Measuring Trends in TIMSS and PIRLS

Ina V.S. Mullis and Michael O. Martin

50th IEA General Assembly

Tallinn, 5-8 October, 2009

Trends in TIMSS and PIRLS

- Measuring trends fundamental to the TIMSS and PIRLS enterprise
- Trend data provide indispensable information for making policy decisions
  - Is the education system moving in the right direction?
  - Are students performing better on some parts of the curriculum than others?
  - Are some groups of students making better progress than others?

Trend Data from TIMSS and PIRLS

Achievement

- Distributions of student achievement – means and percentiles
- Percentages of students reaching International Benchmarks
- Percent correct on individual achievement items
- Relative progress in achievement across cohorts from 4th to 8th grades

Trend Data from TIMSS and PIRLS

Contexts for teaching and learning

- Curriculum – intended and taught
- School climate and resources
- Characteristics of the teaching workforce
- Characteristics of students
- Instructional practices
- Home environment

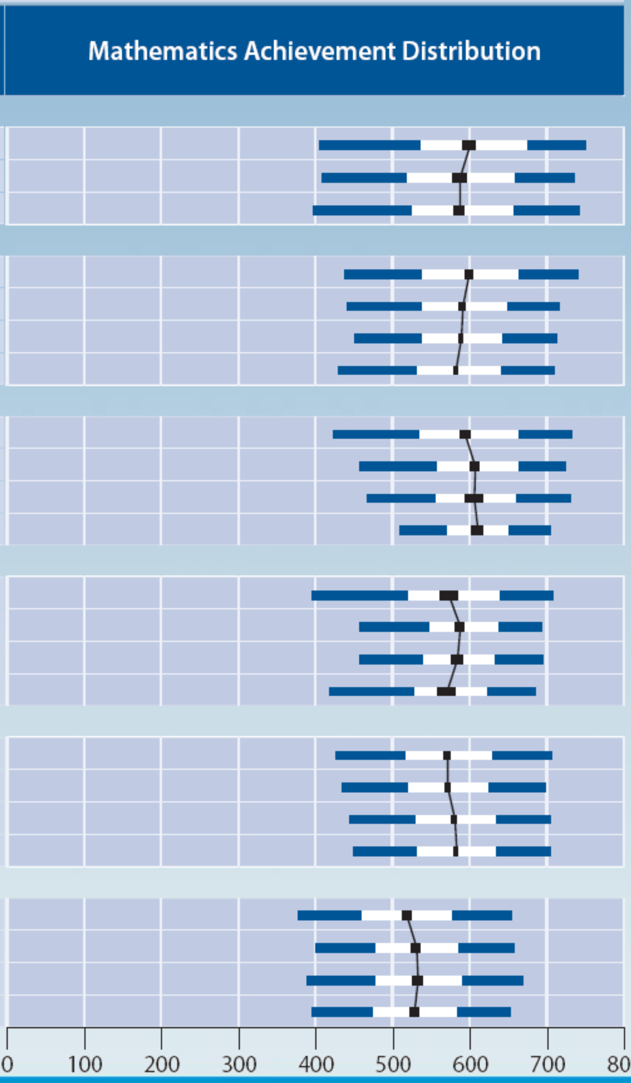
Excerpt from TIMSS 2007 International Report

Exhibit 1.3 **Trends in Mathematics Achievement – 1995 Through 2007 (Continued)**

TIMSS2007 Mathematics 8th Grade

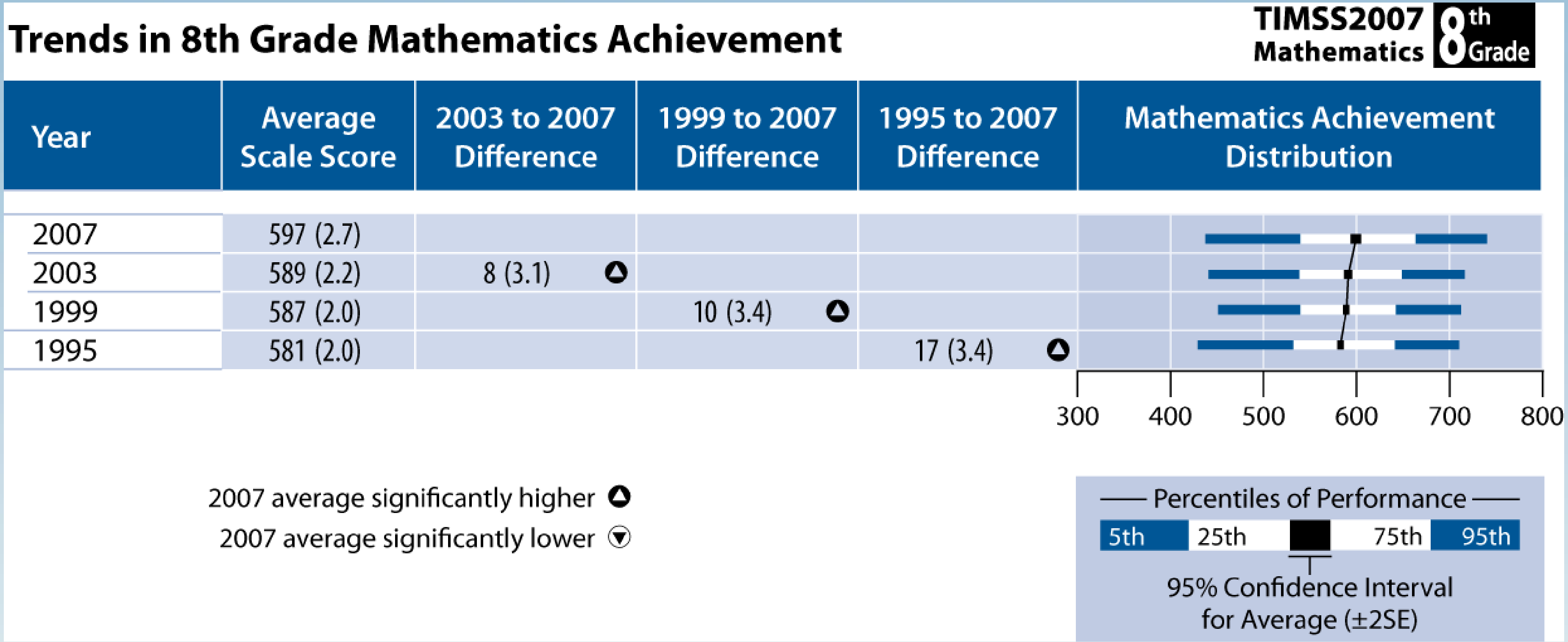
Country	Average Scale Score	2003 to 2007 Difference	1999 to 2007 Difference	1995 to 2007 Difference
Chinese Taipei				
2007	598 (4.5)			
2003	585 (4.6)	13 (6.4) ▲		
1999	585 (4.0)		13 (5.9) ▲	
Korea, Rep. of				
2007	597 (2.7)			
⋈ 2003	589 (2.2)	8 (3.1) ▲		
1999	587 (2.0)		10 (3.4) ▲	
1995	581 (2.0)			17 (3.4) ▲
Singapore				
2007	593 (3.8)			
2003	605 (3.6)	−13 (5.2) ▼		
1999	604 (6.3)		−12 (7.2)	
1995	609 (4.0)			−16 (5.6) ▼
Hong Kong SAR				
† 2007	572 (5.8)			
† 2003	586 (3.3)	−14 (6.6) ▼		
† 1999	582 (4.3)		−10 (7.2)	
1995	569 (6.1)			4 (8.4)
Japan				
2007	570 (2.4)			
2003	570 (2.1)	0 (3.1)		
1999	579 (1.7)		−9 (2.9) ▼	
1995	581 (1.6)			−11 (2.8) ▼
Hungary				
2007	517 (3.5)			
2 2003	529 (3.2)	−12 (4.7) ▼		
1999	532 (3.7)		−15 (5.0) ▼	
1995	527 (3.2)			−10 (4.7) ▼

Mathematics Achievement Distribution



SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007

Example for One Country - Korea



Trends in 8th Grade Mathematics Achievement

TIMSS2007 Mathematics 8th Grade

Year	Average Scale Score	2003 to 2007 Difference	1999 to 2007 Difference	1995 to 2007 Difference	Mathematics Achievement Distribution
2007	597 (2.7)				
2003	589 (2.2)	8 (3.1) ▲			
1999	587 (2.0)		10 (3.4) ▲		
1995	581 (2.0)			17 (3.4) ▲	

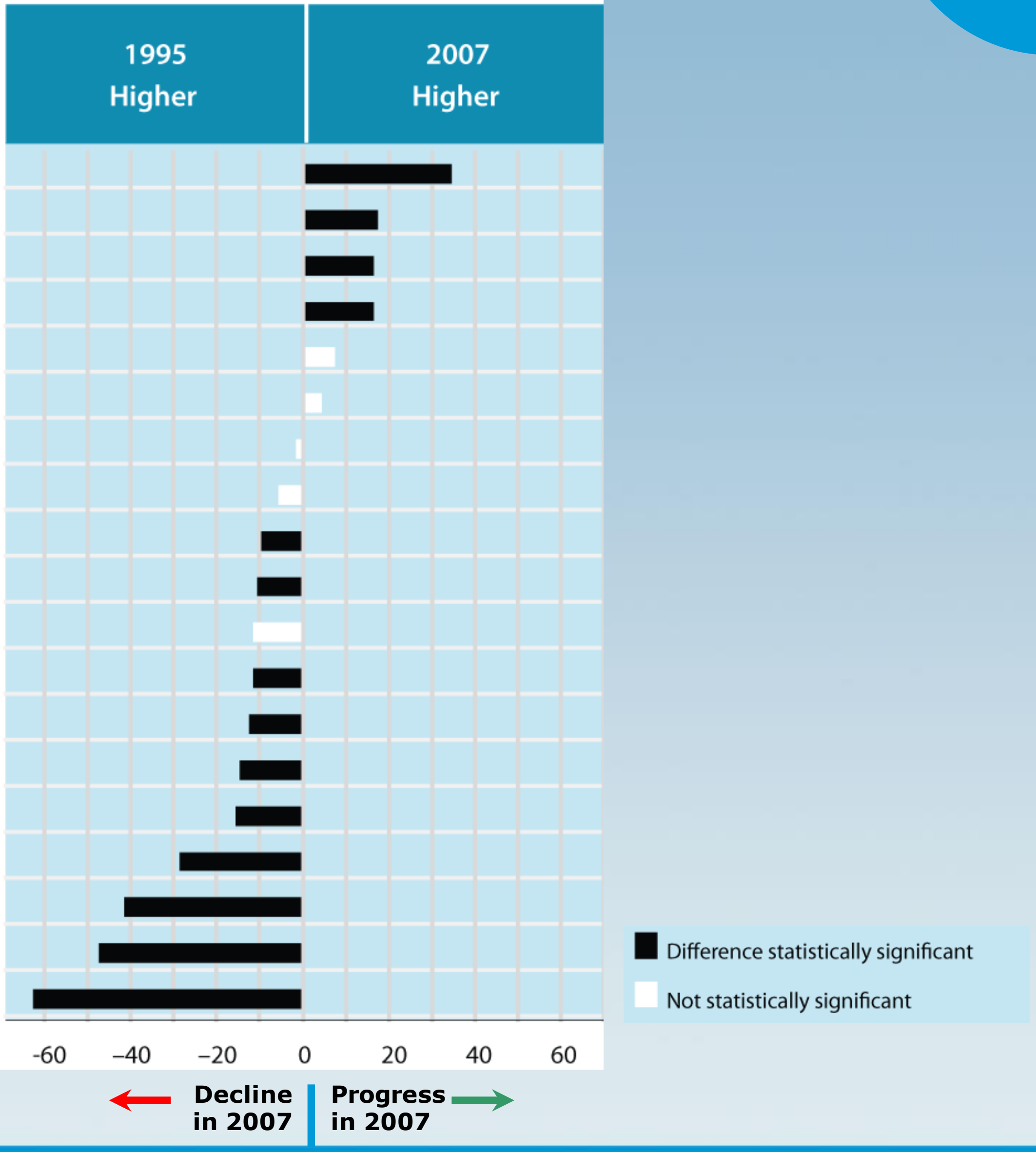
2007 average significantly higher ▲

2007 average significantly lower ▼

300
400
500
600
700
800
Percentiles of Performance
5th
25th
75th
95th
95% Confidence Interval for Average (±2SE)

Trends in Mathematics 1995–2007

Countries	1995 to 2007 Difference
Lithuania	34 (4.7)
Korea, Rep. of	17 (3.4)
England	16 (5.6)
United States	16 (5.5)
Slovenia	7 (3.6)
Hong Kong SAR	4 (8.4)
Cyprus	-2 (2.9)
Scotland	-6 (6.8)
Hungary	-10 (4.7)
Japan	-11 (2.8)
Russian Federation	-12 (6.7)
Romania	-12 (6.2)
Australia	-13 (5.4)
Iran, Islamic Rep. of	-15 (5.6)
Singapore	-16 (5.6)
Norway	-29 (2.9)
Czech Republic	-42 (5.1)
Sweden	-48 (4.8)
Bulgaria	-63 (7.6)

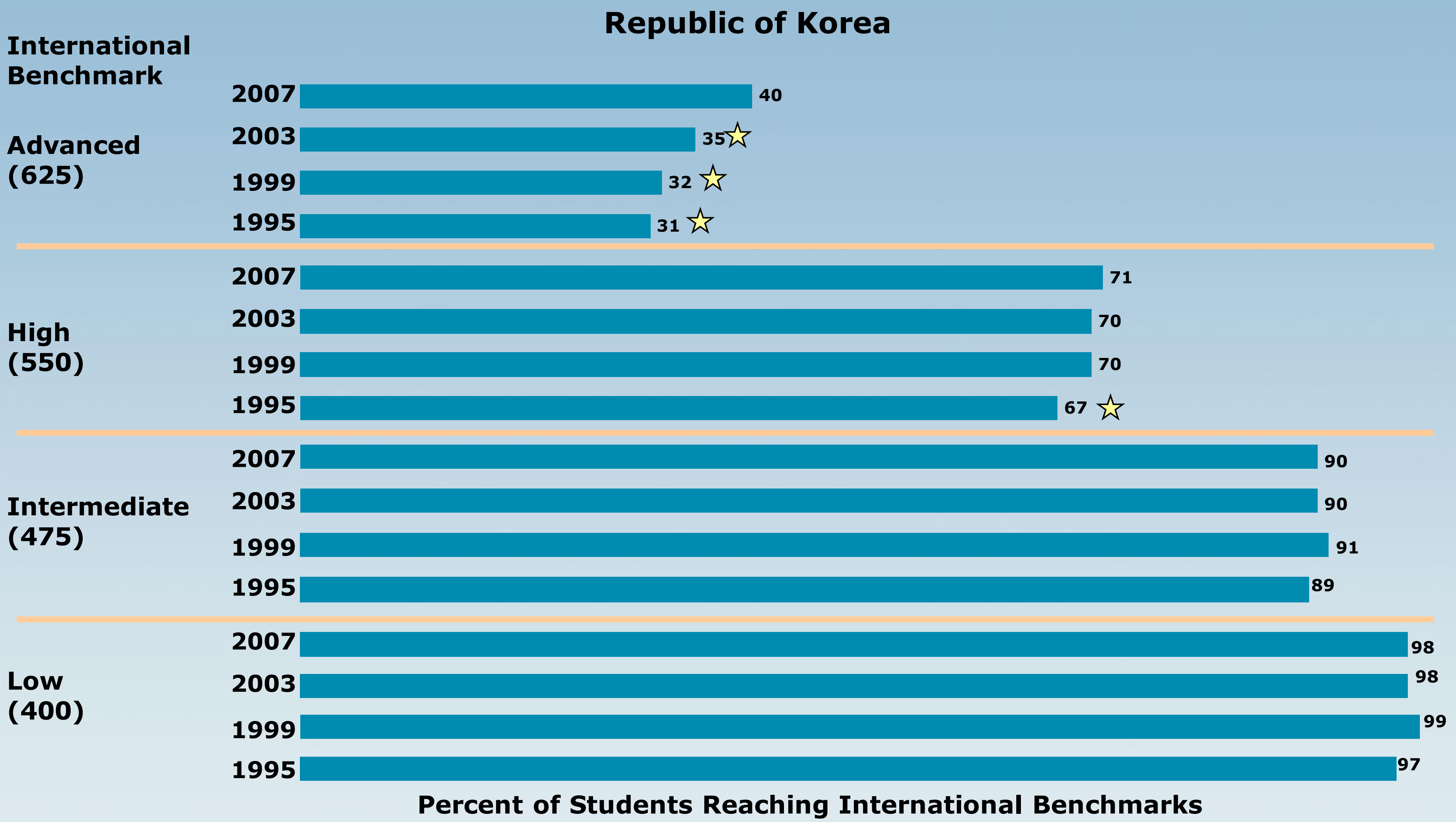


Monitoring Educational Reforms

- Adding another year of school – starting younger

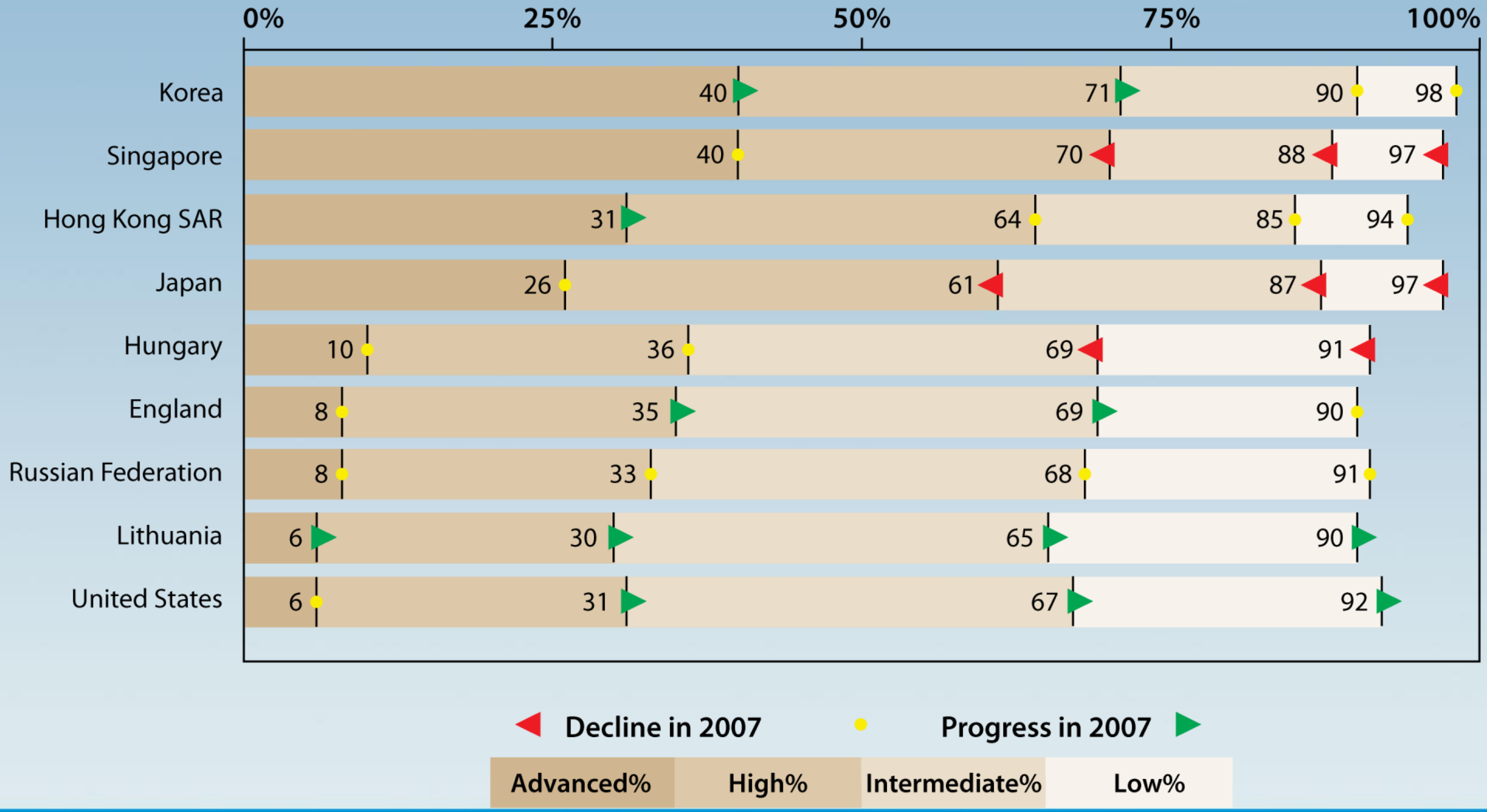
Slovenia – PIRLS		
	2001	2006
Average achievement	502	522
Years of schooling	3	3 or 4
Average age	9.8	9.8

Trends in Performance at the TIMSS International Benchmarks – Mathematics 8th Grade



Trends at TIMSS 2007 Benchmarks: 1995 to 2007

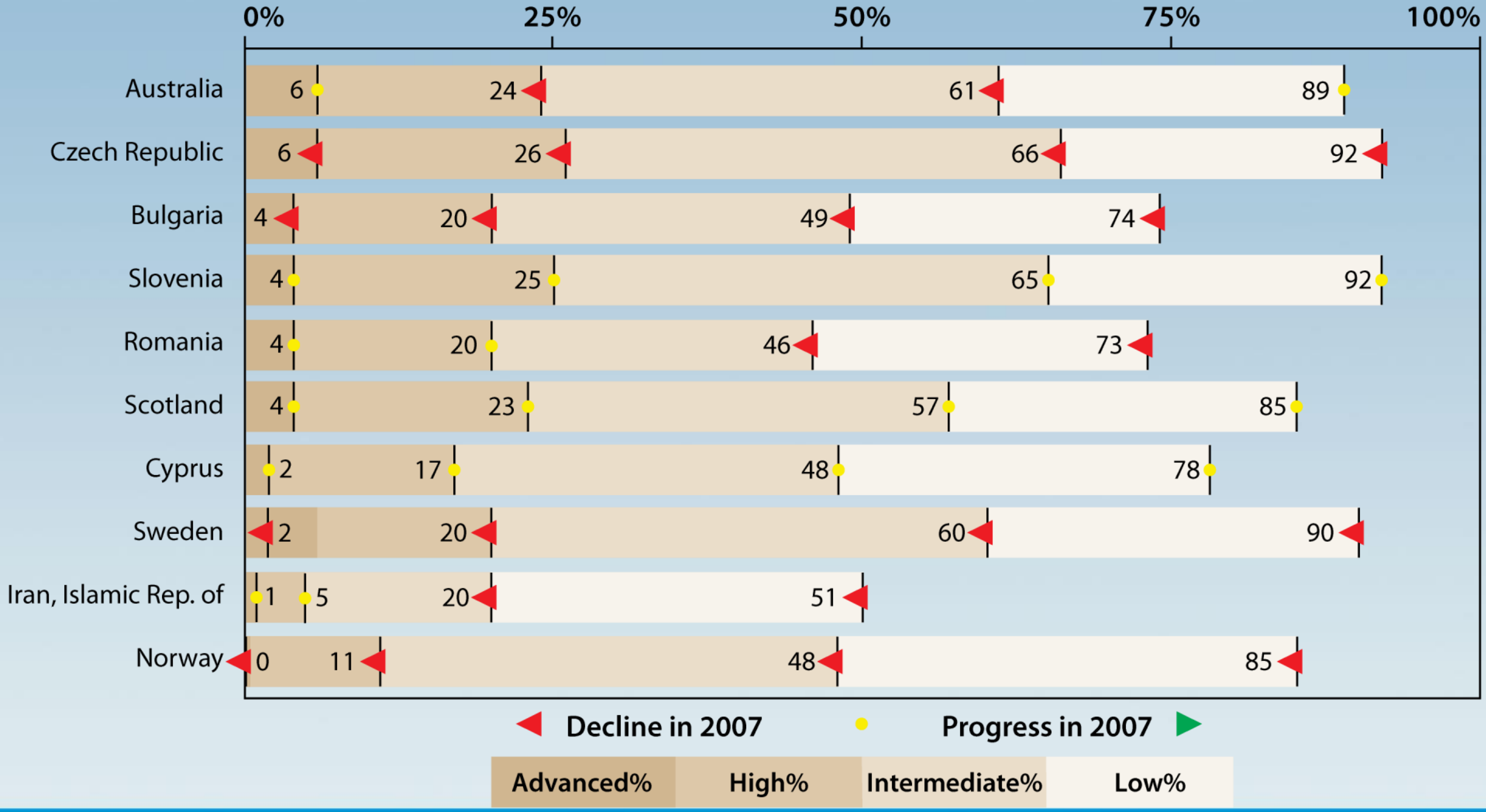
Eighth Grade Mathematics



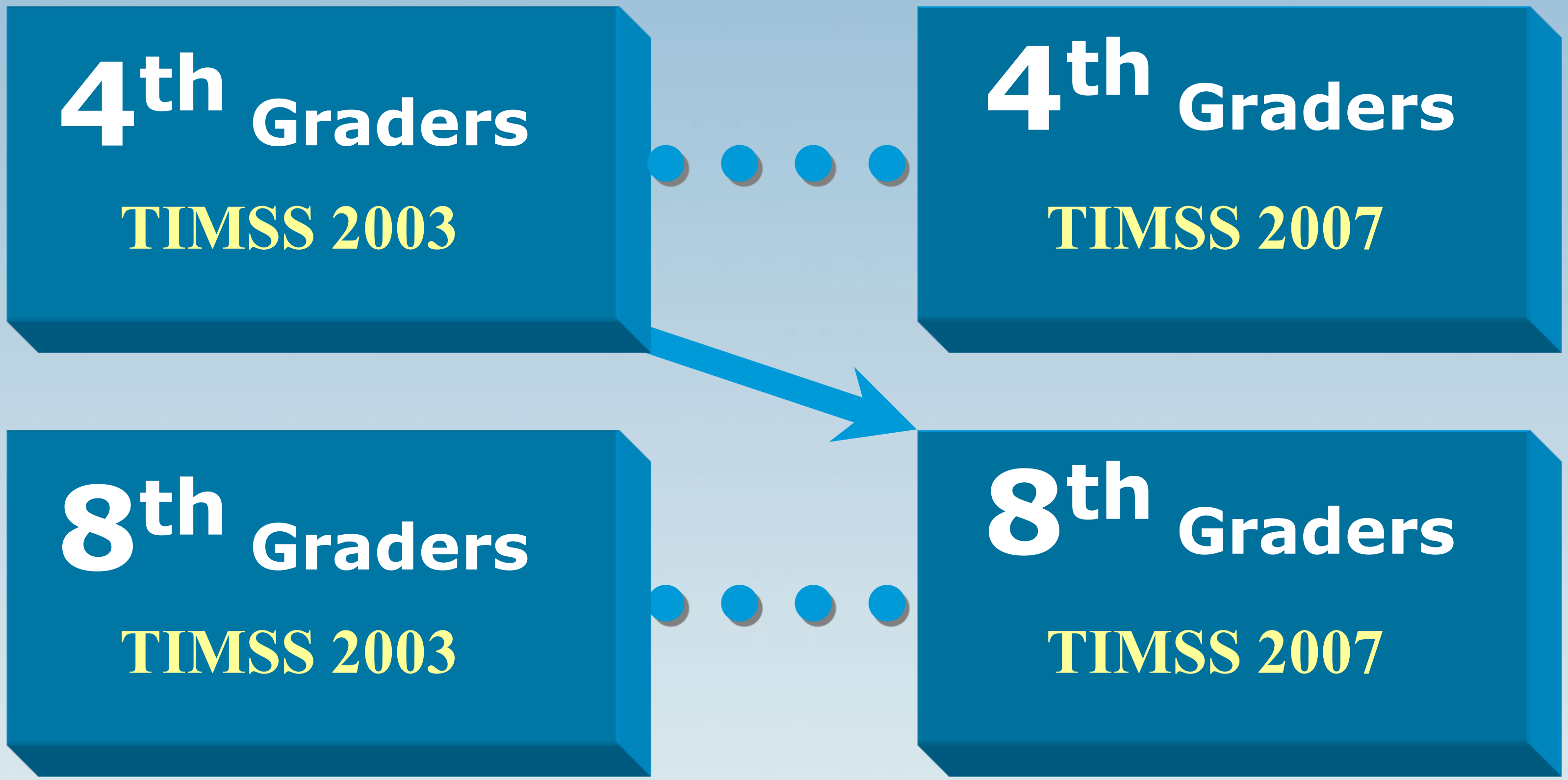
TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College

Trends at TIMSS 2007 Benchmarks: 1995 to 2007

Eighth Grade Mathematics (cont.)



Cohort Comparison Over Time



4th Graders

TIMSS 2003

4th Graders

TIMSS 2007

8th Graders

TIMSS 2003

8th Graders

TIMSS 2007

Measuring Trends Is Challenging!

Part 1

Trend measurement always difficult methodologically

TIMSS and PIRLS methodology based on ETS innovations for NAEP

History of experience with NAEP

Measuring Trends Is Challenging! Evolution of Methodology

- State of the art, circa 1950 – test equating (e.g., SAT in the U.S.)
- State of the art, circa 1970 – NAEP in the U.S. – equivalent populations, median p-values for groups
  - Item based, not based on scores for individual students

Measuring Trends Is Challenging!

- Using median p-values problematic
  - overall country performance improved, while it declined in two of four regions – North and South (migration northwards)
- Exhaustive examination of measures of central tendency
- State of the art, circa 1975 – average p-values to be more robust against demographic shifts

Measuring Trends Is Challenging!

- Using average p-values problematic for trends
  - Cannot change assessment items from cycle to cycle
  - As items are released with each cycle, basis for trend becomes less reliable – fewer and fewer items
- State of the art, circa 1985 – IRT scaling, not dependent on same items

Measuring Trends Is Challenging!

- Using only IRT problematic
  - Saw regression to mean for subpopulations
  - IRT not dependent on assessing same items from cycle to cycle, but does estimate student performance from responses to items
  - IRT requires many items for reliable estimation of student performance...

Measuring Trends Is Challenging!

- State of the art, circa 1995 – IRT with “plausible values” methodology
- Still, the more items, the more reliable the estimates
- TIMSS and PIRLS apply the methodology of IRT with many items to measure trends – which also brings challenges

Measuring Trends Is Challenging!

Part 2

Complications of measuring change in a changing environment

…especially across 60 countries

** Important Lesson **

When measuring change, do not change the measure.

Albert E. Beaton
John W. Tukey

** Extension to Important Lesson **

When measuring change, you sometimes have to change the measure because the world is changing.

Ina V.S. Mullis

Michael O. Martin

Changing World

- Shifting demographics
  - Immigration and emigration (within and across countries)
  - Countries unify or split up (Germany, Yugoslavia)
  - Increasing school enrollments

Changing World

- Methodological advances
  - IRT scaling
  - Image scoring
  - Web based assessment
  - Tailored or targeted testing

Changing World

- Education policies
  - Age students start school (Australia, Slovenia, Russian Federation, Norway)
- Policies for greater inclusion
  - Accommodations for students with learning disabilities and second-language learners
  - Countries adding additional language groups (Latvia, Israel)

Changing World -cont

- Curriculum frameworks
  - Calculator use; performance assessment
- Catastrophic events
  - Natural disasters (earthquakes, hurricanes, tsunamis)
  - Tragic incidents (Lebanon, Palestine)

Changing World -cont

- Contexts and situations for items
  - "Boombox" to "iPhone"
- Changes affecting individual items
  - Graphing calculators in TIMSS Advanced
  - Stimulus materials becoming dated, or too familiar

Assessments Need to Evolve

If don’t change the measure to some extent

- May be making changes anyway since the contexts have changed
- Cannot stay at the forefront of providing high-quality measures
- Cannot provide information on topics policymakers and educators find important

Assessments Need to Evolve

What to do in a changing world?

- Redo previous cycles to match
  - Rescaled 1995
- Bridge study
  - Some students previous procedure and some new
- Different configurations for trend than new
  - Broadening inclusion (e.g., additional language groups)

Assessments Need to Evolve

The evolving design used in TIMSS and PIRLS

- ⅓, ⅓, ⅓ model
- Items from three cycles ago are released and replaced with new
- For 2011, all 1995 and 1999 items released
  - ⅓ will be from 2 cycles ago (e.g., 2003)
  - ⅓ will be from 1 cycle ago (e.g., 2007)
  - ⅓ will be new for 2011

Assessments Need to Evolve

TIMSS and PIRLS resolve tension between

- Maintaining continuity with the past procedures
- Maintaining current relevance in a changing context

Keep Present as Point of Reference

- Link backwards while moving forwards
- Keep substantial portions of assessment constant (e.g., 3 literary and 3 informational passages)
- Introduce new aspects carefully and gradually (e.g., 2 literary and 2 informational passages)
- Plan as trend assessment

In Summary, Measuring Trends

- Is fundamental to educational improvement
- Is extremely complicated
- Needs to use highest methodological standards
- Needs to be done with common sense

Part 3

How TIMSS and PIRLS Meet the Challenges of Measuring Trends

Linking Assessments Over Time in TIMSS and PIRLS

To measure trends in achievement effectively,

- We must have **data** from **successive assessments** on a **common scale**
- TIMSS and PIRLS do this using IRT scaling (with adaptations for large-scale assessment – developed by U.S. NAEP)

IRT Scaling for Measuring Trends

- **Item Response Theory** – useful for measuring trends because it uses items with **known properties** to estimate to students' ability
- The most important property is the **difficulty** of the items – but other properties also
- If we know these item properties are for **successive assessments**, we can use them to estimate students' ability from one assessment to the next, i.e., measure trends

Linking Assessment Data in TIMSS and PIRLS

TIMSS and PIRLS administer assessments repeatedly:

- TIMSS – 1995, 1999, 2003, 2007, 2011…
- PIRLS – 2001, 2006, 2011…

…and report achievement results on common scales

How do we do this?

Linking Assessment Data in TIMSS and PIRLS

- We include **common items** in adjacent assessment cycles, as well as items unique to each cycle
- We use IRT scaling to link the data to a **common scale**
- All we need to do this is to know the **properties** of the items – both the common items and items unique to the assessment

Important Properties of Items

In IRT, the properties of items are known as **item parameters**

- TIMSS and PIRLS use a 3-parameter IRT approach
- Most important parameter: **item difficulty**
- For added accuracy:
  - Parameter for **item discrimination**
  - Parameter for **guessing** by low ability students on multiple-choice items

How Do We “Know” the Properties of the Items?

- Although we have been talking about “known properties,” in fact the parameters of the items are **not** known to begin with
- so item parameters must be **estimated** from the assessment data, building from cycle to cycle
  - Process known as **concurrent calibration**

Item Calibration - Estimating Item Parameters

Generally:

Two-step procedure:

1. Use the student response data to provide estimates of the item parameters
2. Then, use these item parameters to estimate student ability

For trend measurement:

- Repeat with each assessment

IRT Scaling in TIMSS for Trends

Achievement scales established with **TIMSS 1995** data

1. Item Calibration – estimated item parameters from 1995 data
   - Used all items, treated all countries equally
2. Student scoring – using item parameters, gave all 1995 students achievement scores
   - Set achievement scales to have a mean of 500 and a standard deviation of 100

IRT Scaling in TIMSS for Trends
Example: Grade 8 mathematics

In **TIMSS 1999**, we needed to link to the data from 1995 to measure trends. To do this, we needed to know the properties of our items

We had two key components:

- **Items** from 1995 and 1999, one third in common
- **Countries** that participated in 1995 and 1999, 25 in both

IRT Scaling in TIMSS for Trends

Calibrating TIMSS 1995 and 1999 items

	1995 Items only	Common Items	1999 Items only
1995 Data 25,000	⅔ 111 items	⅓ 48 items	
1999 Data 25,000		⅓ 48 items	⅔ 115 items

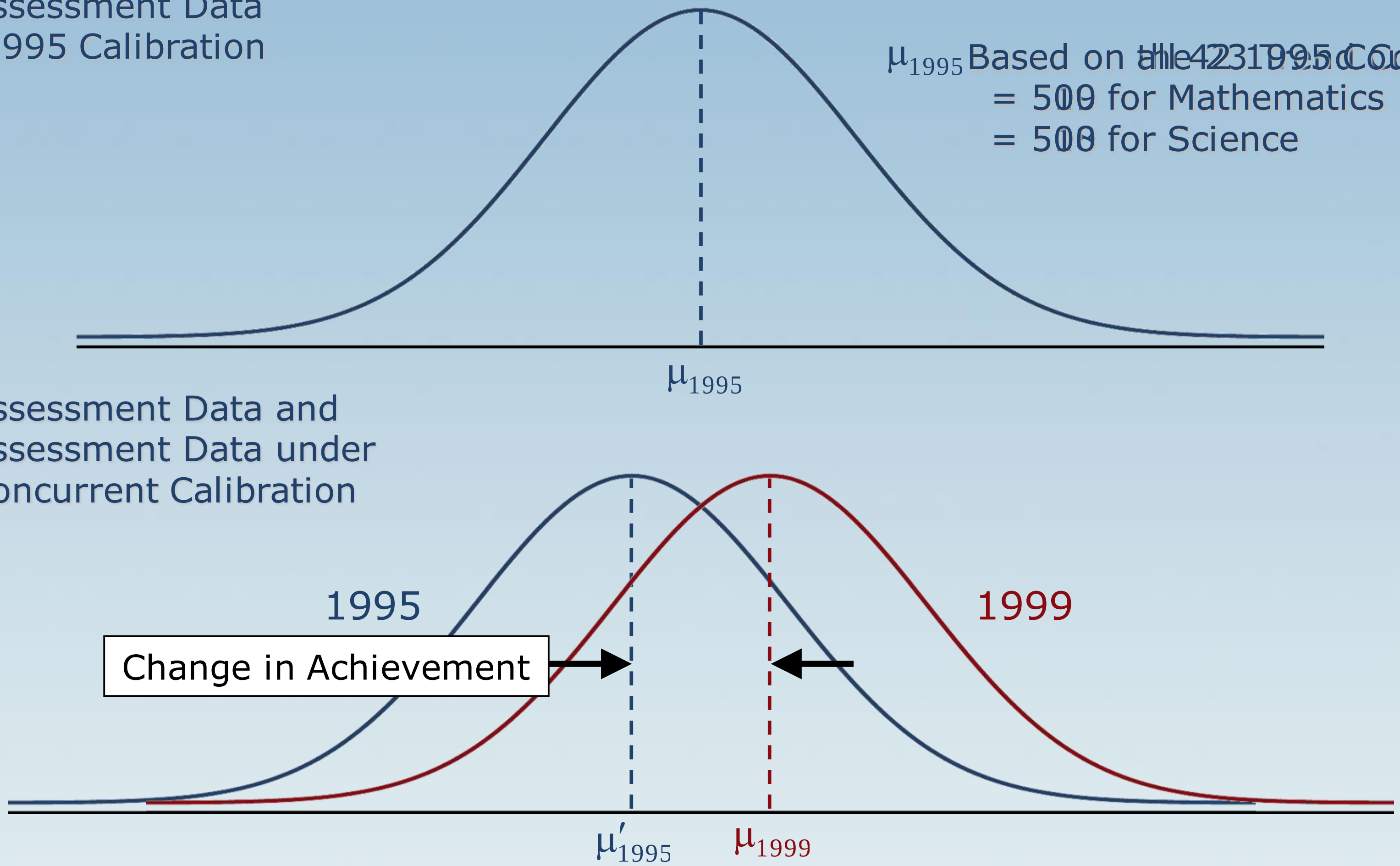
IRT Scaling in TIMSS for Trends

	1995 Items only	Common Items	1999 Items only
1995 Calibration	111 + 48 = 159 items		
1995-1999 Concurrent calibration	111 + 48 + 115 = 274 items		

TIMSS 1995 Items now have two sets of parameters

- but not on the same scale

Placing the 1999 Scores on the 1995 Metric



1995 Assessment Data
under 1995 Calibration

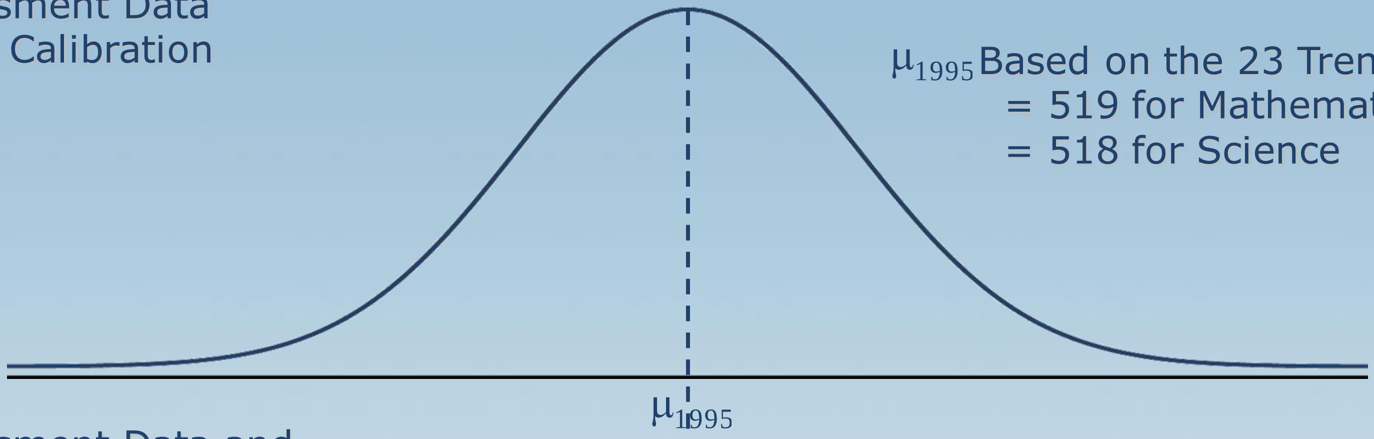
μ_{1995} Based on all 42 1995 Countries
= 500 for Mathematics
= 500 for Science

μ_{1995} Based on the 23 Trend Countries
= 519 for Mathematics
= 518 for Science

1995 Assessment Data and
1999 Assessment Data under
1999 Concurrent Calibration

Placing the 1999 Scores on the 1995 Metric

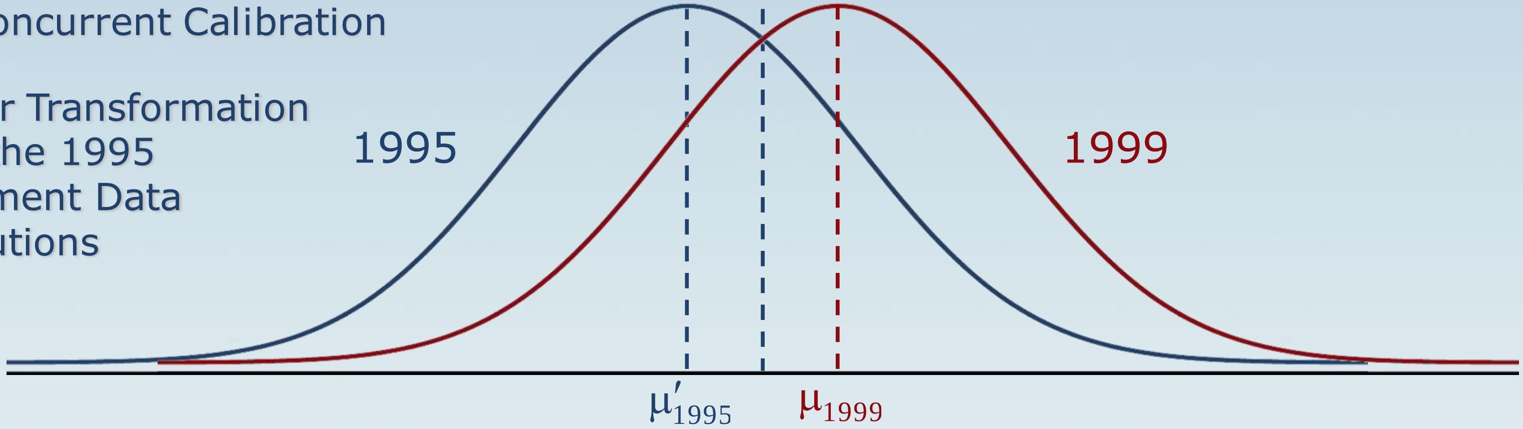
1995 Assessment Data
under 1995 Calibration



μ_{1995} Based on the 23 Trend Countries
= 519 for Mathematics
= 518 for Science

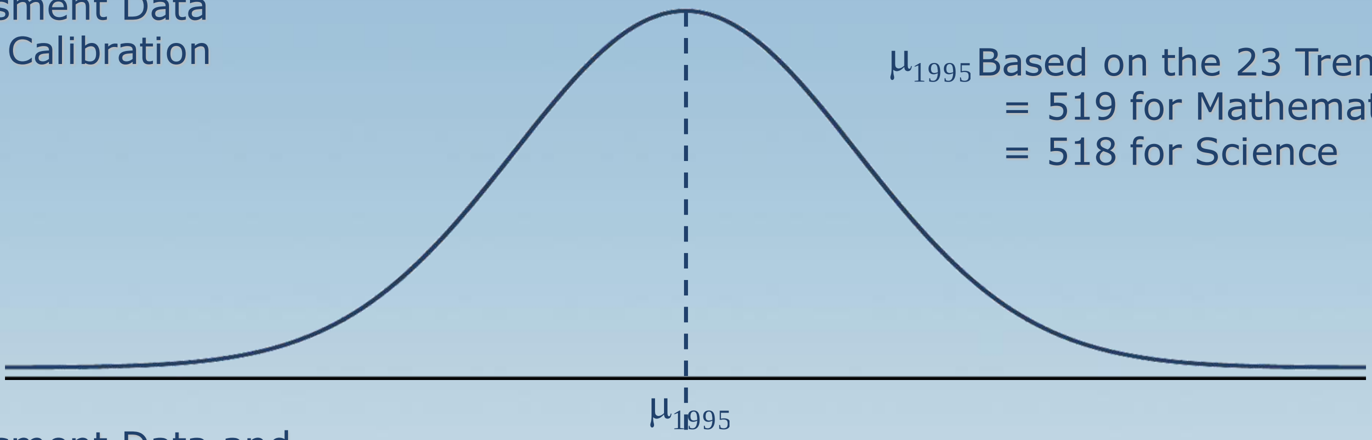
1995 Assessment Data and
1999 Assessment Data under
1999 Concurrent Calibration

A Linear Transformation
Aligns the 1995
Assessment Data
Distributions



Placing the 1999 Scores on the 1995 Metric

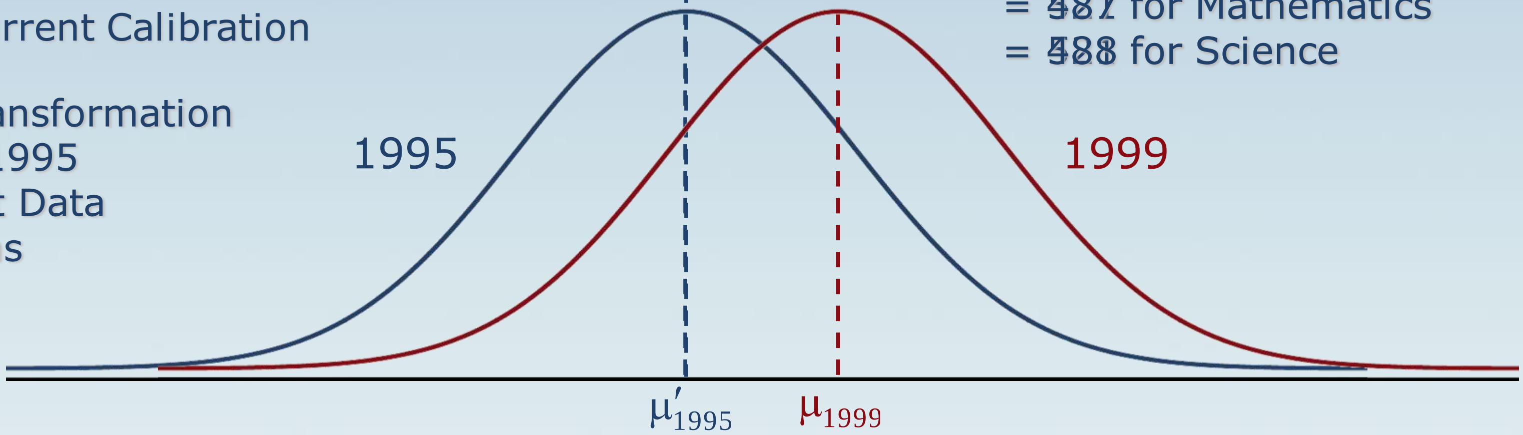
1995 Assessment Data
under 1995 Calibration



μ_{1995} Based on the 23 Trend Countries
= 519 for Mathematics
= 518 for Science

1995 Assessment Data and
1999 Assessment Data under
1999 Concurrent Calibration

A Linear Transformation
Aligns the 1995
Assessment Data
Distributions



μ_{1999} Based on all 38 1999 Countries
= 487 for Mathematics
= 488 for Science

Based on the 23 Trend Countries
= 521 for Mathematics
= 521 for Science

IRT Scaling in TIMSS for Trends

We check our linking:

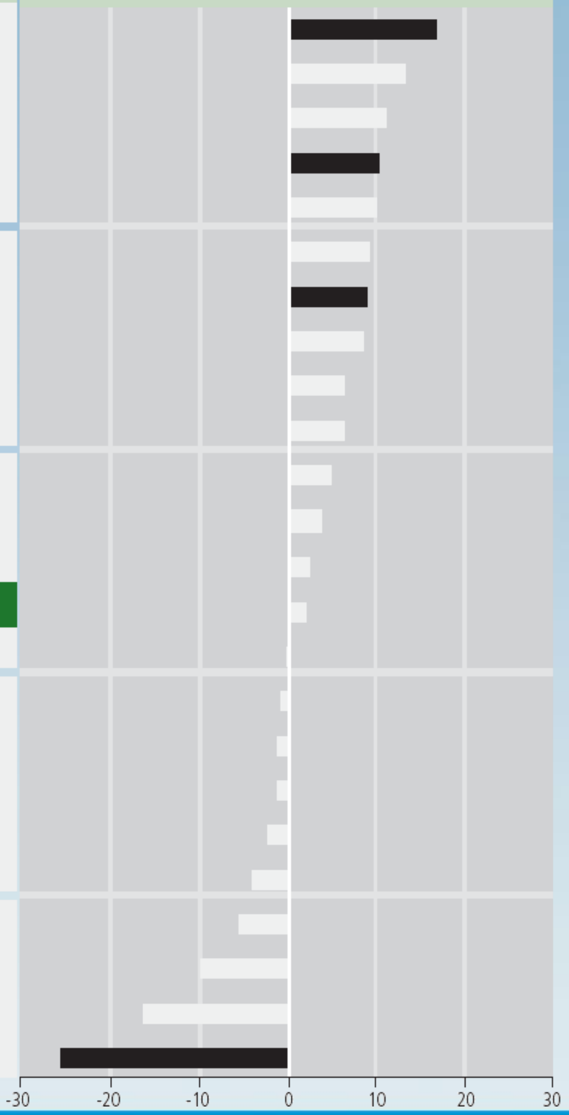
1. We already have scores for **1995** countries using parameters from **1995** item calibration
2. We estimate **new** scores for same 1995 countries using parameters from the **concurrent 1995/1999** calibration

Because the **same student data** are used, the scores should match, and they do, within sampling error

TIMSS 1999 Mathematics

	1995 Average Scale Score	1999 Average Scale Score	1995-1999 Difference
Latvia (LSS)	488 (3.6)	505 (3.4)	17 (5.0)
Hong Kong, SAR	569 (6.1)	582 (4.3)	13 (7.5)
Netherlands	529 (6.1)	540 (7.1)	11 (9.5)
Canada	521 (2.2)	531 (2.5)	10 (3.2)
Lithuania	472 (4.1)	482 (4.3)	10 (6.1)
United States	492 (4.7)	502 (4.0)	9 (6.2)
Cyprus	468 (2.2)	476 (1.8)	9 (2.9)
Belgium (Flemish)	550 (5.9)	558 (3.3)	8 (6.8)
Korea, Rep. of	581 (2.0)	587 (2.0)	6 (2.8)
Australia	519 (3.8)	525 (4.8)	6 (6.1)
Hungary	527 (3.2)	532 (3.7)	5 (4.9)
Iran, Islamic Rep.	418 (3.9)	422 (3.4)	4 (5.2)
Russian Federation	524 (5.3)	526 (5.9)	2 (8.0)
International Avg. §	**519 (0.9)**	**521 (0.9)**	**2 (1.3)**
Slovak Republic	534 (3.1)	534 (4.0)	0 (4.9)
Slovenia	531 (2.8)	530 (2.8)	-1 (3.9)
Romania	474 (4.6)	472 (5.8)	-1 (7.4)
England	498 (3.0)	496 (4.1)	-1 (5.2)
Japan	581 (1.6)	579 (1.7)	-2 (2.2)
Singapore	609 (4.0)	604 (6.3)	-4 (7.4)
Italy	491 (3.4)	485 (4.8)	-6 (6.0)
New Zealand	501 (4.7)	491 (5.2)	-10 (7.1)
Bulgaria	527 (5.8)	511 (5.8)	-16 (8.2)
Czech Republic	546 (4.5)	520 (4.2)	-26 (6.1)

Difference in Average Achievement Between 1995 and 1999



SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

IRT Scaling in TIMSS for Trends

Similar approach for TIMSS 1999 and 2003:

	1995/1999 Items only	Common Items (95,99,03)	2003 Items only
1999 Data	84 items	79 items	
2003 Data		79 items	115 items

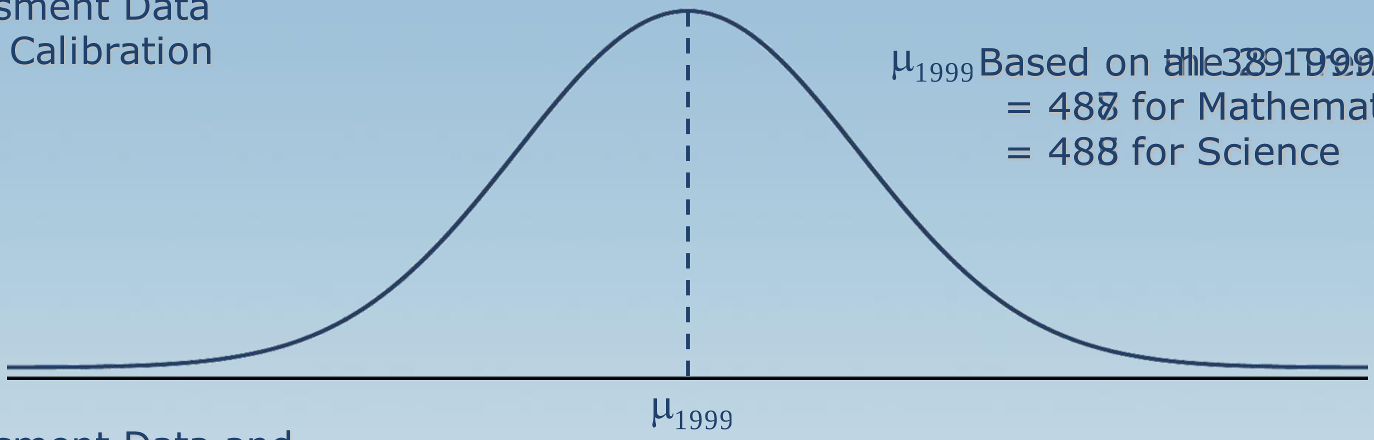
IRT Scaling in TIMSS for Trends

	1995/1999 Items only	Common Items (95,99,03)	2003 Items only
1995/1999 Calibration	84 + 79 = 163 items		
1999/2003 Concurrent calibration	84 + 79 + 115 = 278 items		

TIMSS 1999 Items now have two sets of parameters – but not on the same scale

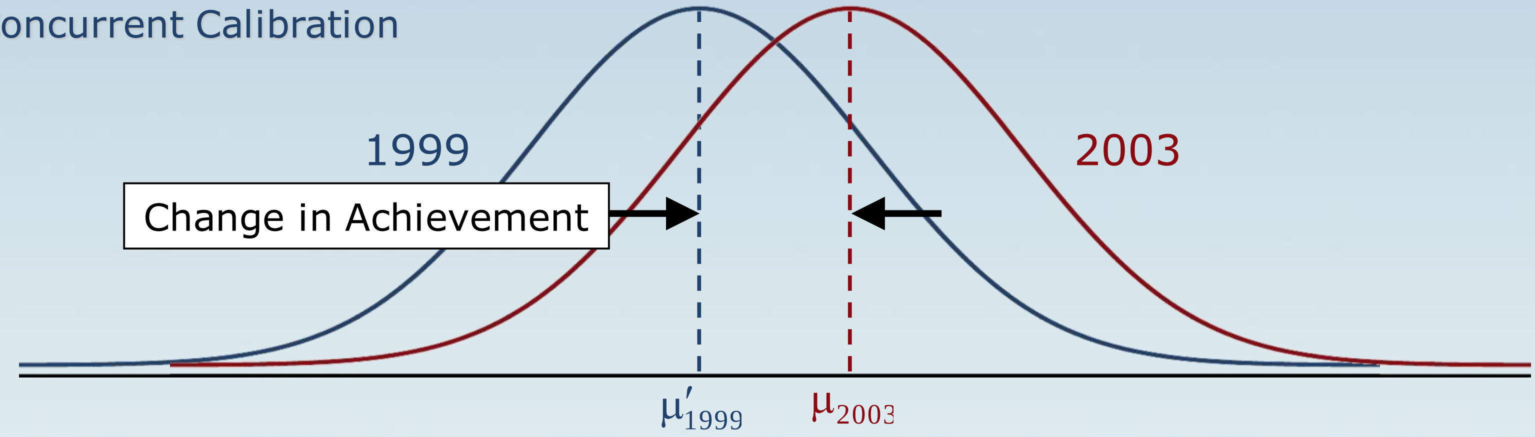
Placing the 2003 Scores on the 1995 Metric

1999 Assessment Data
under 1999 Calibration



μ_{1999} Based on all 38 1999 Countries
= 487 for Mathematics
= 488 for Science

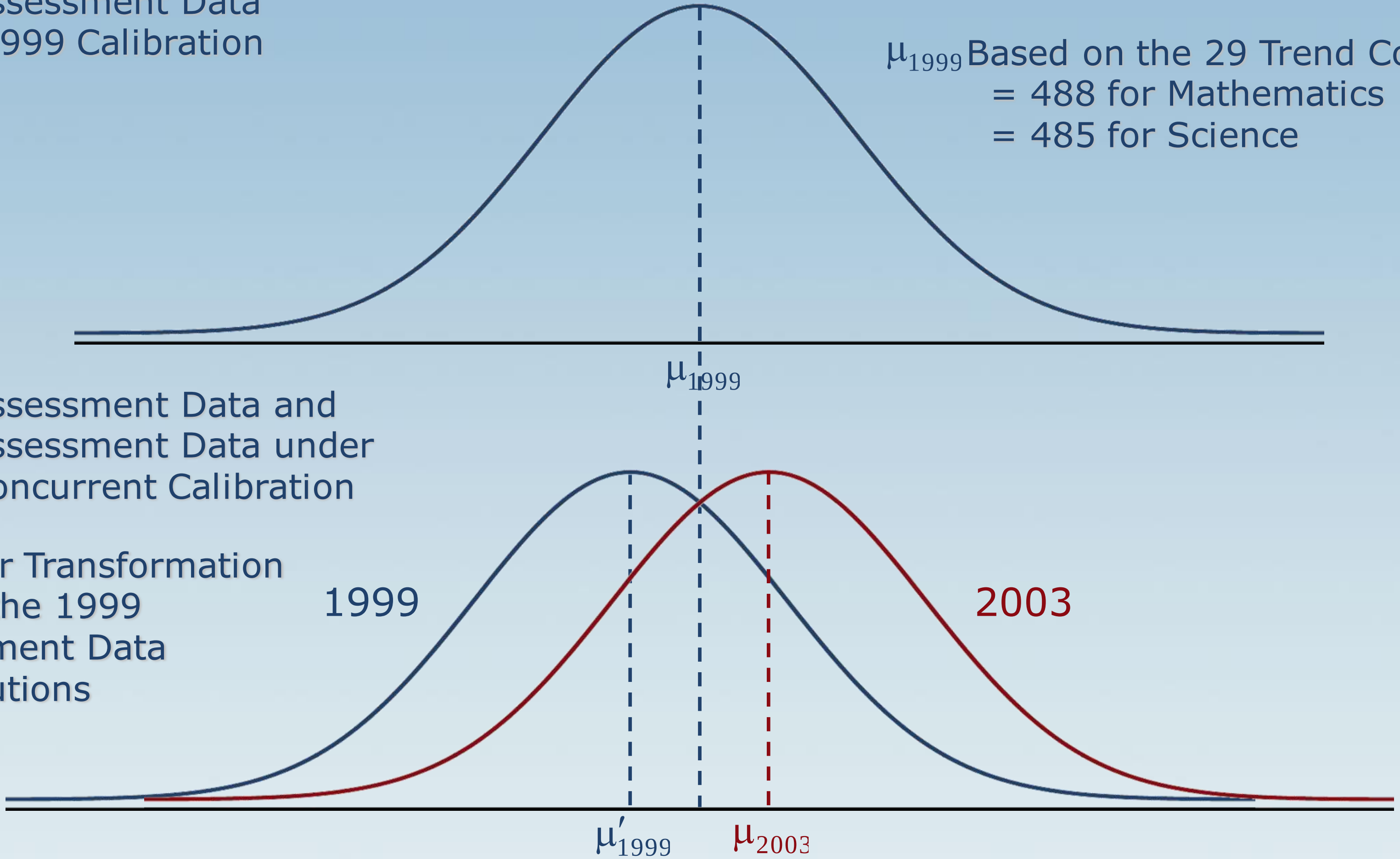
1999 Assessment Data and
2003 Assessment Data under
2003 Concurrent Calibration



Placing the 2003 Scores on the 1999 Metric

1999 Assessment Data
under 1999 Calibration

μ_{1999} Based on the 29 Trend Countries
= 488 for Mathematics
= 485 for Science



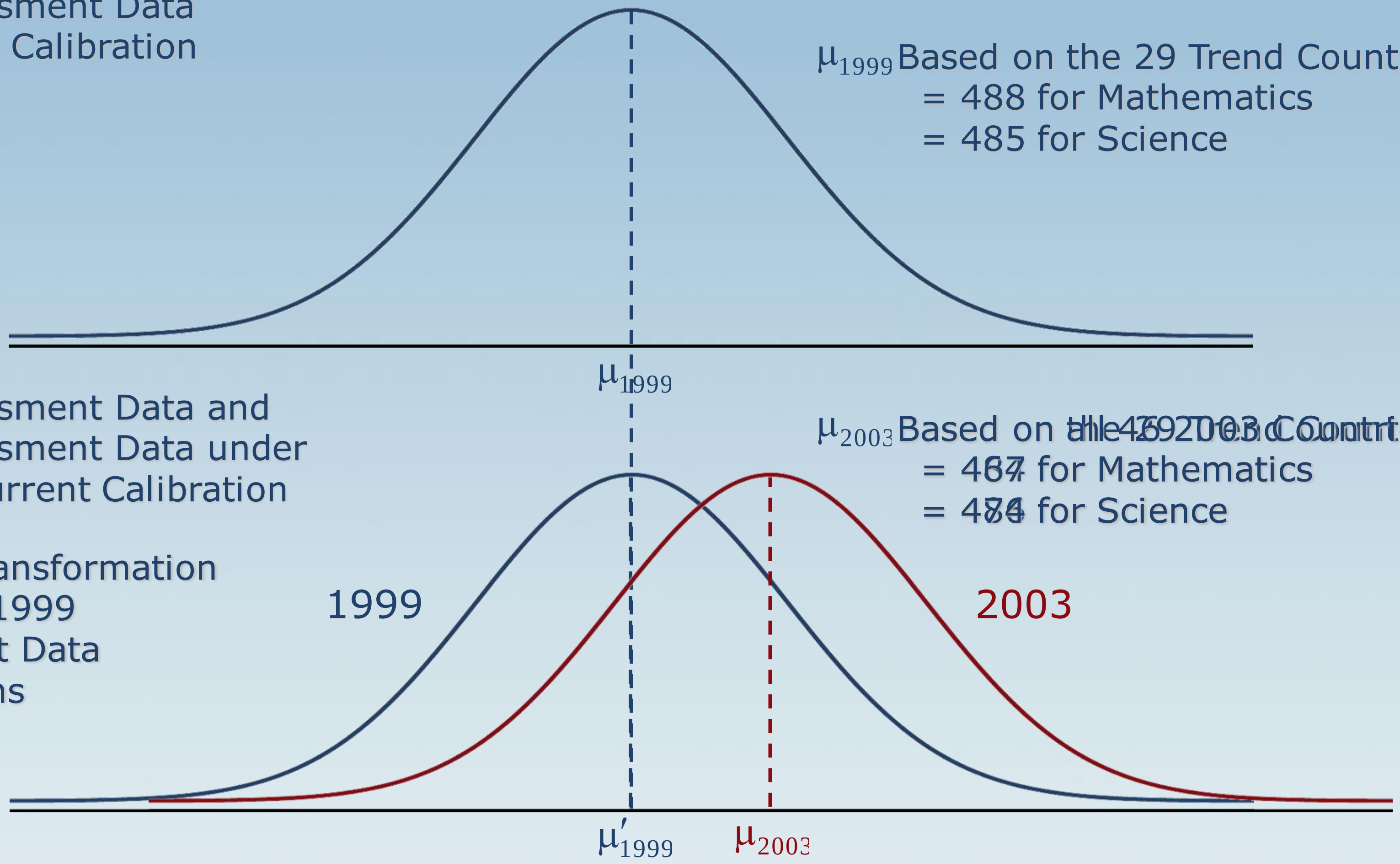
1999 Assessment Data and
2003 Assessment Data under
2003 Concurrent Calibration

A Linear Transformation
Aligns the 1999
Assessment Data
Distributions

Placing the 2003 Scores on the 1999 Metric

1999 Assessment Data
under 1999 Calibration

μ_{1999} Based on the 29 Trend Countries
= 488 for Mathematics
= 485 for Science



1999 Assessment Data and
2003 Assessment Data under
2003 Concurrent Calibration

μ_{2003} Based on all 46 2003 Countries
= 467 for Mathematics
= 474 for Science

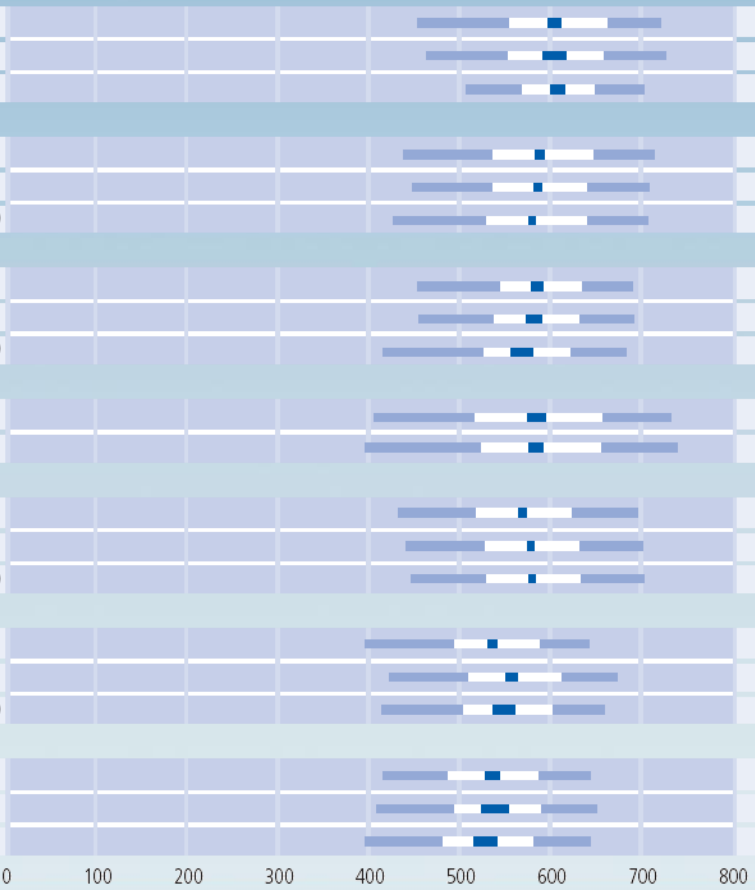
A Linear Transformation
Aligns the 1999
Assessment Data
Distributions

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

Exhibit 1.3: **Trends in Mathematics Achievement**

MATHEMATICS Grade 8

Countries	Average Scale Score	1999 to 2003 Difference	1995 to 2003 Difference	Mathematics Achievement Distribution	Average Age
Singapore					
2003	605 (3.6)				14.3
1999	604 (6.3)	1 (7.2)			14.4
1995	609 (4.0)		-3 (5.4)		14.5
Korea, Rep. of					
2003	589 (2.2)				14.6
1999	587 (2.0)	2 (2.9)			14.4
1995	581 (2.0)		8 (3.0) ▲		14.2
Hong Kong, SAR					
2003	586 (3.3)				14.4
1999	582 (4.3)	4 (5.4)			14.2
1995	569 (6.1)		17 (7.0) ▲		14.2
Chinese Taipei					
2003	585 (4.6)				14.2
1999	585 (4.0)	0 (6.0)			14.2
Japan					
2003	570 (2.1)				14.4
1999	579 (1.7)	-9 (2.6) ▼			14.4
1995	581 (1.6)		-11 (2.6) ▼		14.4
Belgium (Flemish)					
2003	537 (2.8)				14.1
1999	558 (3.3)	-21 (4.1) ▼			14.1
1995	550 (5.9)		-13 (6.5) ▼		14.1
Netherlands					
2003	536 (3.8)				14.3
1999	540 (7.1)	-4 (8.1)			14.2
1995	529 (6.1)		7 (7.3)		14.4



SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Trends Between 2003 and 2007

- Change in assessment design from 2003 to 2007
  - More time to complete each block of items
- Usual concurrent calibration linking probably not enough
  - Need a bridge from 2003 design to 2007 design

Bridging Study

2003 Achievement Booklet	Part 1			Part 2		
Booklet 1	M01	M02	S06	S07	M05	M07
Booklet 2	M02	M03	S05	S08	M06	M08
Booklet 3	M03	M04	S04	S09	M13	M11
Booklet 4	M04	M05	S03	S10	M14	M12
Booklet 5	M05	M06	S02	S11	M09	M13
Booklet 6	M06	M01	S01	S12	M10	M14
Booklet 7	S01	S02	M06	M07	S05	S07
Booklet 8	S02	S03	M05	M08	S06	S08
Booklet 9	S03	S04	M04	M09	S13	S11
Booklet 10	S04	S05	M03	M10	S14	S12
Booklet 11	S05	S06	M02	M11	S09	S13
Booklet 12	S06	S01	M01	M12	S10	S14

- We identified four TIMSS 2003 booklets to be used as bridge booklets in 2007

2007 Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

2007 Bridge Booklet	Part 1			Part 2		
Booklet 5	M05	M06	S02	S11	M09	M13
Booklet 6	M06	M01	S01	S12	M10	M14
Booklet 11	S05	S06	M02	M11	S09	S13
Booklet 12	S06	S01	M01	M12	S10	S14

Bridging Study

- Essentially an insurance policy
- All Trend Countries Administered Four Bridge Booklets
  - Booklets 5, 6, 11 & 12 from TIMSS 2003
- The Bridge Data Are Used to Measure the Effect of Changing the Booklet Design for 2007
  - TIMSS 2003 Booklets Consisted of 6 Blocks
  - TIMSS 2007 Booklets Consist of 4 Blocks

Bridging Study
– Did Design Change Have an Effect?

- Compare average p-values of Bridge Items
  - In Bridge Booklets
  - In TIMSS 2007 Booklets
- Result: average p-values of Bridge Items are slightly higher (i.e., easier) in the TIMSS 2007 booklets
  - 8th Grade: 1.4% for Math, 1.2% for Science
  - 4th Grade: 0.9% for Math, 0.4% for Science

Conclusion: Necessary to incorporate bridge into trend scaling

Calibrating the Items

2003 Achievement Booklet	Part 1			Part 2		
Booklet 1	M01	M02	S06	S07	M05	M07
Booklet 2	M02	M03	S05	S08	M06	M08
Booklet 3	M03	M04	S04	S09	M13	M11
Booklet 4	M04	M05	S03	S10	M14	M12
Booklet 5	M05	M06	S02	S11	M09	M13
Booklet 6	M06	M01	S01	S12	M10	M14
Booklet 7	S01	S02	M06	M07	S05	S07
Booklet 8	S02	S03	M05	M08	S06	S08
Booklet 9	S03	S04	M04	M09	S13	S11
Booklet 10	S04	S05	M03	M10	S14	S12
Booklet 11	S05	S06	M02	M11	S09	S13
Booklet 12	S06	S01	M01	M12	S10	S14

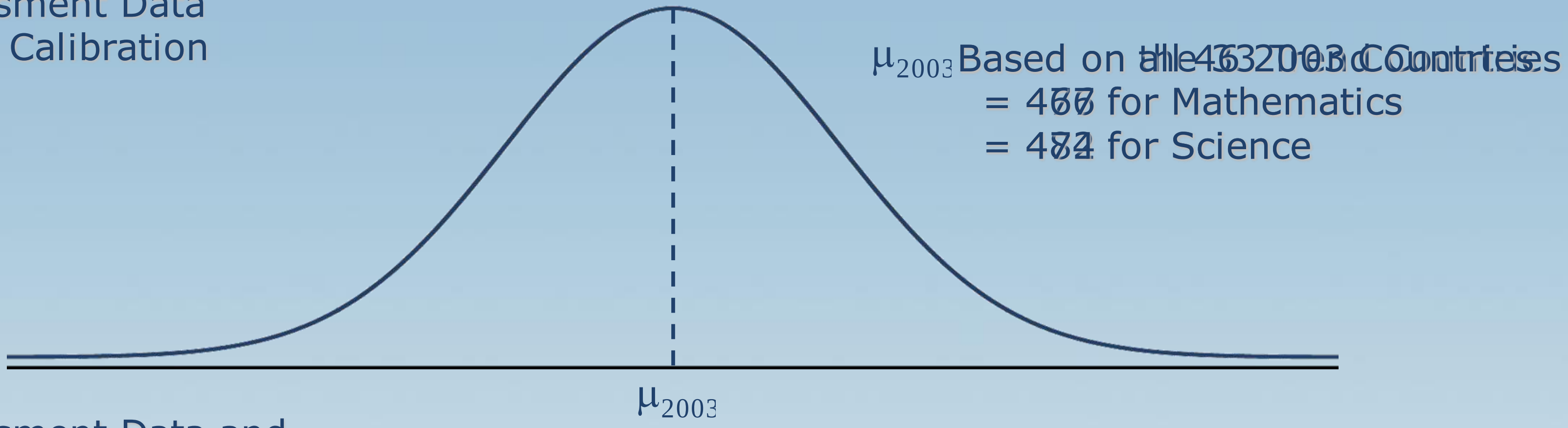
2007 Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

2007 Bridge Booklet	Part 1			Part 2		
Booklet 5	M05	M06	S02	S11	M09	M13
Booklet 6	M06	M01	S01	S12	M10	M14
Booklet 11	S05	S06	M02	M11	S09	S13
Booklet 12	S06	S01	M01	M12	S10	S14

- 2003 Trend and 2007 Bridge – same items, different distributions
- 2007 Trend – treat as different items

Placing the 2007 Scores on the 1995 Metric

2003 Assessment Data
under 2003 Calibration

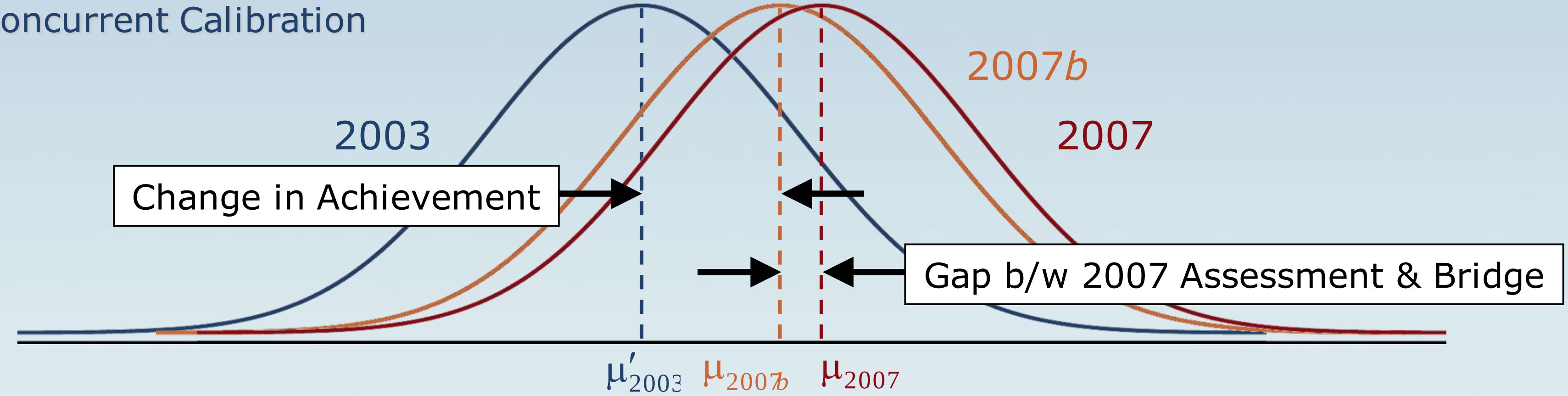


μ_{2003} Based on all 46 2003 Countries
= 466 for Mathematics
= 474 for Science

μ_{2003} Based on the 32 Trend Countries
= 477 for Mathematics
= 482 for Science

μ_{2003}

2003 Assessment Data and
2007 Assessment Data under
2007 Concurrent Calibration



2007*b*

2003

2007

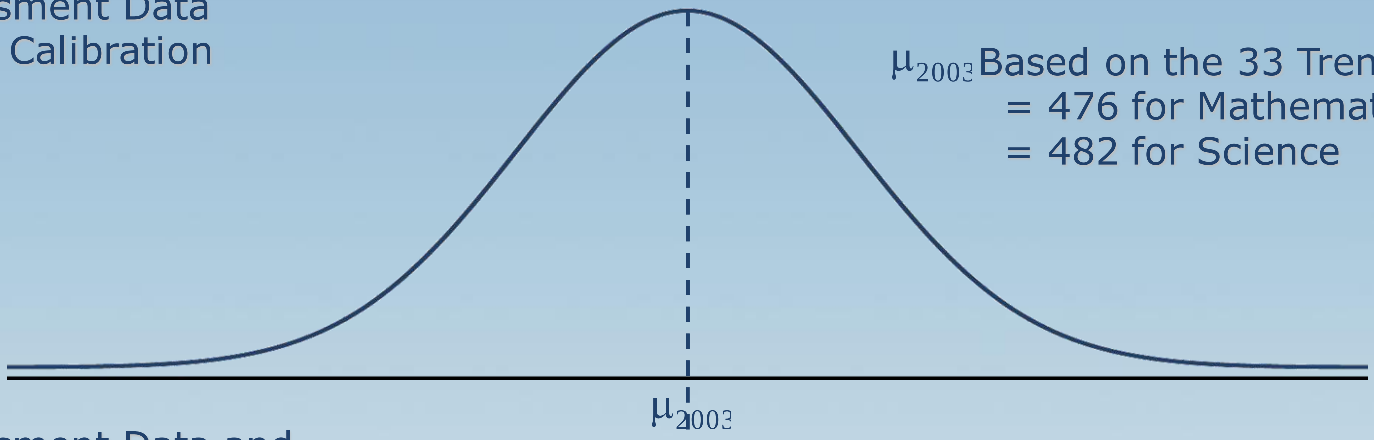
Change in Achievement

Gap b/w 2007 Assessment & Bridge

μ'_{2003} μ_{2007b} μ_{2007}

Placing the 2007 Scores on the 1995 Metric

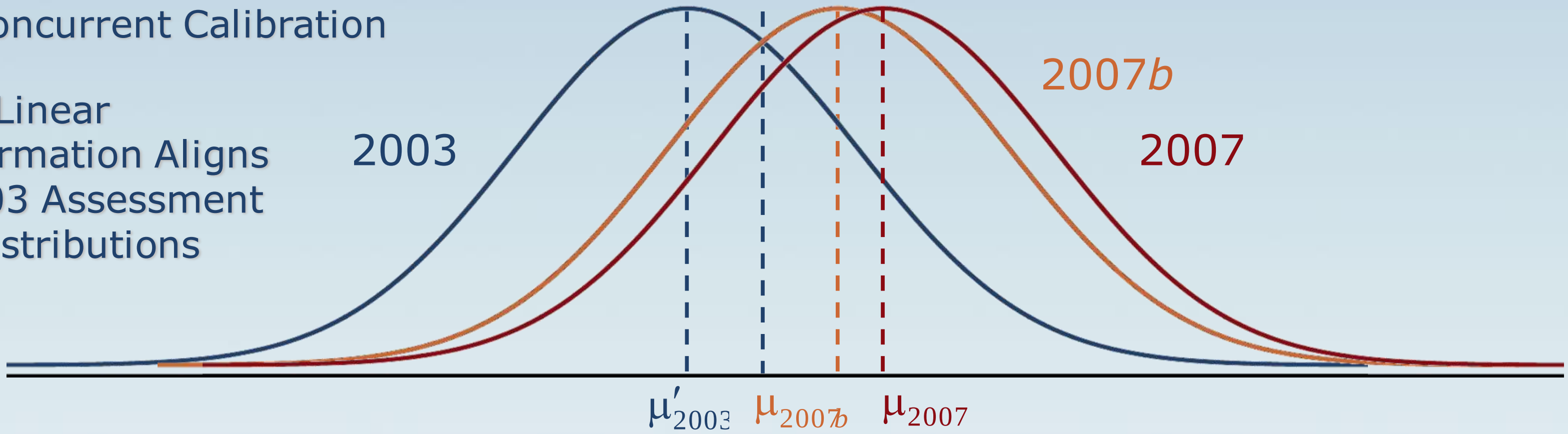
2003 Assessment Data under 2003 Calibration



μ_{2003} Based on the 33 Trend Countries
= 476 for Mathematics
= 482 for Science

2003 Assessment Data and 2007 Assessment Data under 2007 Concurrent Calibration

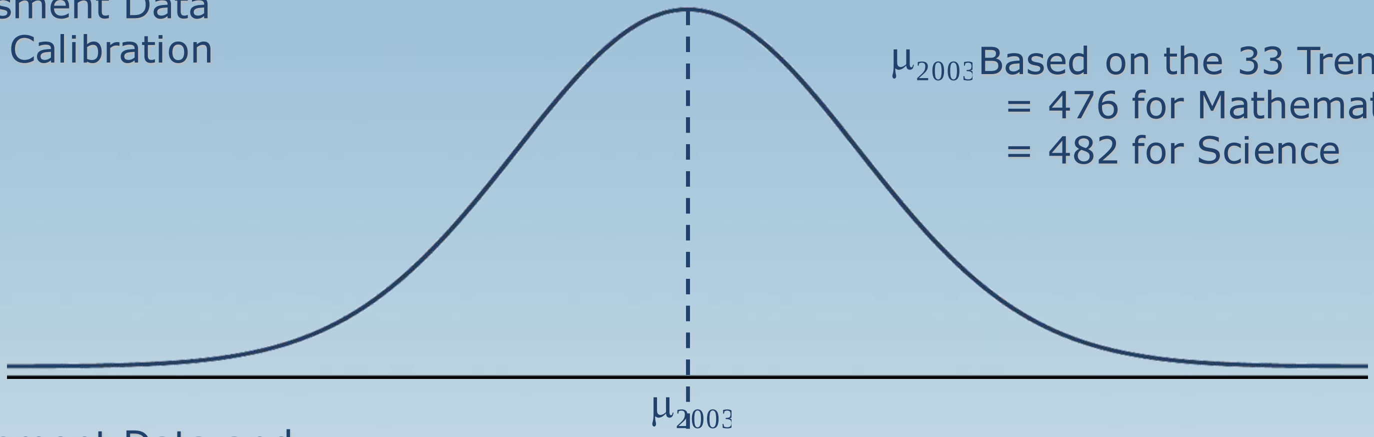
A First Linear Transformation Aligns the 2003 Assessment Data Distributions



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

Placing the 2007 Scores on the 1995 Metric

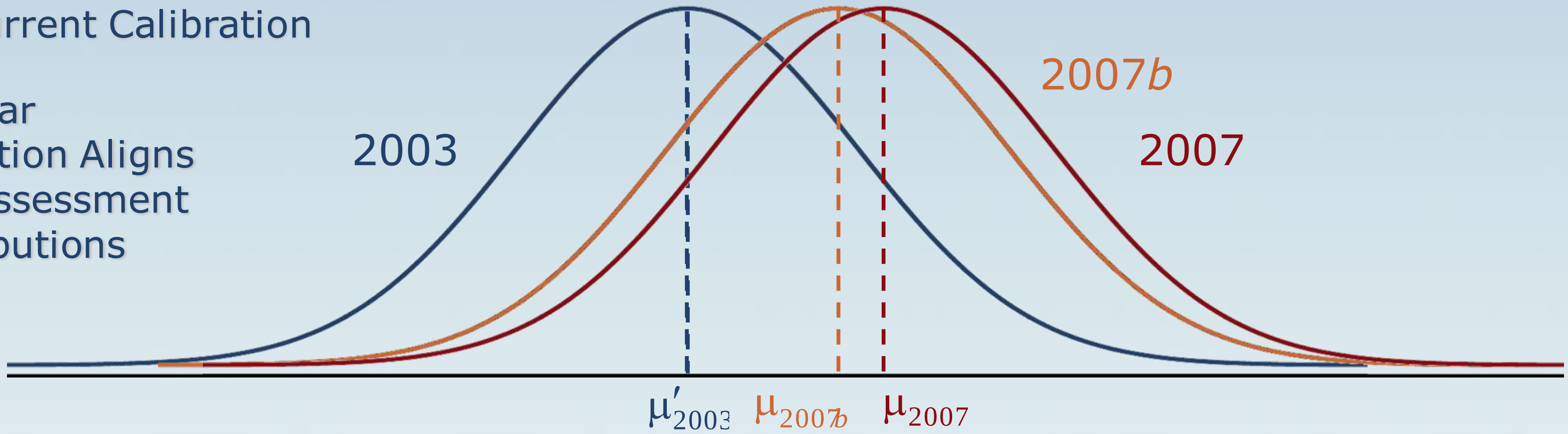
2003 Assessment Data under 2003 Calibration



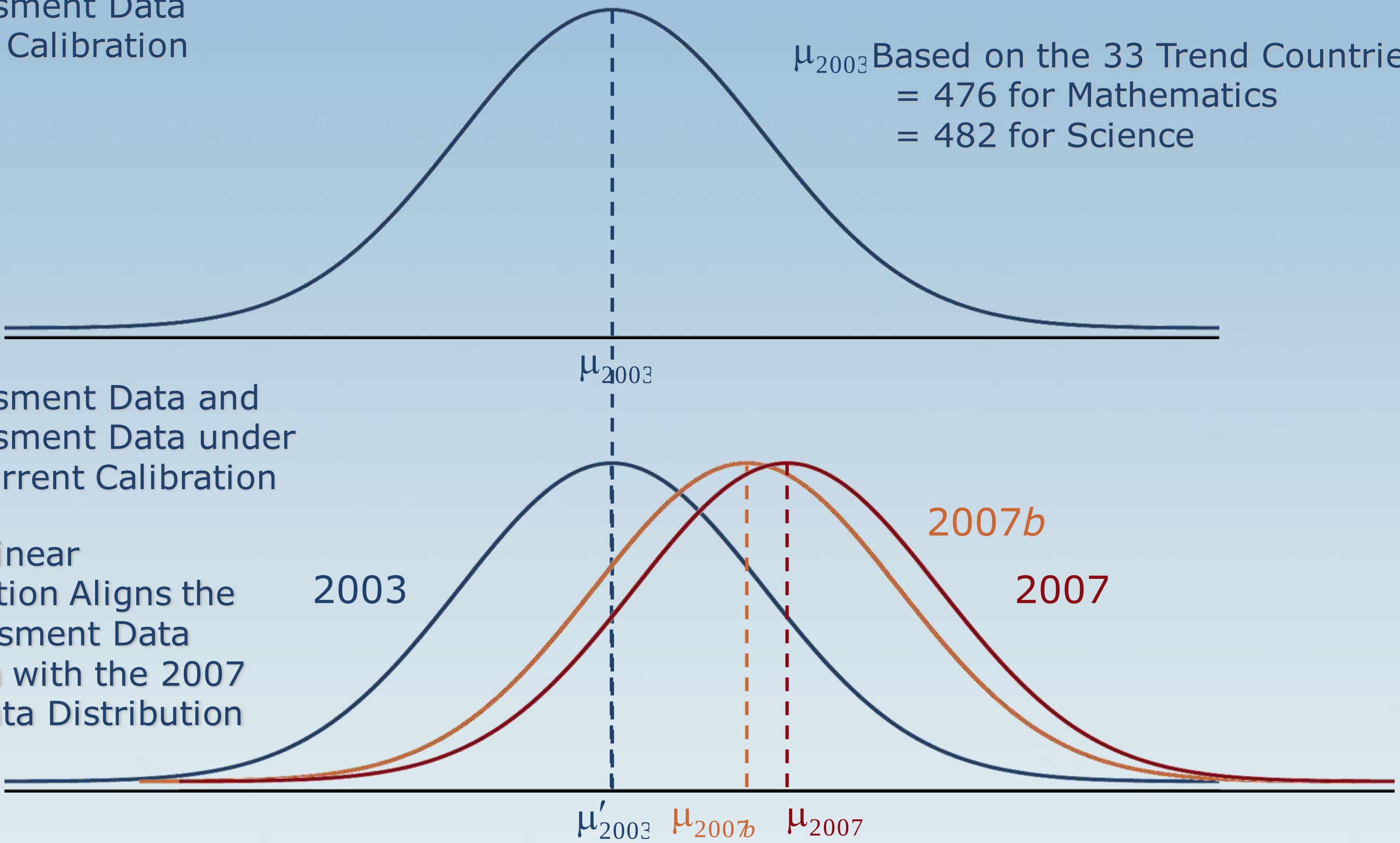
μ_{2003} Based on the 33 Trend Countries
= 476 for Mathematics
= 482 for Science

2003 Assessment Data and 2007 Assessment Data under 2007 Concurrent Calibration

A First Linear Transformation Aligns the 2003 Assessment Data Distributions



Placing the 2007 Scores on the 1995 Metric



2003 Assessment Data under 2003 Calibration

μ_{2003} Based on the 33 Trend Countries
= 476 for Mathematics
= 482 for Science

2003 Assessment Data and 2007 Assessment Data under 2007 Concurrent Calibration

A Second Linear Transformation Aligns the 2007 Assessment Data Distribution with the 2007 Bridging Data Distribution

Placing the 2007 Scores on the 1995 Metric

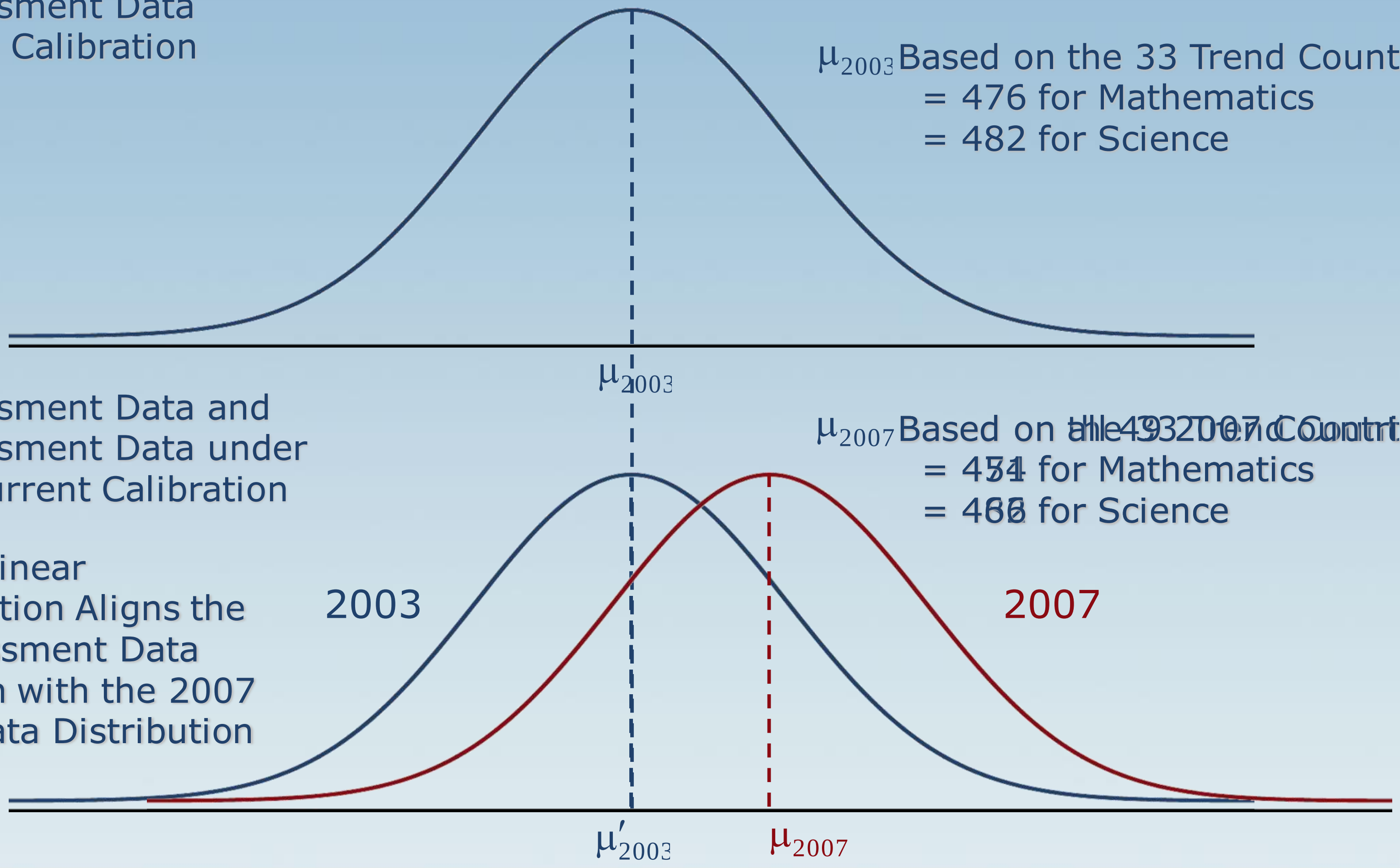
2003 Assessment Data
under 2003 Calibration

μ_{2003} Based on the 33 Trend Countries
= 476 for Mathematics
= 482 for Science

2003 Assessment Data and
2007 Assessment Data under
2007 Concurrent Calibration

μ_{2007} Based on all 49 2007 Countries / Based on the 33 Trend Countries
= 451 / 474 for Mathematics
= 466 / 462 for Science

A Second Linear Transformation Aligns the 2007 Assessment Data Distribution with the 2007 Bridging Data Distribution



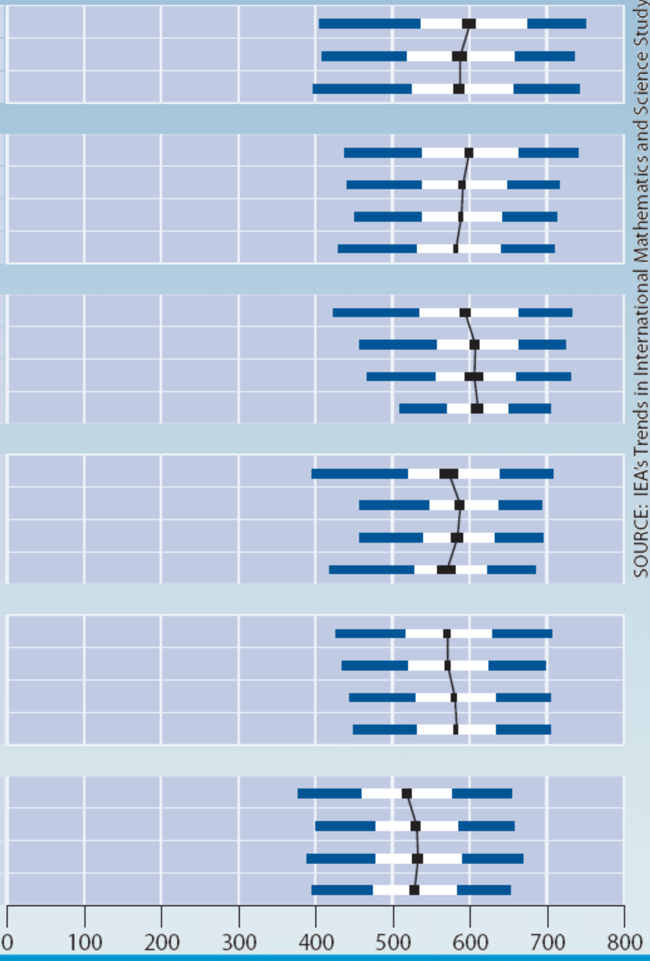
Excerpt from TIMSS 2007 International Report

Exhibit 1.3 **Trends in Mathematics Achievement – 1995 Through 2007 (Continued)**

TIMSS2007 Mathematics 8th Grade

Country		Average Scale Score	2003 to 2007 Difference	1999 to 2007 Difference	1995 to 2007 Difference
Chinese Taipei					
	2007	598 (4.5)			
	2003	585 (4.6)	13 (6.4) ▲		
	1999	585 (4.0)		13 (5.9) ▲	
Korea, Rep. of					
	2007	597 (2.7)			
⧓	2003	589 (2.2)	8 (3.1) ▲		
	1999	587 (2.0)		10 (3.4) ▲	
	1995	581 (2.0)			17 (3.4) ▲
Singapore					
	2007	593 (3.8)			
	2003	605 (3.6)	–13 (5.2) ▼		
	1999	604 (6.3)		–12 (7.2)	
	1995	609 (4.0)			–16 (5.6) ▼
Hong Kong SAR					
†	2007	572 (5.8)			
†	2003	586 (3.3)	–14 (6.6) ▼		
†	1999	582 (4.3)		–10 (7.2)	
	1995	569 (6.1)			4 (8.4)
Japan					
	2007	570 (2.4)			
	2003	570 (2.1)	0 (3.1)		
	1999	579 (1.7)		–9 (2.9) ▼	
	1995	581 (1.6)			–11 (2.8) ▼
Hungary					
	2007	517 (3.5)			
2	2003	529 (3.2)	–12 (4.7) ▼		
	1999	532 (3.7)		–15 (5.0) ▼	
	1995	527 (3.2)			–10 (4.7) ▼

Mathematics Achievement Distribution



SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007

In Summary, TIMSS and PIRLS Linking Methodology Is…

- Very well adapted to the philosophy of measuring trends with gradual, evolutionary changes
- Also deals well with major situational changes
  - Booklet design changes
  - Major framework changes

Measuring Trends in Educational Achievement

Michael O. Martin and Ina V.S. Mullis

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