Measuring Postsecondary Competencies: Lessons from Large-Scale K-12 Assessments

Daniel Koretz
Harvard Graduate School of Education

KoKoHs II International Conference
Berlin
4 April 2016

Premise

- Postsecondary education lacks a culture of assessment and evaluation
- This gives rigorous development of postsecondary assessments great promise
- But positive effects may be undermined if mistakes made in large-scale K-12 testing are repeated

Challenges in large-scale assessment

1. Constructing assessments that measure specific competencies

2. Matching tests to specific uses and inferences

3. Avoiding “function creep”: sukzessive Zunahme von weiteren Funktionen

- 2 & 3 have often been ignored in K-12

- 2 & 3 are already problems in postsecondary assessment

Topics

- Differences in match between test and curricula
- Variations in test use
- Behavioral responses to testing
- The move toward assessing competencies

A representation of validity for comparative assessments

- Both the test and the target of inference are weighted composites of “performance elements”
- Weights describe importance
  - Test weight: sensitivity of score to change in element
    - Many test weights = 0 (omitted content)
  - Inference weight: importance of element to a specific inference
- Validity requires *extrapolating* from the test’s weighted composite to the *particular* weighted target

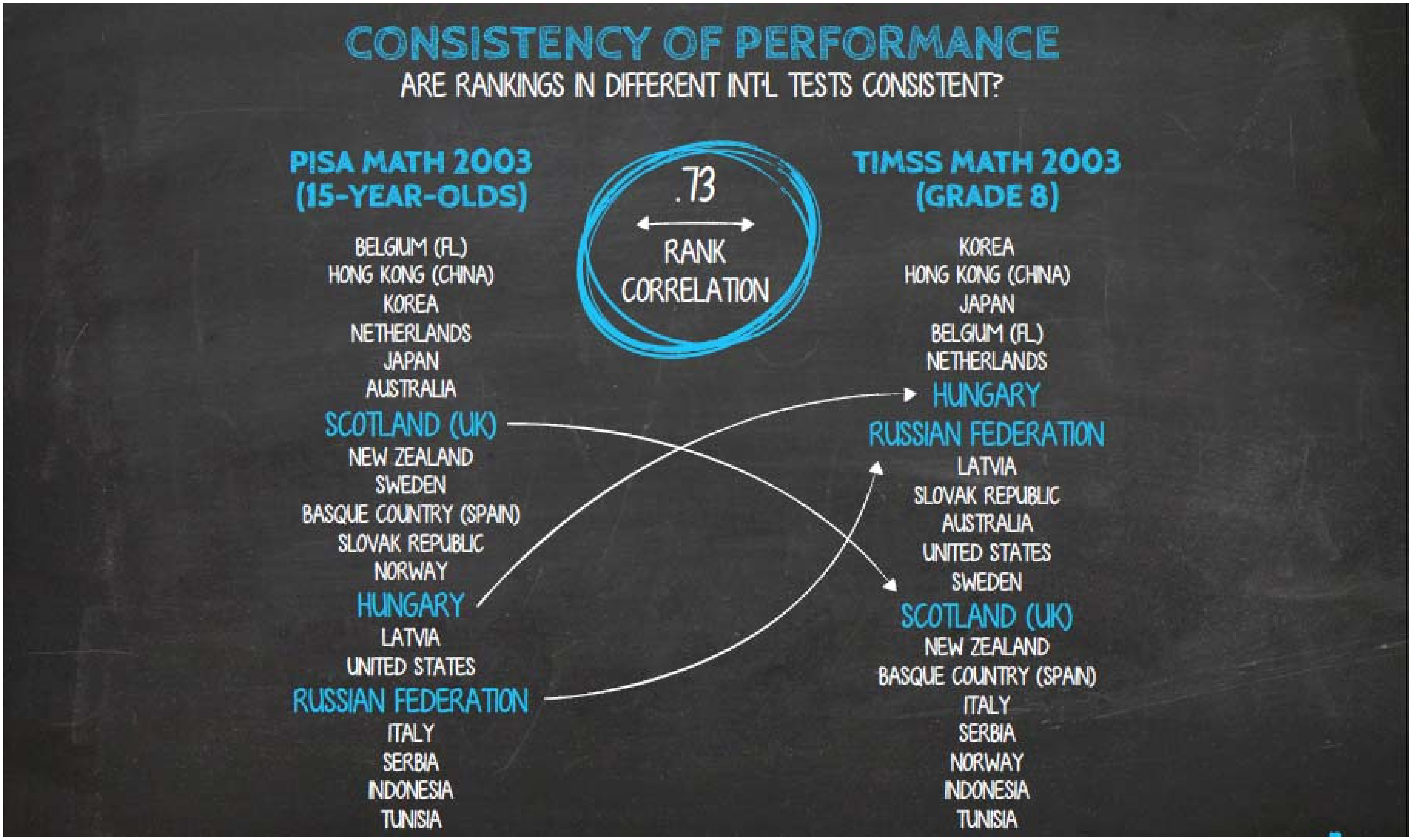
Extrapolation in comparative assessment

- Match between test and target often varies
  - Weights vary across tests
  - Inference weights vary across institutions, programs, & courses

- Well documented (but often ignored) in K-12

- Appears in many contexts:
  - International comparisons
  - Across tests within the US
  - Across groups within one test

TIMSS/PISA: ranks of country means



CONSISTENCY OF PERFORMANCE

ARE RANKINGS IN DIFFERENT INT'L TESTS CONSISTENT?

PISA MATH 2003 (15-YEAR-OLDS)

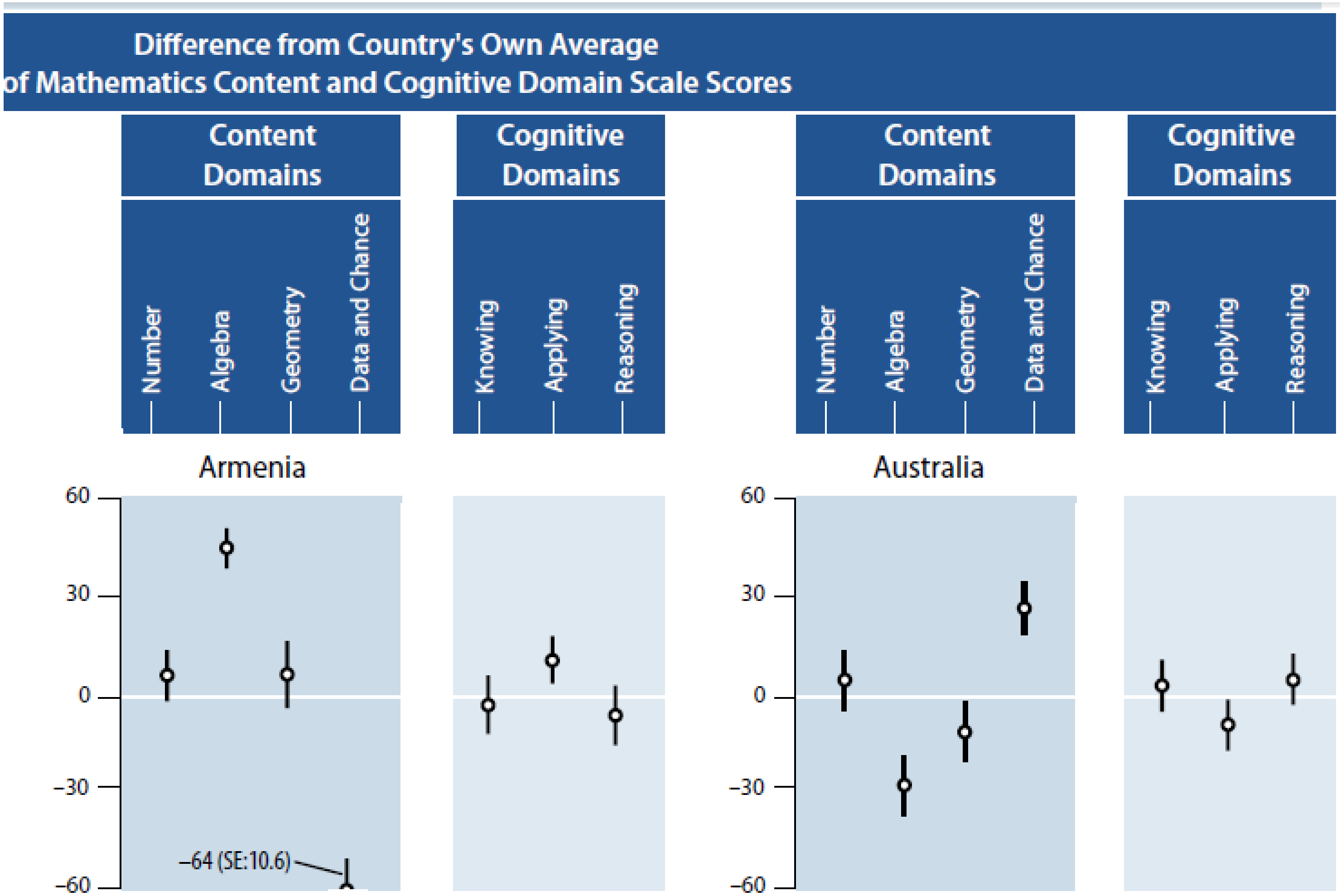
BELGIUM (FL)
HONG KONG (CHINA)
KOREA
NETHERLANDS
JAPAN
AUSTRALIA
SCOTLAND (UK)
NEW ZEALAND
SWEDEN
BASQUE COUNTRY (SPAIN)
SLOVAK REPUBLIC
NORWAY
HUNGARY
LATVIA
UNITED STATES
RUSSIAN FEDERATION
ITALY
SERBIA
INDONESIA
TUNISIA

.73 ↔ RANK CORRELATION

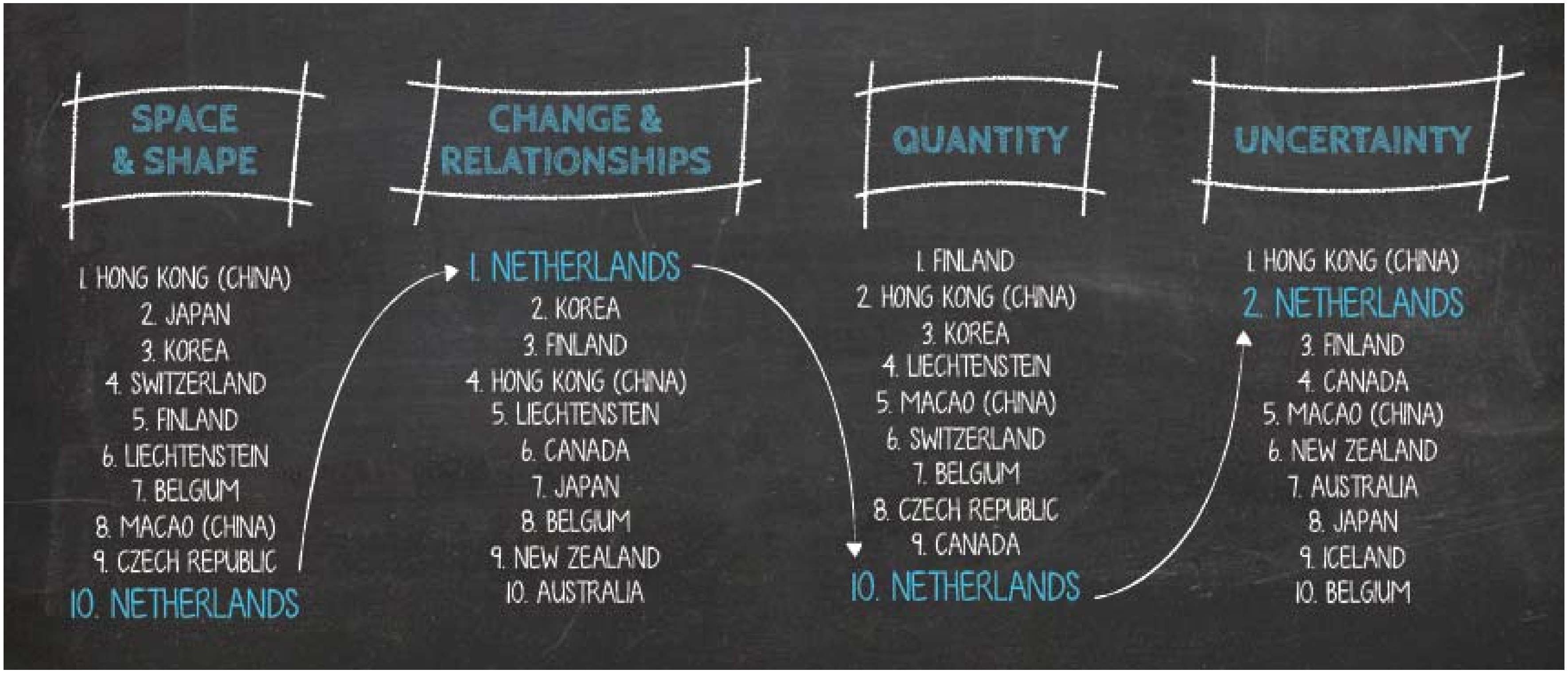
TIMSS MATH 2003 (GRADE 8)

KOREA
HONG KONG (CHINA)
JAPAN
BELGIUM (FL)
NETHERLANDS
HUNGARY
RUSSIAN FEDERATION
LATVIA
SLOVAK REPUBLIC
AUSTRALIA
UNITED STATES
SWEDEN
SCOTLAND (UK)
NEW ZEALAND
BASQUE COUNTRY (SPAIN)
ITALY
SERBIA
NORWAY
INDONESIA
TUNISIA

TIMSS 2007 grade 8: lack of robustness across parts

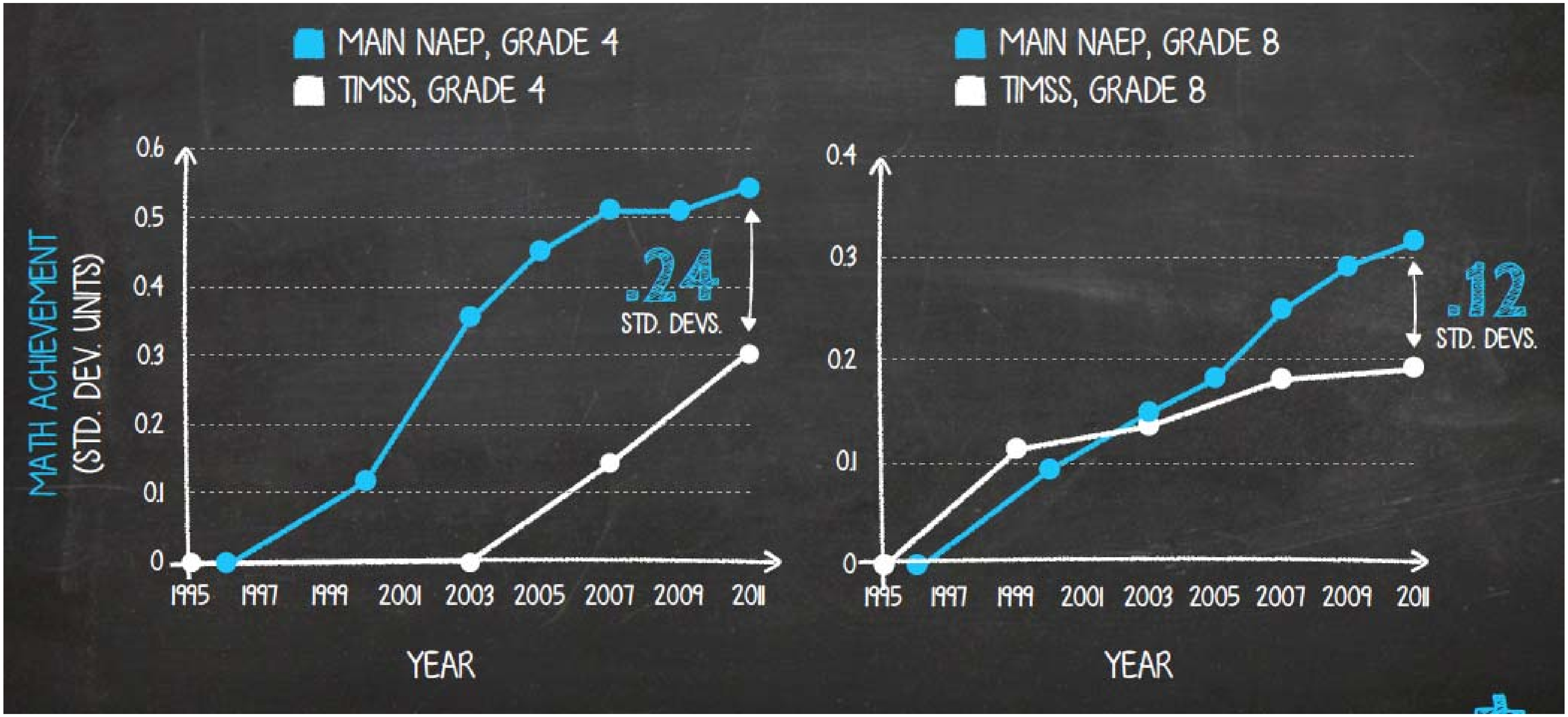


PISA 2003: robustness across strands: top 10 performers by strand

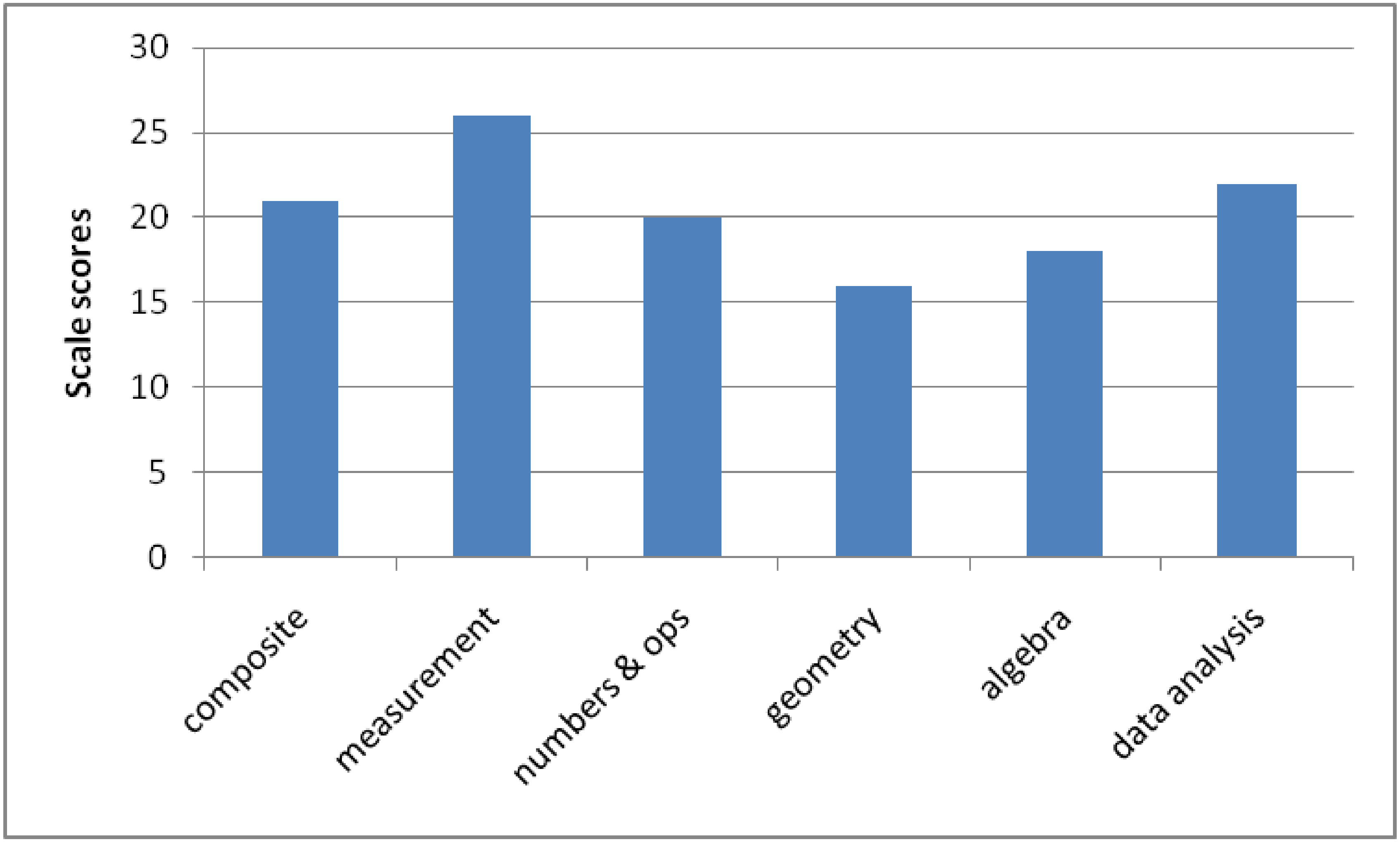


SPACE & SHAPE	CHANGE & RELATIONSHIPS	QUANTITY	UNCERTAINTY
1. HONG KONG (CHINA)	1. NETHERLANDS	1. FINLAND	1. HONG KONG (CHINA)
2. JAPAN	2. KOREA	2. HONG KONG (CHINA)	2. NETHERLANDS
3. KOREA	3. FINLAND	3. KOREA	3. FINLAND
4. SWITZERLAND	4. HONG KONG (CHINA)	4. LIECHTENSTEIN	4. CANADA
5. FINLAND	5. LIECHTENSTEIN	5. MACAO (CHINA)	5. MACAO (CHINA)
6. LIECHTENSTEIN	6. CANADA	6. SWITZERLAND	6. NEW ZEALAND
7. BELGIUM	7. JAPAN	7. BELGIUM	7. AUSTRALIA
8. MACAO (CHINA)	8. BELGIUM	8. CZECH REPUBLIC	8. JAPAN
9. CZECH REPUBLIC	9. NEW ZEALAND	9. CANADA	9. ICELAND
10. NETHERLANDS	10. AUSTRALIA	10. NETHERLANDS	10. BELGIUM

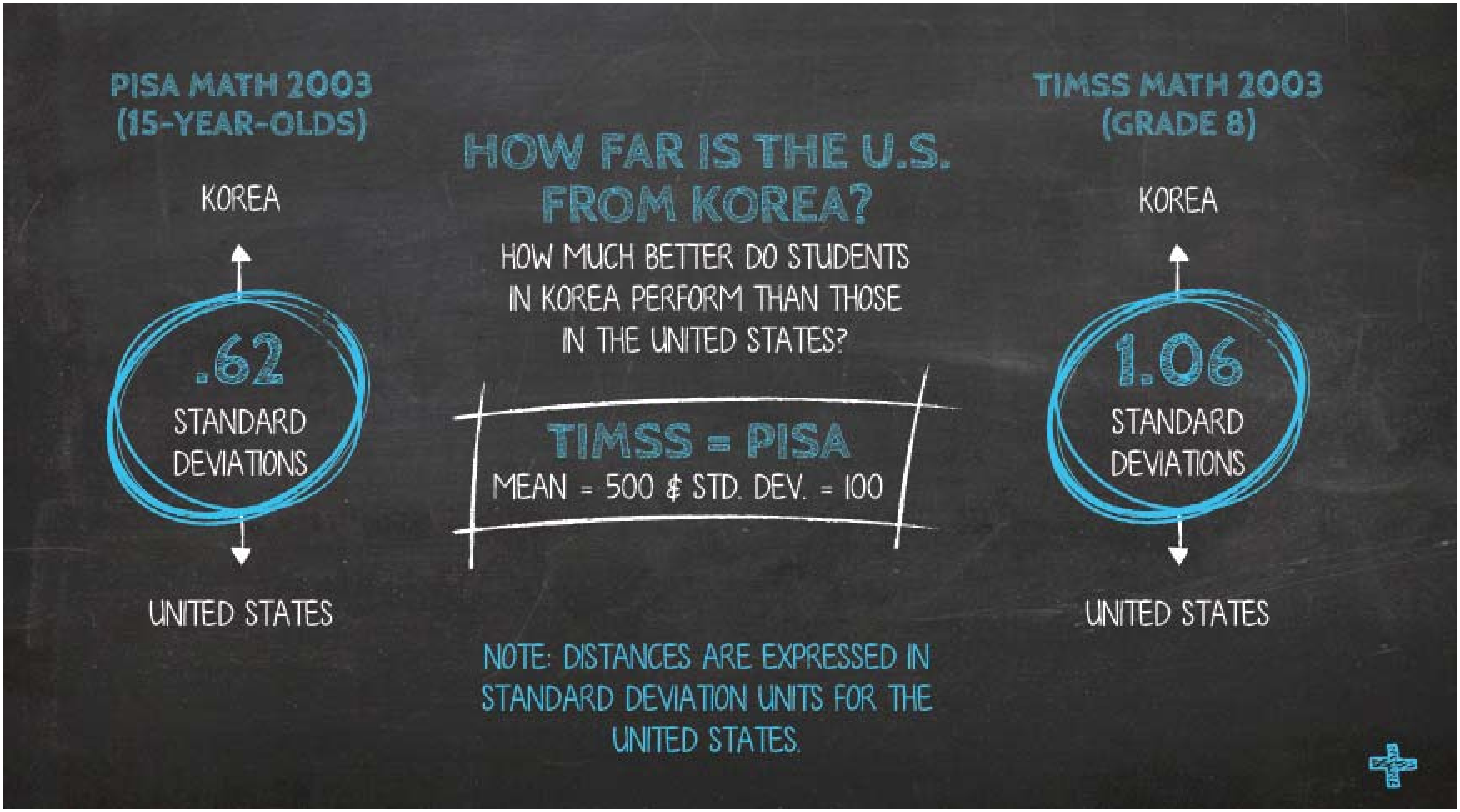
Within-country differences in aggregate trends: US, math, NAEP vs. TIMSS



Within-country variation in subgroup differences: 2009 NAEP grade 4 math, White-Hispanic math



TIMSS/PISA: Korea-US



PISA MATH 2003
(15-YEAR-OLDS)

KOREA

.62
STANDARD
DEVIATIONS

UNITED STATES

HOW FAR IS THE U.S.
FROM KOREA?

HOW MUCH BETTER DO STUDENTS
IN KOREA PERFORM THAN THOSE
IN THE UNITED STATES?

TIMSS = PISA
MEAN = 500 & STD. DEV. = 100

NOTE: DISTANCES ARE EXPRESSED IN
STANDARD DEVIATION UNITS FOR THE
UNITED STATES.

TIMSS MATH 2003
(GRADE 8)

KOREA

1.06
STANDARD
DEVIATIONS

UNITED STATES

Implications of incompleteness and differential match

- Score differences do not fully describe differences in output
- Ranks may not be robust
- When ranks are consistent, estimates of gaps are highly imprecise
- Differences in match are confounded with differences in educational output

The importance of test use

- Choice of use affects numerous aspects of design
  - All designs are compromises, with “trade-offs”
- Choice of us affects intended inferences

A few design trade-offs from choice of uses

- Monitoring of institutions or systems, versus scores for students
  - Matrix sampling
- Summative versus diagnostic
  - Complexity and realism of tasks

A taxonomy of uses

Use	Examples	Stakes	Inference
Fully internal	Instructor-designed tests	None	No comparison
Diagnostic external	traditional NRTs	None	**Limited** comparison
Monitoring	NAEP, ILSAs	Vary	Comparison, (causal)
"Value-added" modeling	High-stakes tests	High	Comparison, causal
Other accountability	High-stakes tests	High	Comparison, causal

Inferences in diagnostic external testing

- The original model of large-scale comparative assessment
- Content based on surveys of many curricula in a decentralized system (similar to TIMSS)
  - Match between test weights and inference weights varied
- Incompleteness and variations in match were recognized
- Warnings from the Iowa testing program:
  - Information from test is necessarily incomplete
  - Treat scores as “specialized, supplementary” information
  - Scores alone are never sufficient to evaluate a program or school

Treatment of incompleteness and variation in match in comparative assessments

- Incompleteness: sometimes noted in passing
- Variations in match: sometimes documented, but are not emphasized
- Validity depends on how well scores support the primary inferences despite incompleteness and variations in match

How severe will these problems be in postsecondary assessment?

- Depends on severity of:
  - Variations in intake and educational goals
  - Uses of scores

- Variations in match will be greater than in K-12
  - Large differences in intake and goals among institutions, programs, and courses

- Function creep is already severe in postsecondary assessment

Examples of "function creep"

- Chicago, Iowa Tests of Basic Skills:
  - Designed as a diagnostic tool
  - Used to hold schools accountable and as a criterion for promotion between grades
- SAT college admissions test:
  - Designed to predict postsecondary performance
  - Used by USED to compare state systems
- PISA:
  - Designed as a monitoring tool
  - Routinely used to support causal inferences about system and school quality

Competing functions for AHELO

- Institutional (instructional?) improvement:
  - "It [is] important to re-emphasize that AHELO is intended as a tool for institutional improvement"
- Comparative monitoring:
  - "The…feasibility of AHELO rests on its capacity to produce valid and reliable results across different countries…"
- Accountability:
  - Justified by "a shift…towards models combining greater autonomy with increased transparency and *accountability*…[which] has led to increased demands for…outcomes assessment."

Function creep with the CLA

“Institutions…have used the CLA to benchmark [1] value-added growth in student learning…[2] compared to that of other institutions…

Student-level metrics provide guidance to students and data to faculty and administrators for [3, 4, 5, 6] making decisions about grading, scholarships, admission, or placement…

Results for graduating seniors may be used as an independent [7] corroboration of the rapid growth of competency-based approaches among colleges.

Graduating seniors use their results…to [8] provide potential employers with evidence of their work readiness.”

http://cae.org/participating-institutions/cla-references/

Advertised uses for ACT CAAP

- Monitor institutional progress over time
- Compare across institutions
- Hold institutions accountable
- Evaluate students' readiness for upper-level courses
- Evaluate readiness for the workplace
- Evaluate group performance in specific content areas
- Evaluate student growth

Behavioral responses to testing

- When scores are important, educators often focus instruction on the tested sample rather than the domain
  - The test defines the curriculum
  - Undesirable test preparation & score inflation often result
  - Severity varies with student and school characteristics
- Only modest pressure is required
- Essential questions for research:
  - How severe will these problems be in postsecondary?
  - How will these vary across institutions and assessments?

Is this different for testing competencies?

- Positive:
  - May expand the range of constructs measured
  - Closer to “criterion behaviors”
  - May encourage desired types of pedagogy
- Neutral
  - Also vulnerable to curriculum displacement & Campbell’s Law
- Negative
  - Costly
  - More limited sampling from domain
  - Limited generalizability
  - Less diagnostic value

Two responses to incompleteness of measures, failures of invariance

- Constrain both inferences and uses appropriately
- Ignore them
  - Typical in K-12 assessment
  - In postsecondary assessment: unclear, but initial signs are troubling

Some recommendations

- **Resist temptation: “Less is more”**
- Avoid too broad an inference
- Avoid function creep
- When using data for instructional or institutional evaluation, combine scores with other data
- When using data for comparative purposes:
  - Explicitly recognize incompleteness
  - Check robustness if possible
  - Avoid spurious precision
- When using for institutional and instructional improvement:
  - Monitor effects