Discussion of "Putting test scores on an interval scale"

Peter van Rijn

ETS Global

PIAAC seminar, Paris, June 14, 2019

Introduction

- Discussion of scale properties of measurement has a long history (Campbell, 1920; Stevens, 1946)
- Discussion of the scale properties of test scores is not new (e.g., Brogden, 1977), but also not gone (Domingue, 2014).
- It is not always commonly agreed that test scores do not have interval scale properties.
- The argument in favor comes from the theory of additive conjoint measurement (Krantz et al., 1971).
- Large-scale educational assessments have two main sources of error: sampling error and measurement error.

Introduction

- The studied case in Bond and Lang (2018) is interesting because they use a relatively straightforward outcome: educational attainment.
- In their case, gaps between groups in early education are expressed in terms of eventual educational attainment.
- They can do this because they have longitudinal data (CNLSY).
- Normally, in testing, it is done the other way around (how well do test scores predict gaps in educational attainment?)
- In PIAAC, educational attainment is already available for (most) participants, so their approach is worth considering.

Some Confusion

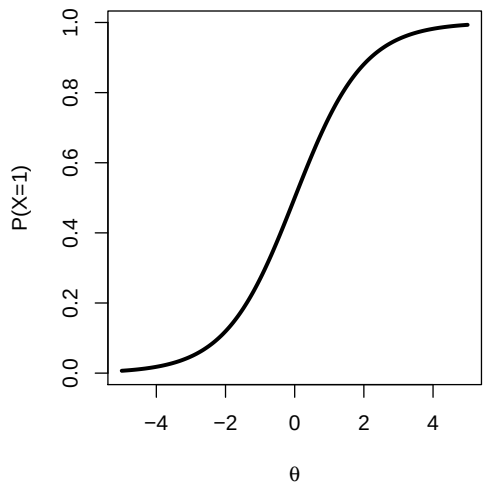
- Bond and Lang (2013) argue that test scores have ordinal scale properties, so monotonic transformations are allowed ('permissible statistical operation' in Stevens' (1946) terminology).
- A sixth order polynomial is applied, but then means and standard deviations of test gaps are evaluated (which are not permissible for interval data, right?)
- It can be argued that if means and standard deviations are used, then monotonic transformations should not be used either.

Measurement Error

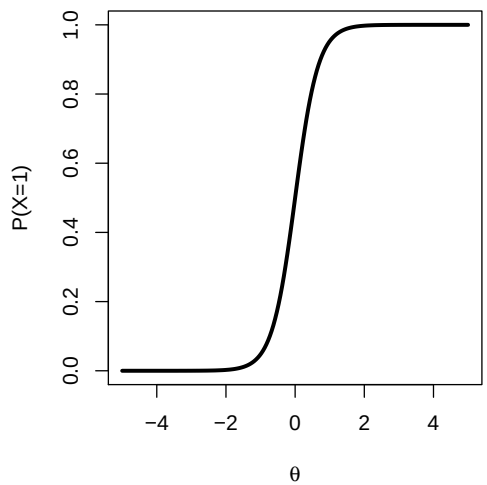
- Large-scale educational assessment are intended to estimate group means (one measurement error).
- It is more complicated to estimate a difference in group means (two measurement errors).
- It is even more complicated to estimate a change in a difference in group means over time (e.g., Harris, 1963; four measurement errors).
- Bond and Lang (2013; 2018) discuss longitudinal data for which measurement error can be even more complex (Mellenbergh & van den Brink, 1998).

Another Paradox

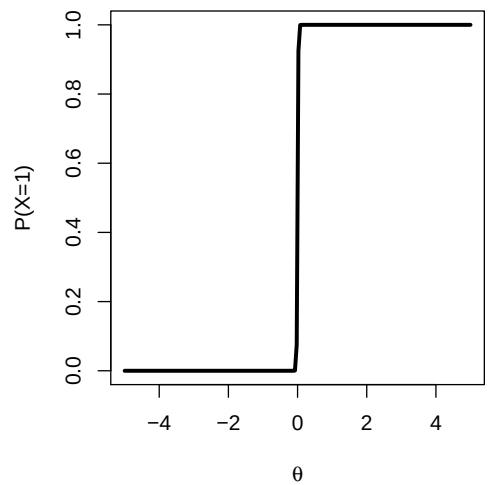
- Rasch model: sum score is sufficient statistic for θ, relates to principle of maximum entropy.
- It has been argued that ability θ in the Rasch (1960) model has interval-scale properties, but if the discrimination increases to infinity the Mokken (1971) model is obtained, which has ordinal-scale properties (van der Maas, 2013):



Interval



Interval



Ordinal?

The van der Linden-Briggs debate

- In 2015 and 2016, Wim van der Linden and Derek Briggs debated on interval scales in educational measurement at conferences of NCME and the psychometric society: "Equal interval scales in educational testing: Attainable goal or myth?"
- Van der Linden's position: The quest for interval scale has been a waste of time. There are many examples in physics and statistics where measurements are expressed on a scale with unequal units. Test scores have norm-referenced and criterion-referenced interpretations that we should not alter.
- Briggs's position: The foundations of magnitudes of differences based on test scores are a matter of debate. Research should be done to establish properties (double cancellation, etc.).

Discussion

- "Test score scales are not equal-interval. Is this a problem?"
- We should not forget how scales were created. For example, PISA scale is norm-referenced (mean of 500, sd of 100 refers to OECD countries in 2000)
- Measurement errors matter, but so do confidence intervals and these are typically (much) larger if there are multiple sources of measurement error.
- Note that PIAAC and PISA data are publicly available (both item-level test and questionnaire data, even process data (e.g., response times))