



A Note on Large Scale Educational Assessments

Assessments, Models, Target(s) of Inference

Matthias von Davier, ETS



Large-Scale Assessments, 1

- NAEP, the National Assessment of Educational Progress, a government mandated assessment in the United States
- Complex student and item sample in grades 4,8 and 12
- Reading, math and other subjects
- Large collection of background variables
- <http://nces.ed.gov/nationsreportcard/>



Large-Scale Assessments, 2

- NAEP served as the model for other assessments, such as:
 - TIMSS – now Trends in Mathematics and Science Study (IEA)
 - PISA – Programme for International Student Assessment (OECD)
 - PIRLS – Progress in International Reading Literacy Study (IEA)
 - ...

Large-Scale Assessments, 3





Large Scale Assessments, 4

Using the picture of my home, are there:

- Neighboring houses of similar size?
- Trees around the houses?
- Cars in the street?
- How many garages in my home?
- How many bedrooms?
- How many computers in the house?



Assessment Purposes, 1

- No individual student scores, only group-level reporting (groups defined by gender, ethnicity, ...).
- Minimum reporting group sizes (rule of 62...).
- Large-scale assessments are good for many purposes
- But: Not appropriate for every purpose! Because of:
 - Cross-sectional design, no growth/change assessment.
 - Non experimental, descriptive, no causal inferences.
 - Vertical scales across many grade levels not easy.
 - Designed for broad construct coverage and group-level per time-point reporting, IRT linking across cycles for trends.



Assessment Purposes, 2

- International Assessments such as PISA, TIMSS and PIRLS inherit these design principles:
 - Goal is to provide assessments for comparisons across participating countries, and across sufficiently(!) large groups within country.
 - International assessments often have much smaller per country sample sizes than national assessments may have per state.



Assessment Purposes, 3

- International Assessments such as PISA, TIMSS and PIRLS are, by design, not national assessments:
 - National assessment systems are better targeted towards national standards.
 - Braun & Qian study shows how NAEP can be used to evaluate standards assessed by state testing programs.
 - International assessments may piggyback better- targeted national options for doing similar evaluations.



Issues in Modeling (international)

1. Coverage of curriculum and choice of measurement model.
2. Target is a diverse sample of countries, developing and industrialized countries.
3. Comparability of measures, national adaptations, differential item functioning.
4. Background data mean different things in different countries (number of bathrooms, refrigerators, goats ... in a household).



Issues in Modeling (general)

1. Level of background variables is not trivial, is the percentage of kids receiving free lunch in a school a school-level variable?
2. Results can be misrepresented, e.g.: Eliminate the correlation of *computer use* and *math proficiency* by conditioning on XYZ variables.
3. Some relevant variables are “latent”: Parental intent to give their children the best possible education versus the actual school choice.



Outlook: Large-Scale Assessments

Change in educational systems, statistical issues, and potential misinterpretations are continuing challenges:

- Better cooperation between content specialists and statisticians, quantitative researchers is necessary.
- Training and research on better using the existing international databases is necessary.
- A model could be the NAEP (again): Establish secondary-analysis grant programs.
- Forum for evolutionary approaches, improving existing techniques and models, as well as more revolutionary approaches.



Outlook: Large-Scale Assessments

- Assessments designed for different purposes can supplement each other:
 - Assessment hooks: align large-scale international assessments with assessments targeting smaller groups (state-level down to school-level reporting).
 - Finally, hook assessments targeting individual scores, assessments for measuring change, and measuring intervention effects into (inter-) nationally comparable scales.