UiO : CEMO – Centre for Educational Measurement
University of Oslo

A Look at the

Most Pressing Design Issues

in International Large-Scale Assessment

Leslie Rutkowski

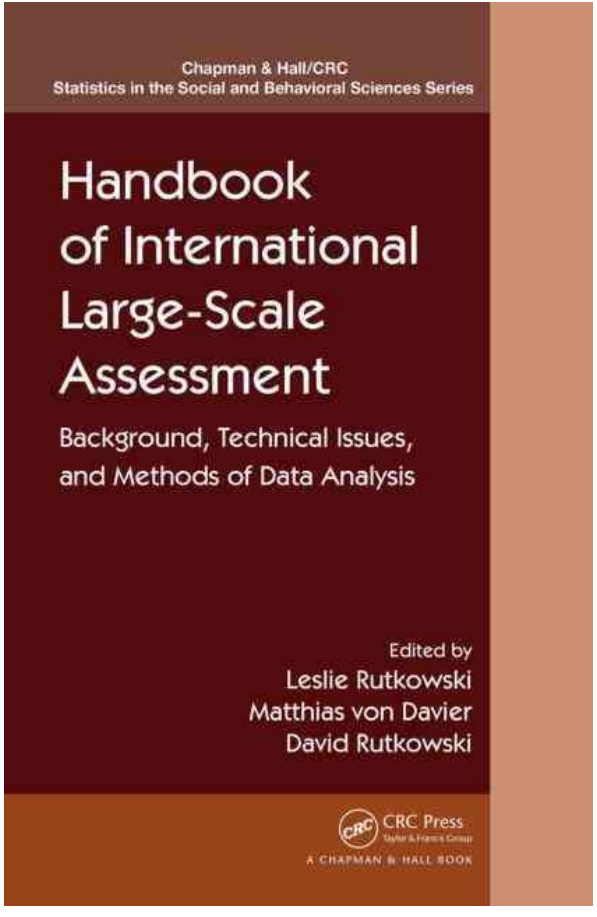
Professor of Educational Measurement

My Task

Offer perspectives on the most pressing design issues in ILSA

My Orientation

Quantitative methodologist with a research interest in ILSA models and methods



An Acknowledgement

ILSA programs have made enormous strides in

- Measurement methodology
- Measurement practice
- Cross-cultural comparison

Current State-of-the-Science

- New platforms
- New tests
- New populations
- New demands on test results

Three Areas with Possible Design Solutions

- Cross-cultural comparability
- Data quality and measurement error
- Desire to make causal inferences

The ILSA Situation

- In TIMSS, PIRLS, PISA, we are dealing with 30-80+ heterogeneous populations
  - OECD / non-OECD
  - Eastern / Western / Northern / Southern hemispheres
  - Many languages, cultures, geographies, religions

Cross-Cultural Comparability

- The degree to which comparisons on a latent variable (e.g., teachers' beliefs) can be validly compared across populations
- Terminology:
  - Differential item functioning
  - Measurement invariance

Cross-Cultural Comparability

- Differences can be because of instrument
- Or a different understanding of the construct
- Ultimately, we risk errors of inference

* Note: this isn't the same as genuine differences in construct (e.g., *problem solving strategies*)

ILSA Situation

- International educational surveys *usually* emphasize achievement
- Much of the psychometric “heavy lifting” is concentrated here
  - Pilot study
  - Field trial
  - Main survey

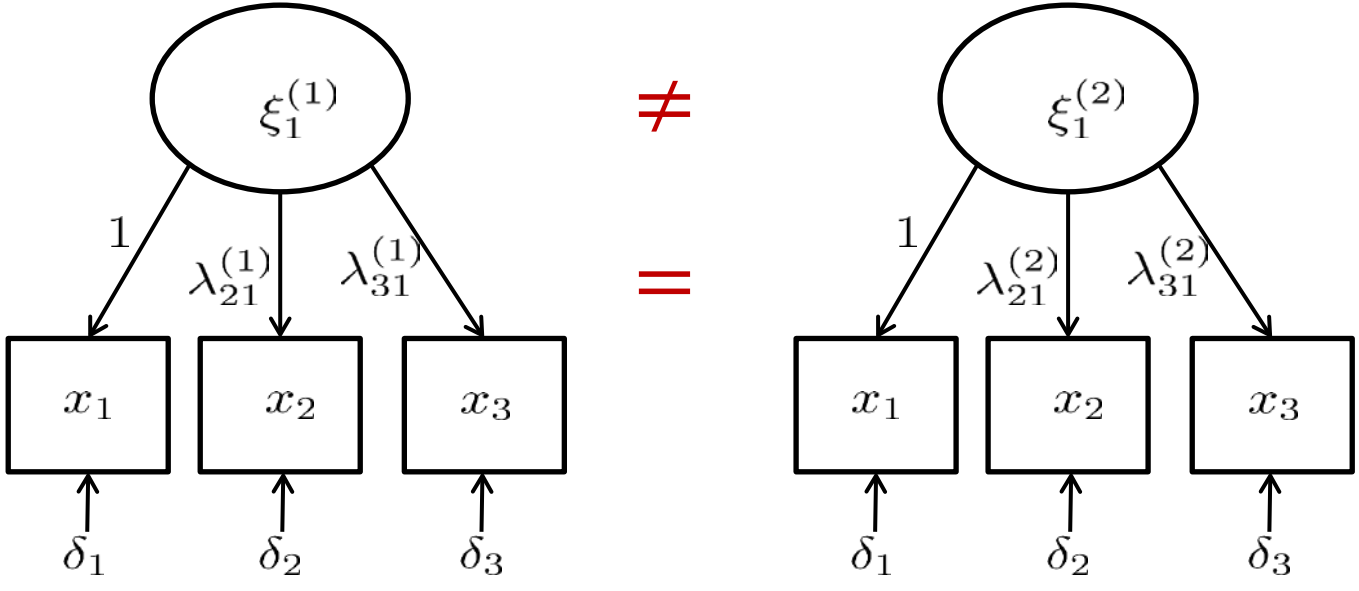
} Are the test items working well in all participating educational systems?

Also “Background”

- “Background” information is increasingly important in it’s own right
  - For contextualizing achievement
  - As outcomes (affective, behavioral, experiential)

- A relatively new but growing emphasis on background/context scale invariance
  - TALIS 2008
  - PISA 2012, 2015
  - TALIS 2013
  - Expected: PISA 2018, TALIS 2018, and so on

From both perspectives, our *measures* should be equivalent



Plus the mean structure: $\tau_i^{(1)} = \tau_i^{(2)}$ for all i

Equivalence Evidence

- We know, from fairly extensive analyses of international assessments (test and background), that equivalence is not forthcoming.
- Given system-level heterogeneity, can we expect to achieve strong equivalence?

One Example

- TALIS 2013 teacher questionnaire

13. In your teaching, to what extent do you feel prepared for the elements below?

Please mark one choice in each row.

			Not at all	Somewhat	Well	Very well
TT2G13A	a)	Content of the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4
TT2G13B	b)	Pedagogy of the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4
TT2G13C	c)	Classroom practice in the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4

- Some cultures/participants don’t use all categories
- It’s reasonable to expect other cultural differences in response styles

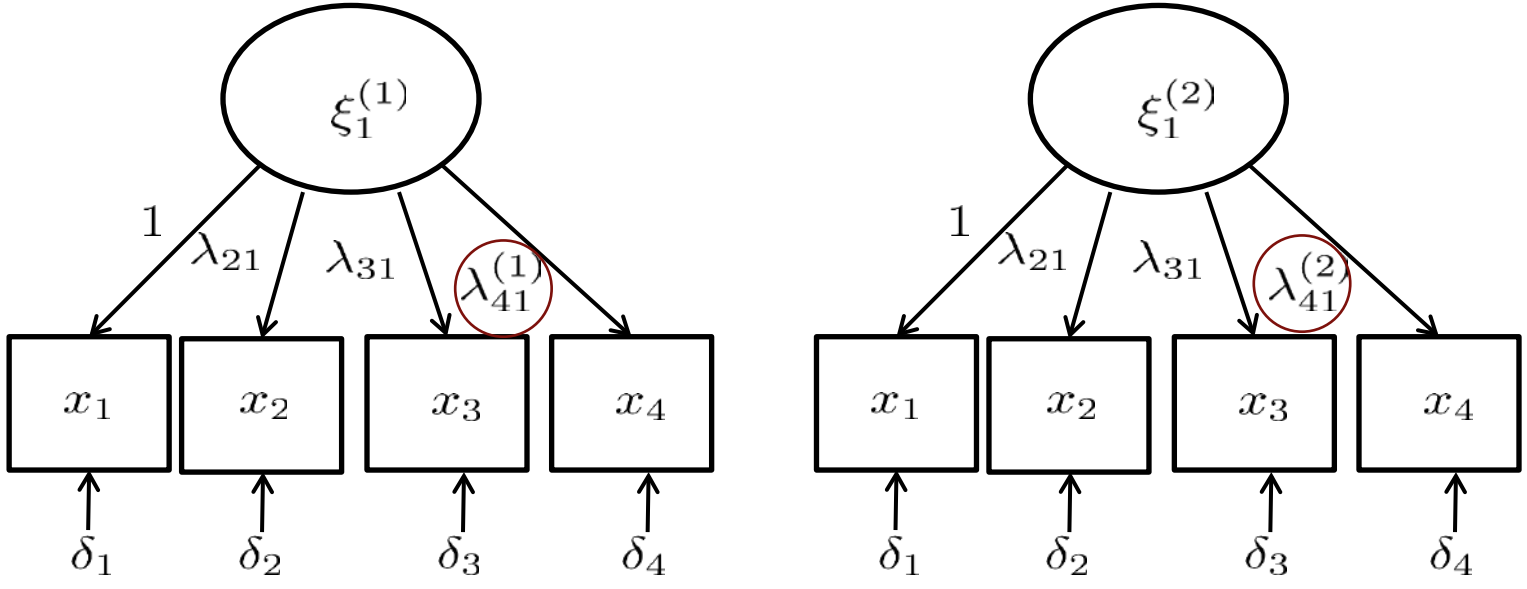
Allowing Heterogeneity In Our Models

- Oliveri & von Davier (2011, 2014) recommend an approach that uses *some* country-specific ***item parameters.***
- In PISA, ***blocks of easy items*** were offered as an option – for countries whose proficiency was expected to be low. (More on this next).
- National options in BQ

For Context (PISA 2012)

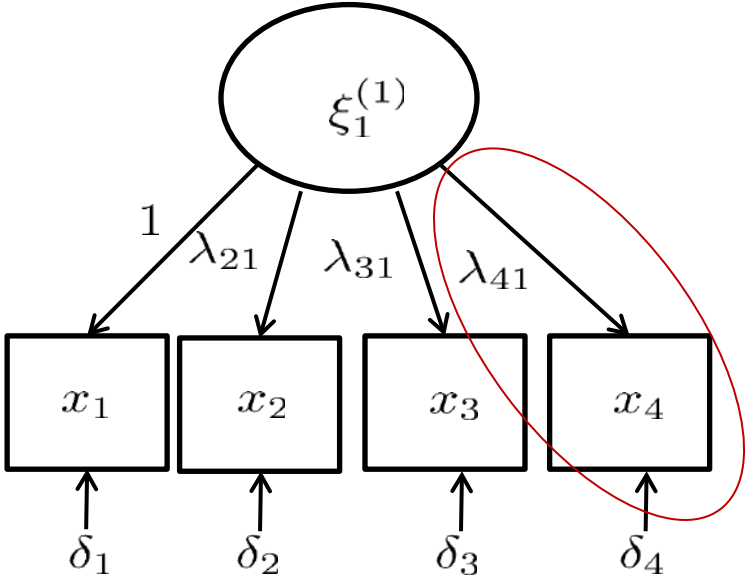
Designation	Booklet ID	Cluster				Standard Booklet Set	Easier Booklet Set
Standard Only	1	M5	S3	**M6A**	S2	X	
	2	S3	R3	**M7A**	R2	X	
	3	R3	**M6A**	S1	M3	X	
	4	**M6A**	**M7A**	R1	M4	X	
	5	**M7A**	S1	M1	M5	X	
	6	M1	M2	R2	**M6A**	X	
	7	M2	S2	M3	**M7A**	X	
Core	8	S2	R2	M4	S1	X	X
	9	R2	M3	M5	R1	X	X
	10	M3	M4	S3	M1	X	X
	11	M4	M5	R3	M2	X	X
	12	S1	R1	M2	S3	X	X
	13	R1	M1	S2	R3	X	X
Easy Only	21	M5	S3	**M6B**	S2		X
	22	S3	R3	**M7B**	R2		X
	23	R3	**M6B**	S1	M3		X
	24	**M6B**	**M7B**	R1	M4		X
	25	**M7B**	S1	M1	M5		X
	26	M1	M2	R2	**M6B**		X
	27	M2	S2	M3	**M7B**		X

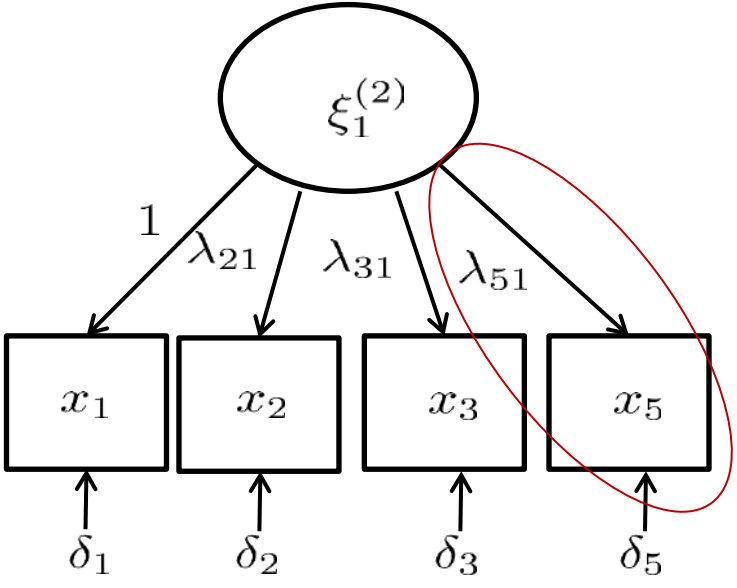
Oliveri & von Davier (2011, 2014)



- This is being implemented in PISA 2015 in both achievement and BQ

PISA "Easy" Booklets





Possible Ceiling Effects

- PISA high performers (Finland, Singapore, Shanghai) tend to have far greater proportion correct on many items
  - PM915Q02: .62 internationally vs. .88 in Singapore
  - PM205Q01: .56 internationally vs. .71 in Finland
  - PM423Q01: .78 internationally vs. .94 in Shanghai

Possible Ceiling Effects

- Proportion of items answered correctly by 80% of examinees
  - Shanghai: .350
  - Finland: .138
- In contrast
  - International: .057
  - Chile: .025 (and they opted for easy booklets)

Tailoring via *Challenging* Booklets

- Along with unique parameters for achievement items
  - Consider “hard” booklets
  - Methods already exist with easy booklet options

And Further Tailoring

- Make better use of national options along with unique parameters
- Offers the possibility of improved local usefulness and cross-cultural comparability

Estimation Methods and Measurement Error

Because of booklet design, methods used to estimate achievement

- Are essentially the same as *missing data imputation*
- Produce population achievement distributions
- Borrow information from student background questionnaires to optimize population and subpopulation differences

Importantly, these methods depend on error-free data

Estimation Methods and Measurement Error

But – as is typical in self-report data – there is measurement error

And that has consequences for estimating sub-population achievement

A Few Examples

PISA 2012 – some countries administered parent questionnaires

One question is the same between parent and student

- *"Have you/Has your child ever repeated a grade?"*
  - *Never*
  - *Once*
  - *More than once*
  - *Missing*

The 2-way contingency tables

UiO : CEMO – Centre for Education
University of Oslo

From PIRLS 2011

Notably, this issue seems to be concentrated in relatively poor countries

Table 3

Correlation Between Parent and Student Report of Number of Books in the Home

	Raw Correlation		Unattenuated Correlation			Raw Correlation		Unattenuated Correlation	
Country	*r*	*SE*	r_{corr}	SE_{corr}	Country	*r*	*SE*	r_{corr}	SE_{corr}
Indonesia	0.25	0.03	0.26	0.01	Dubai	0.48	0.01	0.79	0.02
Kuwait	0.25	0.03	0.35	0.03	Italy	0.49	0.02	0.73	0.03
Azerbaijan. Republic of	0.28	0.02	0.35	0.02	Iran	0.49	0.02	0.69	0.02
Qatar	0.29	0.02	0.46	0.03	Poland	0.49	0.01	0.64	0.02
Malta	0.30	0.02	0.42	0.03	Slovenia	0.49	0.01	0.60	0.02
Morocco	0.31	0.02	0.34	0.02	Hong Kong	0.49	0.02	0.68	0.03
Botswana	0.33	0.03	0.36	0.02	France	0.50	0.02	0.73	0.03
Abu Dhabi	0.36	0.02	0.52	0.03	Belgium (French)	0.50	0.02	0.78	0.03
Trinidad And Tobago	0.38	0.02	0.53	0.03	Finland	0.50	0.02	0.62	0.02
Canada (Alberta)	0.38	0.02	0.52	0.03	Morocco (Grade 6)	0.51	0.02	0.51	0.01
United Arab Emirates	0.40	0.01	0.66	0.02	Spain (Andalucia)	0.51	0.01	0.69	0.02
Oman	0.40	0.01	0.56	0.02	Czech Republic	0.51	0.02	0.68	0.02
Singapore	0.43	0.01	0.57	0.02	Germany	0.51	0.02	0.74	0.03
Canada	0.44	0.01	0.53	0.01	Ireland	0.51	0.02	0.80	0.03
Canada (Quebec)	0.44	0.02	0.57	0.02	Spain	0.51	0.02	0.71	0.02
Norway	0.44	0.02	0.62	0.03	Georgia	0.55	0.02	0.92	0.03
Australia	0.45	0.02	0.63	0.03	Austria	0.55	0.01	0.82	0.03
Canada (Ontario)	0.45	0.02	0.57	0.02	Lithuania	0.55	0.01	0.71	0.02
Saudi Arabia	0.45	0.02	0.78	0.03	South Africa	0.55	0.03	0.75	0.03
Honduras. Republic of	0.45	0.03	0.49	0.02	Portugal	0.56	0.02	0.76	0.03
Netherlands	0.46	0.02	0.65	0.03	Chinese Taipei	0.56	0.01	0.91	0.03
Int. Avg.	0.46		0.66		Sweden	0.57	0.02	0.86	0.03
Northern Ireland	0.47	0.02	0.72	0.03	Croatia	0.57	0.01	0.70	0.02
Israel	0.47	0.02	0.71	0.03	Denmark	0.59	0.01	0.84	0.03
New Zealand	0.47	0.02	0.70	0.03	Romania	0.62	0.02	1.00	0.03
Colombia	0.48	0.02	0.48	0.02	Slovak Republic	0.65	0.02	0.90	0.02
Russian Federation	0.48	0.02	0.59	0.02	Hungary	0.65	0.01	1.00	0.03
					Bulgaria	0.68	0.01	1.00	0.03

Measurement Error

Although it's reasonable to expect some discrepancies between a grade 4 child and his/her parents…

…The degree of disagreement seems beyond what we might expect (if both respondents understand the question)…

…And it's less reasonable that 15 year olds and their parents don't agree on grade repetition.

Improving Measurement of Key Reporting Variables

- Increasingly ILSA data are called upon to inform policy interventions and reform.
- It's all the more important to ensure that *key reporting variables* (varies by country) are measured well.

Improving Measurement of Key Reporting Variables

Plausible design “solutions”:

- Measure from *more* reliable source (records, census)
  - US Census *small area income and poverty estimates*
- Measure from *more* sources (parents, records)

Some sort of small(er) sample validity study should accompany especially parent measures of important variables

Causal Inferences and ILSAs

Over the last 10 years, there has been an enthusiastic push for making *causal inferences* with ILSA data

Of course, it's reasonable that policy makers and researchers want to know *what* can be done to improve educational outcomes

But ILSA data are observational and cross-sectional. So the best we can do is use *quasi-experimental* designs

Causal Inferences

Quasi-experimental designs *seek to approximate* the gold-standard of randomization

Generally, emphasizes the *what if* aspect of a sequence of events – *what if* Jane had gone to pre-school one year earlier

BUT causal inferences hinge on meeting a pretty stringent set of assumptions.

Using D. Rutkowski & Delandshere (2016) validity framework:

- Meeting assumptions is often questionable
- Or untestable (even in randomized studies!)

Validity Perspective

Necessitates a question that is limited and focused

- *Can a counseling intervention reduce drop out rates among at-risk 15-year-olds?*

Policy makers want broader inferences

- *How can we improve graduation rates among at-risk 15-year-olds?*

ILSAs generally don't target either sort of research question – the data are consequent to (even) broader aims

Designing Causal Questions

Kaplan (2016) recommends defining a limited, focused set of causal questions early in development

- Should be integrated into framework
- Must consider and include context to (hopefully) isolate the cause

Rubin (2007) argues that identifying important variables is *non-trivial*. And should be based on substantive expert opinion.

Designing a Stronger Causal Foundation

Longitudinal component – repeated measures over time on same individuals (or a subset)

- ***Not a silver bullet*.** But puts inferences on stronger footing.

TIMSS is a natural place to pilot this

- Measures 4th and 8th graders (natural 4 year lag) → randomly equivalent populations measured at two time points.
- Would require (at least) a sufficient set of items to link over time.

Summary

ILSAs are used in more and different ways than ever before

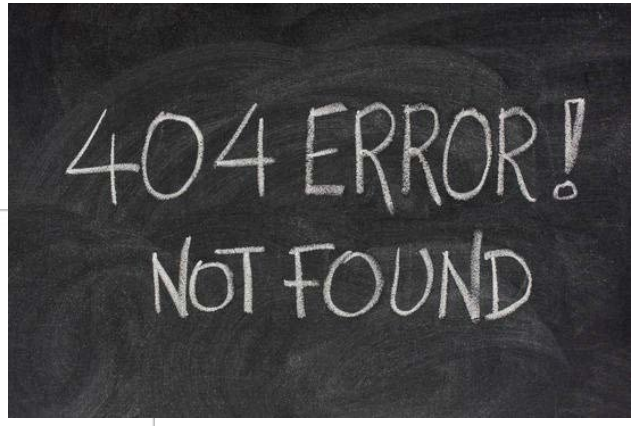
They serve as the evidence base upon which to

- Draw causal inferences
- Motivate reforms
- Implement interventions
- Celebrate or shame

Summary

ILSAs are

- Carefully developed
- Ambitious
- With great potential to measure and compare



But they are ultimately

- Observational
- Cross-sectional
- Subject to cultural differences
- And error prone

Summary

In their current forms, they are being pressed into service beyond their capacity

Although my proposal (while not cheap or easy) can potentially improve matters

We will always have fallible data that should be used within their capacity

Thank you.

leslie.rutkowski@cemo.uio.no