

# Impact of test score scaling model on regression and multilevel estimates

**Maciej Jakubowski**

*Faculty of Economic Sciences, University of Warsaw*

*Directorate of Education, OECD*

# Outline

---

- main empirical question
- scaling student test scores in PISA and TIMSS
- previous research
- PISA: comparing regression estimates obtained with Rasch scores and plausible values
- TIMSS: comparing regression estimates obtained with 1PL and 3PL student scores

# Main empirical question

---

- International surveys of student literacy use different scaling methodologies
- while theoretical differences between methods are relatively well established, little is known about the impact of scaling methodology on results from secondary analysis
- this research asks whether a choice of scaling methodology affects estimates obtained from linear regression, multilevel regression, and quantile regression

# International surveys of achievement/skills

---

- **PISA** 2000, 2003, 2006, 2009
  - reading, mathematics, science
  - 15-year-olds
- **TIMSS** 1995, 1999, 2003, 2007
  - mathematics and science
  - 4th and 8th graders
- **PIRLS** - reading
- **CivED** - civic education
- **IALS/ALL/PIAAC** - adult literacy

# survey concept and measurement of achievement

---

- different goals (curriculum vs. labor market skills)
- differently defined domains (measurement frameworks)
- different country populations (age vs. grade)
- different groups of countries (OECD+partners vs. all willing to participate)
- similar measurement methodology
  
- IRT models used to scale student responses
- Rotating test booklets
- Plausible values imputation

# PISA scaling model

---

- Simple Rasch model is used to calibrate items
- final scores are estimated as plausible values with multidimensional partial credit population model
- several domains are scaled simultaneously
- all information from the student background questionnaire is used for PV imputation
- 5 imputed values
- BRR weights to account for survey design

# TIMSS scaling model

---

- similar plausible values model is used to estimate final student scores
- in TIMSS 1995 this model was based on the 1PL IRT model
- in TIMSS 1999 new 3PL model was introduced
- student scores from TIMSS 1995 were rescaled with the new model to obtain comparable trend estimates
- 5 imputed values
- Jackknife method to account for survey design

# which surveys are analyzed in this paper?

---

- PISA 2000 public datasets contain plausible values as well as Rasch WLE student literacy estimates
- TIMSS 1995 student scores were re-estimated with the new model; 1PL and 3PL student scores are available
- we compare publicly available differently estimated student scores
- no attempt to re-estimate IRT models

## Previous research:

# "International surveys of educational achievement: how robust are the findings?"

---

- Brown, Micklewright, Schnepf, Waldmann, 2007, Journal Of The Royal Statistical Society Series A, vol. 170(3)
  - TIMSS 1995 and TIMSS 1999 data
  - 1-parameter and 3-parameter IRT models give different results
  - discrepancies are small if mean scores is considered
  - crucial if score dispersion is analyzed, especially for low-performing countries
- Mean achievement as measured by different surveys is highly correlated
- SD or different measures of score variation differ importantly for some low achieving countries, but are quite coherent for developed countries

# PISA: Rasch scores vs. plausible values

---

- reading literacy scores are compared only, because these were measured with the highest precision (ca. 100 items)
- Rasch WLE scores vs.. 5 plausible values
- 43 countries
- three methods:
  - linear regression (BRR weights)
  - multilevel regression (no weighting),
  - quantile regression (BRR weights)

# PISA - estimated models

---

- model 1

$$y = b_0 + b_1 * oth\_lang$$

- model 2

$$y = b_0 + b_1 * oth\_lang \\ + b_2 * hisei + b_3 * mean\_hisei \\ + b_4 * female + b_5 * joyread$$

# linear regression

---

	mean_hisei			oth_lang		
	wle	pv		wle	pv	
pv1	4.20	3.96	4.20	-27.68	-26.54	-28.13
	2.43	2.22	2.42	628.85	540.85	609.92

- WLE coefficients downward biased
- smaller variance of WLE coefficients
- SE smaller by 1-5% with WLE comparing to PV
- correlations: 0.98-0.99

# multilevel regression

---

	mean_hisei			oth_lang		
	pv1	wle	pv	pv1	wle	pv
mean	4.18	3.93	4.18	-21.98	-21.48	-22.47
SD	1.62	1.55	1.62	19.42	18.12	18.99

- WLE coefficients downward biased
- smaller variance of WLE coefficients
- SE smaller by 1-5% with WLE comparing to PV
- correlations: 0.98-0.99

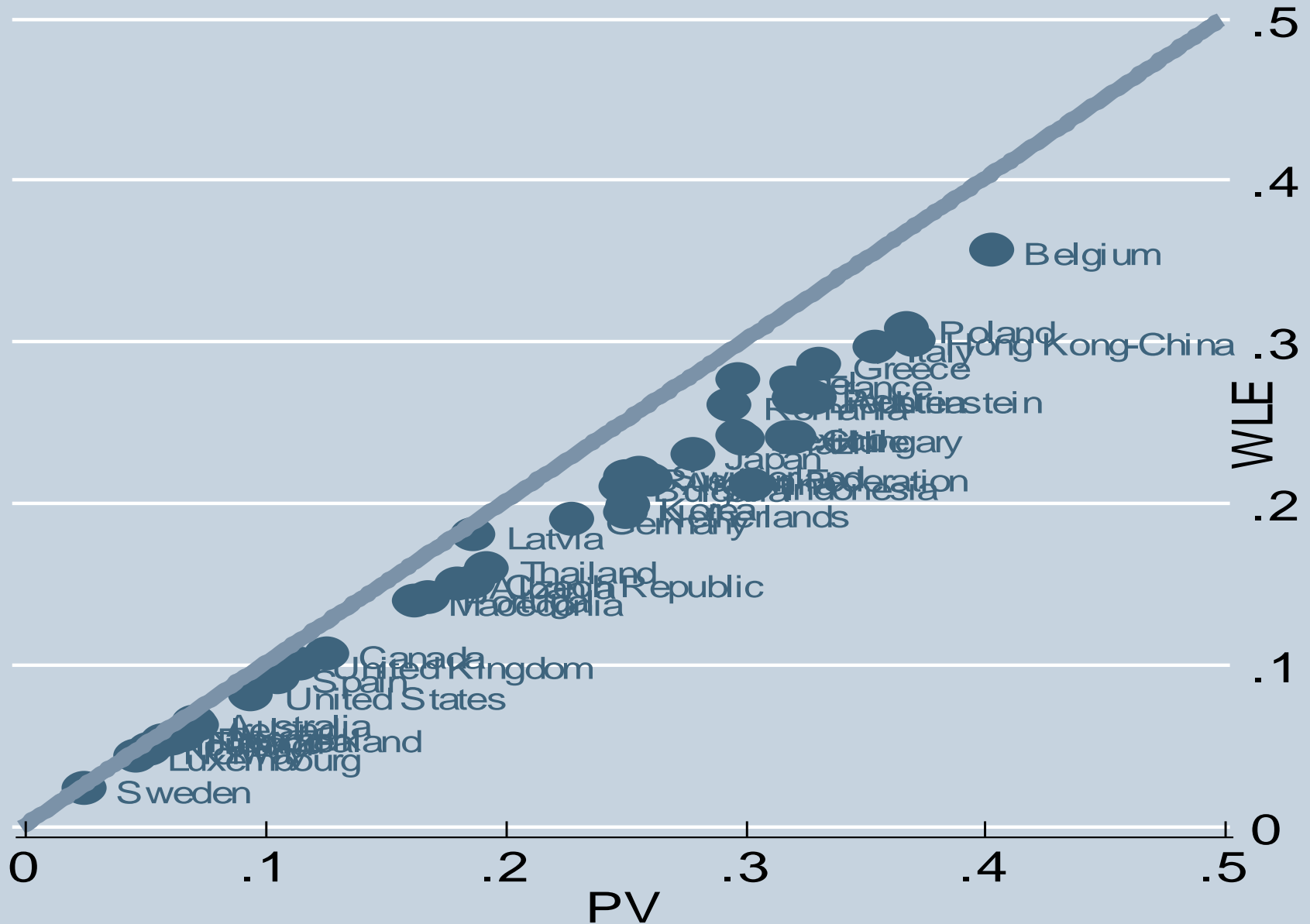
# multilevel regression variance decomposition

---

	icc	within (SD)	between (SD)
pv1	0.22	67.0	34.0
wle	0.18	71.0	31.7
pv	0.22	66.9	33.9

- intraclass correlation and between school variance are downward biased with WLE
- within school variance is upward biased with WLE (higher measurement error)

# between school variance

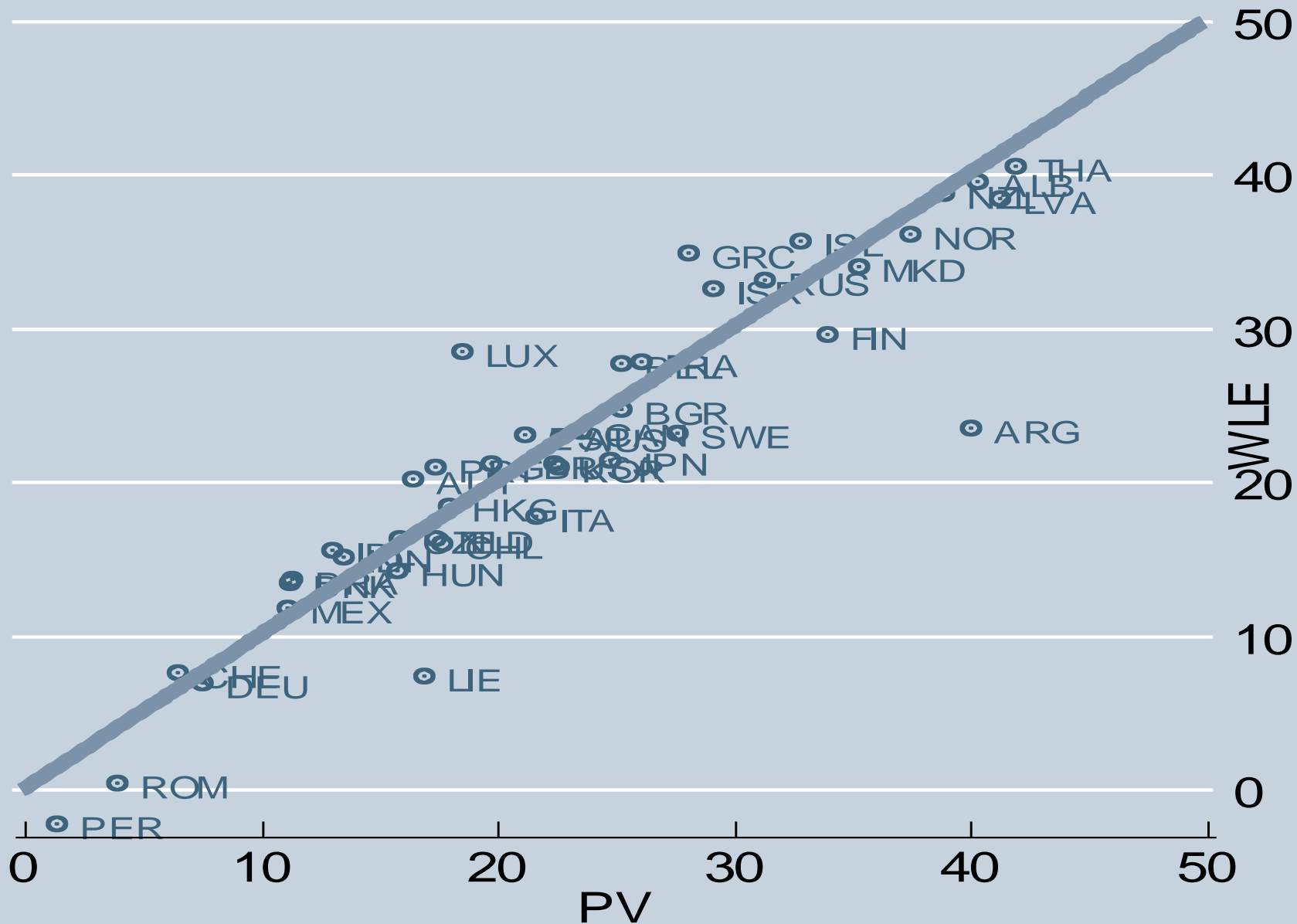


# quantile regression

---

- no clear pattern of differences between coefficients
- correlations 0.8-0.9
- smaller correlations for extreme quantiles (0.05, 0.10, 0.90, 0.95)

# female 0.1 quantile



# TIMSS: 1PL vs. 3PL model

---

- estimated linear regression model

$$y = b_0 + b_1 * migrant \\ + b_2 * no\_of\_books \\ + b_3 * higher\_grade$$

- 1PL coefficients smaller than 3PL
- SE smaller for *migrant* and *no\_of\_books*, but higher for *higher\_grade*

**Thank you!**

**Questions? Ideas?**

[mjakubowski@uw.edu.pl](mailto:mjakubowski@uw.edu.pl)

[maciej.jakubowski@oecd.org](mailto:maciej.jakubowski@oecd.org)