

# **A Global Dataset on Education Quality**

Nadir Altinok, Noam Angrist and Harry Anthony Patrinos

January 2018

# Table of Contents

1. **Overview**
2. Motivation
3. Literature
4. This Paper's Contribution
5. Methodology
6. The Database
7. Results
8. Robustness Checks
9. Limitations and Future Research

# Overview

- **Largest, most current globally comparable panel dataset of education quality**
  - 130+ countries from 1965-2015 in 5-year intervals
  - 100+ developing countries
  - 90.9% of global population
- **Constructed by linking standardized, psychometrically-robust international and regional achievement tests**
  - Multiple linking methods, standard errors to quantify reliability
  - Compare distributions, in addition to country-level mean scores
  - Disaggregate by gender, socio-economic status, rural/urban, immigration status

# Table of Contents

1. Overview
- 2. Motivation**
3. Literature
4. This Paper's Contribution
5. Methodology
6. The Database
7. Results
8. Robustness Checks
9. Limitations and Future Research

# Motivation

- **Shift from a focus on *quantity* of schooling to *quality* of schooling**

- Strong link between quality of schooling and growth (e.g. Hanushek and Woessmann, 2012)



- **Need for globally comparable database on education quality**

- e.g. Barro-Lee global analogy (years of schooling) for quality of schooling (global)

- **Leverage emergence of International Standardized Achievement Tests (ISATs)** such as PISA and TIMSS - psychometrically constructed, robust and standardized international testing regimes to measure cognitive skills

- Include 60-70 countries every 3-4 years
- However often exclude developing countries and only started in the mid 1990s



- **Include more countries, in particular developing countries**

- Include more of global distribution of learning
- Have the most potential to gain from a quality education in terms of development

- **Expanded longitudinal dataset**

- There is much to learn from a rich array of achievement tests over time, although disparate and not standardized

# Table of Contents

1. Overview
2. Motivation
- 3. Literature**
4. This Paper's Contribution
5. Methodology
6. Results
7. Robustness Checks
8. Limitations and Future Research

# Literature

We build on a literature aiming to produce credible estimate of cognitive skills over time and across countries.

- **Barro (2001)** – regression to link ISATs
- **Hanushek and Kimko (2000)** – link tests over time using *NAEP*
- **Altinok and Murseli (2007)** – link regional tests with international using *doubloon countries* to include developing countries
- **Hanushek and Woessmann (2012)** – link education quality with growth
- **Angrist, Patrinos and Schlotter (2013)** and **Altinok et al. (2014)** – largest, most recent datasets on education quality including developing countries
- **Hanushek and Woessmann (2015)** – equate *variation* in addition to *levels* on ISATS

# Table of Contents

1. Overview
2. Motivation
3. Literature
- 4. This Paper's Contribution**
5. Methodology
6. The Database
7. Results
8. Robustness Checks
9. Limitations and Future Research



# This Paper's Contributions

- (1) Largest, most inclusive and recent dataset on education quality
- (2) Inclusion of multiple methods of linking for more reliable estimates
- (3) Distributional information on educational performance beyond mean scores
- (4) Disaggregation by gender, urban/rural, immigration status, language
- (5) Inclusion of standard errors
- (6) Robustness Tests

# Table of Contents

1. Overview
2. Motivation
3. Literature
4. This Paper's Contribution
- 5. Methodology**
6. The Database
7. Results
8. Robustness Checks
9. Limitations and Future Research

# Assessments Used

## (1) International Standardized Achievement Tests (ISATs)

- *Pre 1990s* – FIMS, FISS, SIMS, SISS, SRC, RLS, MLA and IAEP
- *Post 1990s*
  - **PISA** – OECD, 72 countries/areas, 15-year olds, math, reading, science, every 3 years
  - **TIMSS** – IEA, 63 countries/areas, grade 4 and 8, math and science, every four years
  - **PIRLS** - IEA, 60 countries/areas, grade 4, reading, every 5 years

## (2) Regional Standardized Achievement Tests (RSATs)

- **SACMEQ** - East and Southern Africa, 14 countries, grade 6, math, reading and science, every ~5 years
- **LLECE** – Latin America and the Caribbean, 15 countries, grades 3 and 6,
- **PASEC** – Francophone Africa, 10 countries, grades 2 and 5, math and reading

# Assessments Used – full table

<i>No</i>	<i>Year</i>	<i>Organization</i>	<i>Abbr.</i>	<i>Subject</i>	<i>Countries/ Areas</i>	<i>Grade/Age</i>	<i>Included</i>
1	1959-1960	IEA	Pilot Study	M,S,R	12	7,8	
2	1964	IEA	FIMS	M	12	7, FS	■
3	1970-71	IEA	SRC	R	15	4,8, FS.	
4	1970-72	IEA	FISS	S	19	4,8, FS.	■
5	1980-82	IEA	SIMS	M	19	8, FS	■
6	1983-1984	IEA	SISS	S	23	4,8, FS	■
7	1988, 1990-91	NCES	IAEP	M,S	6, 19	4,7-8	■
8	1990-1991	IEA	RLS	R	32	3-4,7-8	■
9	Every four years since 1995 (latest round is 2015)	IEA	TIMSS	M,S	45, 38, 26, 48, 66, 65, 65	3-4,7-8, FS	■
10	1992-97	UNESCO	MLA	M,S,R	72	6,8	■
11	1997, 2006, 2013	UNESCO	LLECE	M,S,R	13, 16 (only 6 for science)	3,6	■
12	1999, 2002, 2007	UNESCO	SACMEQ	M,R	7, 15, 16	6	■
13	1993-2001,2002-2012, 2014	CONFEMEN	PASEC	M,R	22 (before 2014), 10	Until 2014: 2,5 After 2014: 3, 6	■
14	Every five years since 2001 (latest round is 2011)	IEA	PIRLS	R	35, 41, 55	4	■
15	Every three years since 2000 (latest round is 2015)	OECD	PISA	M,S,R	43, 41, 57, 74, 65, 71	Age 15	■

# Assumptions

## **(1) Same underlying population**

- Given we are using sample-based ISATs and RSATs and equate using overlapping countries (“doubloon” countries, this assumption is satisfied if the population tested is similar and participation rates reach a certain threshold or non-participation is random

## **(2) Tests measure similar proficiencies**

- We link across precise dimensions such as subject and schooling level (primary vs. secondary) to increase proficiency overlap.

## **(3) Differences are test-fixed effects not country-fixed effects**

- We address this assumption by equating using an average across countries that participate in both tests. The reliability of the equating exercise is enhanced with an increase in the number of countries that take both tests being equated.

# Overarching Linking Intuition

The foundation for our approach is to index across a given pair of international and regional achievement tests with results from countries that participate in both (“doubloon” countries)

To link results over time, we perform a similar procedure using the United States as an anchor since it has participated in all IEA assessments since 1965 as well as a consistently administered national assessment, the National Assessment of Educational Progress (NAEP).

# Specific Linking Methods

- **Mean:** constant adjustment between mean scores of “doubloon countries” from the two assessments.
  - Assumption: assumes similar distributions, which is often unlikely

$$\text{linking}_Y^m(x) = y = x - \mu(X) + \mu(Y)$$

- **Linear:** linear adjustment using means and standard deviations with a linear relationship
  - Assumption: allows for differences in difficulty along test score scales; assumes similar standard deviations across tests and over time, which is unlikely

$$\text{linking}_Y^l(X) = y = \sigma(Y) \left[ \frac{x - \mu(X)}{\sigma(X)} \right] + \mu(Y)$$

# Specific Linking Methods

- **Pseudo-Linear:** mean linking with a coefficient based on means and not only an additive translation:
  - Enables more over-time comparably since not sensitive to standard deviation changes over time.

$$\text{linking}_Y^{pl}(X) = y = \frac{\mu(Y)}{\mu(X)} x$$

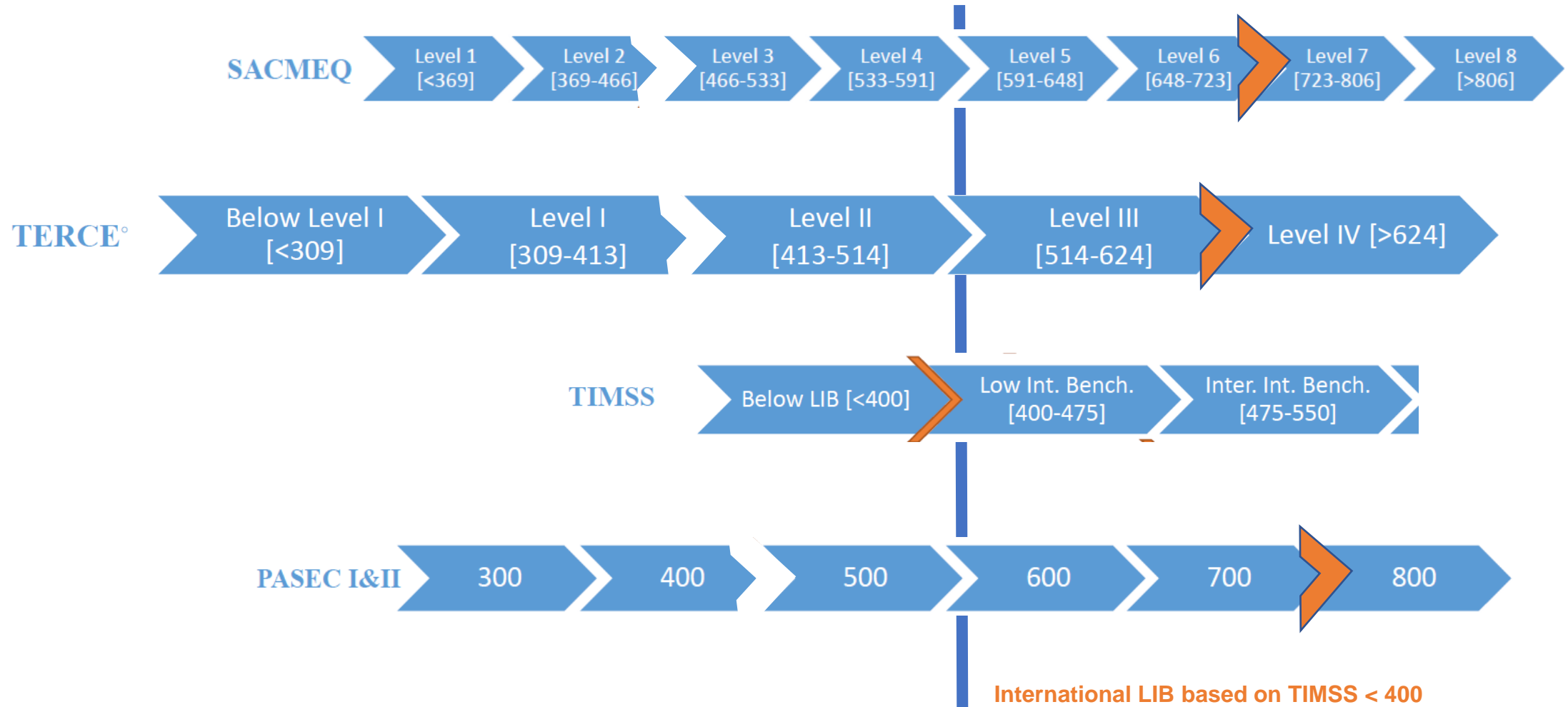
- **Pre-smoothed Equipercentile:** anchor based on equivalent percentiles rather than mean scores (take inverse of cumulative distribution); smooth distribution where no score so no discontinuities
  - Best used when X and Y differ nonlinearly in difficulty (e.g. on high vs. low scores).

$$\text{linking}_Y^e(X) = G^{-1}[F(x)]$$



# Threshold definitions example

## Comparison of benchmarks in math - primary education



# Summary

1. Mean Scores

Pseudo  
-linear  
method

2. Percent of Students Reaching **Minimum** Proficiency Threshold

3. Percent of Students Reaching **Intermediate** Proficiency Threshold

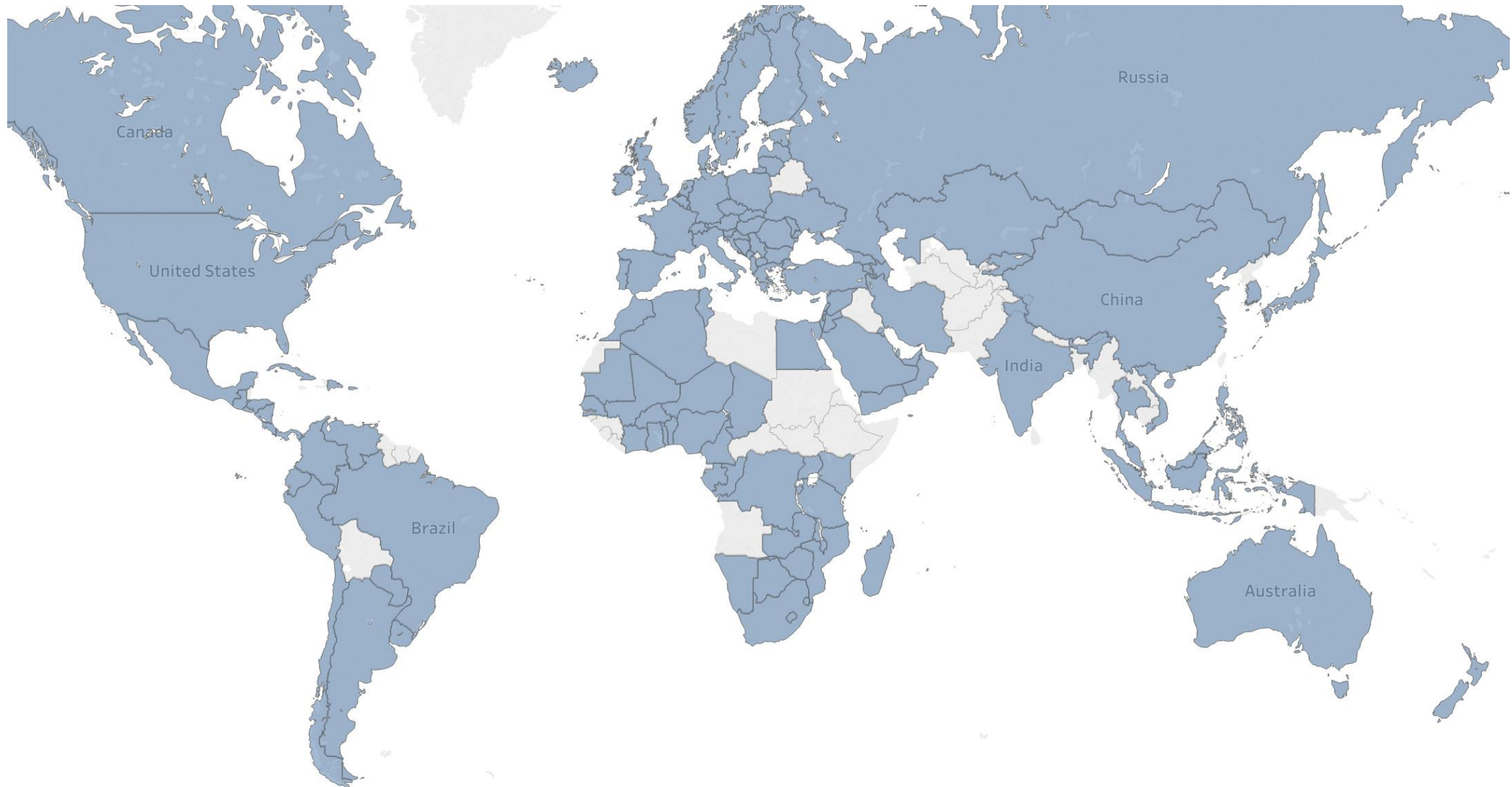
4. Percent of Students Reaching **Advanced** Proficiency Threshold

Equi-  
percentile  
method

# Table of Contents

1. Overview
2. Motivation
3. Literature
4. This Paper's Contribution
5. Methodology
- 6. The Database**
7. Results
8. Robustness Checks
9. Limitations and Future Research

**1965-2015:**  
131 countries, 163 areas, 90.9% of global population



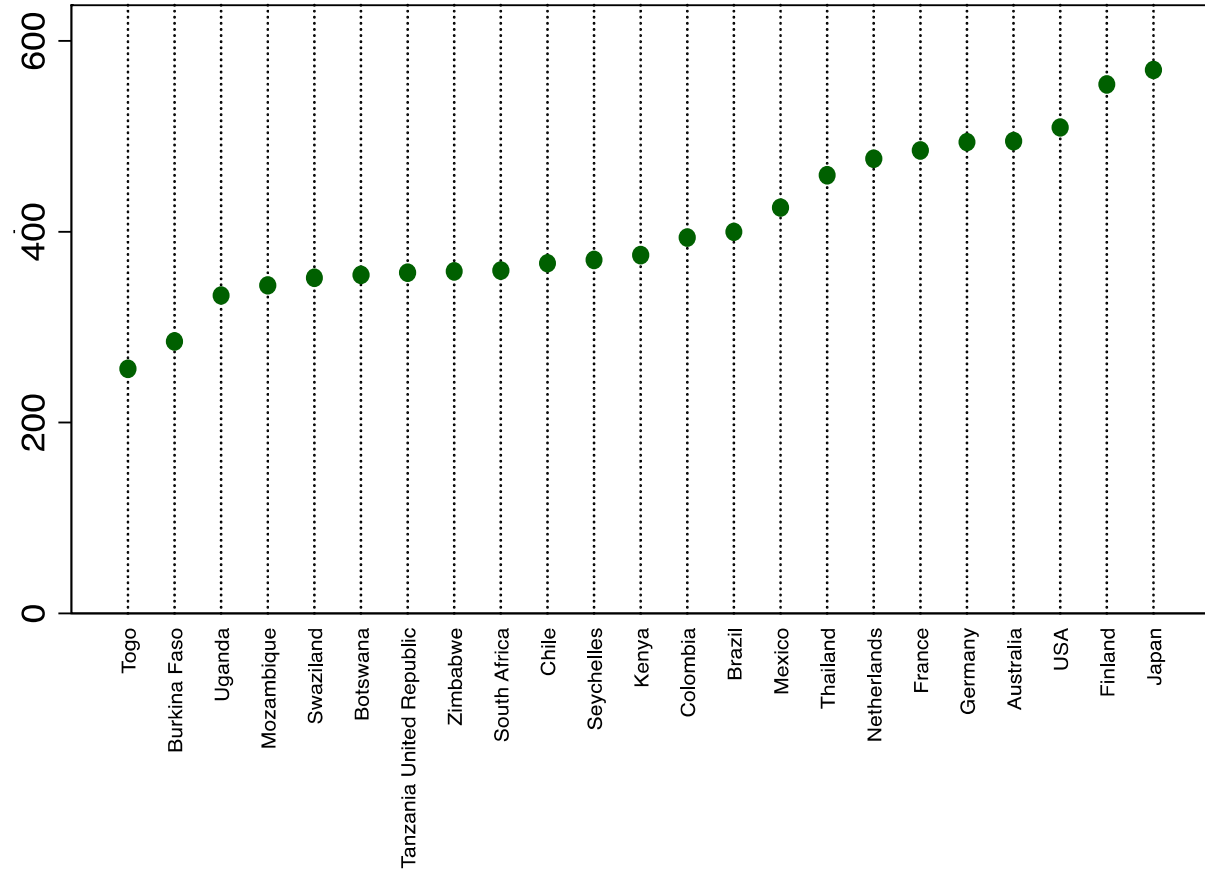
# Table of Contents

1. Overview
2. Motivation
3. Literature
4. This Paper's Contribution
5. Methodology
6. The Database
- 7. Results**
8. Robustness Checks
9. Limitations and Future Research

## Insights / Takeaways

- Developing Countries Cluster at the Bottom of a Global Scale
- Sub-Saharan Africa trails the global distribution, followed by Latin America

### Mean Scores for Selected Developed and Developing Countries (1965-2015)

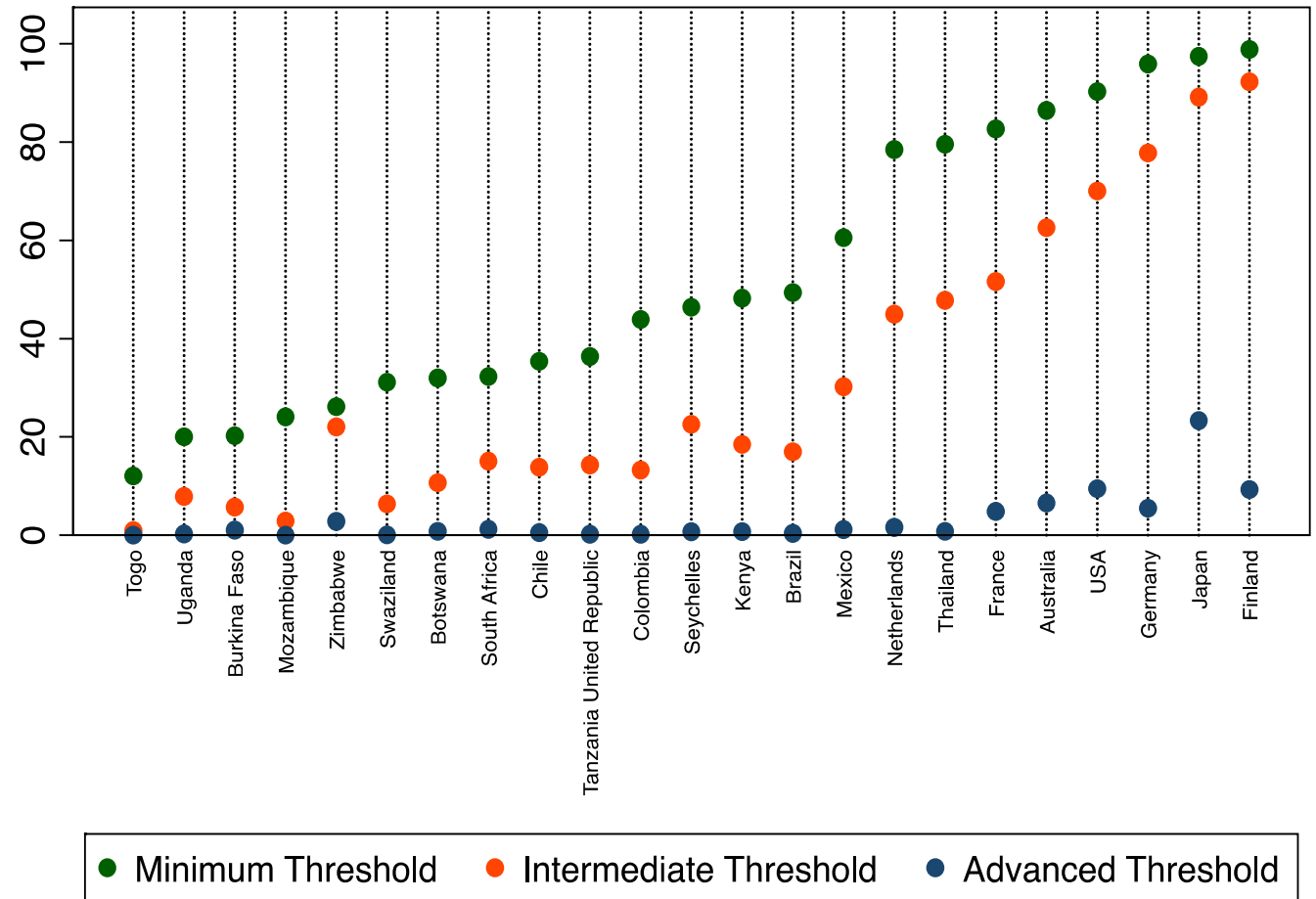


## Insights / Takeaways

We know developing countries do worse – we **quantify** the gap, up to 5x:

- Less than 50% of kids reach the minimum thresholds in developing countries (vs. 86% in developed countries)
- The percent of students reaching intermediate proficiencies in developed countries is often greater than the percentage of students reaching basic thresholds of proficiency in developing countries

### Percent of Students Reaching Minimum, Intermediate and Advanced Proficiency Benchmarks for Selected Developed and Developing Countries (1965-2015)

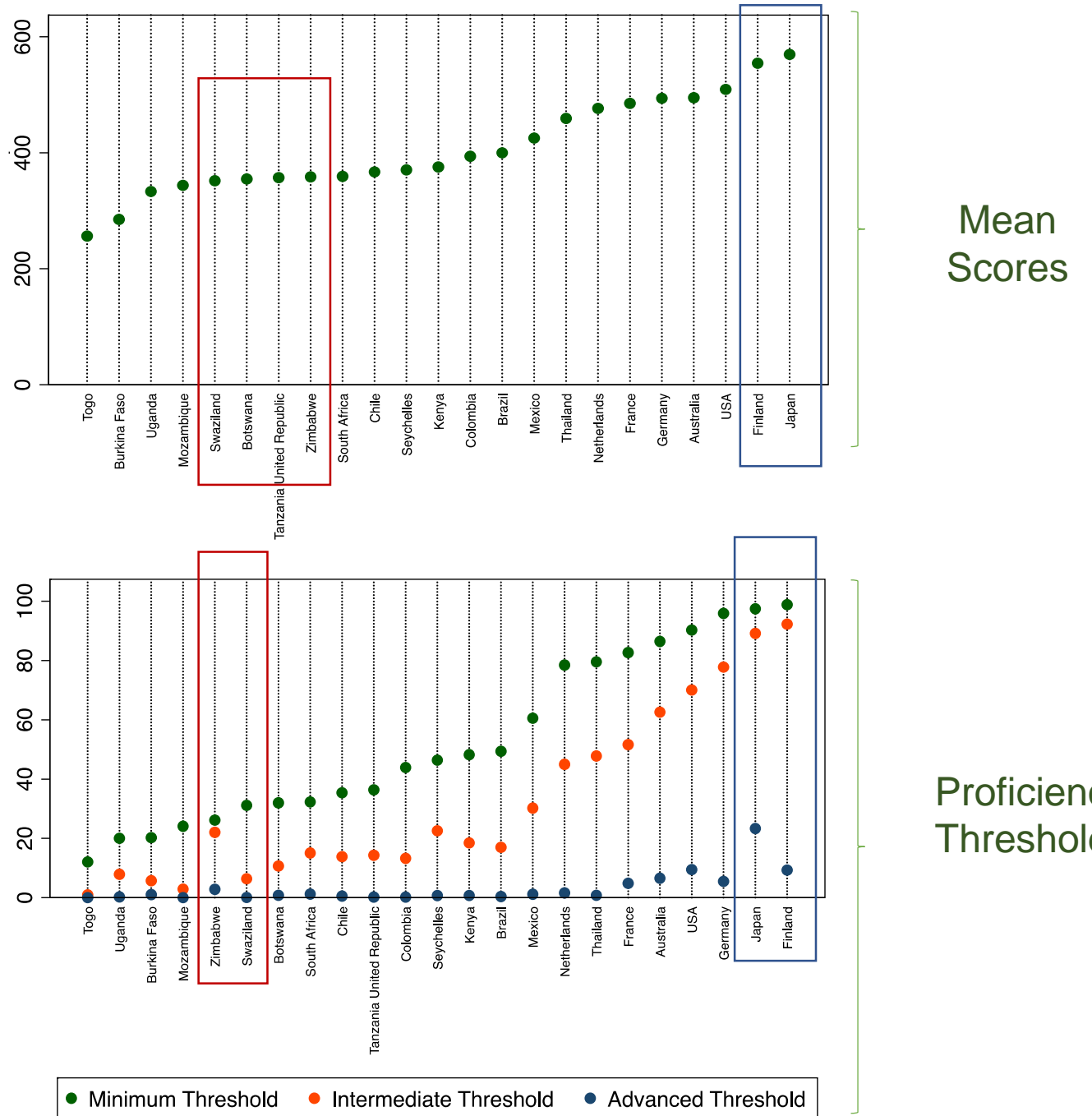


## Insights / Takeaways

Distributional performance reveal insights not captured by mean scores – for example differences in potential “innovation” / technological frontier education vs. mass education

*Examples:*

- Japan vs. Finland
- Zimbabwe vs. Swaziland

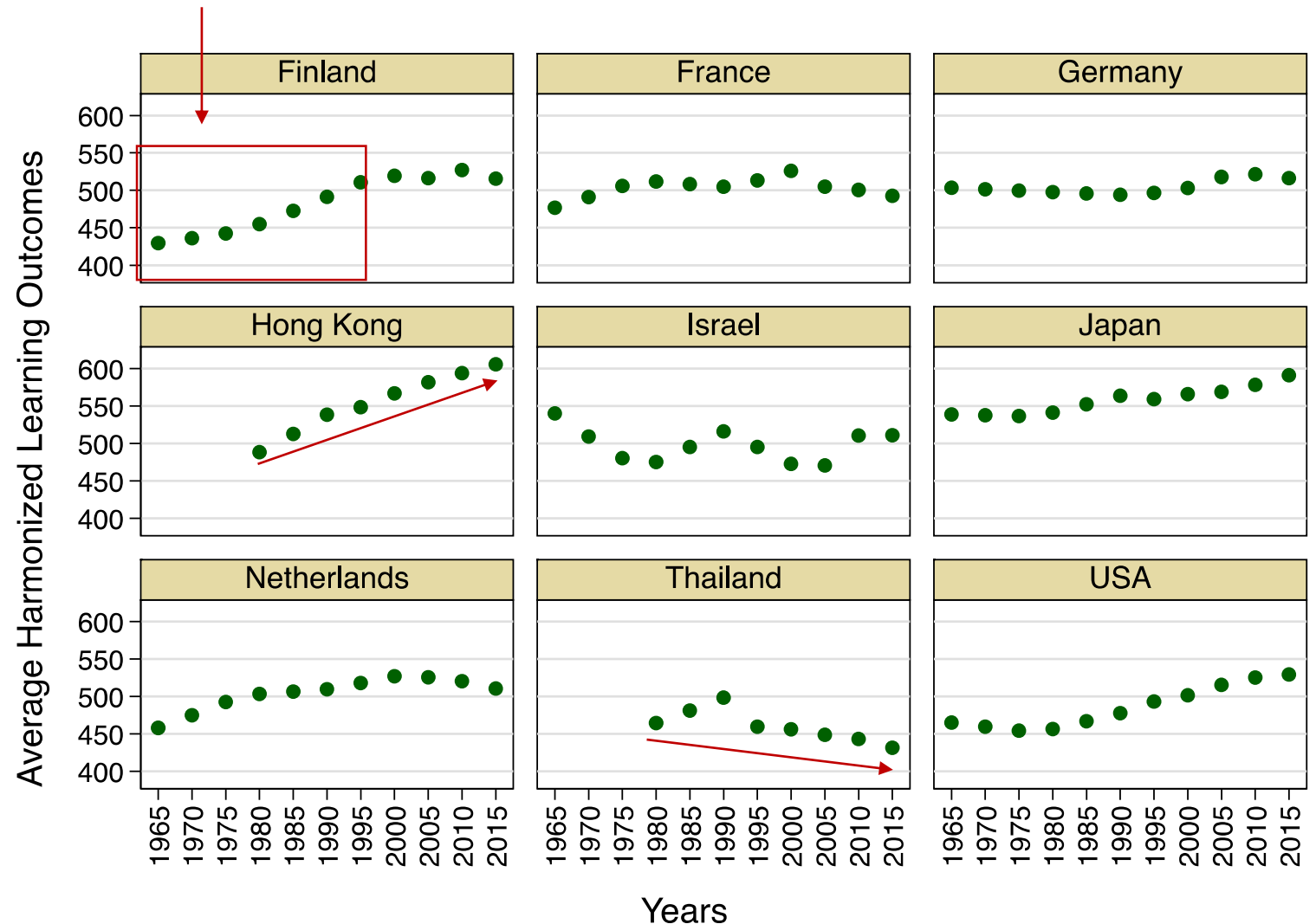




Finnish progress occurs during reforms of 1960s and 1970s (comprehensive school reform of 1972-1977), long before PISA 2000s, so current theories likely don't account for much.

Select economies reveal stories of successful versus failed education policy reform via longitudinal data on education quality (e.g. *Finland, Hong Kong vs Thailand*)

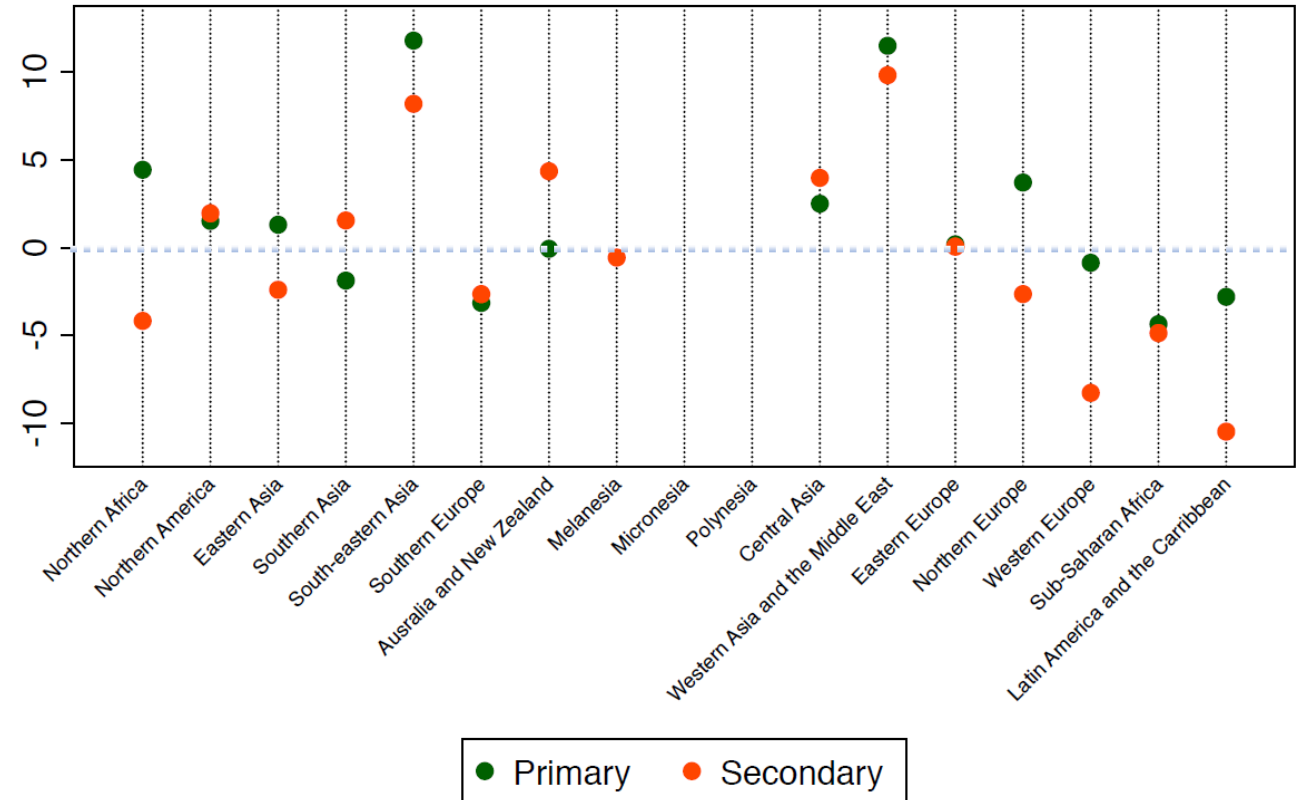
- Longitudinal data reveals scope for further understanding
- Availability of data necessary first step but certainly not enough



## Insights / Takeaways

- Gender premiums are surprisingly small and vary widely by region
- Girls do best in the Middle East and Southern Asia, potentially driven by a selection effect.

Female Learning Premium (1965-2015)  
By region



# Learning is strongly associated with growth 3x higher for developing vs. developed countries



# Table of Contents

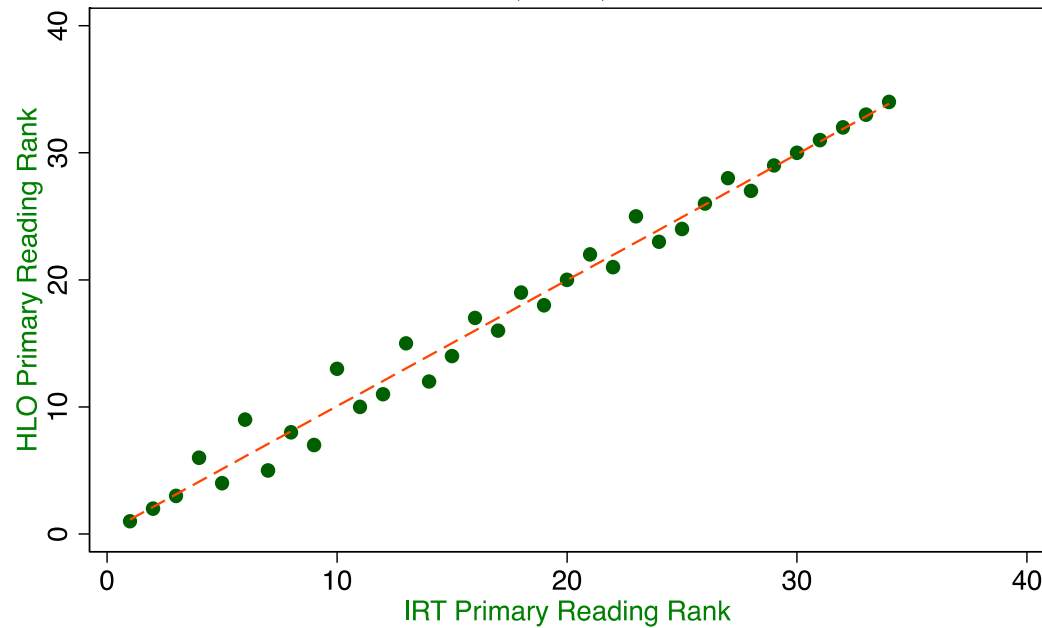
1. Overview
2. Motivation
3. Literature
4. This Paper's Contribution
5. Methodology
6. The Database
7. Results
- 8. Robustness Checks**
9. Limitations and Future Research

# Item Response Theory (IRT): Comparison to *LINCS* project where there is item overlap

*Overall Pearson Correlation of .98+ 2005,2010*

Ranks - Primary Reading Scores

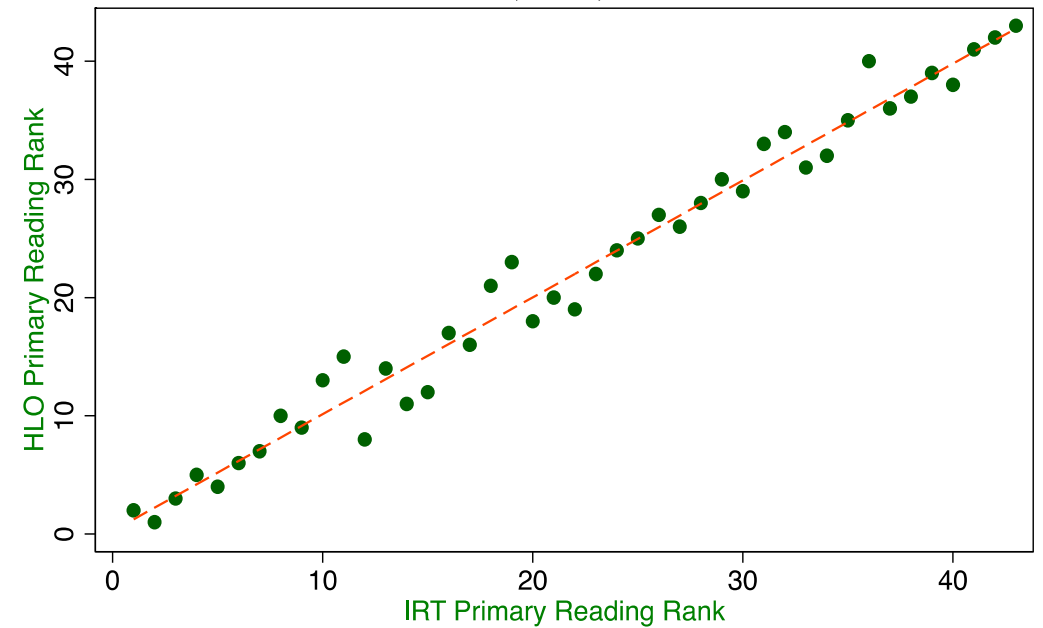
HLO vs. IRT, 2005, 34 countries



Spearman Rank Correlation Coefficient: .9914438502673799

Ranks - Primary Reading Scores

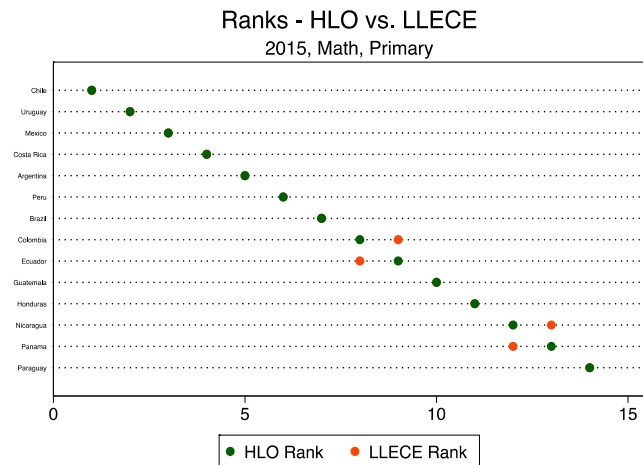
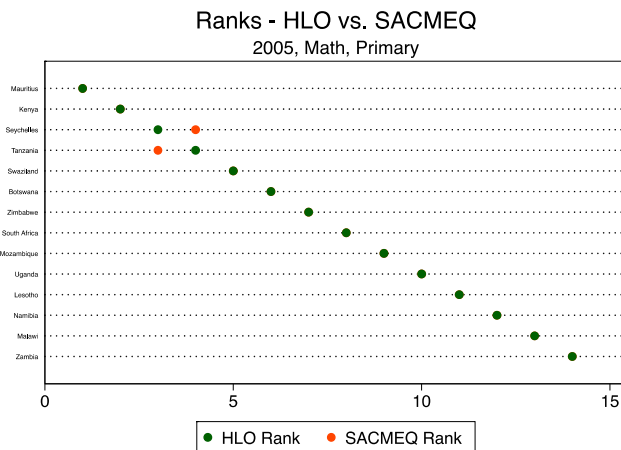
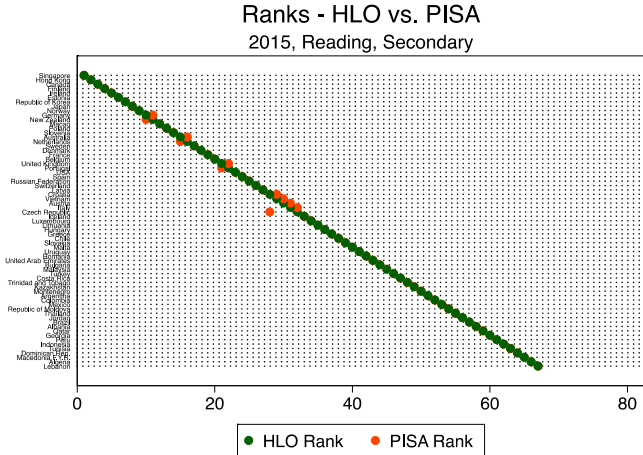
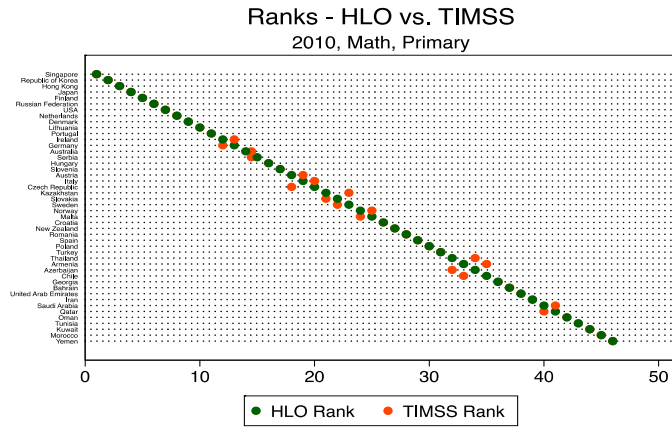
HLO vs. IRT, 2010, 43 countries



Spearman Rank Correlation Coefficient: .9885231048021746

# Original Scores:

Comparison of ranks based on both international and regional;  
often direct overlap, marginal rank changes (avg rank change is 0)



# Table of Contents

1. Overview
2. Motivation
3. Literature
4. This Paper's Contribution
5. Methodology
6. The Database
7. Results
8. Robustness Checks
- 9. Limitations and Future Research**

# Utility of Dataset

- **Monitor global education policy goals**
  - There is a movement towards a global, uniform test of ability – until then, there is a need to link various disparate testing regimes to monitor progress on global goals such as SDG 4 and *Education For All*
- **Uncover causal links and drivers of education quality and development**
  - Dependent variable credibly comparable across countries and over time
  - Exploit larger, more inclusive panel dataset to better control for country and time factors
    - Cross-sectional as well as longitudinal (PISA and TIMSS only extend back to mid 1990s)
  - Easily merge-able with other measures of interest: GDP, HDI, etc.



# Limitations

- **Potential selection effects** – e.g. enrollment, retention, makes it harder to capture “value added” learning
- **Doubloon countries** – limited number of doubloon countries, making the reliability of the index tenuous
- **Standard Deviation and Means are often arbitrary** – converting into meaningful units is difficult
- **Precision vs. Coverage trade-off**
  - Grade, Subject, Years
  - Disaggregation (e.g. Gender, Urban/Rural, SES, etc)
- **Proficiency threshold definitions (min, intermediate, advanced)** - either skews relevance for developed or developing countries

# Future Research

- Consistently update every few years
  - to including more developing countries, enhance longitudinal dimension, increase overall robustness as more “double-counted countries” are included, keep current, upgrade methodology
- Enhance country coverage (include EGRA/EGMA, ASER)
- Build up dataset to enable better capturing of “value-added” learning
  - Link quality data to quantity data – e.g. Average Years of Schooling (Barro-Lee or UIS)
  - Account for selection effects via enrollment and retention
- Include additional measures of education quality
  - E.g. “returns to schooling” data
- Explore causal links between education quality and development outcomes

# Contact

**Nadir Altinok**, University of Lorraine | [nadir.altinok@univ-lorraine.fr](mailto:nadir.altinok@univ-lorraine.fr) | [@AltinokNadir](https://twitter.com/AltinokNadir)

**Harry Patrinos**, World Bank | [hpatrinos@worldbank.org](mailto:hpatrinos@worldbank.org) | [@hpatrinos](https://twitter.com/hpatrinos)

**Noam Angrist**, University of Oxford | [nangrist@bsg.ox.ac.uk](mailto:nangrist@bsg.ox.ac.uk) | [@angrist\\_noam](https://twitter.com/angrist_noam)