7-2007

# Effects of item positions on their difficulty and discrimination : A study in PISA Science data across test language and countries

Luc T. Le
*ACER*, Luc.Le@acer.edu.au

# Effects of item positions on their difficulty and discrimination- A study in PISA Science data across test language and countries

**Luc T. Le**
*Australian Council for Educational Research, 19 Prospect Hill Road (Private Bag 55) Camberwell VIC 3124, Australia; email: le@acer.edu.au*

## Abstract

This study was based on a four-cluster rotation design of 13 linked test booklets from PISA 2006 science data. It investigated effects of item positions on their difficulty and discrimination parameter estimates obtained from one and two-parameter IRT Partial Credit models. The analyses were done separately for 57 test language groups from 53 countries with a total of about 340,000 students.

The results revealed that for all of the test language groups the items tended to become more difficult when they were located later in the test. However, a high linear relationship between the item difficulty estimates by the four cluster locations was found. Moreover, open-ended items seemed to show more change than items in other formats. There were small variations in the cluster locations for the item point-biserial discrimination and the item discrimination parameter from the two-parameter Partial Credit model across the test language groups.

## 1. Introduction

### Outline of PISA

The PISA study (Programme for International Student Achievement) is a very extensive 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 15-year-old students' literacy in reading (in the mother tongue), mathematics and science with regard to their 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. Fifty-seven countries participated in PISA 2006 where science was the main focus. The test was translated (and/or adapted) into more than 40 different languages equivalent to the English and French source versions developed by the PISA consortium. In the PISA study, the cognitive items were organised into different test booklets by a linked design (see OECD, 2004). Each student was assigned a test booklet randomly. Then, in the scaling process their responses were analysed mainly based on the IRT *partial credit model* (Masters, 1982) with additional adjustment to compensate for the test booklet effects (see PISA technical reports: OECD, 2002; OECD, 2004).

### IRT partial credit model and its extension

Item response theory (IRT) models are now widely used in analysing and constructing educational and psychological tests (Hambleton, 1983; Lord, 1980). The central element of item response theory is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying

ability. The Partial Credit Model (PCM or one-parameter Partial Credit Model) has been developed for polytomous scored items (Masters, 1980). The model can be described by a mathematical probability function:

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x} (\theta - b_i - \tau_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta - b_i - \tau_{ij})}, \; x = 0, 1, 2, ..., m_i \qquad (1)$$

where $P_{ix}(\theta)$ denotes the probability of a person with ability level $\theta$ (on the latent dimension) to score $x$ on item $i$ with $m_i + 1$ ordered categories 0, 1,…, $m_i$. $\tau_{ij}$ denotes a step parameter, standing for the event that the person responded to category *j rather than j-1* ($\tau_{i0} \equiv 0$). The item parameter $b_i$ gives the location of the item on the latent continuum. This parameter is also known as "item difficulty".

By adding a discrimination parameter $a_i$ into the PCM, Muraki (1992) expanded this model as *two-parameter partial credit model* or *generalized partial credit model (GPCM)*

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x} a_i (\theta - b_i - \tau_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} a_i (\theta - b_i - \tau_{ij})}, \; x = 0, 1, 2, ..., m_i \qquad (2)$$

Comapred with the traditional test theory, the item parameter invariance is a very robust property within IRT models. The *invariance* means that the values of the parameters are identical in different samples or across different conditions of interest (Lord, 1980).

The necessary condition for the absolute invariance of the parameters here is that the model needs to fit perfectly with the data. In other words, "invariance only holds when the fit of the model to the data is exact in the population" (Hambleton et *al*., 1991, p. 23). However, this ideal condition is never matched in practice. There is always some degree of variance of the item parameters across examinee groups, especially in complex designs. For example, different linked test forms can be administered at the same time and/or the sample can include different examinee groups according to their demographics, study backgrounds, cultures and ability levels. Recent research in PISA showed some variation of the item difficulty by test language (Grisay et *al*. 2006), country (Le, 2006a) and gender (Le, 2006b).

From a model aspect, it is expected that the more parameters added the better the model-data goodness-of-fit statistics are (Fitzpatrick et *al*., 1996; Harris, 1989; Hambleton, 1983; Hambleton & Cook, 1983). However, it may not hold true for the stability of the item parameter estimates across the examinee groups. Item parameter invariance may not be guaranteed by the mere fact that an IRT model fits to individual data sets (van der Linden & Hambleton, 1997; Engelhard, 1994).

In particular, the effects of item position on examinee performance have been shown in several studies. Some of them found no differences in item difficulty for different test position (Rubin & Mott, 1984; Klein & Bolus, 1983; Zwick, 1991), but

others found evidence that items are more difficult when they appear later in tests (Walz et *al*., 2000; Wise et *al*., 1989).

### Research questions

In this paper, three key research questions have been addressed:

- Can items become more difficult when they are located towards the end of the test?
- How does item discrimination vary based on the items' positions in the test?
- How do the estimates of the item discrimination (slope) parameter from GPCM change when items are located towards the end of the test?

## 2. Method

### Data

PISA cycle 3 data were collected in 2006. The data used in this study included 57 test language groups from 53 participating countries (28 OECD and 25 non-OECD)[1], where each of the PISA test booklets were responded to by at least 100 students from each of the test language groups. A test language group included all examinees in a country sample who were using that test language. There were approximately 340,000 students in the analysis with about 49.5% males and 50.5% females.

**Table 1. Allocation of item clusters to test booklets in PISA 2006**

| Booklet | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| 1 | S1 | S2 | S4 | S7 |
| 2 | S2 | S3 | M3 | R1 |
| 3 | S3 | S4 | M4 | M1 |
| 4 | S4 | M3 | S5 | M2 |
| 5 | S5 | S6 | S7 | S3 |
| 6 | S6 | R2 | R1 | S4 |
| 7 | S7 | R1 | M2 | M4 |
| 8 | M1 | M2 | S2 | S6 |
| 9 | M2 | S1 | S3 | R2 |
| 10 | M3 | M4 | S6 | S1 |
| 11 | M4 | S5 | R2 | S2 |
| 12 | R1 | M1 | S1 | S5 |
| 13 | R2 | S7 | M1 | M3 |

[1] Data from other 4 countries were not included in this study due to their late submission.

The PISA cognitive items were organised by clusters and arranged into 13 linked test booklets. Table 1 presents the structure of this cluster design. There were seven science clusters (S1-S7) with a total of 104 items, four mathematics clusters (M1-M4) and two reading clusters (R1-R2). According to this design, each cluster, and therefore each item, appeared in four booklets in different locations.

Five variables or dimensions classifying the science item characteristics were defined in the PISA framework (OECD, 2006). The detailed number of items in each category is provided in Table 2.

**Table 2. Item classification and frequency**

| Dimension | Number | Per cent | Dimension | Number | Per cent |
|---|---|---|---|---|---|
| | | | **Science** | | |
| **Item Focus** | | | **Knowledge** | | |
| Global | 27 | 26.0 | *Of Science* | | |
| Personal | 26 | 25.0 | EASS | 12 | 11.5 |
| Social | 51 | 49.0 | LIVS | 22 | 21.2 |
| **Item Context** | | | PHYS | 24 | 23.1 |
| ENV | 20 | 19.2 | *About Science* | | |
| FRO | 27 | 26.0 | SENQ | 23 | 22.1 |
| HAZ | 17 | 16.3 | SEXP | 19 | 18.3 |
| HEA | 27 | 26.0 | STEC | 4 | 3.8 |
| NAT | 10 | 9.6 | | | |
| Other | 3 | 2.9 | | | |
| **Item Competency** | | | **Item Format** | | |
| EPS | 50 | 48.1 | CMC | 27 | 26.0 |
| ISQ | 26 | 25.0 | CR | 4 | 3.8 |
| USE | 28 | 26.9 | MC | 38 | 36.5 |
| | | | OR | 35 | 33.7 |
| **Total** | 104 | 100 | **Total** | 104 | 100 |

*Focus*: Situations relating to the self, family and peer groups (*Personal*), to the community (*Social*) and to life around the world (*Global*).

*Context*: Life situations involving science and technology: Environment (ENV), Frontiers (FRO), Hazards (HAZ), Health (HEA), and Natural resources (NAT).

*Competency*: Explaining phenomena scientifically (EPS), identifying scientific questions (ISQ) and using scientific evidence (USE).

*Scientific knowledge*: Both "knowledge of science" and "knowledge about science". "Knowledge of science" includes *Physical systems* (PHYS), *Living systems* (LIVS), and *Earth and space systems* (EASS); while "Knowledge about science" refers to *Scientific enquiry* (SENQ), *Scientific explanations* (SEXP) and *Science and technology* (STEC).

*Item format:* The 2006 PISA test consisted of four types of cognitive items: (1) multiple choice (MC); (2) closed response (CR), which is short verbal or numerical response, with a clear correct answer; (3) complex multiple choice (CMC) that is a series of true/false or yes/no choices, with one answer to be chosen for each element in the series; and (4) open-ended response (OR). Most of the OR items required markers. In the IRT analysis, the data of MC and CR items were recoded as

dichotomous (0 and 1) while data from the other item types were recoded as partial credit (0 and 1 or 0, 1 and 2).

### Analysis

*Calibrating items*: For each test language group and each of the booklets, item difficulty parameter estimates (from PCM) were obtained using Conquest (Wu et *al*., 1997) and item discrimination/slope parameter estimates (from GPCM) were obtained using MULTILOG (Thissen et *al*., 1997). Both software packages use the same estimation algorithm *EM* (Bock & Aitken, 1981).

*Model identification:* In each of the calibration, the sample ability mean of 0 is set up in Conquest, and the mean of 0 and standard deviation of 1 is set up for the sample ability in MULTILOG.

## 3. Results

### Stability of item difficulty estimates from the PCM

Table 3 shows the distribution of the mean difference and the correlation of the item difficulty estimates obtained from each pair of the four cluster positions across the test language groups. The table indicates that, in general, items gradually become more difficult along with the designed cluster locations. The biggest gap is the difference between the first and fourth clusters and the smallest gap is between the first and second and then between the second and third. Interestingly, this pattern occurs consistently for all individual language groups.

Moreover, the Pearson correlations between the item difficulty estimates by the four cluster positions are very high. The highest ones are for two consecutive clusters (mean r12= mean r23 =0.97), while the lowest correlation is for the first and fourth clusters (mean r14=0.94). This demonstrates a very high linear relationship among the four estimates for each item.

**Table 3. Mean difference of item difficulty estimates by cluster locations across the test language groups**

| | Difficulty difference (in logits) | | Correlation | |
|---|---|---|---|---|
| **Clusters** | **Mean** | **Std. Deviation** | **Mean** | **Std. Deviation** |
| Between 1st and 2nd | 0.09 | 0.05 | 0.97 | 0.02 |
| Between 1st and 3rd | 0.19 | 0.07 | 0.96 | 0.01 |
| Between 1st and 4th | 0.37 | 0.09 | 0.94 | 0.02 |
| Between 2nd and 3rd | 0.10 | 0.04 | 0.97 | 0.02 |
| Between 2nd and 4th | 0.28 | 0.08 | 0.96 | 0.02 |
| Between 3rd and 4th | 0.18 | 0.05 | 0.96 | 0.02 |

Table 4 gives the average percentage number of the items where they are significantly harder at the the first cluster position or harder at the fourth cluster position, respectively at the 0.05 level. The last row of the table shows that on

average, only 2% of the items are more difficult at first cluster position, and that 52% the items are significantly more difficult at the fourth cluster position. The difference is not statistically significant for 46% of the items.

**Table 4. Average percentage of items by their difficulty difference at the first and fourth cluster positions**

| Item category | Harder at 1st cluster (%) | Harder at 4th cluster (%) | No significant difference (%) |
|---|---|---|---|
| **Focus** | | | |
| Global | 2.8 | 47.2 | 50.0 |
| Personal | 2.8 | 49.8 | 47.4 |
| Social | 0.9 | 55.7 | 43.4 |
| **Context** | | | |
| ENV | 0.5 | 49.6 | 49.8 |
| FRO | 3.8 | 45.0 | 51.3 |
| HAZ | 0.4 | 68.8 | 30.8 |
| HEA | 2.2 | 49.8 | 48.0 |
| NAT | 1.2 | 51.4 | 47.4 |
| **Competency** | | | |
| EPS | 2.6 | 45.3 | 52.1 |
| ISQ | 2.0 | 50.5 | 47.5 |
| USE | 0.5 | 65.5 | 34.0 |
| **Science Knowledge** | | | |
| About science | 1.3 | 60.8 | 38.0 |
| Of science | 2.4 | 45.1 | 52.5 |
| **Format** | | | |
| CMC | 3.7 | 37.5 | 58.8 |
| CR | 1.3 | 29.8 | 68.9 |
| MC | 1.3 | 49.7 | 49.0 |
| OR | 1.1 | 68.4 | 30.5 |
| **Overall** | **1.9** | **52.0** | **46.1** |

*Significant level: 0.05

With respect to the item format, the table shows that 68% of the OR items are significantly harder at the fourth position than at the first position. This figure is much higher than that of items with other formats (MC: 50%, CMC: 37% and CR: 30%).

Regarding the item contents, the change of item difficulty from the first to fourth positions is as follows:

Focus: The difference between the three categories is small;

Context: HAZ items tend to change more than other items;

Competency: USE items tend to change more than other items;

Science knowledge: The "knowledge about science" items tend to change more than the "knowledge of science" items.

## Stability of item point-biserial discriminations

Table 5 demonstrates a summary of the mean of item point-biserial discriminations (PB) by the four cluster positions across the test language groups. There is a very small change of the PB here. On average, the PB mean from the fourth cluster position is only 0.02 larger than the PB mean from the first cluster. The difference between them is less than 0.05 for all test language groups.

**Table 5. Mean of item point-biserial discriminations by cluster locations across the test language groups**

| Statistics | 1st | 2nd | 3rd | 4th | Different between 1st and 4th |
|---|---|---|---|---|---|
| Mean | 0.41 | 0.43 | 0.43 | 0.43 | 0.02 |
| Std. Deviation | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 |
| Min | 0.34 | 0.33 | 0.32 | 0.31 | -0.05 |
| Max | 0.47 | 0.50 | 0.49 | 0.49 | 0.05 |

## Stability of item discrimination parameter from the GPCM

This section presents the variation of the item discrimination estimates obtained from Multilog (GPCM) by the four cluster positions. Table 6 shows the means of the differences as well as correlation means between them across the test language groups. On average, the highest correlation is from two consecutive clusters (r12=0.76, r23 =0.77, r34=0.74), while the lowest one is from the two farthest clusters (r14=0.70). These figures, however, are much lower than the corresponding correlations for item difficulty estimates presented above, suggesting that linear relationship between the the item discrimination/slope estimated at the four cluster possitions is lower than that for the item difficulty.

**Table 6. Mean difference of item discrimination estimates by cluster locations across the test language groups**

| | Difference | | Correlation | |
|---|---|---|---|---|
| Clusters | Mean | Std. Deviation | Mean | Std. Deviation |
| Between 1st and 2nd | 0.09 | 0.09 | 0.76 | 0.10 |
| Between 1st and 3rd | 0.11 | 0.08 | 0.73 | 0.14 |
| Between 1st and 4th | 0.18 | 0.14 | 0.70 | 0.14 |
| Between 2nd and 3rd | 0.03 | 0.11 | 0.77 | 0.15 |
| Between 2nd and 4th | 0.09 | 0.12 | 0.74 | 0.13 |
| Between 3rd and 4th | 0.06 | 0.14 | 0.74 | 0.17 |

Table 7 provides the average percentage number of the items where their discrimination estimates are significantly larger at the first cluster position or at the fourth cluster position. The table shows that on average, only 1% of the items have significantly larger discrimination estimates at first cluster position, and that 10% the items have significantly larger discrimination estimates at the fourth cluster position. The difference is not statistically significant for 89% of the items.

**Table 7. Average percentage of items by their discrimination difference at the first and fourth cluster positions**

| Item category | Large at 1st cluster (%) | Larger at 4th cluster (%) | No significant difference (%) |
|---|---|---|---|
| **Focus** | | | |
| Global | 1.0 | 8.7 | 90.3 |
| Personal | 2.4 | 8.9 | 88.7 |
| Social | 0.7 | 11.8 | 87.5 |
| **Context** | | | |
| ENV | 2.0 | 7.9 | 90.1 |
| FRO | 1.2 | 9.9 | 88.8 |
| HAZ | 1.0 | 12.8 | 86.2 |
| HEA | 0.8 | 8.4 | 90.8 |
| NAT | 0.9 | 17.9 | 81.2 |
| **Competency** | | | |
| EPS | 1.0 | 9.5 | 89.5 |
| ISQ | 1.4 | 10.6 | 88.0 |
| USE | 1.4 | 11.3 | 87.2 |
| **Science Knowledge** | | | |
| About science | 1.5 | 10.6 | 87.9 |
| Of science | 0.9 | 10.0 | 89.1 |
| **Format** | | | |
| CMC | 1.3 | 7.5 | 91.2 |
| CR | 3.5 | 3.1 | 93.4 |
| MC | 1.5 | 8.5 | 89.9 |
| OR | 0.5 | 15.1 | 84.4 |
| **Overall** | **1.2** | **10.3** | **88.5** |

*Significant level: 0.05

## 4. Summary and discussion

The main findings drawn from this study are: (1) The items themselves become more difficult when they are located towards the end of the test; (2) the estimates of the item difficulty from the four cluster positions are very highly correlated to each other; (3) OR items tend to increase their difficulty more than items with other formats; (4) changing of the item difficulty could also be influenced by item content; (5) there is a small variation in the item point-biserial discrimination; and (6) the mean of the item discrimination/slope parameter from GPCM demonstrates a small change when the items are located in different positions. The estimates of the this parameter by the four positions are moderately correlated to each other.

Findings from this study suggests that, in a test equating process for linked test forms, the different locations of the items in the different forms should be taken into account to maintain test scale stability.

Results from this study also give a practical caution of fatigue or speededness effects in a test design. When designing tests, the key is to find a resonable balance between having enough items to reliably measure standards and ensuring that most students have been given enough time to complete the test or do not become overly fatigued towards the end of the test.

This study may suffer the following limitation. In some circumstances the assumption of equivalent ability mean by booklets may not be totally fulfilled. However, according to the PISA sampling design it is expected that the difference is small and does not significantly affect the overall findings.

## References


Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

Engelhard, G., Jr. (1994). Historical views of the concept of invariance in measurement theory. In M.Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. *2*, pp. 73-99). Norwood, NJ: Ablex.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, *33*(3), 291-314.

Grisay, A., de Jong J., Gebhart, E., Berezner, A., & Halleux-Monseur, B. (2006, April). *Translation equivalence across PISA countries*. Paper presented in the annual meeting of American Educational Research Association, San Francisco CA.

Hambleton, R. K. (1983). Application of item response models to criterion-referenced assessment. *Applied Psychological Measurement*, *7*(1), 33-44.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. London: Academic Press, Inc.

Hambleton, R. K.,Swaminathan, H., &Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*, 35-41.

Klein, S. P., & Bolus, R. (1983). *The effect of item sequence on bar examination scores*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.

Le, L. T. (2006a, April). *Analysis of Differential Item Functioning*. Paper presented in the annual meeting of American Educational Research Association, San Francisco CA.

Le, L. T. (2006b, July). *Investigating Gender Differential Item Functioning across Countries and Test Languages for PISA Science items*. Paper presented at the 5th conference of International Test Commission, Brussels.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

OECD (2002). *PISA 2000 Technical Report* – OECD. Paris.

OECD (2004). *PISA 2003 Technical Report.* OECD. Paris.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. OECD, Paris.

Rubin, L. S., & Mott, D. E. (1984). *The effect of the position of an item within a test on the item difficulty value.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog* (version 7) [Computer sotware]. Lincolnwood, IL: Scientific Software International.van der Linden,W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students*. Minneapolis, MN: National Center on Educational Outcomes, Report *34*.

Wise, L. L., Chia, W. J., & Park, R. (1989). *Item position effects for test of word knowledge and arithmetic reasoning*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wu, M. L., Adams, R. J. and Wilson, M. R. (1997). *Conquest: Generalised Item Modelling Software*. ACER: Australian Council for Educational Research, Australia.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practices, 10*(3), 10-16.