Theoretical Implications of Equating Methodologies

Dr Juho M Looveer
NSW Department of Education and Training
9th National RoundTable
Crowne Plaza, Sydney, Australia
7-9 November 2004
At the July 2002 Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) meeting, Council referred technical issues associated with the testing and reporting of nationally comparable data against the literacy and numeracy benchmarks to the Performance Measurement and Reporting Taskforce (PMRT), including:

i) a common equating method
to be used by all jurisdictions ...
Simple Basic Notions

\[ P(X = x: \beta, \delta, \kappa) = \frac{e^{[\kappa_x + (\beta - \delta)]}}{\sum_{k=0}^{m} e^{[k\kappa_x + x(\beta - \delta)]}} \]
Unidimensionality

\[
\text{Number} + \text{Measurement} + \text{Space} = \text{Numeracy}
\]

Adapted from Conference Proceedings (Tognolini, 1996)
The I.C.C. / T.C.C.

- The Item Characteristic Curve (ICC) and Test Characteristic Curve are theoretical curves.

- Ideally, all points lie on the ICC (TCC).
Ideal ICC

Rasch Item Characteristic Curve

P(x)

x
Usual ICC

Rasch Item Characteristic Curve

Theoretical Curve
Equating

- A process that enables item difficulty and student ability from different tests to be plotted on a single scale

- Basic Strategies (using Rasch methodology):
  - Parameter Shift (items or persons)
  - Anchoring
  - Concurrent
Parameter Shift Equating

- Two tests are scaled separately
- Means are calculated
  (for persons or items)
- All items or persons from one test are shifted so that the means for the two tests are equal
Parameter-Shift Equating

Two separate curves / tests
Parameter-Shift Equating

Two separate tests on one graph
Anchoring - Equating

- One test is calibrated
- Item locations or person measures from the first test are applied when calibrating the other test
- This method needs either common items or common persons
Concurrent Equating

- All data is processed together
  i.e. scales are co-calibrated

![Rasch Item Characteristic Curve](image)
Implications

Concurrent Equating and Anchoring
- Tries to fit all of the data to the one curve
- “makes the rope thicker”

Parameter Shift
- Means are aligned
- Allows for different shaped curves
  i.e. a different construct
National Benchmarking in Australia

- “minimum level for further progress in education”
  - i.e. a cut-point towards the lower end of the scale
Use of a Cut-Score

Rasch Item Characteristic Curve

Cut-score = 0.1
Problems

- Problem – large changes in % above or below the cut-scores
  - Within states, across time
  - Across states

- Could it be due to differences in the equating?

- Could it be due to differences in the curves?
Concurrent Equating

Rasch Item Characteristic Curve

Cut-score = 0.1
Parameter Shift Equating

Rasch Item Characteristic Curve

Cut-score = 0.11
Parameter Shift Equating

Rasch Item Characteristic Curve

Cut-score = 0.11
Cut-score = 0.06
Research

- Question –

Would different results (mean score, percentage meeting benchmark) have been obtained had a different equating procedure been used?
Comparing Software

Quest Vs RUMM Y7 Writing 2001

--- Items ---

<-- Item difficulty -->
Comparing Methods

Concurrent vs Parameter Shift

Y7 Writing 2001
“Some software led to significant differences in item locations, depending on the equating methodology used”
The practical purpose of equating is to put scores onto scales with a similar range of scores.

There may be educational reasons for having different constructs in the tests being equated.

Parameter-shift method can lead to larger differences in %ages above or below cut-scores, particularly at ends of a distribution.
Issues

- Misfit of items (or persons) can effect the scale
  - Which misfitting persons / items should be deleted from the equating?

- Long-term trends are more important than year-to-year fluctuations
Discussion