AIR®
Advancing Evidence.
Improving Lives.

Overview of NCES Large-Scale Assessments With a Closer Look at NAEP and TIMSS

Martin Hooper

American Institutes for Research

AERA Professional Development Course| June 2022

What are large-scale assessments?

Tests that focus on measuring and monitoring what populations know and can do in academic subject areas

- Populations are usually certain ages/grade in cities, states, countries
- Subpopulation measurement (gender, SES, race/ethnicity) is also prioritized
- Academic subject areas include mathematics, science, reading, civics, computer literacy, etc.
- National assessments like NAEP for the United States allow for trend monitoring and comparisons between states and large cities
- International assessments like TIMSS, PIRLS, and PISA allow for country-level trend monitoring and between-country comparisons

Advancing Evidence.
Improving Lives.

Closer look at National Assessment for Educational Progress (NAEP)

National Assessment of Educational Progress (NAEP)

- Congressionally mandated in the United States - began in 1969
- A comprehensive nationally representative assessment of what U.S. students know and can do in a broad range of academic subjects
- Designed to produce national, state, and large urban school district results, not a testing program for individual students or schools
- Administered by the National Center for Education Statistics (NCES)
- State assessments since 1990, urban district results (on trial basis) since 2002
- Policy oversight and guidance by the National Assessment Governing Board (NAGB)

Q: Which of the following statements is FALSE?

A. NCES is charged with administering NAEP assessments

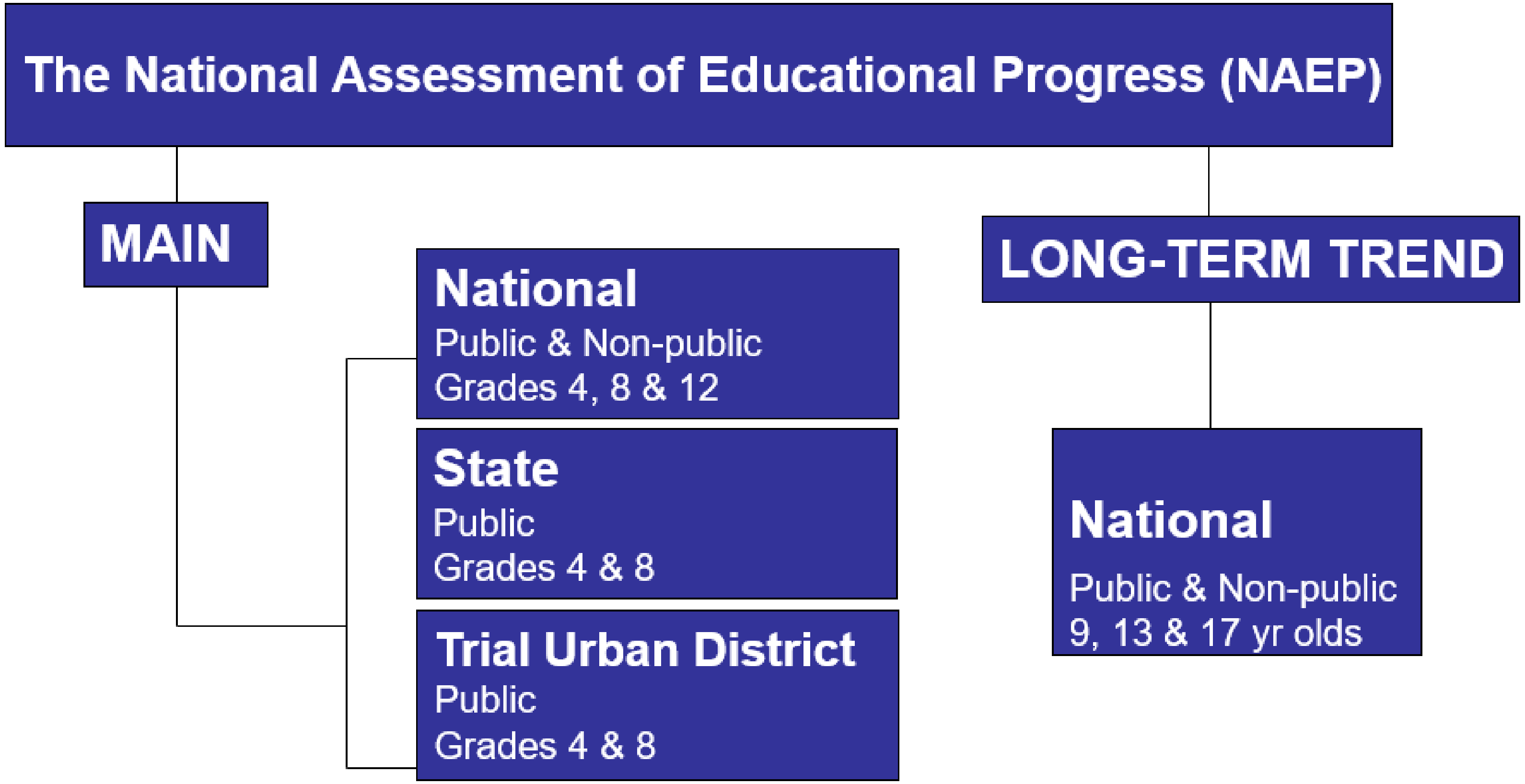
B. The NAEP Governing Board (NAGB) guides NAEP policy

C. NAEP is congressionally mandated

D. NAEP is not a large-scale assessment

QUIZ

NAEP Components



The National Assessment of Educational Progress (NAEP)

MAIN

National
Public & Non-public
Grades 4, 8 & 12

State
Public
Grades 4 & 8

Trial Urban District
Public
Grades 4 & 8

LONG-TERM TREND

National
Public & Non-public
9, 13 & 17 yr olds

Main NAEP: National Assessments

- Reports statistical summaries of student achievement and factors related to educational performance for the nation and specific subgroups of the population

- Includes students drawn from both public and nonpublic schools and reports results for student achievement at grades 4, 8, and 12

- Follows subject-area frameworks developed by NAGB

Main NAEP: State Assessments

- First reported performance at the state level in 1990
- Assesses public school students only at grades 4 and 8, in selected subjects, reading, mathematics, science, and writing
- Allows states to monitor progress over time in the selected subject areas and allows comparisons with other states and with the nation
- Follows subject-area frameworks developed by NAGB that are identical, by grade and subject, to those for the national assessment

Frequency of Assessments

NAEP conducts national and state assessments at least once every two years in reading and mathematics in grades 4 and 8 (beginning in 2003).

- Due to COVID, NAEP was not administered in 2021 but instead administered in 2022.

Other NAEP subjects (writing, science, history, geography, civics, economics, and arts) in 4th, 8th and 12th grades will continue to be assessed on a voluntary basis at regularly scheduled intervals.

TEL assessment was administered in 2014 first for grade 8 only.

NAEP Reporting

Achievement levels

Scale scores

Disaggregated data by:

- Subgroups (e.g., race/ethnicity, gender, SD, ELL)
- SES: National School Lunch Program eligibility, parental education level (for grade 8 only)
- Geography (national, state and regional comparisons; school location)
- Background information (i.e., school, teacher and student questionnaires)

NAEP Achievement Levels

NAEP reports by scale scores and achievement levels

- Levels set by NAGB and designed to describe what students should know and be able to do

NAEP reports three achievement levels at each grade:

- ***NAEP Basic*** denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade
- ***NAEP Proficient*** represents solid academic performance for each grade assessed
- ***NAEP Advanced*** represents superior performance

How NAEP Data are Used

- NAEP provides a credible independent standard that educators, legislators and the general public can use to gauge the effectiveness of education policies.
- NAEP data enable policymakers to examine the relative emphasis of state standards.
- NAEP data are used by research scientists who study education policy issues and the development of student skills and abilities.
- Although numerous data tools are available, NAEP student-level data is restricted-use and use must be licensed.
- The NAEP primer data used for the examples in this training is publicly-available.

Q: Which comparison cannot be made with NAEP data?

A. Comparing 2019 8th grade mathematics scores between California and Massachusetts

B. Comparing 2019 4th grade reading scores between New York City and Chicago

C. Comparing 2019 4th grade reading scores between the United States and Canada

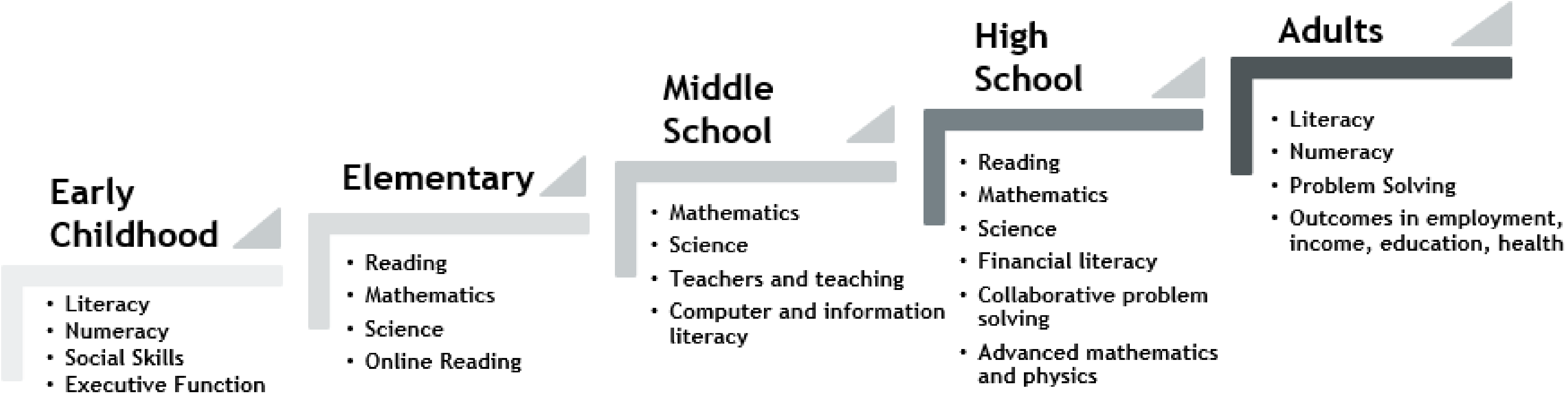
D. Comparing 2019 8th grade mathematics scores for the United States with 2017 mathematics scores for the United States

QUIZ

Advancing Evidence.
Improving Lives.

International Large-Scale Assessments (ILSAs)

NCES International Studies Across the Lifespan



Early Childhood

- Literacy
- Numeracy
- Social Skills
- Executive Function

Elementary

- Reading
- Mathematics
- Science
- Online Reading

Middle School

- Mathematics
- Science
- Teachers and teaching
- Computer and information literacy

High School

- Reading
- Mathematics
- Science
- Financial literacy
- Collaborative problem solving
- Advanced mathematics and physics

Adults

- Literacy
- Numeracy
- Problem Solving
- Outcomes in employment, income, education, health

IEA		PIRLS TIMSS	ICILS TIMSS	TIMSS	
OECD	IELS		TALIS	PISA*	PIAAC

*PISA Young Adult Follow-Up Study (PISA YAFS), a longitudinal extension of PISA 2012, was released in June 2021. *EdSurvey* 2.7 can be used to analyze PISA YAFS.

Uses of International Assessments

Provide information for international achievement comparisons

- *How does the performance in one country compare with that of other countries?*

Examine trends in achievement

- *How does one country’s achievement increase or decrease over time?*

Improve education by informing policy, research, and practices

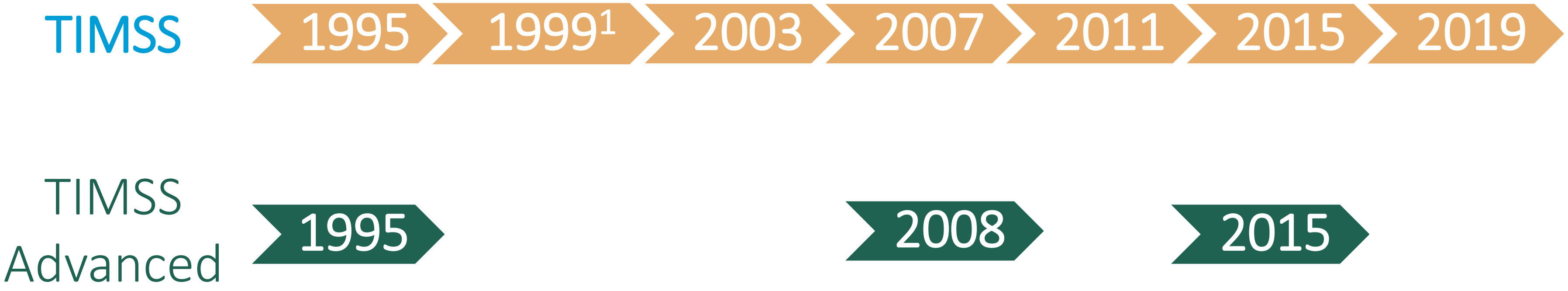
- *What factors are associated with educational achievement? What can we learn from others about what works (and what doesn't)? What could be adopted by or adapted?*

AIR®

Advancing Evidence.
Improving Lives.

Closer look at Trends in International Mathematics and Science Study (TIMSS)

TIMSS and TIMSS Advanced Administration Cycle



TIMSS

1995 1999[1] 2003 2007 2011 2015 2019

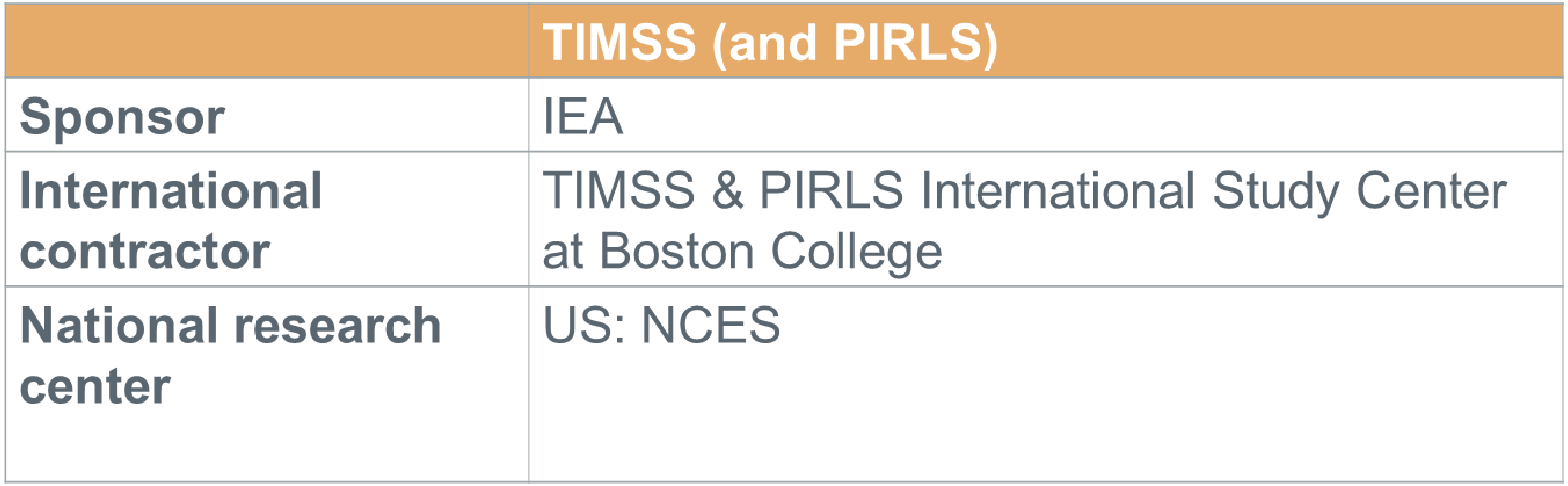
TIMSS Advanced

1995 2008 2015

TIMSS: Mathematics and science assessment at grades 4 and 8

TIMSS Advanced: Advanced mathematics and physics assessment in the final year of secondary school

Sponsorship and Management



	TIMSS (and PIRLS)
Sponsor	IEA
International contractor	TIMSS & PIRLS International Study Center at Boston College
National research center	US: NCES

Complex Sampling :Why Do We Use Samples?

Impossible to test everyone on everything

- Too many people
- Too many items
- Too expensive

Not necessary to test everyone on everything, e.g.,

- Blood sample
- Soup sample

ILSA context:

- Some students are tested on some items
- Results should be seen in the context of the student and item sample design

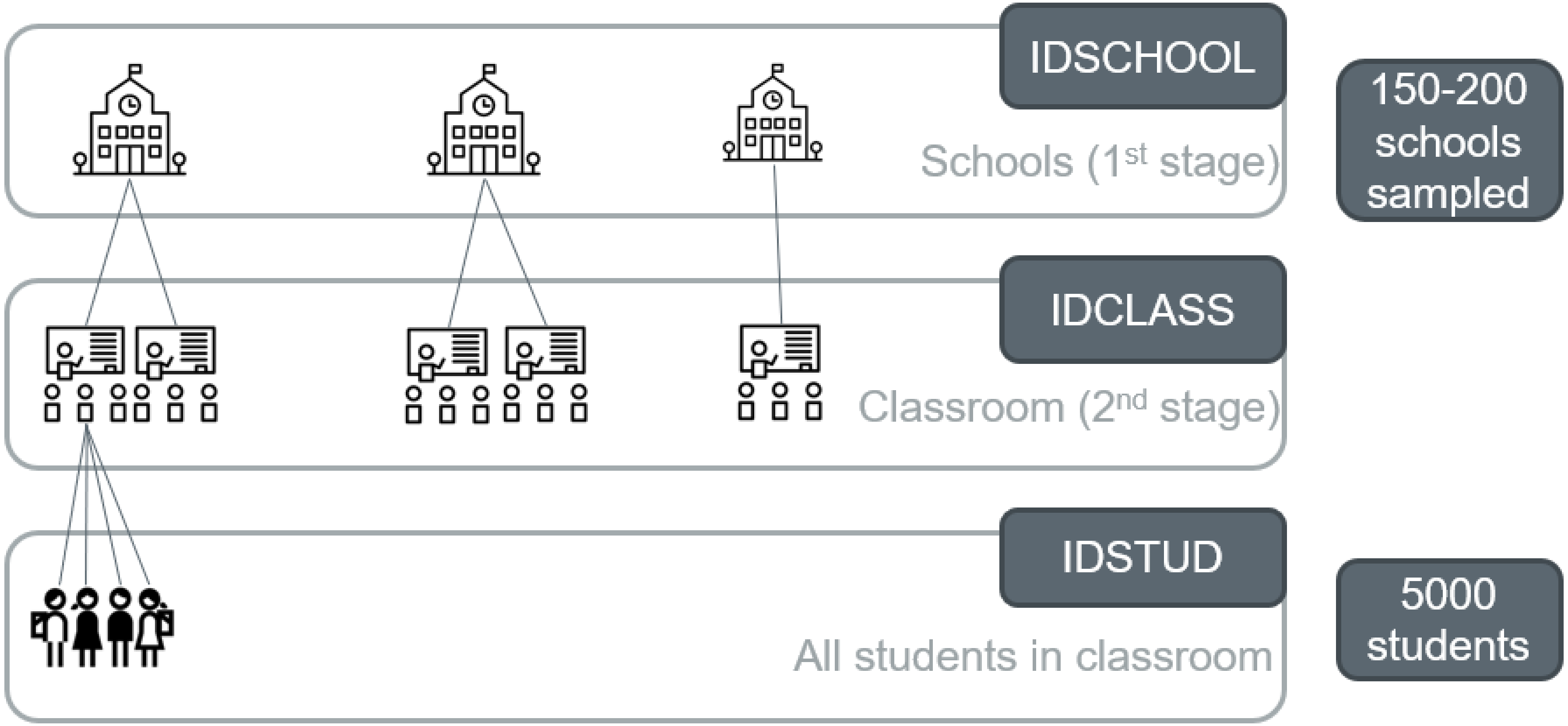
Source: Sabine Meinck, Ph.D., IEA, Design of International Large-Scale Assessments and Implications for Multilevel Modeling

Complex Sample Designs

Why “complex”?

- Multiple stages of sampling
- Homogenous clusters are sampled
- Stratification is used
- Selection probabilities differ for different sampling units

Multiple Sampling Stages



Sampling Schools in TIMSS

- Schools are sampled with **probability proportional to their size** from the list of all schools in the population
  - Larger schools are more likely to be selected than small ones
  - Students in large schools with many students at the target grade have a lower probability of selection than classes in smaller schools that have just one or two classes
  - Result: the overall probability of selection of students is more similar across different size schools
- Each stratum can have a different sampling rate. Strata could include location and race/ethnic composition, etc.
  - If we sample some schools at higher rates than others, aren't we introducing bias in our results?
- Full classes are chosen at random within each school's assessed grade (4 or 8)
  - All students in a class participate

Implications for Large-Scale Data Analysts

- Many statistical software packages assume the data being analyzed come from a simple random sample with independent observations
- But the usual statistical software usually cannot estimate variances accurately when observations are not independent
    - TIMSS's clustering of observations in schools and classes prevents them from being independent
    - Statistical software exists for "complex samples"
        - NCES's EdSurvey!

Stratification and Weights

- To account for unequal probabilities of selection, sampling weights should be used in all statistical calculations where inferences are made to populations
  - Otherwise, population estimates of means or percentages will be biased
  - Most standard software have ways to specify weights

TIMSS Student Weight

- Students are assigned sampling weights to adjust for over- or under-representation of particular groups in the final sample
- **Student weight is the inverse of the probability of selection**
- Students with higher weight values are representing more people in the target population
- Use of sampling weights is necessary for computation of sound, representative estimates
- Weights adjust for nonparticipation
- Sum of the overall student weights equals the number of students in the target population

Probabilities and Weights

- **Student weight is the inverse of the probability of selection**
- Suppose a school a probability of selection of 0.1 and each student within a school has a probability of selection of 0.2. What is the students’ probability of selection?

School prob	Student within school prob	Joint Prob
1/10	2/10	2/100

- Student weight = inverse probability
  - 100/2 = 50

Q: If a school has a 0.2 probability of selection and once school is selected the student has a 0.05 probability of selection, what is the students weight?

A. 100

B. 200

C. 20

D. 500

Sampling Weights

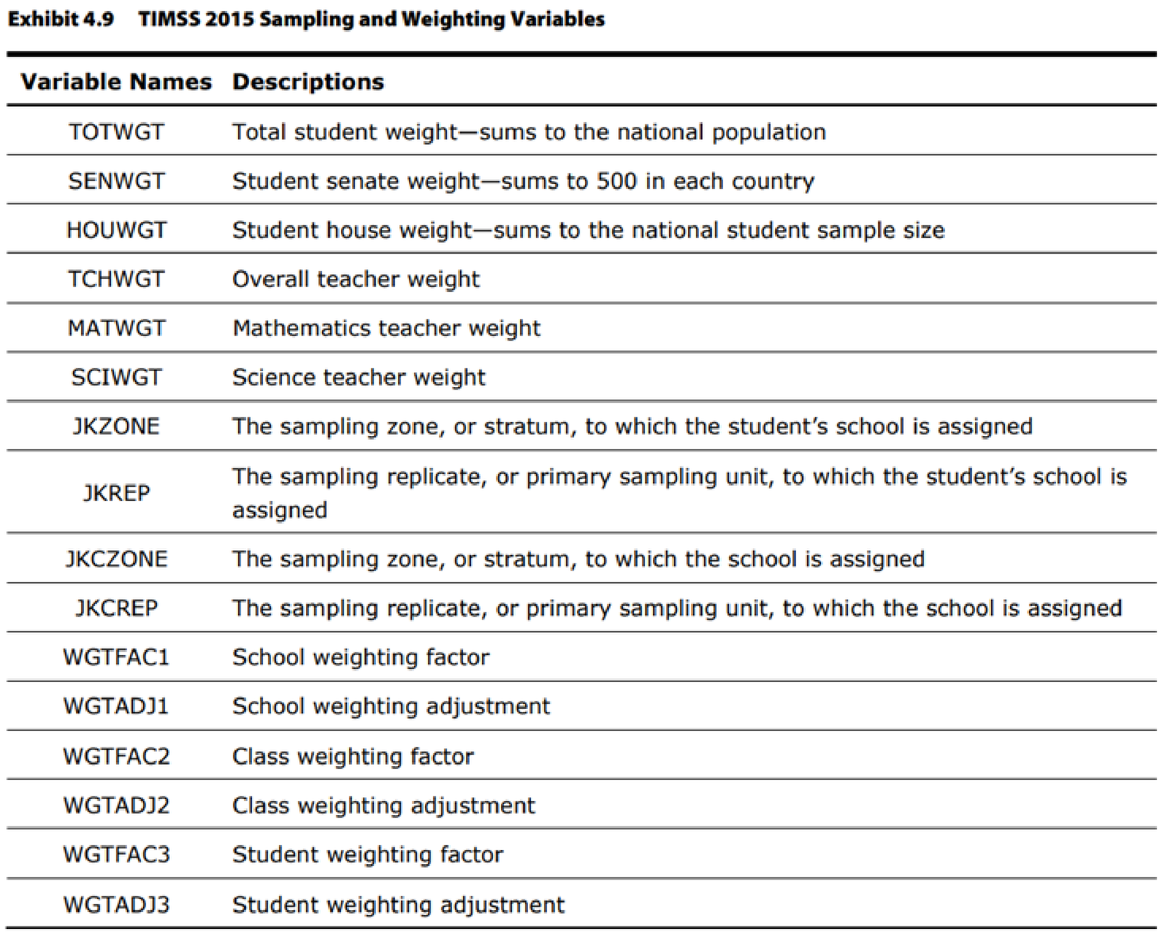


Exhibit 4.9 TIMSS 2015 Sampling and Weighting Variables

Variable Names	Descriptions
TOTWGT	Total student weight—sums to the national population
SENWGT	Student senate weight—sums to 500 in each country
HOUWGT	Student house weight—sums to the national student sample size
TCHWGT	Overall teacher weight
MATWGT	Mathematics teacher weight
SCIWGT	Science teacher weight
JKZONE	The sampling zone, or stratum, to which the student's school is assigned
JKREP	The sampling replicate, or primary sampling unit, to which the student's school is assigned
JKCZONE	The sampling zone, or stratum, to which the school is assigned
JKCREP	The sampling replicate, or primary sampling unit, to which the school is assigned
WGTFAC1	School weighting factor
WGTADJ1	School weighting adjustment
WGTFAC2	Class weighting factor
WGTADJ2	Class weighting adjustment
WGTFAC3	Student weighting factor
WGTADJ3	Student weighting adjustment

Single-level analysis

Student background and achievement data

- TOTWGT, SENWGT, HOUWGT (student-level)

Teacher background data

- MATWGT, SCIWGT (student-level)

Multilevel analysis

Need to construct weighting variables at different levels using weight factor and adjustment

The impact of clustering

- TIMSS’s clustering of observations in schools prevents them from being independent
- To account for cluster dependencies, variance estimates should be combined from separate estimates from within clusters
  - Otherwise, estimates of variances will be too small (and biased)
- Variances are essential in statistical tests of significance
  - Biased variances could make differences between scores appear significant when in fact they are not

How to Handle Clustering:

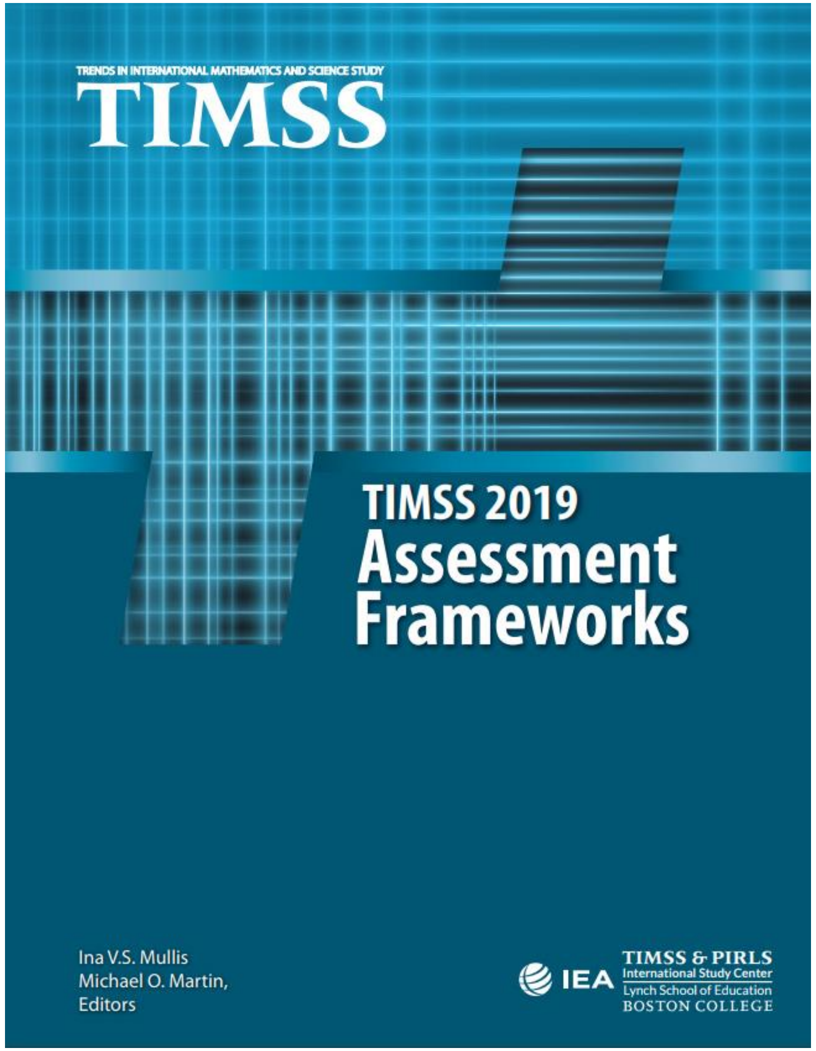
Three recommended methods

- Replication methods. TIMSS and NAEP uses jackknife repeated replication
- Hierarchical linear models
- Taylor series approximations

TIMSS Frameworks

- **Content** dimension that specifies the mathematics/science content to be assessed

- **Cognitive** dimension that specifies the thinking processes to be assessed



TRENDS IN INTERNATIONAL MATHEMATICS AND SCIENCE STUDY

TIMSS

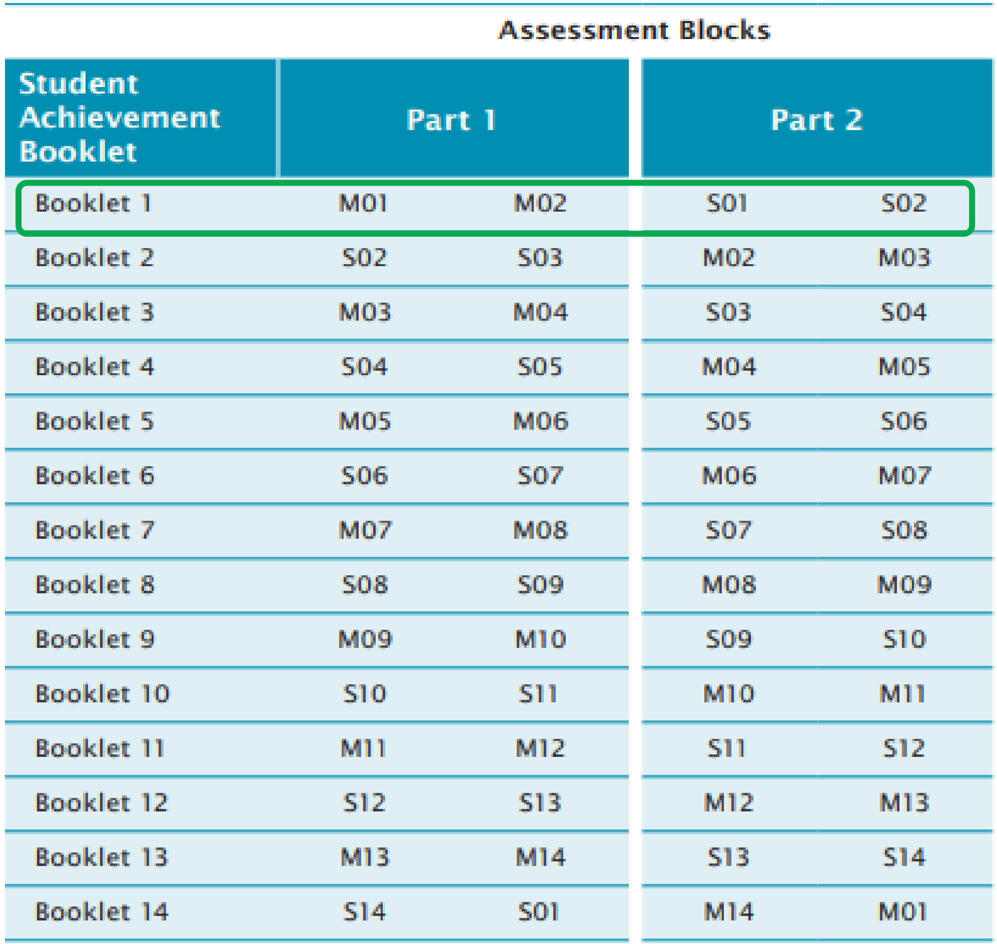
TIMSS 2019
Assessment
Frameworks

Ina V.S. Mullis
Michael O. Martin,
Editors

IEA

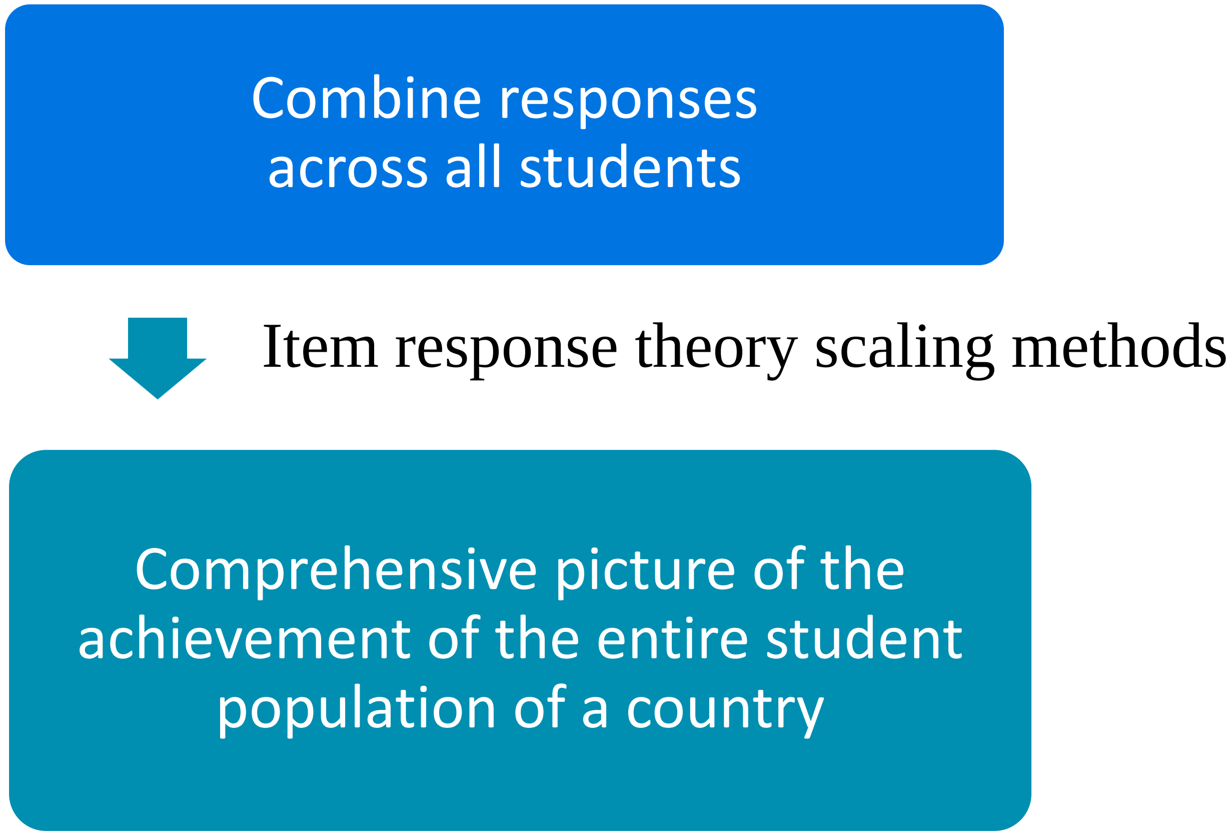
TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE

Complex Assessment Design



	Assessment Blocks			
Student 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

TIMSS uses a matrix-sampling approach



Combine responses across all students

Item response theory scaling methods

Comprehensive picture of the achievement of the entire student population of a country

TIMSS Achievement Scales

Range

0 – 1000

Standard deviation

100

Centerpoint

500

(set to correspond to the mean of overall achievement in 1995)

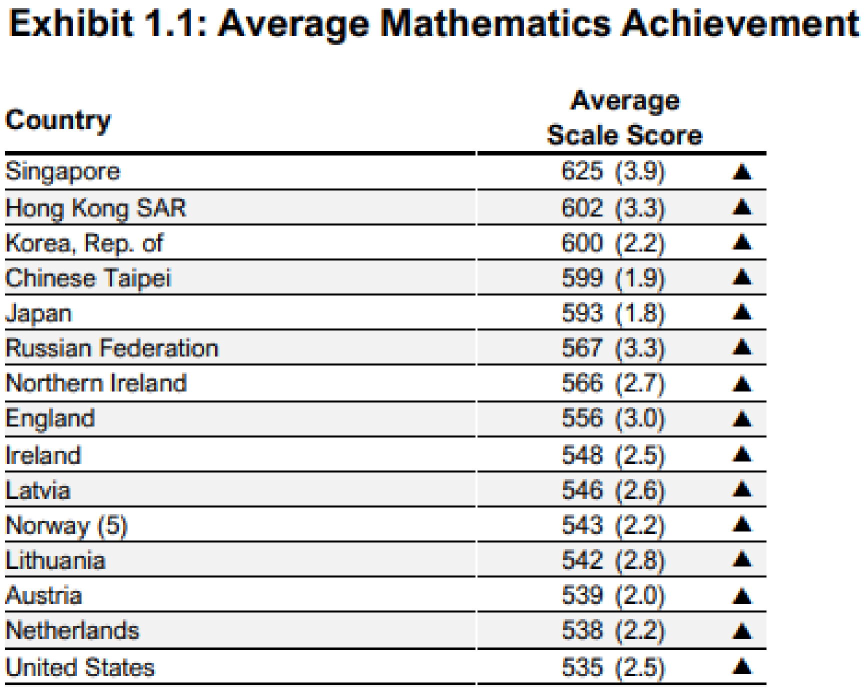
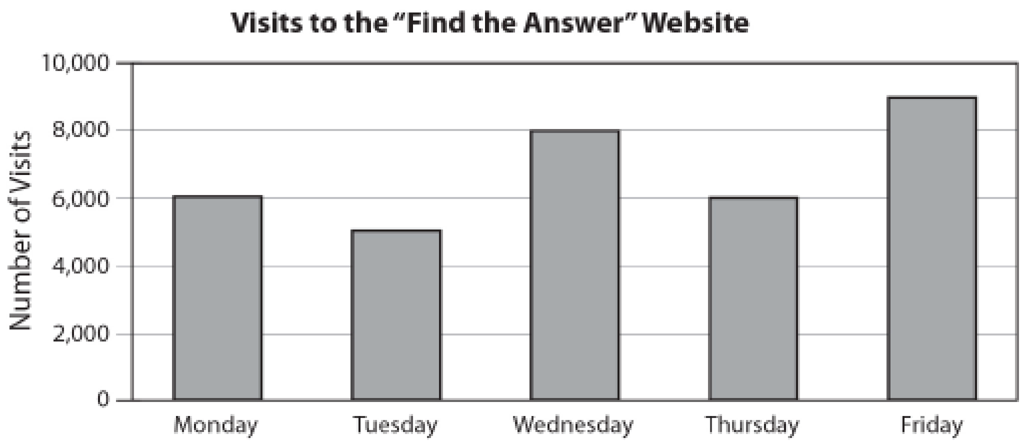


Exhibit 1.1: Average Mathematics Achievement

Country	Average Scale Score	
Singapore	625 (3.9)	▲
Hong Kong SAR	602 (3.3)	▲
Korea, Rep. of	600 (2.2)	▲
Chinese Taipei	599 (1.9)	▲
Japan	593 (1.8)	▲
Russian Federation	567 (3.3)	▲
Northern Ireland	566 (2.7)	▲
England	556 (3.0)	▲
Ireland	548 (2.5)	▲
Latvia	546 (2.6)	▲
Norway (5)	543 (2.2)	▲
Lithuania	542 (2.8)	▲
Austria	539 (2.0)	▲
Netherlands	538 (2.2)	▲
United States	535 (2.5)	▲

Example TIMSS Grade 4 Math Assessment Item



The chart shows the number of visits to the "Find the Answer" website.

How many visits were there on Wednesday?

Answer: 8,000

Content Domain: Data Display

Cognitive Domain: Knowing

Proficiency Level: Low

Percentage of students answering correctly:

- U.S. average: 90%
- International average: 84%

Questionnaires (and Example Components)

Student

- Home educational resources (socioeconomic status)
- Self-perceptions, beliefs, and attitudes about learning

Teacher

- Teacher’s years of experience
- Teacher’s job satisfaction

School (Principal)

- School climate and safety
- School resources and technology

Curriculum (Country Representatives)

- National Policies

Resources

IEA TIMSS International Database

IEA International Database Analyzer (IDB Analyzer) and its tutorials

NCES International Data Explorer (IDE)

NCES *EdSurvey* R statistical package

NCES *Dire* R statistical package

NCES Distance Learning Dataset Training Modules (DLDT)

NCES International Activities Program website

TIMSS and PIRLS International Study Center website

IEA website

ILSA Gateway

OECD PISA website

Advancing Evidence.
Improving Lives.

Questions?

Martin Hooper, PhD
American Institutes for Research
mhooper@air.org