Investigating stability of item parameters in one and two-parameter partial credit models on PISA Science cognitive data

Luc T Le, ACER
Investigating stability of item parameters in one and two-parameter Partial Credit models on PISA Science cognitive data

Luc Le
Australian Council for Educational Research, Australia

Outline of PISA
The PISA study (Programme for International Student Achievement) is a very large worldwide survey conducted by the Organisation for Economic Co-operation and Development (OECD). It was first conducted in 2000 and has been repeated every three years since. PISA assesses literacy in reading (in the mother tongue), Mathematics and Science. In these areas PISA mainly assesses 15-year-old students’ capacities to use their knowledge and skills in order to meet real-life challenges, rather than merely looking at how well they have mastered a specific school curriculum. The test has been translated (and/or adapted) into more than 40 different test languages equivalent to the English and French source versions developed by the PISA consortium. In each cycle, one of the assessment areas is the major domain consecutively. Particularly Science was the major domain in 2006 (cycle 3) and is going to the major domain again in 2015 (cycle 6).

PISA test design and IRT models
In PISA study, cognitive items have been organised into different test form booklets by a linked design (for example, see OECD, 2014). And students have been assigned to do the test booklet randomly. The item formats in PISA include multiple-choice and open-ended (short-answer or extended response).

As item response theory (IRT) models have been widely used in analysing and constructing educational and psychological tests. PISA data have been analysed based on the IRT partial credit model (Masters, 1982) which was an extension of the simple logistic Rasch model (Rasch, 1960/1980).

The Partial Credit Model (PCM) or one-parameter Partial Credit Model developed for polytomous scored items. Correspondingly, by adding a discrimination parameter into each item Muraki (1992) generalised this model as two-parameter partial credit model or named as generalized partial credit model (GPCM)

From the model formula it is expected that the more parameters added the more likely the model-data goodness-of-fit statistics achieves better (Fitzpatrick, Link, Ito, & Sykes 1996; Harris, 1989). However, it may be not true for the stability of the item parameter estimates across examinee groups. Item parameter invariance may not be guaranteed by the mere fact that an IRT model is fit to data (van der Linden & Hambleton, 1997; Engelhard, 1994).

Choosing an appropriate model for the test data is an essential element in assuring their quality. Together with fit statistics, the stability or invariant level of the item parameter estimates across examinee groups is an important criterion for selecting an appropriate IRT model for the data set.

Study objectives
This study focuses on investigating the stability of item parameter estimates from PCM and GPCM respectively, by the test form booklets across the country and language groups.

From that, it could suggest whether the item discrimination parameter should be needed in the item calibration model for large and heterogeneous samples of an international test like PISA, particularly for PISA Science 2015.

Methodology, Methods, Research Instruments or Sources Used
Data:
PISA cycle 3 data collected in 2006 were used. There were 57 countries participated in this study with more than 40 test languages. There were 108 cognitive science items organised in 13 linked test booklets. Each item was located in 4 booklets.

Analysis:
Calibrating items: In a country/test language group, item parameter estimates by each of the booklets were obtained with the PCM (Masters, 1982) and GPCM (Muraki, 1992) using two software packages: CONQUEST (Adams, Wu, & Wilson, 2012) and MULTILOG (Thissen, Chen, & Bock, 2003).

Equating scales: to make item parameter estimates comparable across the booklets for each country/test language group, it was designed to equate the item parameter estimates from the different booklets into the same scale.

Computing variance: after equating into the same scale, the variance of the parameter estimates across the booklets for each item was computed.

Examining variance of item parameter estimates: the mean as well as the patterns of the of item parameter estimates across the countries within each model were examined. The variances by the two models were compared.

Conclusions, Expected Outcomes or Findings
Several results were drawn from this study that would lead some discussions, in particular, whether the discrimination parameter is needed. There was a significant variation of item difficulty across the test forms in both estimation models. However there was a very small variation in the item point-biserial discrimination and item discrimination/slope parameter from GPCM. The clear difference in item discrimination between multiple-choice items and open-ended items would be concerned.

Findings from the study would provide some practical assistance in the selection of IRT models for PISA. However, the choice of IRT models could affect ability estimation and thus the comparisons between the examinee groups as well as the categorization of examinees at various proficiency levels.

References


**Intent of Publication**