Ensuring Comparative Validity

Quality Control in IEA Studies

Michael O. Martin and Ina V.S. Mullis

49th IEA General Assembly
Berlin, 6-9 October, 2008

IEA’s Mission: Provide Internationally Comparable Data of High Quality for Improving Education

- Data about student achievement
  - Reading, mathematics, science, civics and citizenship, computer and information literacy
- Data about the contexts for teaching and learning
  - Key factors influencing achievement
  - Educators and policy makers

"Internationally Comparative Data of High Quality"

Providing [internationally comparative] data of high quality

- Requires 100% attention to doing high quality work
- With quality assurance steps along the way
- Classic attributes of high quality achievement data
  - Reliable
  - Valid

Reliability

Extent to which instrument measures consistently what it does measure

- Instrument is the same
- Environment for using instrument is the same
- Person responds to the instrument in the same way
- Instrument is scored in the same way

To ensure that comparisons are made based on "real" achievement and not impacted by extraneous factors

Necessary, but not sufficient for good measurement…

Validity

- Extent to which inferences drawn from results can be supported by evidence
- Requires unified agreement
  - about how the construct has been conceptualized and articulated... e.g., is this mathematics?
  - on how it has been operationalized... e.g., do these items measure mathematics?
- That is, does a student with a high score on the "mathematics" test actually know a lot of mathematics? What evidence do you have?

But… what about "Internationally Comparative"?

- Our curricula are different!
- Our languages are different!
- Our school systems are organized differently!
  - Duration of compulsory schooling
  - Percentage of students attending school ("elites")
  - Stages of schooling (e.g., Primary 1-5, etc.)
  - Different age of entry
  - Different promotion and retention policies

Comparative Validity - Validity in an International Context

- Classic concerns still apply
- In addition, we need to ensure that data are internationally comparable
- Inferences made about achievement differences between countries can be substantiated

Thinking about Comparative Validity in the Context of TIMSS and PIRLS

Discuss the TIMSS and PIRLS procedures for developing the achievement tests as an illustration of how IEA addresses comparative validity as well as reliability and validity traditionally

Steps in Ensuring Comparative Validity of the TIMSS and PIRLS Achievement Data

- Assessment Framework
- Test development
- Translation Verification
- Target Population
- Sampling
- Data Collection

Steps in Ensuring Comparative Validity of the TIMSS and PIRLS Achievement Data (cont.)

- Constructed response scoring
- Database construction
- Achievement scaling
- Reporting achievement data

Comparative Validity in Test Development - Assessment Frameworks

Different curricula?

Define construct in detail

- TIMSS
  - – Content and cognitive domains
- PIRLS
  - – Purposes and processes

Assessment Frameworks (cont.)

Developed through widespread collaboration with participating countries

- Literature reviews, current perspectives
- Surveys to align assessments with countries' curricula
- Iterative reviews by NRCs
  - Within country, in plenary
- Iterative reviews by experts – SMIRC, RDG

Assessment Frameworks (cont.)

Updated with each assessment cycle

- Incorporate fresh perspectives
- Accommodate new countries
- Evolve across time

Item Development and Review

In accordance with Framework

- Assess topics/content in framework
- Ambitious frameworks require many items for adequate measurement
  - Each domain requires sufficient representation
- Trend measurement also requires many items
  - Items have to be released and replaced with each cycle
- TIMSS and PIRLS have **lots** of items!

Item Development and Review

- Developed in proportion to the emphases agreed in Framework
- According to decisions about item format
  - 50% multiple choice; 50% constructed response
- With scoring guides, if constructed response
- According to careful plan for measuring trends
  - Approximately one-half trend, one-half new

Field Test

Essential for confirming appropriateness **and** comparability of items - *different languages?*

- Twice what is needed (more or less)
- Translation by each country
  - IEA provides guidelines and instructions
- Translation verification
  - IEA verifies each translation
  - Issues referred to NRCs for resolution
- Layout verification by TIMSS & PIRLS ISC
- Countries check final printed booklets

Field Test (cont.)

About 50% of TIMSS & PIRLS items are in constructed response format

- Each constructed response item has its own tailored scoring guide (nearly 400 for TIMSS 2007)
- Scoring training materials prepared for each constructed response item
  - Scoring guide
  - Anchor or exemplar papers
  - Practice papers
- Scoring training conducted

Field Test (cont.)

Data Collection a **National** responsibility

- TIMSS & PIRLS ISC develops manuals describing standardized procedures
  - School Coordinator Manual
  - Test Administrator Guide
- IEA DPC checks and processes data
- TIMSS & PIRLS ISC conducts item analyses
  - Difficulty
  - Discrimination
  - Scoring reliability

Finalizing Item Selection

- Task Force and TIMSS & PIRLS ISC makes initial recommendation about items to retain
- Field test data and initial recommendation reviewed by expert committees – SMIRC, RDG
- Field test data and expert committee recommendation about item selection reviewed by the NRCs from participating countries
- Assessment items adopted by NRCs

Test-Curriculum Matching Analysis (TCMA)

How well does the TIMSS assessment match your curriculum?

- Each country identifies the TIMSS items that fit its curriculum
- Analyze achievement based on these items
  - Little evidence of changes in relative achievement across countries

Comparative Validity in Data Collection, Analysis, and Reporting

- Are target populations comparable?
- Was sampling conducted properly?
- Are translations comparable?
- Were the tests administered appropriately?
- Was scoring done correctly?
- Are the data comparable?
- Are the achievement results comparable?

Comparable Target Populations?

Different school system organizations?

In TIMSS & PIRLS,
Amount of Instruction –> Years of Schooling

- PIRLS: 4 years of schooling, counting from 1st year of primary -> (4th grade)
- TIMSS: 4 & 8 years of schooling (4th & 8th grade)
- Based on ISCED definitions

TIMSS and PIRLS: Grade based assessments for improving education

Why grade and not age as the basis?

- Better for improving education!

- Education is organized by grade, so grade-based data easier to use for implementing reforms
- Amount of instruction, not maturation, the primary determinant of achievement
  - Students learn through instruction, not simply by growing older

Comparable Target Populations? -cont.

- Has country chosen correct grade?
- Are all students included in definition?
  - Generally yes, for most countries
  - If less than 100%, annotated in International Reports
- Are exclusions kept to a minimum?
  - Generally yes, for most countries
  - If more than 5%, annotated in International Reports

Sampling Conducted Correctly?

TIMSS & PIRLS Requirements

- Random sampling design – authorized by Statistics Canada
- Accurate school sampling frame
  - School sampling by Statistics Canada
- Accurate classroom sampling
  - Use of WinW3S mandatory

Sampling Conducted Correctly? -cont.

TIMSS & PIRLS goals for sampling participation

- Participation rates for schools and students
  - 100% !!!
- Sampling precision goals
  - Percentages ±5%
  - Means ± .1 S.D.
- Usually 150 schools and one or two classes per school (Approx 4,500 students)

Sampling Conducted Correctly? -cont.

- Procedures acceptable and fully documented?
  - Review by Statistics Canada and Sampling Referee
  - If procedures not acceptable, reported in appendix
- Acceptable participation rates? (At least 85% schools, 85% students)
  - Generally yes, for most countries
  - Others annotated in International Reports or below a line
- Population coverage and participation rates published in International and Technical reports

Translations Comparable?

- Has country correctly translated all test booklets?
  - IEA Secretariat verifies each translation
  - Issues referred to National Research Coordinator for resolution
- Do test booklets conform to international layout?
  - TIMSS & PIRLS ISC verifies final layout before printing

Tests Administered Correctly?

- How do we verify that data collection procedures have been followed?
  - IEA Secretariat and TIMSS & PIRLS ISC conduct program of international quality control monitoring
  - IEA Secretariat recruits Quality Control Monitor (QCM) in each country
  - Training sessions are conducted for QCMs
  - The QCM visits a sample of 15 schools at each grade; records observations and interviews school coordinator and test administrator

Tests Administered Correctly? -cont.

- TIMSS & PIRLS ISC analyzes and reports results in the technical report
  - Generally QCM reports very positive
  - Data collected according to procedures specified in manuals, with very few exceptions
- Country also conducts quality control observations at 15 schools
- NRCs complete online Survey Activities Report

Constructed-response Item Scoring Done Correctly?

- Scoring training conducted separately for Southern Hemisphere and Northern Hemisphere countries

- Training materials updated, based on field test experience
  - Scoring guides refined
  - Enhanced sets of example responses and practice papers

Constructed-response Item Scoring Done Correctly? –cont.

How do we know the scoring was done well?

- Monitor reliability through double scoring
  - Within country current assessment (200 responses per item)
  - Within country across trend assessments (200 responses per item are scanned from previous assessment and delivered via computer for rescoring with current assessment)
  - Across countries current assessment (200 responses per item from English-speaking countries delivered via computer)

Constructed-response Item Scoring Done Correctly? –cont.

What happens if an item is not reliably scored?

- Vast majority of items have high scoring reliability
- Items with less than 70% agreement for within-country or trend reliability are removed from scaling
  - Extremely rare
- Scoring reliability data for all countries documented in technical reports

Are the Data Comparable?

- IEA DPC provides data entry software and variable codebooks to standardize data preparation
- DPC provides extensive training seminars
- DPC checks each country’s data files for internal consistency and accuracy
- DPC interacts with countries to resolve data issues
- DPC creates database and sends to TIMSS & PIRLS ISC and Statistics Canada for analysis and reporting

Are the Data Comparable? -cont.

- Statistics Canada creates sampling weights based on data and previous sampling information
  - Compares estimated population size using weights against estimate from sampling frame
  - Interacts with countries to resolve issues
- Creates final weights, including adjustments for non-response, for analysis and reporting

Are the Data Comparable? -cont.

Initial review of item statistics, before scaling

- TIMSS & PIRLS ISC reviews achievement item statistics – **every** item for **every** country
- Investigates items with poor discrimination or unreliable scoring – sometimes caused by a translation or printing error
- Rare (½ of 1% of item instances), but such “faulty” items are not included in scaling achievement results for that country

Are the Data Comparable? -cont.

Review of **item-by-country** interactions

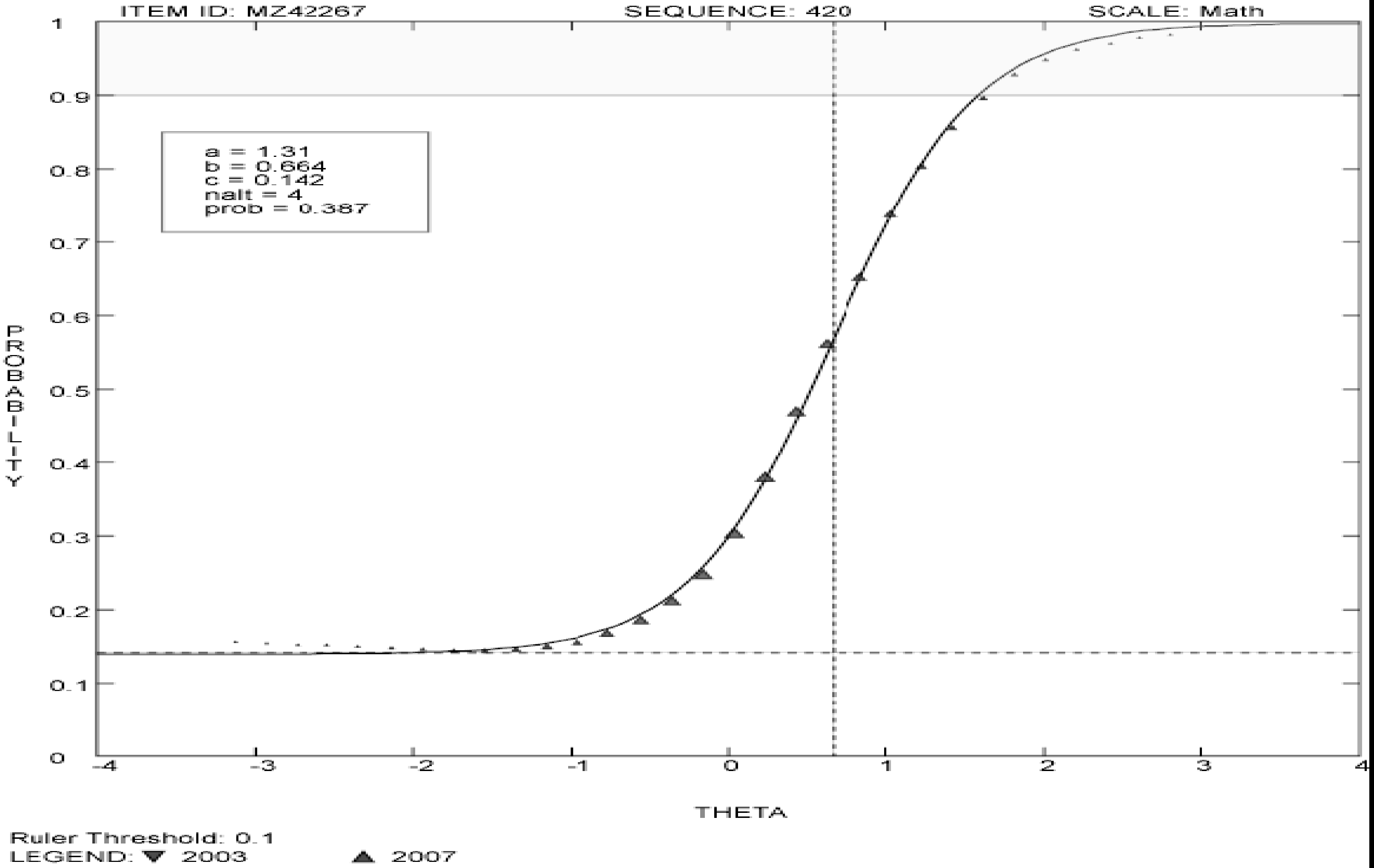
- For each item, examine each country's performance on the item in light of its overall performance
  - Outliers may be due to translation, printing, etc.
- For trend, compare item-by-country interaction patterns for both assessments (e.g., TIMSS 2003 and 2007)
  - If different, may delete that item for that country for trend

Are the Scaled Achievement Results Comparable?

Use **IRT scaling** to summarize achievement data by modeling item difficulty and discrimination – **one** scale for all countries

- Scaling procedure fits a model to each item, the better the fit, the more accurate the result
- Check fitted model against observed data for each item
  - Typically any item issues were discovered during initial review

2007 TIMSS MATHEMATICS ASSESSMENT -- GRADE 8
PARSCALE - Univariate -FREED PRIOR - RUN # 02



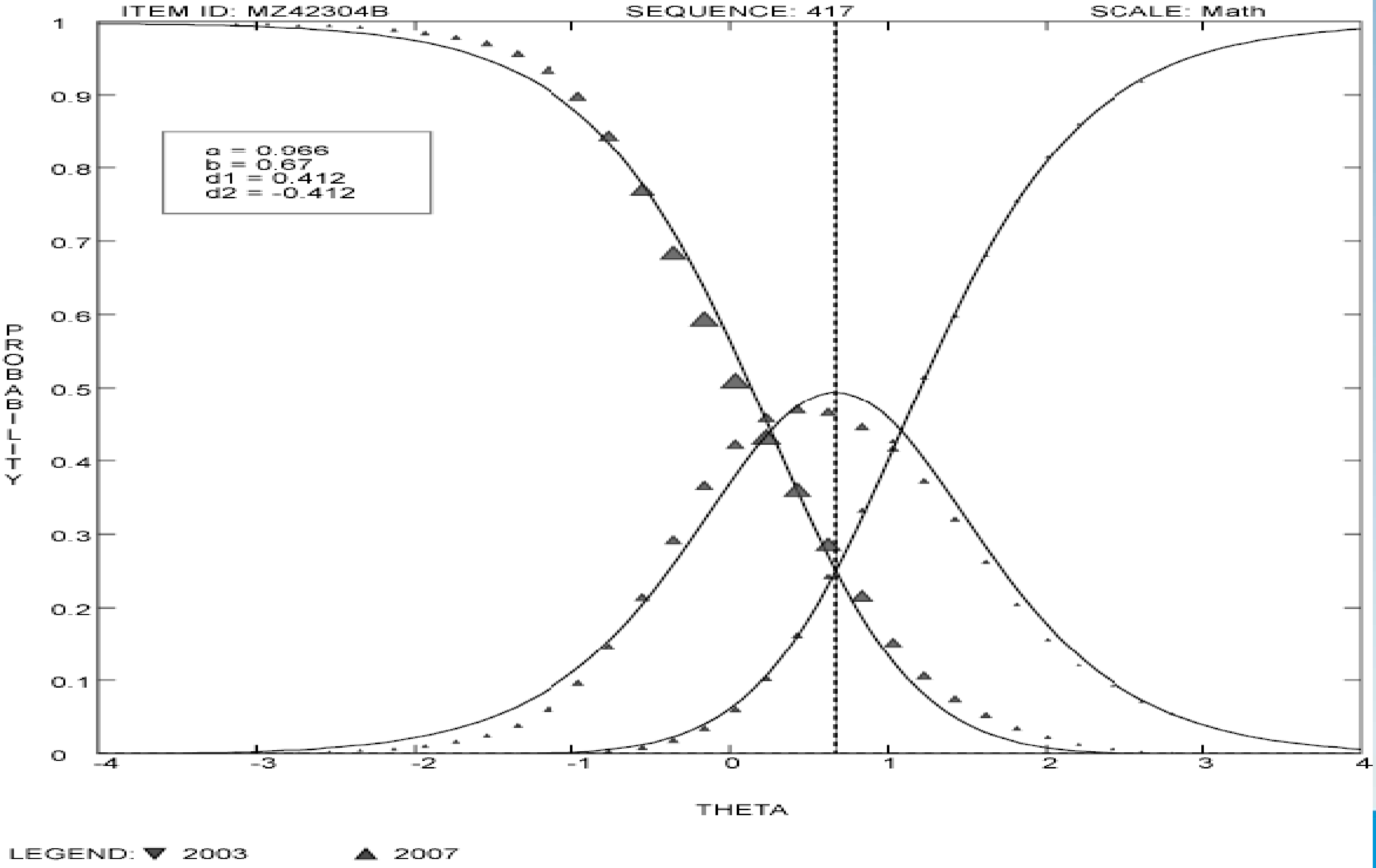
ITEM ID: MZ42267 SEQUENCE: 420 SCALE: Math

a = 1.31
b = 0.664
c = 0.142
nalt = 4
prob = 0.387

Ruler Threshold: 0.1
LEGEND: ▼ 2003 ▲ 2007

SOURCE: PARSCALE 3.1 RUN DATE: 04/18/2008 TIME: 09:16:21
PARPLOT SUN/UNIX Online Version 2.5

2007 TIMSS MATHEMATICS ASSESSMENT -- GRADE 8
PARSCALE - Univariate -FREED PRIOR - RUN # 02



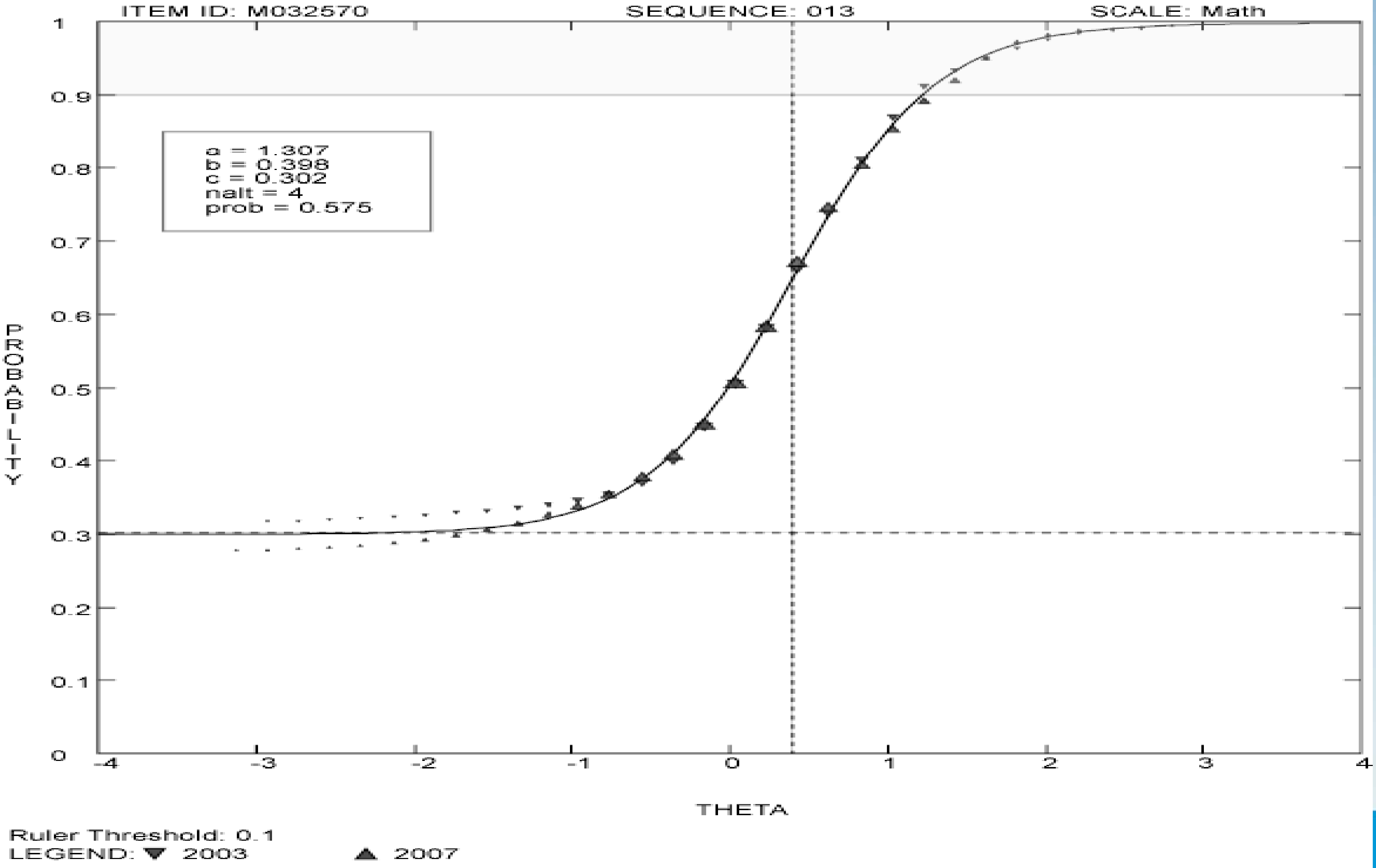
SOURCE: PARSCALE 3.1 RUN DATE: 04/18/2008 TIME: 09:16:21
PARPLOT SUN/UNIX Online Version 2.5

Are the Scaled Achievement Results Comparable? –cont.

For trend items,

- Data scaled together, e.g., TIMSS 2003 and 2007
- Item fit plotted separately to ensure that the item is a good fit to both sets of assessment data

2007 TIMSS MATHEMATICS ASSESSMENT -- GRADE 8
PARSCALE - Univariate -FREED PRIOR - RUN # 02



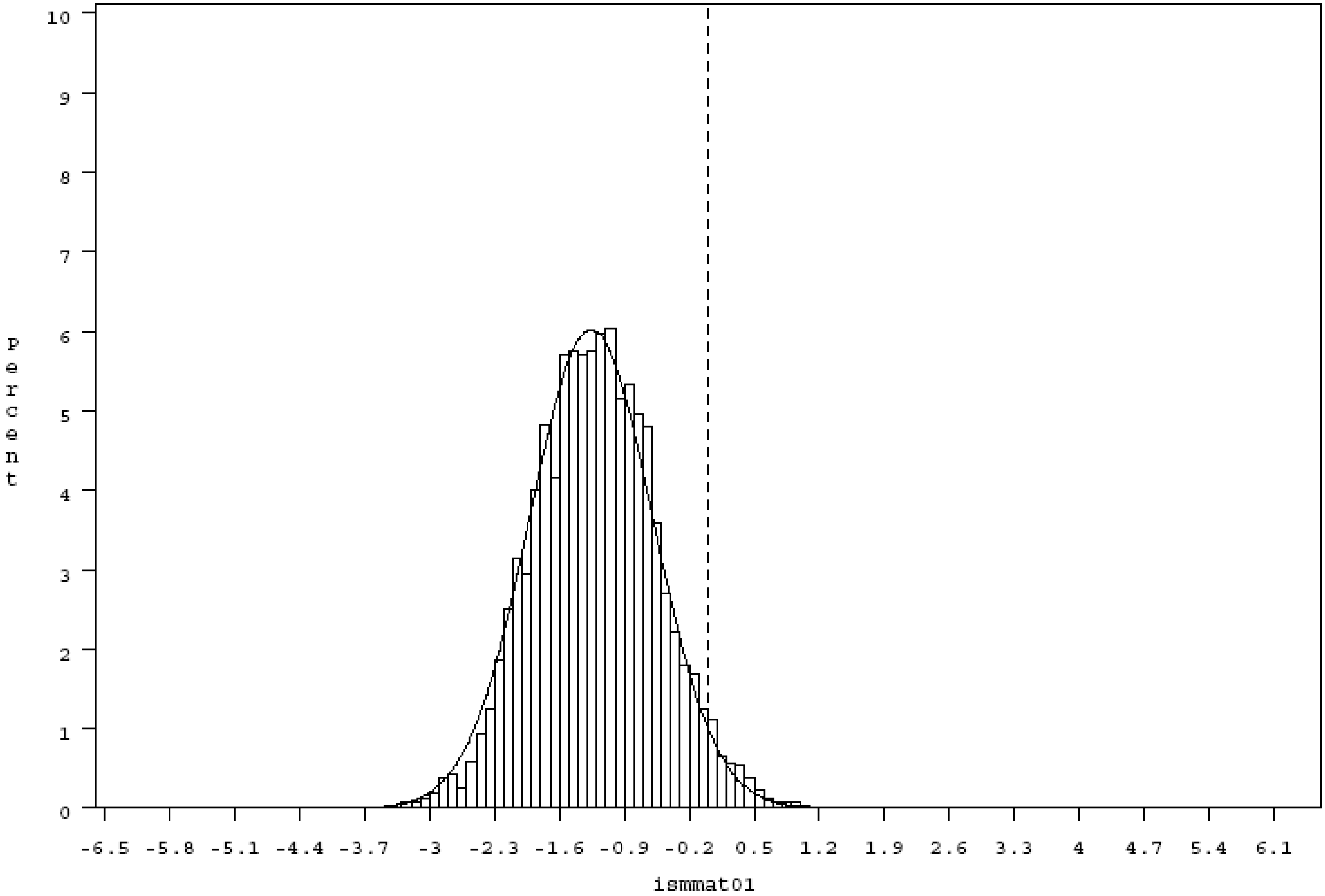
SOURCE: PARSCALE 3.1 RUN DATE: 04/18/2008 TIME: 09:16:21
PARPLOT SUN/UNIX Online Version 2.5

Are the Scaled Achievement Results Comparable? –cont.

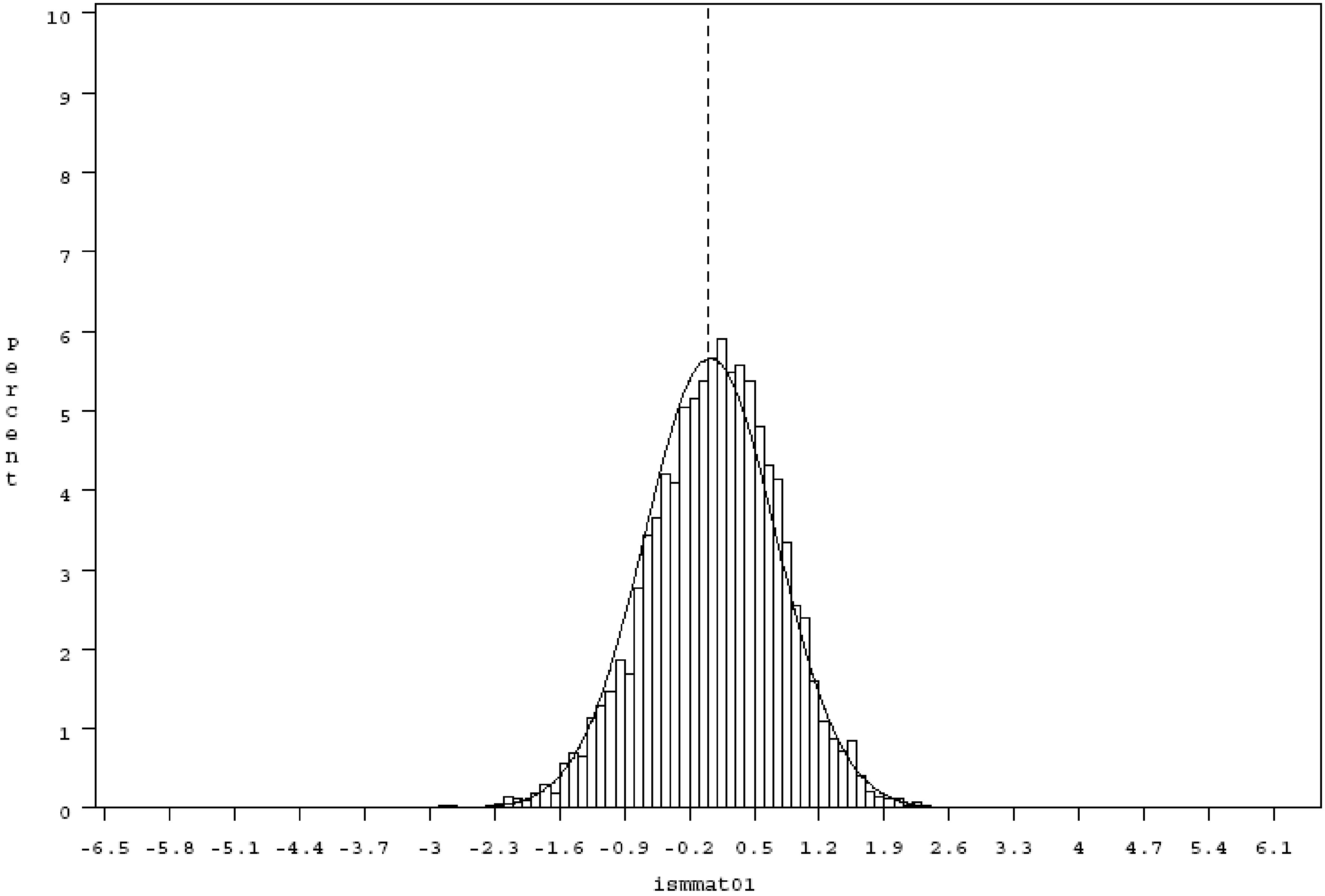
Now that we have item parameters – difficulty and discrimination – we can place students on the scale, i.e., produce student achievement scores (plausible values)

- Done separately for each country
- Done separately for each achievement scale, e.g., for TIMSS 2007, 30 scales
- Each achievement distribution for each country checked separately

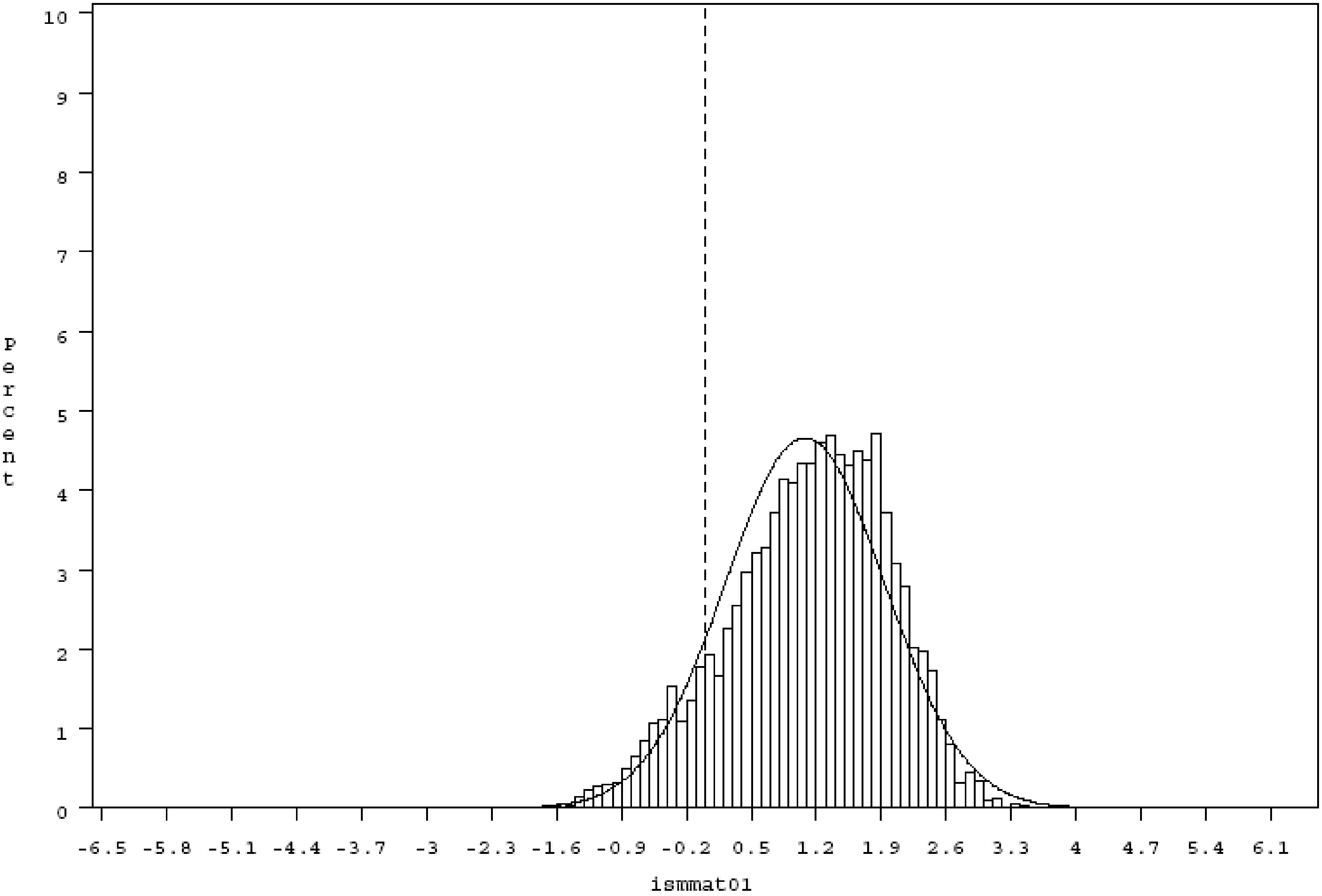
Ach. Hist. T07 – G8 Mathematics PV1 – [redacted]



Ach. Hist. T07 – G8 Mathematics PV1 – [redacted]



Ach. Hist. T07 – G8 Mathematics PV1 – [redacted]



Are the Scaled Achievement Results Comparable? –cont.

Scaling generally is very successful

- For most TIMSS and PIRLS countries, achievement score distributions are very satisfactory, and provide an excellent basis for analysis and reporting
- Plots provide a good quality control check

Are Achievement Results in the TIMSS & PIRLS International Reports Comparable?

- All reported statistics accompanied by standard errors
- Tests of statistical significance performed for many differences
  - Between countries, across assessments
- Annotations for countries not fully meeting sampling guidelines
- Achievement results presented in context

Why Do We Go to All This Trouble?

- To provide evidence of the comparative validity of the TIMSS & PIRLS achievement data
- So that the data can be trusted for important decision making based on comparisons among countries
- So that TIMSS & PIRLS data can form the basis for evidence-based policy making

Ensuring Comparative Validity

Quality Control in IEA Studies

Michael O. Martin and Ina V.S. Mullis

49th IEA General Assembly
Berlin, 6-9 October, 2008

IEA
TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College