Validity Issues in Assessment of Learning Outcomes in Large Scale Educational Surveys: An Examination of Booklet Effect in PISA

Zahid Mehmood

Introduction

- PISA project
- Research Questions
- Modeling Framework
- Results
- Conclusions
- Implications

PISA: Programme for International Student Assessment

- OECD funded educational survey across more than 60 countries
- 3 year cycle; Mathematics, Reading & Scientific Literacy
- Cognitive tests & Background Questionnaires
- Aims:
  - to assess student knowledge & skills
  - to assess background factors influencing student performance

Research Questions

- 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?

PISA 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,
- Each of the PISA test booklets were responded to by at least 100 students from each of the test language groups.
- There were approximately 340,000 students in the analysis with about 49.5% males and 50.5% females.

Item Cluster Design

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

PISA Booklet Design

The cluster (coded 0 to 3) position was used as proxy for the item position in the analyses.

Cluster Position	1	2	3	4	5	6	7	8	9	10	11	12	13
0	M1	R1	S1	R3	R4	R5	R6	R2	M2	S2	M3	R7	S3
1	R1	S1	R3	R4	M2	R6	M3	M1	S2	R5	R7	S3	R2
2	R3	R4	M2	S2	R5	R7	S3	S1	R6	M3	R2	M1	R1
3	M3	R7	S3	R2	M1	R3	R4	R6	R1	S1	M2	S2	R5

R: Reading, M: Mathematics, S: Science

Item Freq. and Distribution

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

Abbreviations

- *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).

GPCM Model

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

- where *Pix(θ)* denotes the probability of a person with ability level θ (on the latent dimension) to score *x* on item *i* with *mi* + 1 ordered categories 0, 1,…, *mi*. τ*ij* denotes a step parameter, standing for the event that the person responded to category *j rather than j-1*
- The item parameter *bi* gives the location of the item on the latent continuum. This parameter is also known as “item difficulty”.

Analysis Software

- *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.

Results

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

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

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

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

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

Conclusions

- The items themselves become more difficult when they are located towards the end of the test;
- The estimates of the item difficulty from the four cluster positions are very highly correlated to each other;
- OR items tend to increase their difficulty more than items with other formats;

Conclusions

- Changing of the item difficulty could also be influenced by item content;
- There is a small variation in the item point-biserial discrimination; and
- The mean of the item discrimination/slope parameter from GPCM demonstrates a small change when the items are located in different positions.

Implications

- 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.
- When designing tests, the key is to find a reasonable 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.

Thanks for Listening

&

Soft Questions