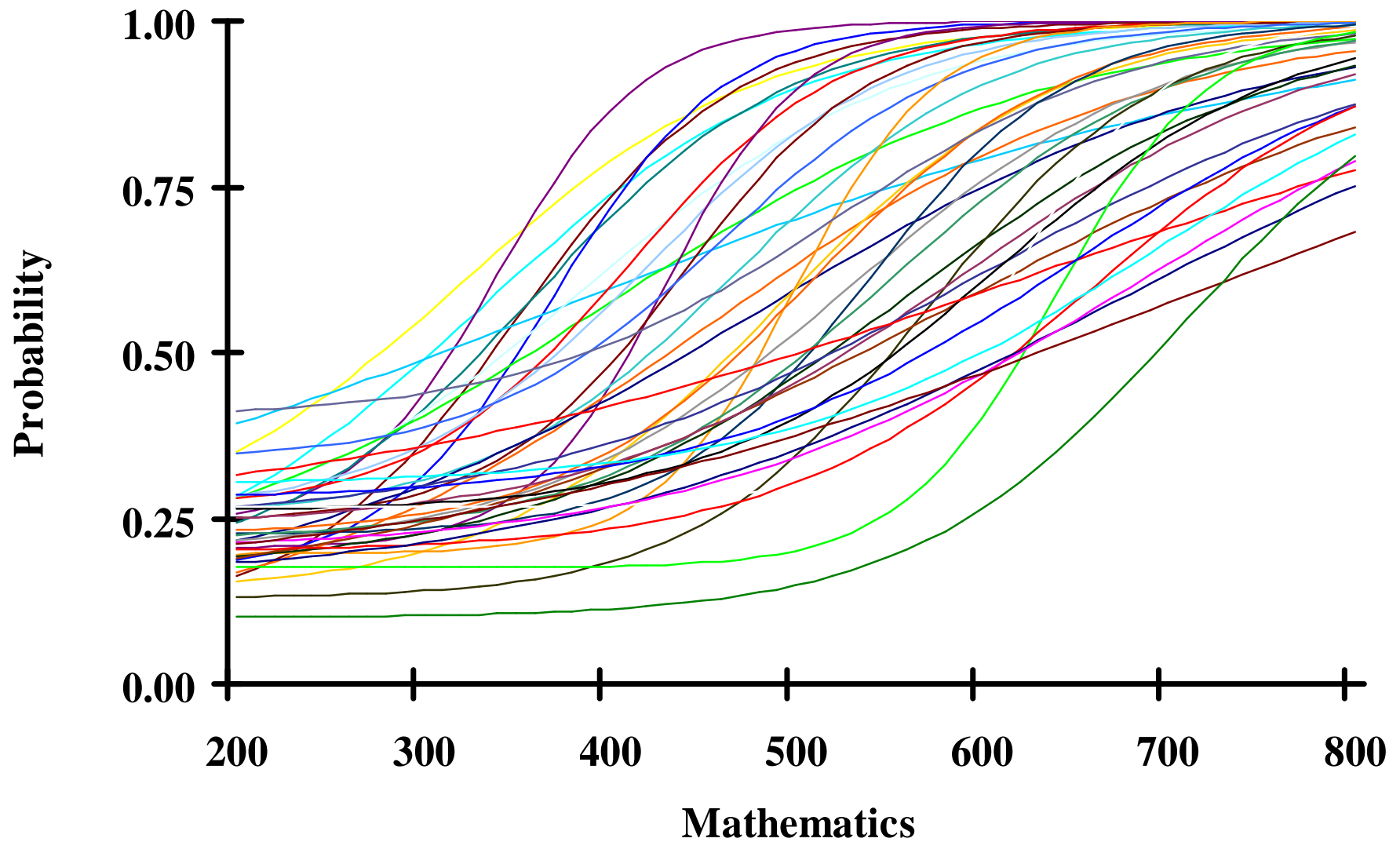
2. Surprisingly, test score reporting attracts very little attention!
  - Name one research study?
  - Without clear and meaningful reporting of information, the other steps are of less value!
  - Also, on this topic, more than other technical topics, many persons think they are experts!

- Many methodologies for improving our score reports (e.g., focus groups, think aloud, experimental studies, etc.)
- There are design principles (e.g. use of color, amount of white space, removing “chart junk”, etc.)
- Improving the meaning of scores (see our work with the SAT, and Mark’s examples yesterday with Wright maps, and PISA example).

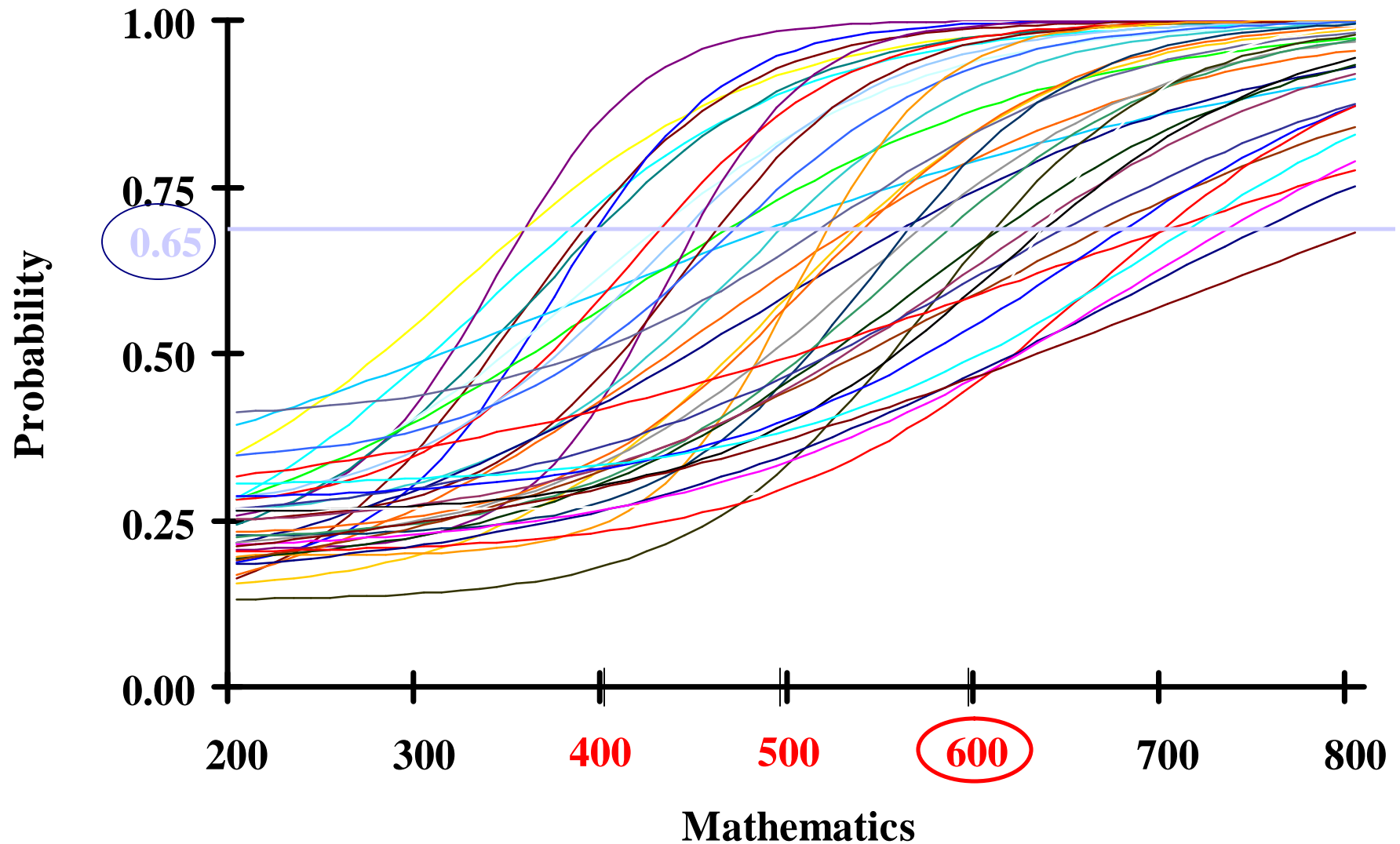
# Item Characteristic Curves for 60 Items



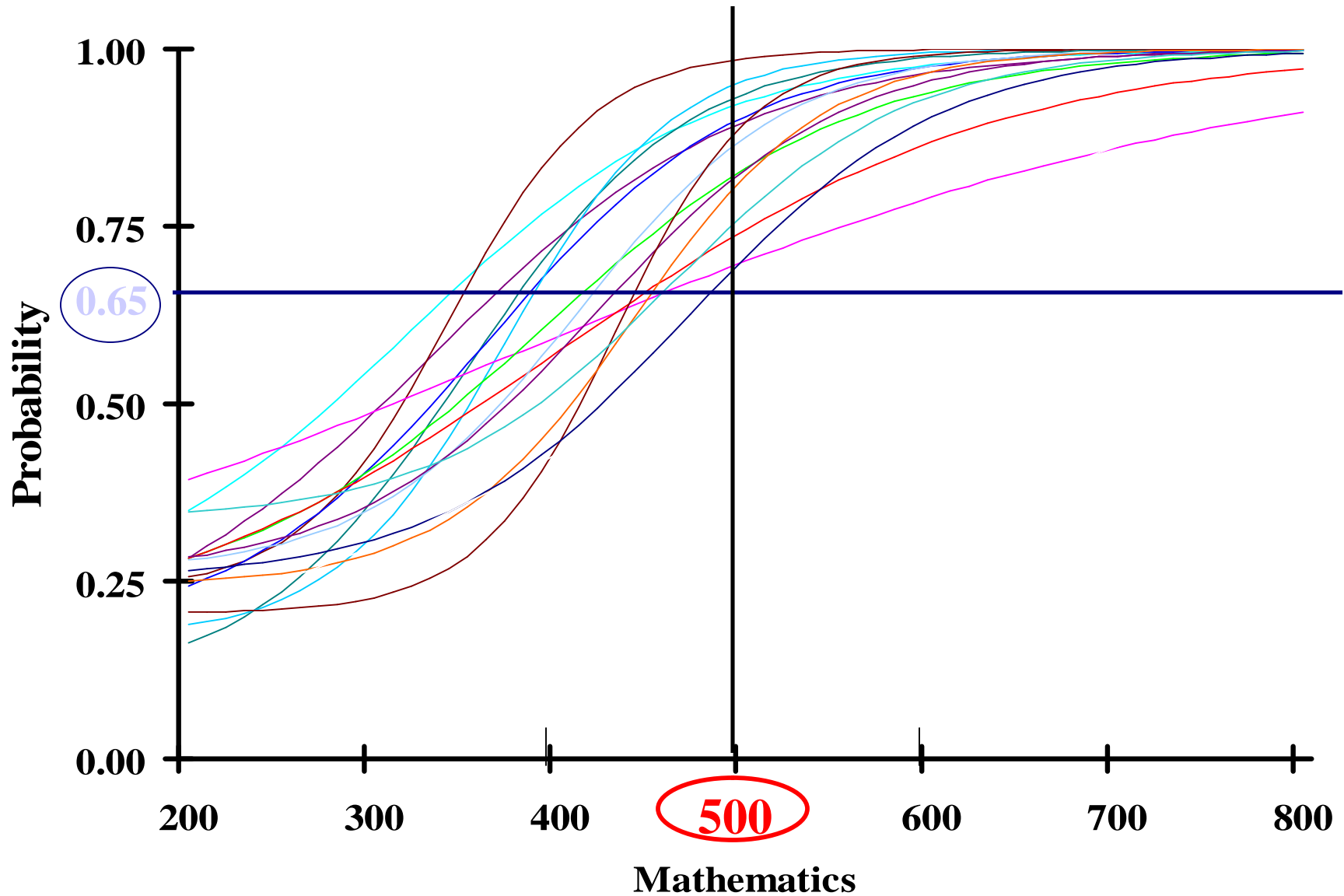
# Making Score Scales More Meaningful



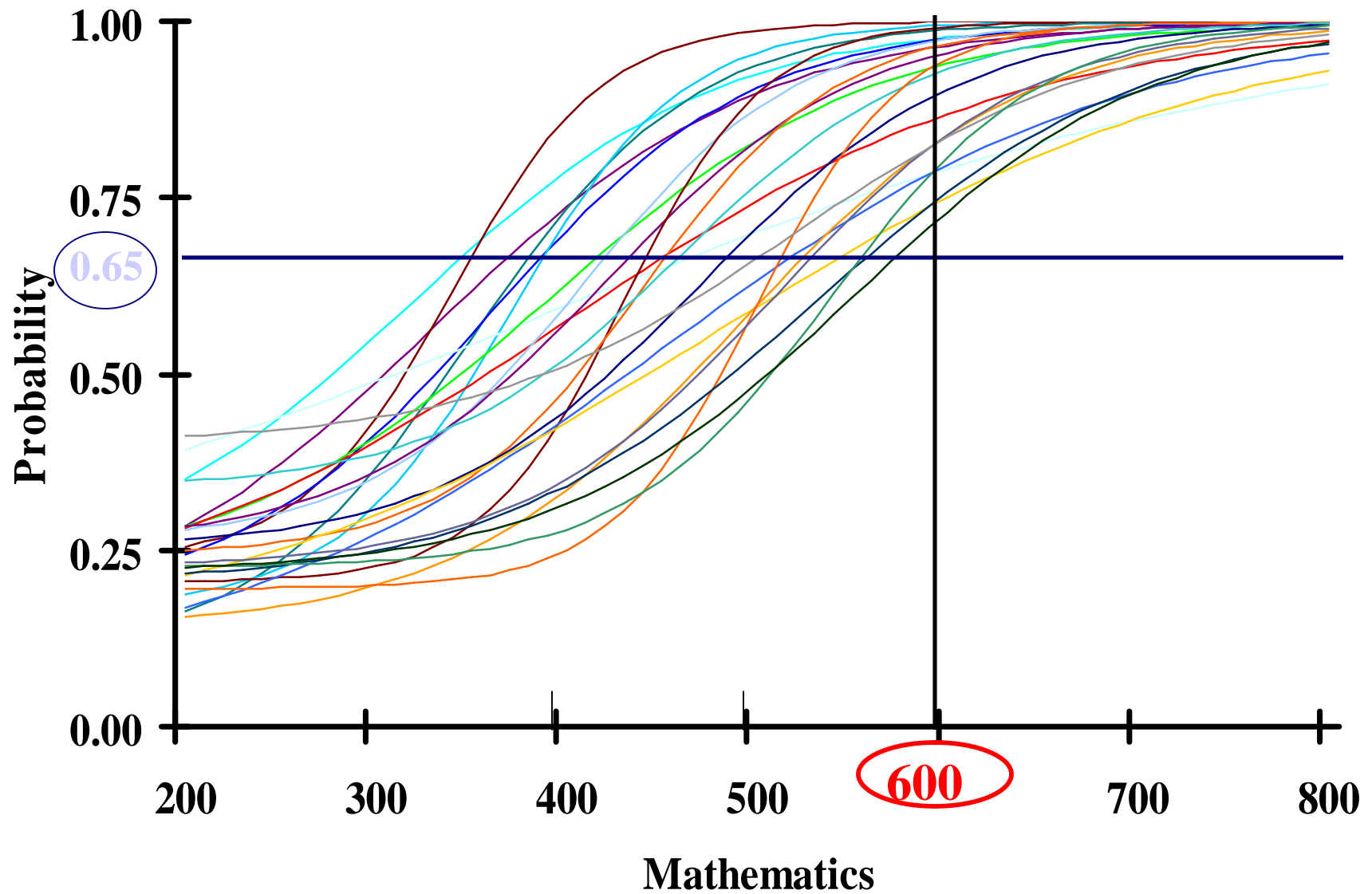
# Making Score Scales More Meaningful



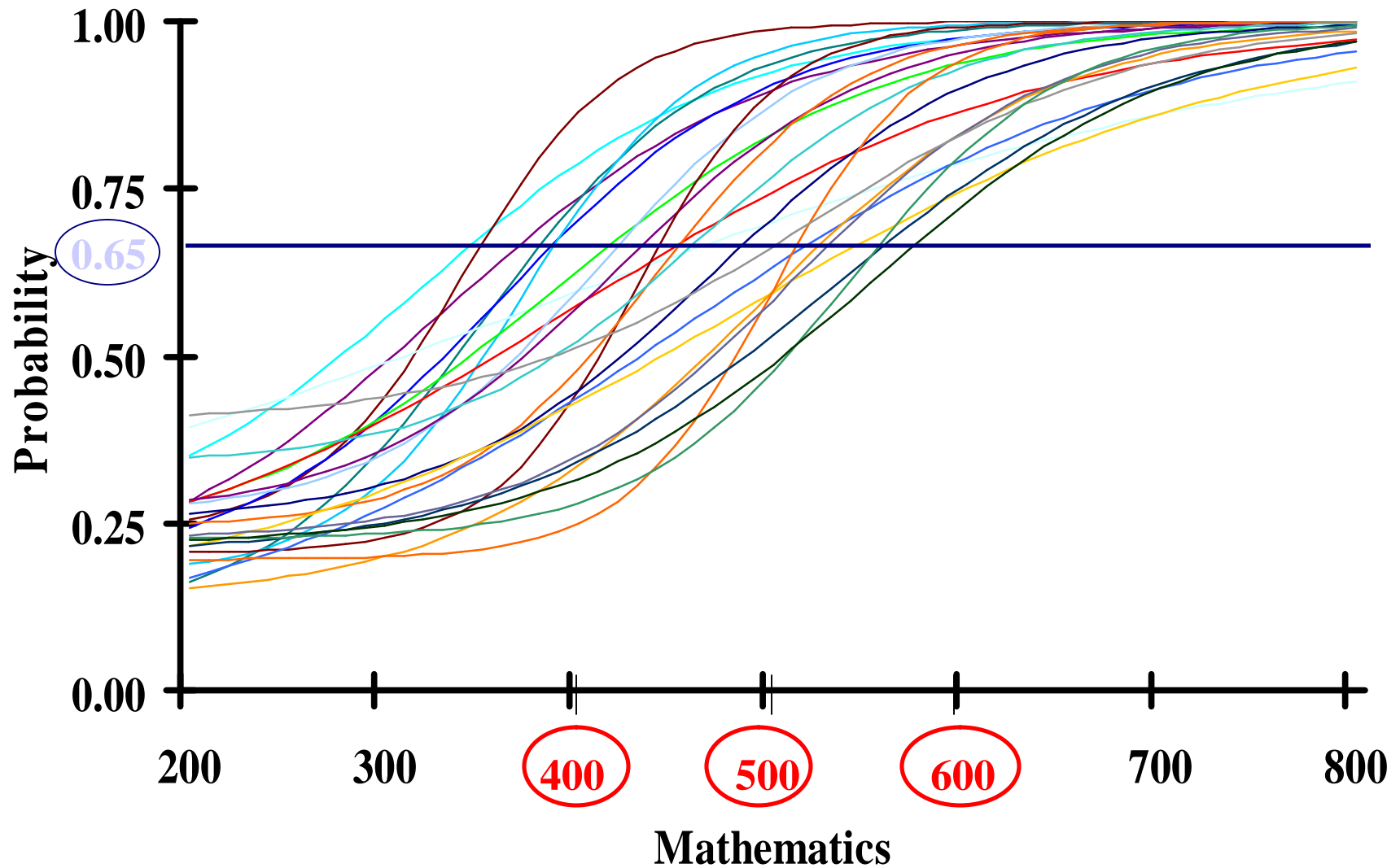
# Making Score Scales More Meaningful



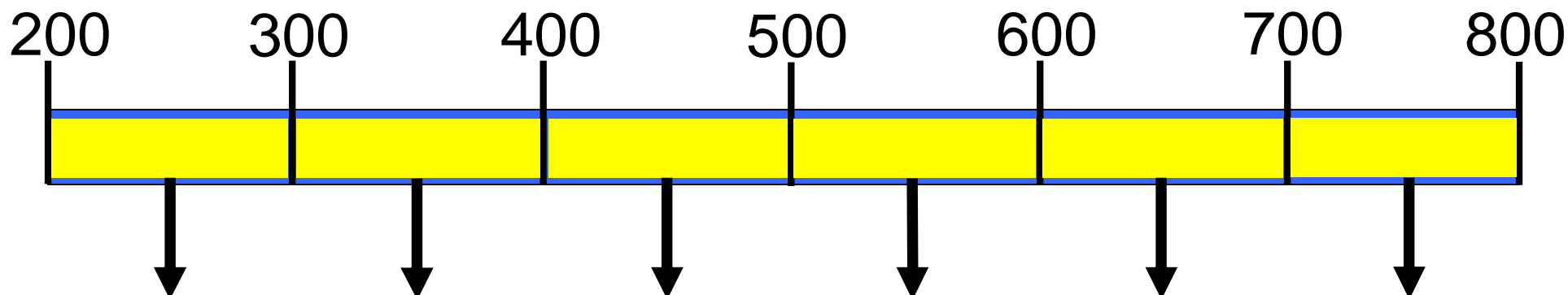
# Making Score Scales More Meaningful



# Making Score Scales More Meaningful



# Meaning of the Mathematics Scale



**Level 700-790:** Students at this level have the ability to apply insight, reasoning, and problem solving strategies to solve a wide range of problems both within and across the content areas. For example, they can solve problems involving newly-defined functions in more than two variables and can solve conditional probability problems by constructing and analyzing a table of possible outcomes.

# Conclusions

- Fred Lord, one of the greatest psychometricians of the 20<sup>th</sup> century noted that about the only time equating was justified was when equating was not needed—since the conditions were so very stringent—strictly parallel forms, near identical content, equal reliabilities at all points on the reporting scale.
- **Implication:** Try not to change the variables that impact on performance.

# Conclusions

- I remain confident that solutions can be found for permitting state to state comparisons.
- At the same time, major and important constraints on the state testing programs will be needed:
  - test administrations will need to be standardized (including timing of the test in the school year)—preferably one mode of test administration.

# Conclusions

--**best** if the **same** test is used across states.

--constructed response tasks may need to be more focused/would like to control test dimensionality (a complication for equating when it is present).

--research on CBT/CAT to study the mode of administration effect. Try to minimize through the use of only one mode if possible.

## **Conclusions (Continued)**

- careful attention to the choosing of common items (if new forms used each year) between years that are representative of the test content (this is where the gain from year to year is being estimated) and carefully monitor the scoring of CR items across time and state.
- if vertical scaling, commit substantial research to it first.

# Research Studies

- Multidimensional modeling of data, and applications to both score equating and reporting (may be the resolution of the multidimensionality introduced with CR items, and may enhance the kind of score reporting that can be done).
- Score reporting (what do users want and need, and how best to communicate the information to maximize understandability and usability--e.g. linking results to instruction).

# Research (continued)

- Viability of vertical equating for scaling test results across grade levels. (How feasible is it and what are the consequences of less than perfect model fit? Do multidimensional models provide a better solution, and how hard are they to work with?)

- Lots of important work ahead for psychometricians; but educators, policy makers and psychometricians need to work together
- psychometricians need to learn to respect the practicalities of solving applied problems in the testing field, and
- Educators and policy makers need to have more respect for technical topics that impact on testing program validity and the time needed to solve the problems!