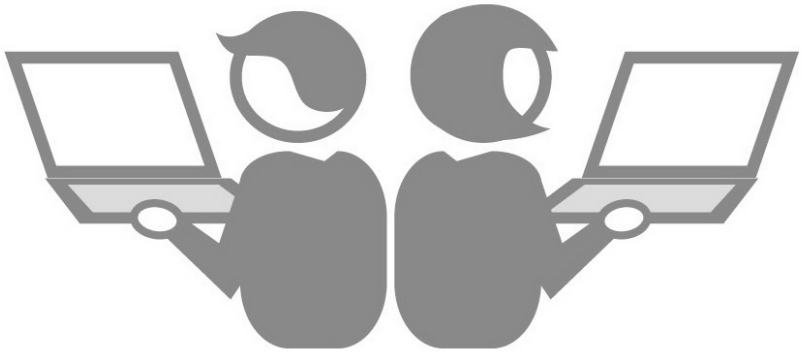
Does Process Data Add Value to the Analysis of International Large-Scale Assessment Data?



Dihao Leng

TIMSS & PIRLS International Study Center

66th IEA General Assembly Meeting

Madrid, Spain, October 8th, 2025

IEA

TIMSS & PIRLS
BOSTON COLLEGE

Acknowledgements

- This research was supported by the Doctoral Dissertation Fellowship from the Lynch School of Education and Human Development at Boston College.
- The work was carried out at the TIMSS & PIRLS International Study Center at Boston College.
- The author expresses deep gratitude to the IEA and participating countries for making public-use data available and for making collaborative international research possible.
- Special thanks to my dissertation chair, Matthias von Davier, as well as my committee members and coauthors for their invaluable guidance and support.

This study aims to advance the goal of creating more equitable and less demanding assessments for students by exploring how process data can enhance educational measurement.

Table of Contents

Introduction

- Event data / Log data: time-stamped records of every action a test-taker performs while interacting with a computer-based assessment system, typically stored in log files.

item_id	item_title	variable_id	startTime	endTime
39102	TheAmazingOctopus	SCROLLBAR_PAGE_6	2021-03-26 01:19:58.902	NULL
39102	TheAmazingOctopus	QUESTION_OPEN	2021-03-26 01:20:05.056	NULL
39102	TheAmazingOctopus	QUESTION_NEXT	2021-03-26 01:21:40.226	NULL
39102	TheAmazingOctopus	NULL	2021-03-26 01:21:49.518	NULL
39102	TheAmazingOctopus	QUESTION_SAVE	2021-03-26 01:21:49.520	NULL
39103	EnjoymentSurvey_AmazingOctopus	DIRECTIONS_END	2021-03-26 01:21:55.141	NULL

- Process data: derived from log data, reflecting students' test-taking processes.

(Bergner & von Davier, 2019；Bezirhan et al., 2021; Goldhammer & Zehner, 2017; He & von Davier, 2015)

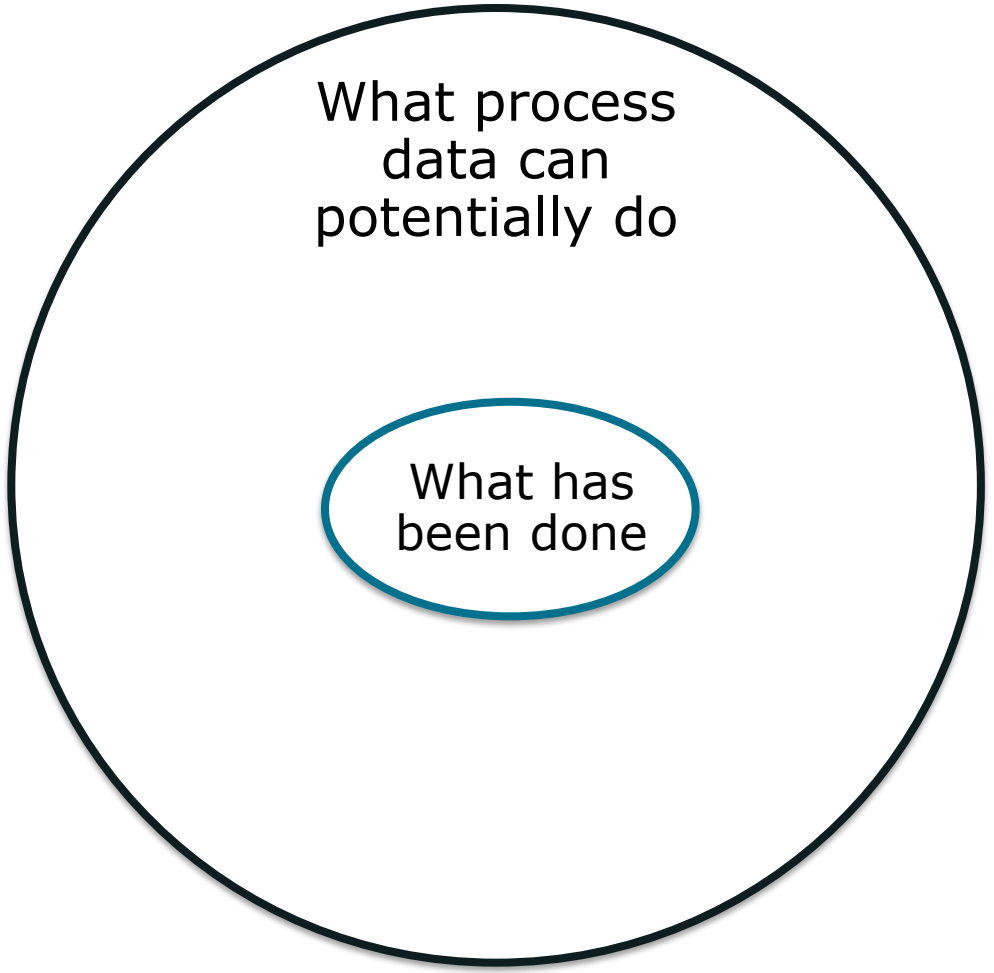
Introduction: Common Uses of Process Data in Educational Assessments

- Understanding response processes and explaining outcomes
- Examining group differences in response processes and outcomes
- Scoring at the item- or task-level
- Improving achievement estimation
- Supporting quality assurance and validation
- Informing assessment design

(Greiff et al., 2015; Bezirhan et al., 2021; Bergner & von Davier, 2019; van der Linden, 2011; Yamamoto & Lennon, 2018; Wang & Xu, 2015)

Introduction

- However…



Introduction: Challenges in Collecting and Maintaining Process Data

- Resource intensive: Requires careful planning, adapted software infrastructure, and extra computing
- Ethical concerns: Privacy and accountability issues

Does Process Data Add Value to the Analysis of International Large-Scale Assessment Data?

Overview of The Three Papers

	Paper 1	Paper 2	Paper 3
Title	Examining Gender Differences in TIMSS 2019 Using a Multiple Group Hierarchical Speed-Accuracy-Revisits Model	Leveraging Process Data and Variable Selection for Achievement Estimation in Large-Scale Assessments	Gender Differences and Types of Test-Taking Behaviors in PIRLS 2021
Research Question(s)	To what extent do mathematics ability, response speed, revisit propensity, and the relationships among them differ between boys and girls?	Can integrating process data and variable selection reduce model complexity in achievement estimation while maintaining measurement precision, compared to the conventional PCA-based approach?	• What classes of test-taking behaviors can be identified among students? • How do gender gaps in reading achievement differ across these behavioral classes? • How would the overall gender gap change if the distribution of the behavioral classes among boys were the same as that of girls?
Use of Process Data	• Understanding Response Processes • **Uncovering Group Differences** • Proposed model can also be used for other purposes such as quality assurance	**Improving the Estimation of Achievement**	• Understanding Response Processes • **Uncovering Group Differences**

Educational Measurement
ISSUES AND PRACTICE
NCME
Check for updates

Educational Measurement: Issues and Practice
April 2024, Vol. 0, No. 0, pp. 1–12

Examining Gender Differences in TIMSS 2019 Using a Multiple-Group Hierarchical Speed-Accuracy-Revisits Model

Dihao Leng, Ummugul Bezirhan, Lale Khorramdel, Bethany Fishbein, and Matthias von Davier, *Boston College, Chestnut Hill, United States*

Abstract: *This study capitalizes on response and process data from the computer-based TIMSS 2019 Problem Solving and Inquiry tasks to investigate gender differences in test-taking behaviors and their association with mathematics achievement at the eighth grade. Specifically, a recently proposed hierarchical speed-accuracy-revisits (SAR) model was adapted to multiple country-by-gender groups to examine the extent to which mathematics ability, response speed, revisit propensity, and the relationship among them differ between boys and girls. Results across 10 countries showed that boys responded to items faster on average than girls, and there was greater variation in boys' response speed across students. A mixture distribution of revisit propensity was found for all country-by-gender groups. Both genders had moderate to strong negative correlations between mathematics ability and response speed, supporting the speed-accuracy tradeoff pattern reported in the literature. Results are discussed in the context of low-stakes assessments and in relation to the utility of the multiple-group SAR model.*

Keywords: gender gaps, mathematics achievement, process data, test-taking behaviors, TIMSS

Gender gaps in achievement on international large-scale assessments have been extensively reported across countries and over time. These achievement gaps vary in direction and magnitude depending on the particular country and domain assessed. Specifically, girls consistently have higher average reading achievement across assessments and countries (Borgonovi, 2022; Mullis et al., 2017; Smith & Wilhelm, 2009), while the gender achievement gap is smaller overall and less consistent in mathematics assessments (Ghasemi & Burley, 2019; Meinck & Brese, 2019; Mullis et al., 2020).

Previous studies have investigated gender differences associated with gender stereotypes (Igbo et al., 2015; Spencer et al., 1999; Tsui, 2007), specific cognitive abilities (Ganley et al., 2014; Maeda & Yoon, 2013), test-taking motivation (DeMars et al., 2013; Eklöf, 2010), and attributes of the test including item format (Beller & Gafni, 2000; Schwabe et al., 2015; Solheim & Lundetræ, 2018) and differential item functioning (Holland & Wainer, 1993; Khorramdel et al., 2020). These investigations have mainly used traditional achievement or item response data. Differences in test-taking behaviors are assumed to be a potential source of observed gender gaps (Baldiga, 2014; Bezirhan et al., 2021; Peng et al., 2014; Stenlund et al., 2017), possibly moderating other factors, but have been historically difficult to measure using data from paper-based assessments.

With the move of large-scale assessments from paper- to computer-based administration, additional data on students' test-taking behaviors become available and can be examined in relation to examinees' ability (Bezirhan et al., 2021; He et al., 2019; von Davier et al., 2019). Low-level event data stored in log files contain detailed traces of human-computer interactions such as keystrokes and mouse clicks; derived variables or aggregates of event data can be further generated, providing information about students' cognitive or behavioral processes during a test (e.g., Bergner & von Davier, 2019; Demirkaya et al., 2023; Goldhammer & Zehner, 2017). For example, students' response times can be derived from lower-level timestamp data, supporting inferences about how students allocate their speed during a test. This type of data, referred to as *process data*, has the potential to provide deeper insights into examinees' response processes beyond traditional paper-based assessments, where only students' responses and scores are available (e.g., Bezirhan et al., 2021; He & von Davier, 2015).

In computer-based assessments, students can usually navigate the assessment at their own pace and revisit previous screens. While some students visit each item only once, others may visit items multiple times to review items, change answers, or strategically allocate the time spent on the test. In this manner, item revisit behavior provides additional information about students' test-taking behaviors beyond response accuracy and response time.

Sophisticated modeling approaches have been developed to model response time and item revisit behaviors (Bezirhan et al., 2021; Jeon et al., 2017; van der Linden, 2007, 2011; van der Linden & Jeon, 2012). In particular, Bezirhan et al. (2021) proposed the hierarchical speed-accuracy-revisits (SAR) model and found a strong positive correlation between

Under review after revision. Please do not cite or circulate without the authors' permission.

Leveraging Process Data and Variable Selection for Achievement Estimation in Large-Scale Assessments

Dihao Leng, Ummugul Bezirhan, Matthias von Davier
TIMSS & PIRLS International Study Center, Boston College

Abstract

The operational scaling in large-scale assessments (LSAs) comprises two stages: item response theory and latent regression modeling. Principal component analyses (PCA) are routinely performed before latent regression modeling for dimension reduction, but this approach retains too many principal components (PCs), threatening numerical stability. This study proposes a new approach: adding process variables to the usual contextual variables and replacing PCA with variable selection for latent regression modeling. We found that using Lasso, random forests, and ultimately gradient boosting for variable selection led to measurement precision similar to or higher than the PCA approach but with considerably fewer covariates; including process variables into latent regression models consistently improved measurement precision. Integrating process data and variable selection yielded the highest measurement precision while achieving parsimony: The latent regression models with the 150 most important variables, including process variables, outperformed those with 330 PCs in France and 270 PCs in the Republic of Korea. The findings suggest that our proposed approach can effectively solve the overparameterization problem in LSA scaling while preserving measurement precision.

APPLIED MEASUREMENT IN EDUCATION
https://doi.org/10.1080/08957347.2025.2486032
Routledge
Taylor & Francis Group
Check for updates

Gender Differences and Types of Test-Taking Behaviors in PIRLS 2021

Dihao Leng and Matthias von Davier

TIMSS & PIRLS International Study Center, Boston College

ABSTRACT

Girls consistently demonstrate higher average reading achievement across countries and assessments by a large margin, yet existing theories inadequately explain this gap. This study uses PIRLS 2021 process data from Finland, Germany, Singapore, and the United Arab Emirates to examine test-taking behavior as a window into gender gaps in reading achievement. Using latent class analysis, three types of test-taking behaviors emerged: "Rapid," "Challenged," and "Engaged." The "Rapid" class in all four countries had more boys, while the "Engaged" class contained more girls. Using BCH and three-step methods to further disaggregate the gender gap reveals that girls in the "Rapid" class outperformed boys across countries, while the average gender achievement gap almost disappeared in the "Engaged" class, except in Singapore. The gender gap in reading achievement may diminish by a mild to moderate extent if more boys were engaged. Findings highlight the importance of addressing disengagement issues in low-stakes assessments.

1. Introduction

Girls have been consistently outperforming boys in reading achievement on average, regardless of country, culture, or education system, as observed in multiple international large-scale assessments (Borgonovi, 2022; Mullis et al., 2009, 2017, 2023; OECD, 2016, 2019, 2023; Solheim & Lundetræ, 2018). One of the most recent pieces of evidence comes from the Programme for International Student Assessment (PISA) 2022 (OECD, 2023). Out of all 81 OECD countries and regions tested, girls outperformed boys in reading in 79, with an average difference of 24 points across OECD countries. A similar advantage for girls was reported in Progress in International Reading Literacy Study (PIRLS) 2021, where girls showed higher reading achievement than boys in 51 of the 57 countries tested, with girls averaging an 18-point advantage over boys (Mullis et al., 2023).[1]

Where do these gender gaps in reading originate? Numerous research efforts have been made to answer this question (e.g., Disenhaus, 2015; Lafontaine & Monseur, 2009; Logan & Johnston, 2010, Marinak & Gambrell, 2010; McElvany & Schwabe, 2019; Schwabe et al., 2015; Solheim & Lundetræ, 2018). The explanations fall into two broad categories: actual gender differences in reading ability and differences in how boys and girls engage with tests.

Common explanations in the first category include gender differences in brain functions or inherent cognitive tendencies (Logan & Johnston, 2010; Spironelli et al., 2010), attitudes toward reading (Logan & Johnston, 2009; Muntoni et al., 2021; Retelsdorf et al., 2015), and reading frequency (Jabbar & Warraich, 2022; Logan & Johnston, 2009). The motivational and behavioral differences may be further attributed to factors such as gender stereotypes (Muntoni et al., 2021; Retelsdorf et al.,

CONTACT Dihao Leng dihaoleng@outlook.com TIMSS & PIRLS International Study Center, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

[1]The PIRLS reporting scale was established during the first assessment cycle with a mean of 500 and a standard deviation of 100 score points.

IEA
TIMSS & PIRLS
BOSTON COLLEGE

Paper 1:

Examining Gender Differences in TIMSS 2019 Using a Multiple Group Hierarchical Speed-Accuracy-Revisits Model

Leng, D., Bezirhan, U., Khorramdel, L., Fishbein, B. and von Davier, M. (2024), Examining gender differences in TIMSS 2019 using a multiple-group hierarchical speed-accuracy-revisits model. *Educational Measurement: Issues and Practice. Advance online publication.* https://doi.org/10.1111/emip.12606

Introduction: Gender Gaps in ILSAs

- Gender gaps in achievement on international large-scale assessments (ILSAs) are observed across countries, assessments, and over time.
- They vary in direction and magnitude depending on the particular country and domain assessed.
- Girls consistently outperform boys in reading assessments on average (Borgonovi, 2021; Mullis et al., 2017; Smith & Wilhelm, 2009).
- The gender achievement gap is smaller and more inconsistent in mathematics assessments (Ghasemi & Burley, 2019; Meinck & Brese, 2019; Mullis et al., 2020).
- Gender gaps have been a recurring theme in the reports of major ILSAs and have been of considerable interest to numerous countries since the inception of TIMSS and PIRLS (Mullis et al., 2003, 2009, 2017, 2023; OECD, 2009, 2018, 2019, 2023).

Introduction: Gender Gaps are Related to…

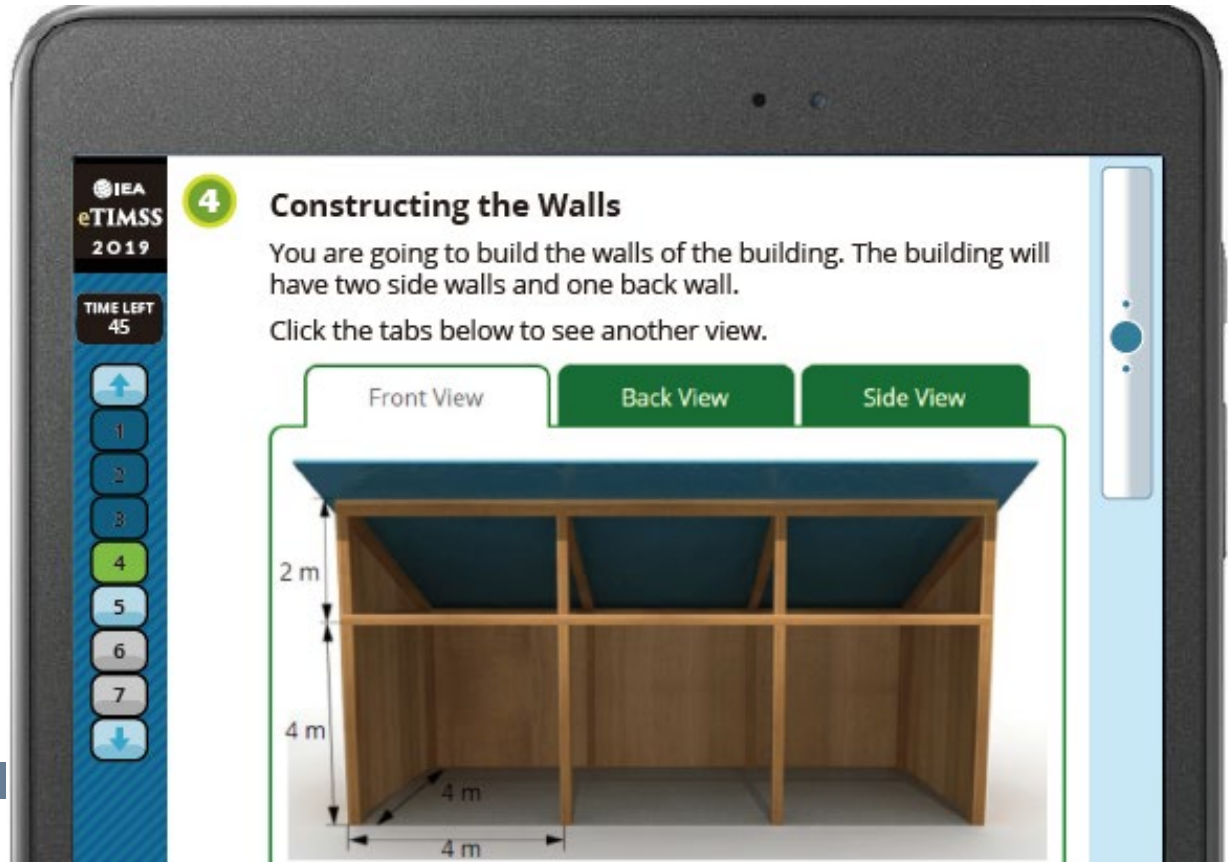
- Gender stereotypes (Igbo et al., 2015; Spencer et al., 1999; Tsui, 2007)
- Specific cognitive abilities (Ganley et al., 2014; Maeda & Yoon, 2013)
- Test-taking motivation (DeMars et al., 2013; Eklöf, 2010)
- Item types (Beller & Gafni, 2000; Solheim & Lundetræ, 2018)
- Differential item functioning (Holland & Wainer, 1993; Khorramdel et al., 2020)
- **Test-taking behaviors** (Baldiga, 2014; Bezirhan et al., 2021; Peng et al., 2014; Stenlund et al., 2017)

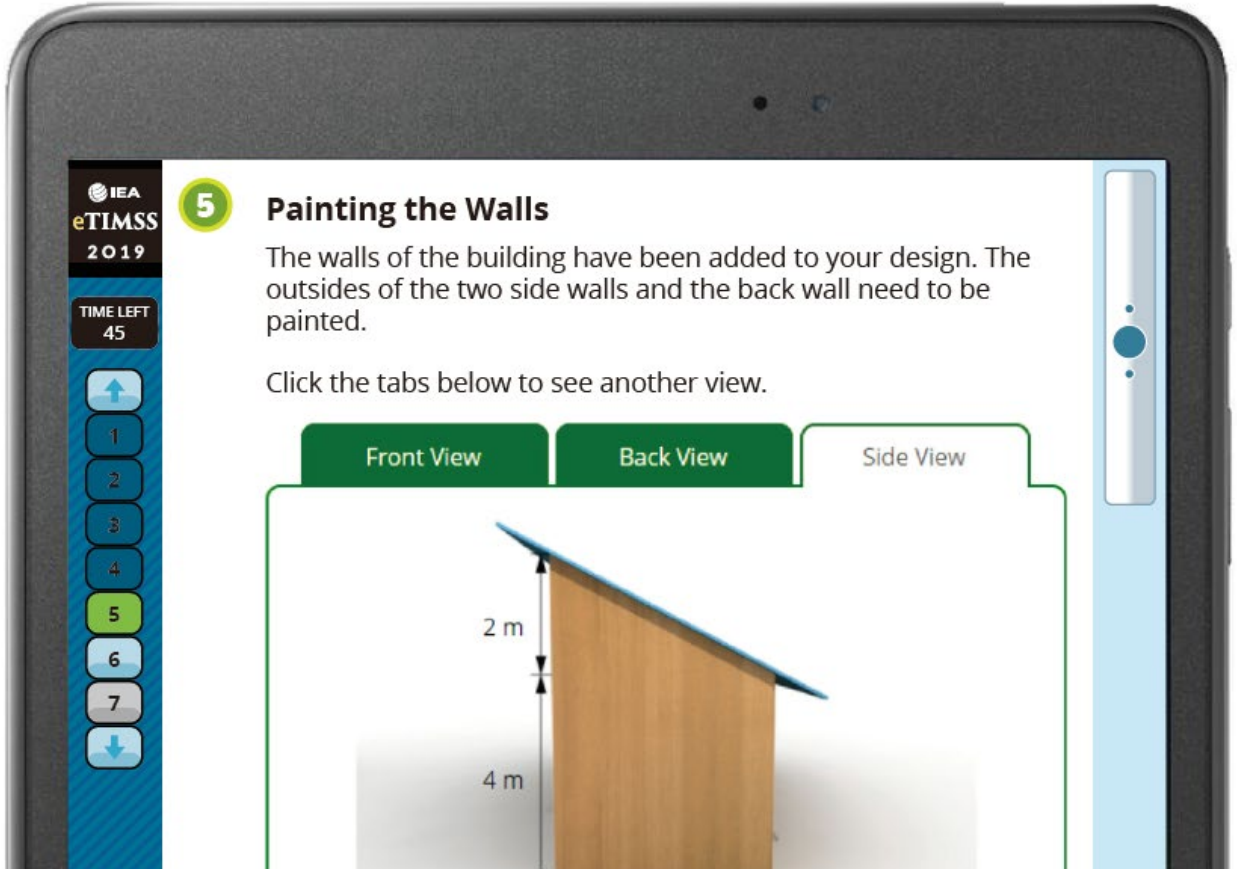
Research Question:

- To what extent do mathematics **ability**, response **speed**, **revisit propensity**, and the **relationships** among them **differ between boys and girls**?

Data

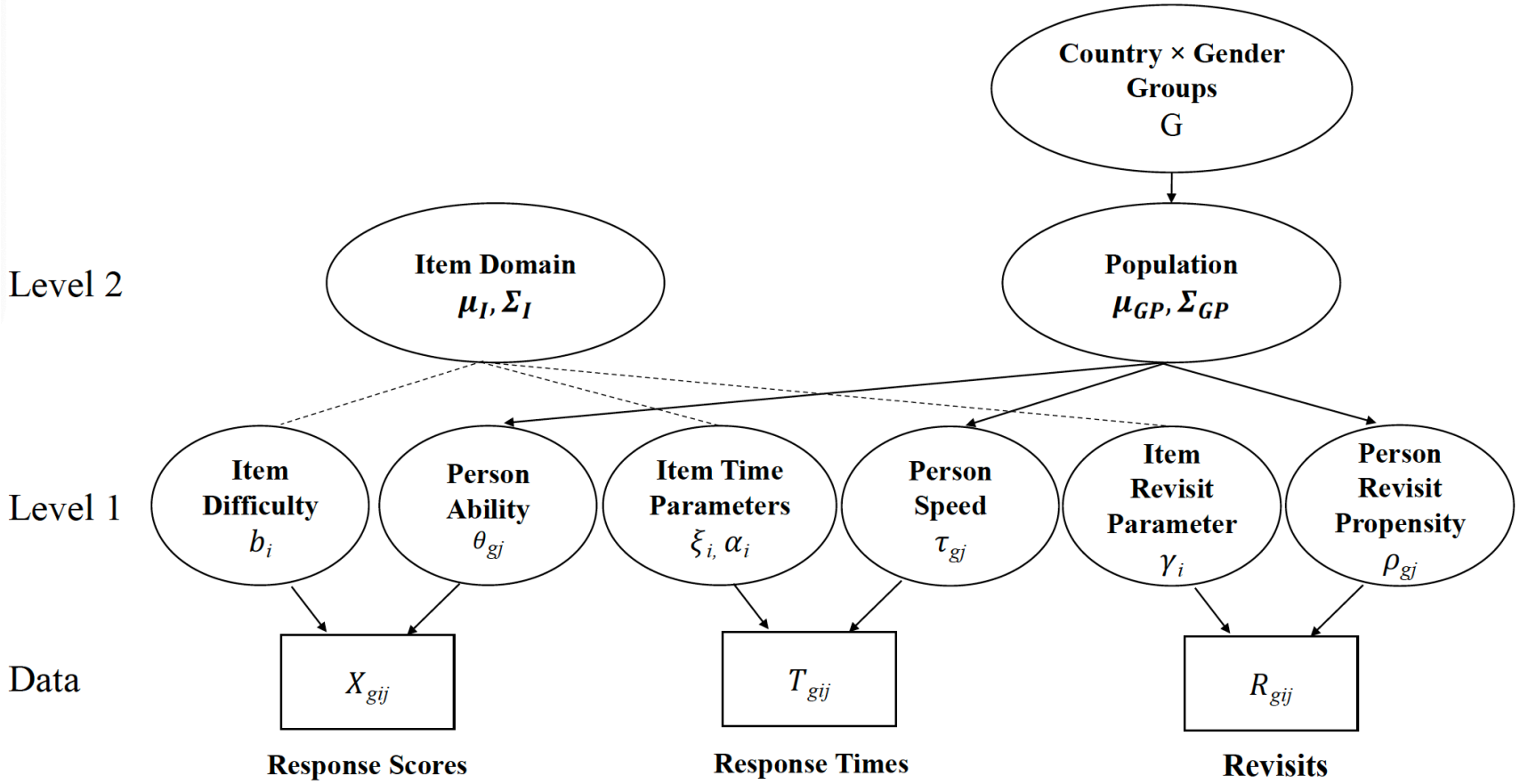
- TIMSS 2019 eighth-grade mathematics Problem Solving and Inquiry (PSI) tasks
- 10 Countries: Georgia, Hungary, Israel, Italy, Korea, Malaysia, Portugal, Qatar, Singapore, Turkey
- N = 6,042



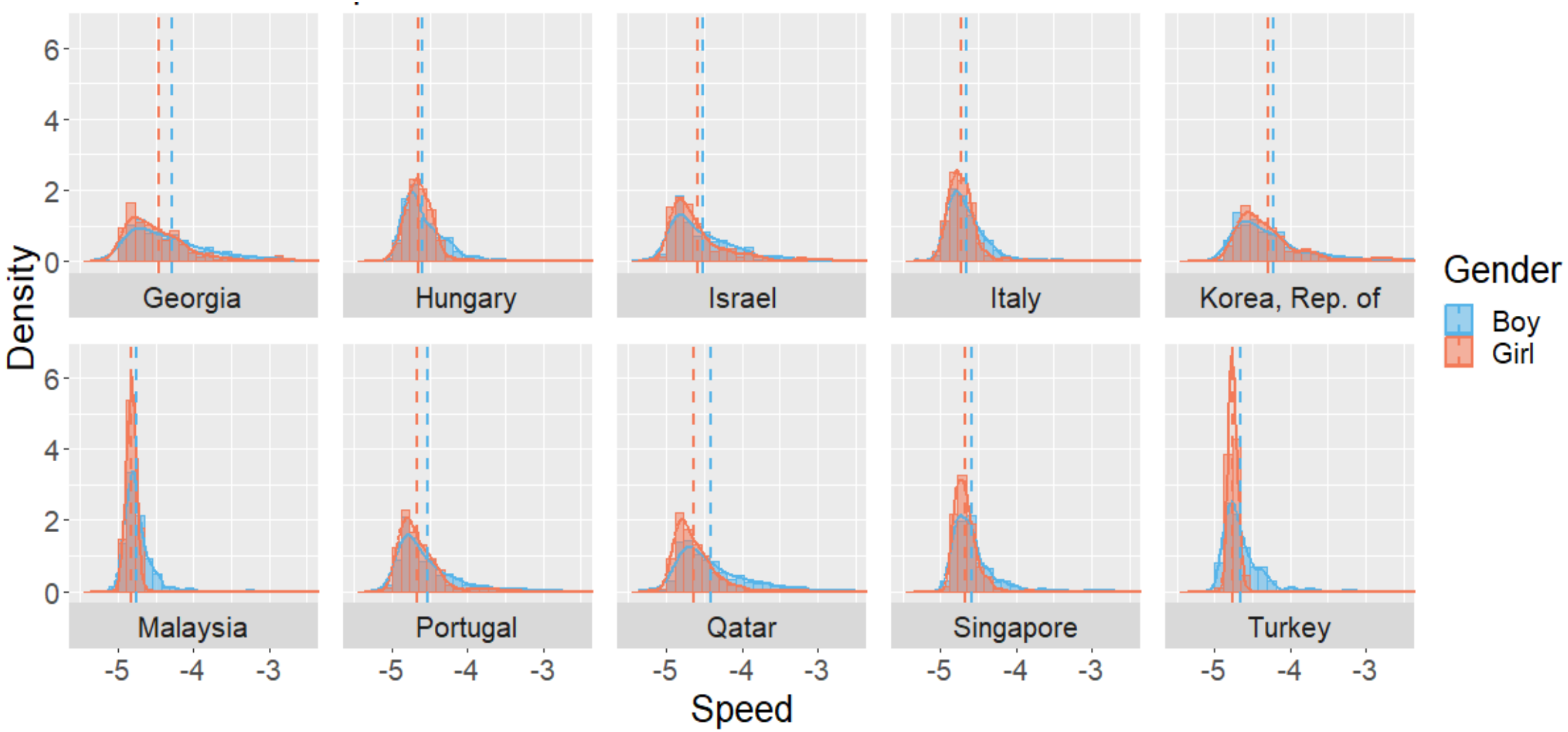


Methods

- The hierarchical fixed-effects multiple-group speed-accuracy-revisits (MSAR) model extended from the speed-accuracy-revisits (SAR) model (Bezirhan et al., 2021):

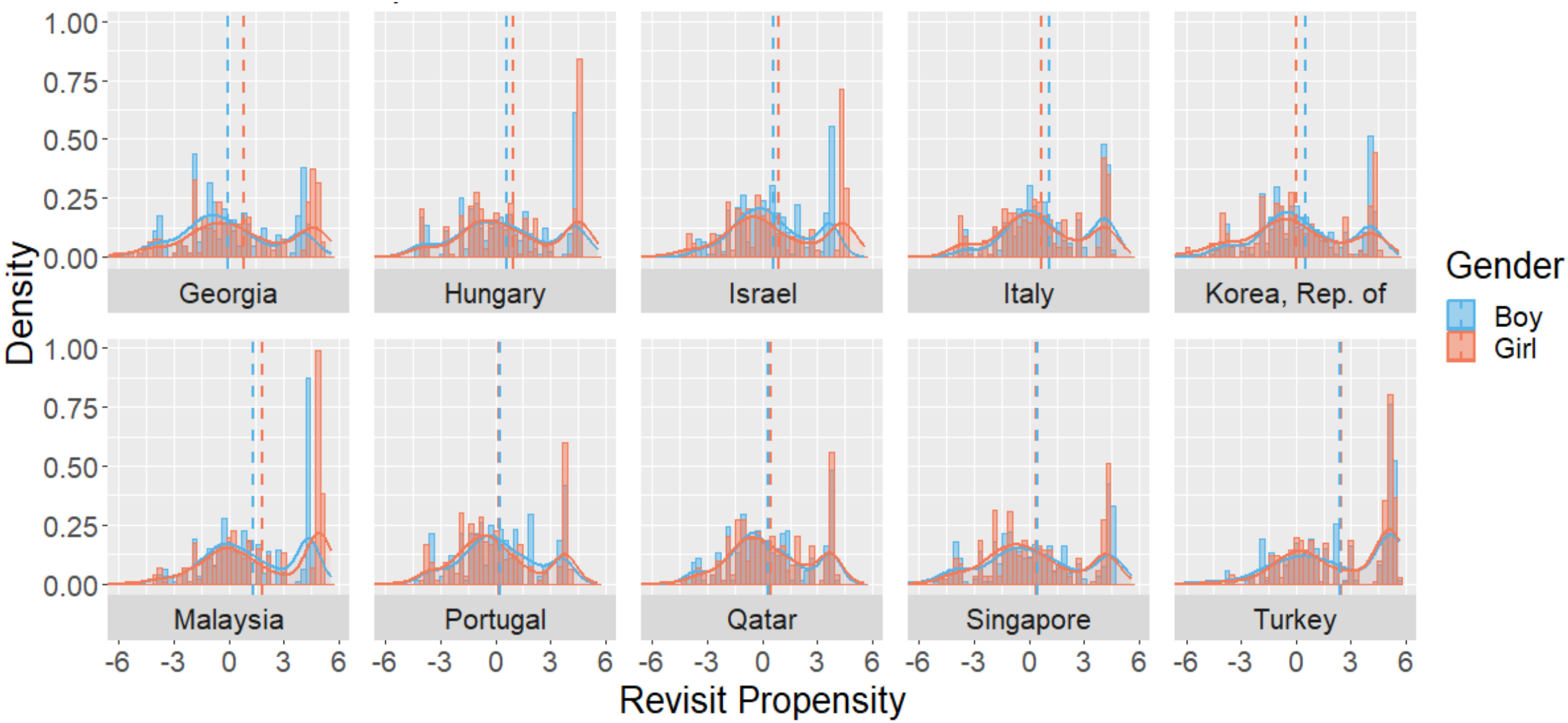


Results: Response Speed Distribution by Country and Gender



Dashed lines: population mean estimate

Results: Revisit Propensity Distribution by Country and Gender



Dashed lines: population mean estimate

Discussion

- Boys responded faster than girls; their speed distributions were more spread out.
- All country-by-gender groups showed a mixture distribution of revisit propensity.
- Across genders and countries, mathematics ability and response speed were negatively correlated.
- Correlations between revisit propensity and the other two latent variables were generally weak, with a few exceptions.
- **This paper showcases the value of process data for understanding gender differences in test-taking behaviors in mathematics.**
- **The flexibility of the MSAR model enables use across assessment cycles, domains, and grouping variables.**

Paper 2:

Leveraging Process Data and Variable Selection for Achievement Estimation in Large-Scale Assessments

Leng, D., Bezirhan, U., & von Davier, M. (accepted). Leveraging Process Data and Variable Selection for Achievement Estimation in Large-Scale Assessments. *Journal of Educational Measurement.*

Introduction: Achievement Estimation in LSAs

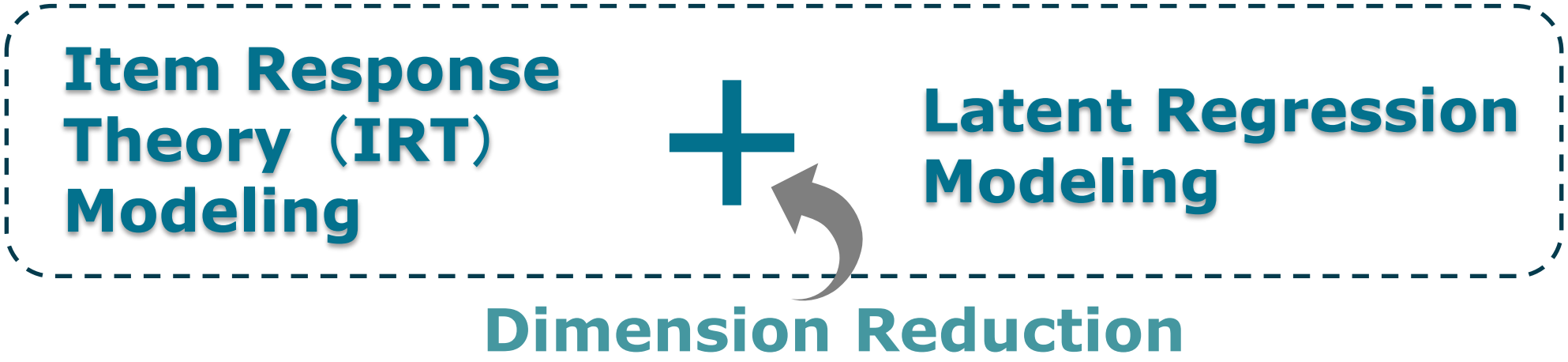
Item Response Theory (IRT) Modeling + **Latent Regression Modeling**

Major assumption: the posterior distribution of **achievement** is statistically associated with both **contextual variables** (e.g. gender, educational resources at home, attitudes towards mathematics) and **observed item responses**.

Latent regression:
Achievement = intercept + $\beta_1$gender+ $\beta_2$class average achievement + …

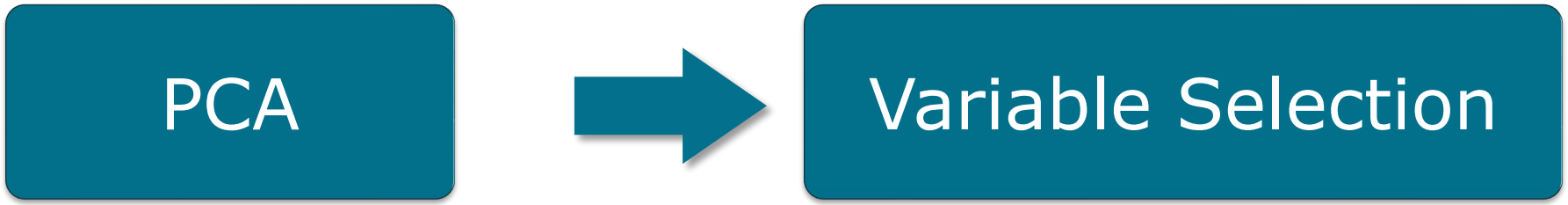
What if: Contextual Variables → Contextual + Process Variables

Introduction: Achievement Estimation in ILSAs



Operational Method: Principal Component Analysis (PCA) —— Overparameterization

What if:



Research Question:

Can integrating **process data** and **variable selection** **reduce model complexity** in achievement estimation while **maintaining measurement precision**, compared to the conventional approach?

Research Design: Two-Way Comparisons

Compare the measurement precision and model complexity as a result of

Contextual Variables	vs	Contextual + Process Variables
PCA	vs	Variable Selection

Data and Sample

TIMSS 2019 eighth grade mathematics and science assessments

Countries	France		Republic of Korea (KOR)	
Subject	Mathematics	Science	Mathematics	Science
# students	4,420	4,422	4,406	4,406
# items	231	238	231	238
# item screens	204	220	204	220
# contextual variables	**963**	**963**	**637**	**637**

Process Variables

- Response speed: estimated from a log-normal model (Schnipke & Scrams, 1997; van der Linden, 2006, 20007)

- Revisits: two dummy variables at student level
  - Reference category: no revisit on any screen
  - Rev_1 = 1: revisited some but not all screens
  - Rev_2 = 1: revisited all screens

Results: Variable Selection

Out of Lasso, random forests, and gradient boosting, gradient boosting almost always led to the smallest test set RMSE when the models used a similar number of variables.

→ **Gradient boosting** was chosen as the final variable selection method.

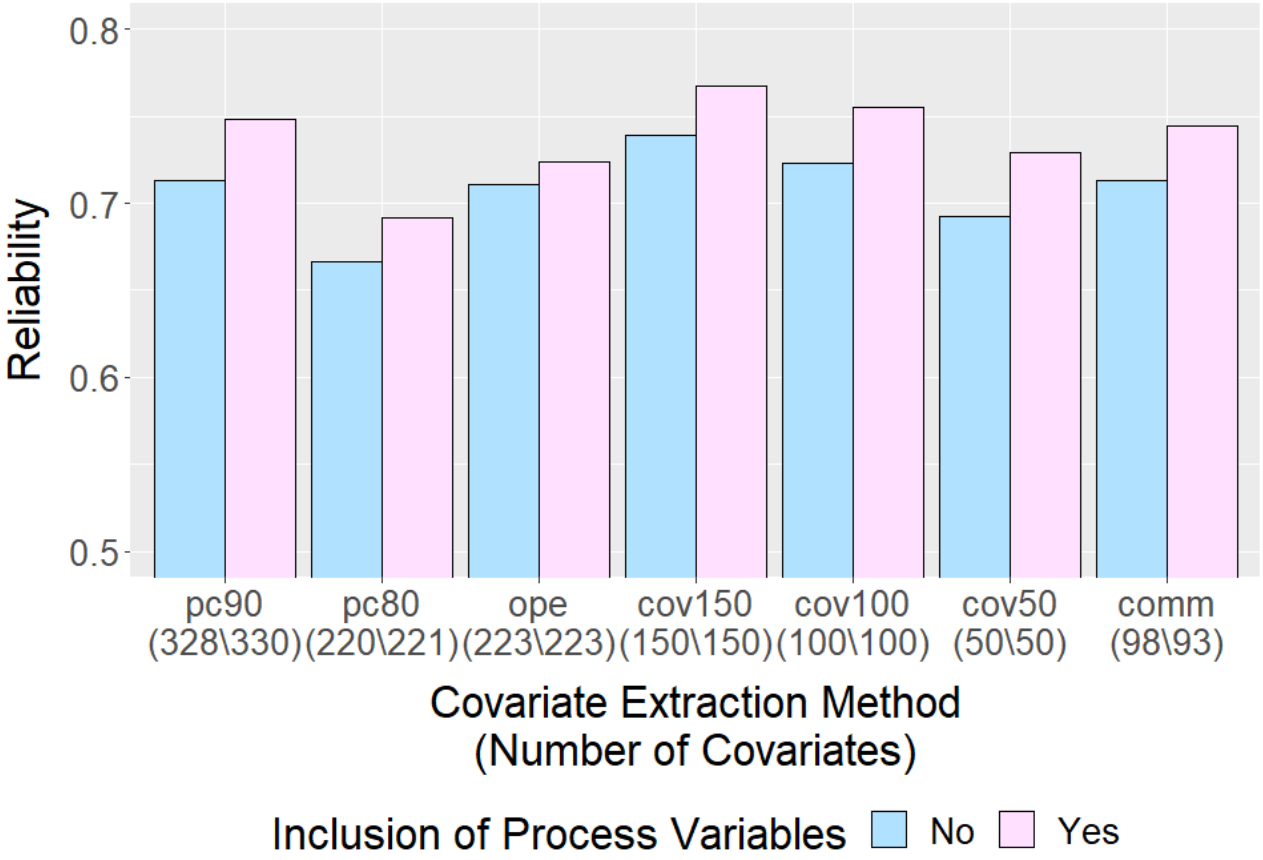
Process variables were consistently identified as important predictors of achievement.

- In France, *response speed* was the second most important predictor.
- In Republic of Korea, *response speed* was the most important predictor; *revisit* was among the 20 most important predictors.

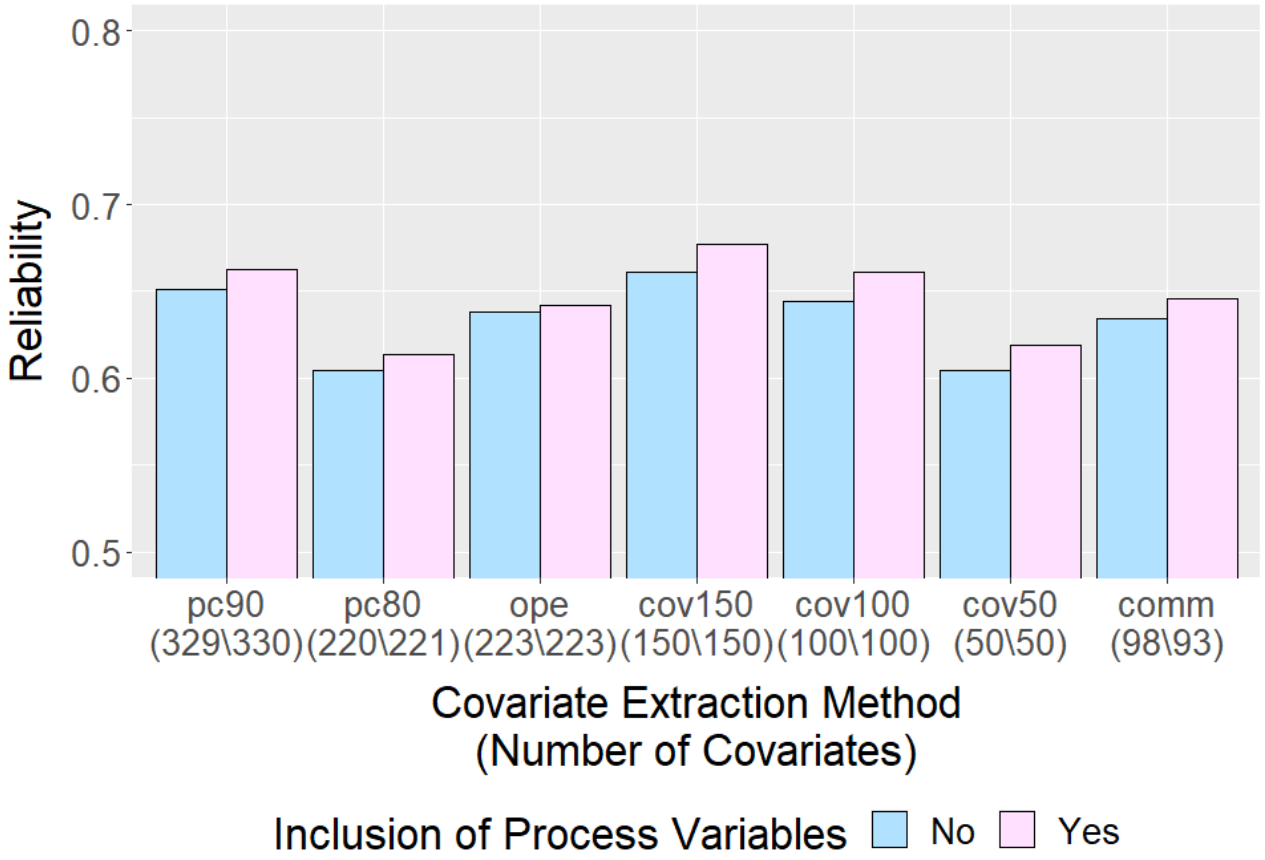
(Variable importance ranking was generated based on gradient boosting.)

Measurement Precision: France

Mathematics

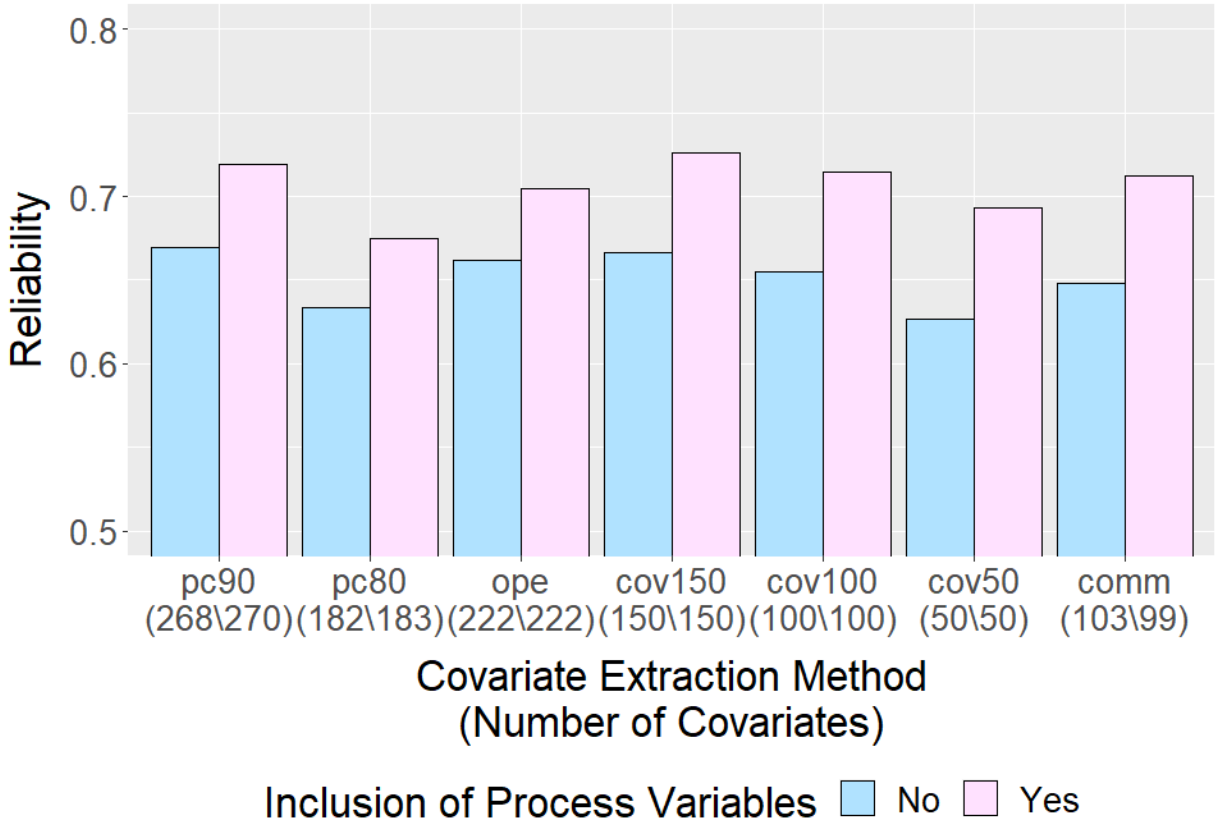


Science

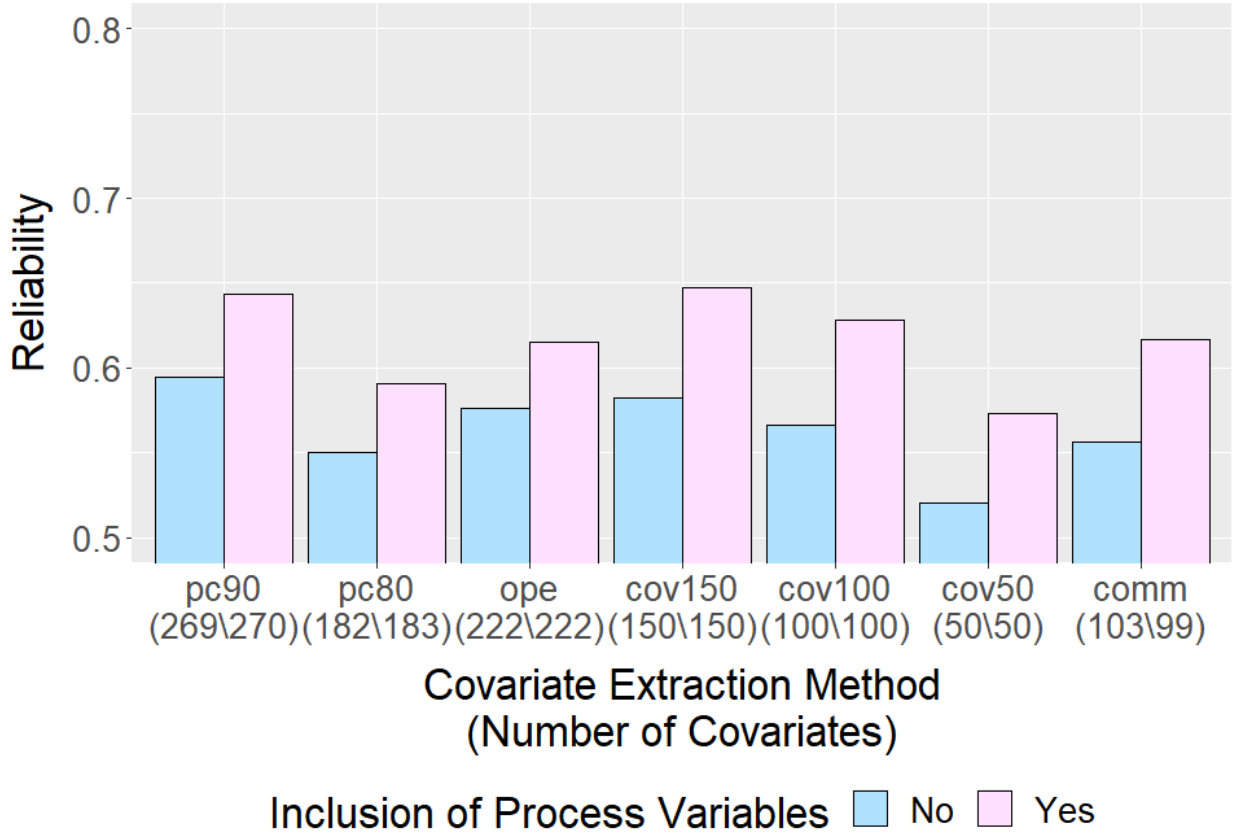


Measurement Precision: Republic of Korea

Mathematics



Science



Discussion

- Paper 2 uses process data for improving achievement estimation.
- We propose combining process data with machine learning–based variable selection in latent regression modeling.
- **Results: maintained (or improved) measurement precision while substantially reducing model complexity.**
- Log data aggregated into a small set of process variables → effectively addressing sparsity and high dimensionality in raw log data.
- Broader significance: **offers a path for future ILSAs to leverage expanding high-dimensional data effectively.**

Paper 3:

Gender Differences and Types of Test-Taking Behaviors in PIRLS 2021

Leng, D., & von Davier, M. (2025). Gender Differences and Types of Test-Taking Behaviors in PIRLS 2021. *Applied Measurement in Education*, 1-19.https://doi.org/10.1080/08957347.2025.2486032

Research Questions

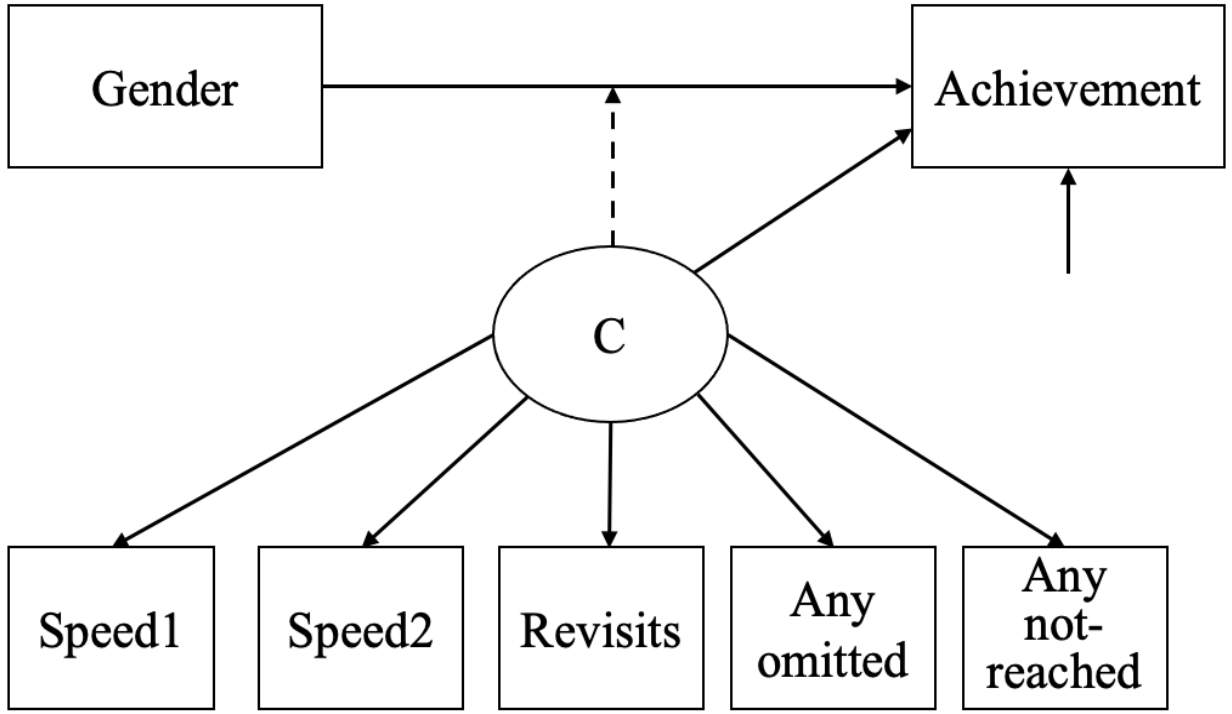
- What classes of test-taking behaviors can be identified among students?
  - How are these behavioral classes distributed among boys and girls?

- How do gender gaps in reading achievement differ across these behavioral classes?

- How would the overall gender gap change if the distribution of the behavioral classes among boys matched that of girls?

Methods: Test-Taking Behavior Classification and Gender Achievement Gaps by Class

The Bolck-Croon-Hagenaars (BCH) and three-step latent-class methods are chosen.



The secondary model:
a moderation model

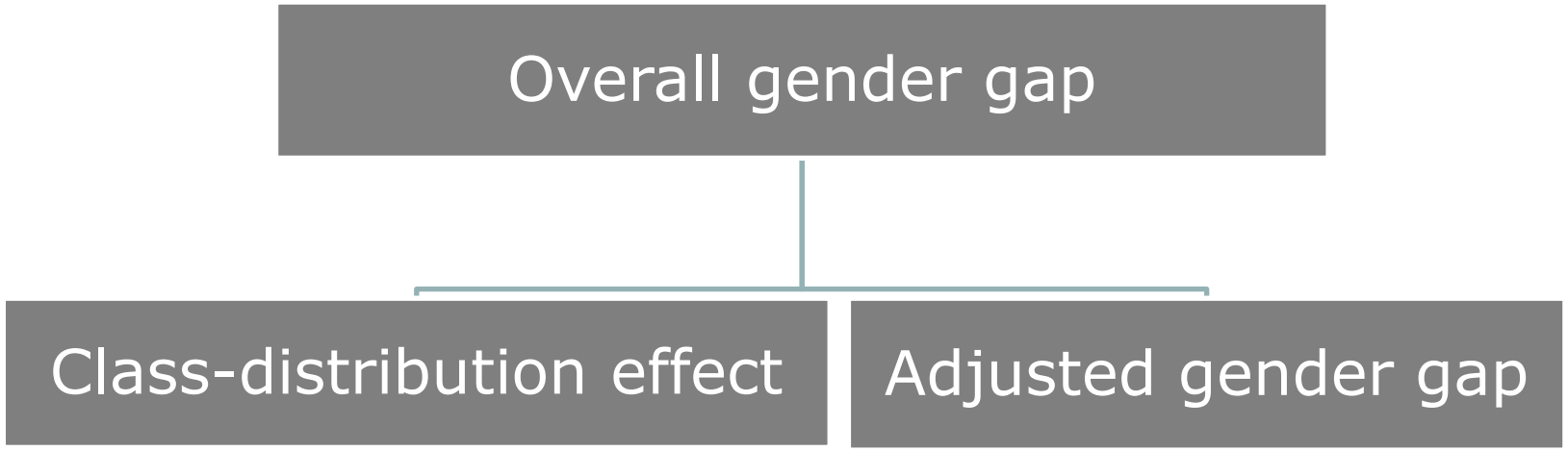
$$Achievement = a_m + b_m \times G + e,$$

where G denotes gender with G = 0 representing girls and 1 for boys; a_m and b_m vary by class.

(Asparouhov & Muthén, 2014 ; Asparouhov & Muthén, 2021; Bakk et al., 2013; Bolck et al., 2004; Vermunt, 2010)

Methods: Standardized Boy Mean and Adjusted Gender Gap

Partitioning analysis
(Beaton & Chromy, 2007)



Hypothetical example:
Girls: 70% engaged, 30% disengaged
Boys: 60% engaged, 40% disengaged ⟶ 70% engaged, 30% disengaged

Results: Classification

Class 1: "Rapid" (36–48%)
- Highest response speed
- Tended to finish all items
- Rarely revisited

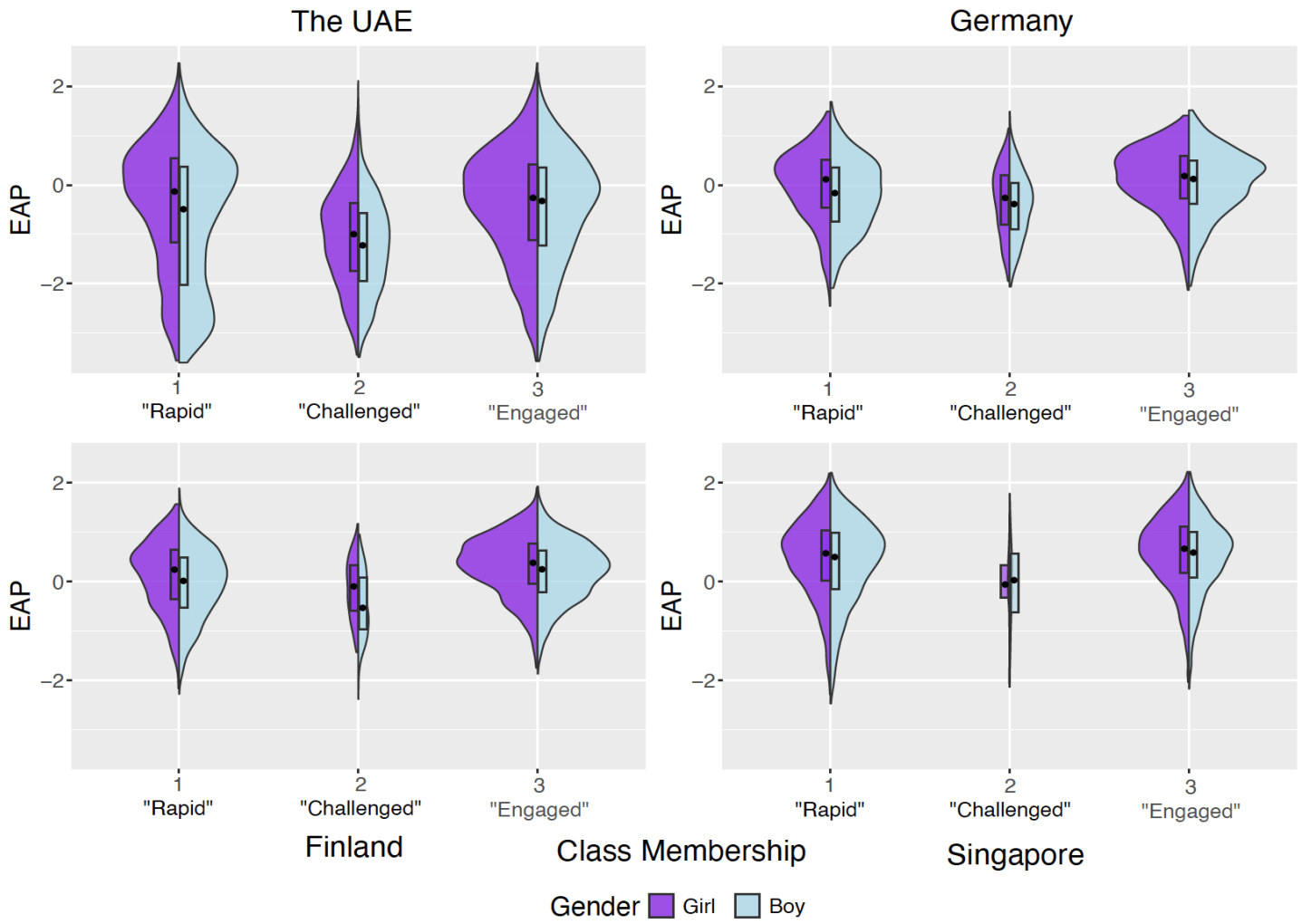
Class 2: "Challenged" (<20%)
- Most likely to have omitted / not reached
- Slow to medium speed

Class 3: "Engaged" (43-49%)
- Medium / slower speed
- Most likely to revisit

Results: Average Posterior Classification Probabilities by Gender

Country	The UAE		Germany		Finland		Singapore	
Class	% Girls	% Boys	% Girls	% Boys	% Girls	% Boys	% Girls	% Boys
Class 1: "Rapid"	32.7	40.5	40.5	43.0	38.6	45.2	44.0	52.1
Class 2: "Challenged"	20.0	18.7	13.7	16.4	12.4	10.6	2.2	3.3
Class 3: "Engaged"	47.4	40.8	45.8	40.5	49.0	44.1	53.8	44.6

Results: Achievement Distribution by Gender, Conditional on the Most Likely Class



UAE, Germany, Finland:

- Largest gender gap in "Rapid" class

All four countries:

- Minimal gap in "Engaged" class

Results: Adjusted Gender Achievement Gap

Achievement	The UAE	Germany	Finland	Singapore
Girl Mean	-0.454	0.006	0.200	0.557
Boy Mean	-0.702	-0.144	0.041	0.437
Standardized Boy Mean	-0.677	-0.124	0.052	0.454
Raw Gender Gap (Boy Mean – Girl Mean)	-0.248	-0.150	-0.159	-0.120
Adjusted Gender Gap (Standardized Boy Mean – Girl Mean)	-0.223	-0.129	-0.147	-0.103
1 – Adjusted Gender Gap / Raw Gender Gap	**10.1%**	**14.0%**	**7.5%**	**14.2%**

Note. The standardized boy mean is the average boy achievement under the hypothetical scenario that boys had the same distribution of test-taking behavior classes as girls.

Discussion

- Three classes of test-taking behaviors were identified among students taking the PIRLS 2021 assessment in Finland, Germany, the UAE, and Singapore: “Rapid”, “Challenged”, and “Engaged”.
- In all four countries, more boys fell into the “Rapid” class, while the “Engaged” class contained more girls.
- Girls in the “Rapid” class significantly outperformed boys across all four countries; no significant gender achievement differences were observed in the “Engaged” class in three countries.
- Gender played a lesser role in reading achievement when students were engaged, highlighting the importance of addressing disengagement issues in low-stakes assessments.

Discussion and Conclusion

Does Process Data Add Value to the Analysis of International Large-scale Assessment Data?

- process data offers unique insights beyond what achievement and traditional background variables can provide.
- Although this dissertation focused on ILSA data, many of these benefits extend to national studies, regional large-scale assessments, and educational and psychological assessments in general.

Limitations

- Caution must be used when generalizing the results.
- All three studies are exploratory and lack a strong theoretical framework.
- Process variables are subject to measurement error and uncertainty.

Implications for Future Research

- Using process data beyond response times can yield richer insights.
- Integrating process data with methods from other fields opens new opportunities.
- Exploring different ways to derive process variables may preserve richer information and reveal more nuanced behaviors.
- More in-depth research is needed to understand the implications of the findings.

Thank You!

Dihao Leng

TIMSS & PIRLS International Study Center

66th IEA General Assembly Meeting
Madrid, Spain, October 8th, 2025