Macrothink Institute™

Journal of Studies in Education
ISSN 2162-6952
2018, Vol. 8, No. 3

Random Item Response Model Approaches to Evaluating Item Format Effects

Yongsang Lee

Korea Institute for Curriculum and Evaluation

Korea, Republic of

Inyong Park

Korea Institute for Curriculum and Evaluation

Korea, Republic of

Received: July 16, 2018 Accepted: August 13, 2018 Published: August 13, 2018

doi:10.5296/jse.v8i3.13387 URL: https://doi.org/10.5296/jse.v8i3.13387

Abstract

The PISA 2006 science assessment is composed of open response, multiple-choice, and constructed multiple choice items. The current study introduced the random item response models to investigate the item format effects on item difficulties, and these models include the linear logistic test model with random item effects (i.e., the LLTM-R) and the hierarchical item response model (i.e., the hierarchical IRM). In this study these models were applied to the PISA 2006 science data set to explore the relationship between items' format and their difficulties. The empirical analysis results in the PISA 2006 science assessment first find that the LLTM-R and the hierachical IRM provides equivalent item difficulty estimates compared with those from the Rasch model and the LLTM, and also clearly show that the item difficulties are substantially affected by item formats. This result implies that item difficulties may be different to each other depending on the item format although they deal with the same content.

Keywords: item format, LLTM-R, Hierarchical IRM, PISA 2006 science assessment

1. Introduction

Over the last few decades, international studies of student achievement have become tremendously popular, and these studies include the Program for International Student Assessment (i.e., PISA). PISA is an internationally standardized assessment that was jointly developed by participating countries and is administered to fifteen-year-olds in schools to identify key demographic, social, economic, and educational factors affecting student performance in reading, mathematics, and science. Each PISA data collection effort assesses one of these three subject areas in depth, and science literacy was the subject area assessed in depth in PISA 2006 (OECD, 2006a, 2006b).

The PISA 2006 Science Assessment was constructed using three item formats: open response (OR), multiple choice (MC), and complex multiple choice (CMC) items. Depending on the item format, item difficulty may vary although the items deal with the same content (Bolger & Kellaghan, 1990; DeMars, 1998, 2000; Garner & Engelhard, 1999; Hellekant, 1994). If there are any significant item format effects, they should be considered at the stage of assessment design. In order to ensure the assessment's quality and provide useful guidelines for designing and revising PISA assessment items, the effect of item format on item difficulty should thus be investigated.

In order to provide an analysis tool for item properties (such as item format), various item response models have been proposed (Butter, De Boeck, & Verhelst, 2004; De Boeck, 2008; Fischer, 1973, 1983; Janssen, Schepers, & Peres, 2004; Mislevy, 1988). One of these models is the linear logistic test model (LLTM; Fischer, 1973, 1983). In the linear logistic test model, item difficulty is expressed as a linear function of item property (e.g., item format) effects, and thus how each item property affects item difficulty can be explored in this framework. Since this model does not specify any random effects with respect to item difficulty, the underlying assumption of this model is that a certain number of item properties fully explain item difficulty. This assumption is, in fact, significant, but may not be valid in reality. There are always certain factors, which we may not notice, affecting item difficulty. In order to deal with this assumption, recent research has introduced the random item model, which is the LLTM with random item effects (De Boeck, 2008). By introducing random item effects, this model tries to quantify the uncertainty; and, consequently, in this model, item difficulty is expressed as a linear combination of this uncertainty, item property values, and their effects.

Another approach to evaluate the item format effects is the hierarchical item response model (Janssen, Tuerlincks, Meulders, & De Boeck, 2000; Janssen et. al., 2004). This model was originally developed to be applied to criterion-referenced measurement, where items measuring the same criterion can be grouped. Since the PISA items were constructed using three item formats (i.e., OR, MC, and CMC), items can be grouped into these item formats. The hierarchical item response model (IRM) specifies the item variance in the model, and partitions this variance into within-group and between-group variances as the traditional hierarchical linear (HLM) model does (Raudenbush & Bryk, 2002). The between-group variance is often interpreted as the group effect in the HLM model; and, thus, one can evaluate the item format effect by estimating this group effect using the hierarchical IRM.

Since the LLTM-R and the hierarchical IRMs specify the random variances for the item parameters, these models are called the random item IRMs.

The purpose of this study is twofold. First, the current study introduced item response model approaches to investigate the item format effect. The current study especially focuses on the random item response models in addition to the fixed item response model. Second, the study examined how the format of items in the PISA 2006 science assessment contributes to item difficulty. In order to tease out the effect of item format and test whether any relationship between item format and difficulty exists, three IRMs were used. For the models, the current study applied the LLTM, the LLTM-R, and the hierachical IRMs to the PISA 2006 data and compared the results across models.

2. The Linear Logistic Test Model

Schools have extensively used large-scale educational assessments to inform educational stake holders about students' academic performance, and have used IRMs to utilize the information gleaned from large-scale assessments. Traditional IRMs, such as the Rasch model, have provided descriptive information about students’ proficiency and item difficulty. Its main concerns are how to measure students’ proficiency or item difficulty, and how to estimate them accurately. Psychometric achievements, however, enable us to explain what factors cause certain levels of student proficiency and item difficulty, and how those factors affect these parameters. One of these achievements is the linear logistic test model (LLTM; Fischer, 1973, 1983), and is expressed in Equation 1.

$$\eta_{pi} = \theta_p - \sum_{k=0}^{K} \gamma_k X_{ik} \tag{1}$$

In this mathematical expression, η_{pi} indicates the log odds of the response of person p to item i; θ_p indicates a person ability parameter for person p; γ_k indicates the effect of item property k; and X_{ik} indicates the value of item i on item property k. Since the LLTM model replaces the item parameter $\beta_i$with a linear combination of γ_k and X_{ik}, item parameter β_i can be calculated with values of γ_k and X_{ik}. With the LLTM model, one has the opportunity to identify item factors that create differences in item difficulty. Although traditional research on item difficulty has focused on identifying these factors, the approaches tend to be qualitative; and thus quite limited in quantifying the relationship between the item's difficulty and its properties (i.e, factors affecting item difficulty). The LLTM is the first item response model which is designed to identify significant properties and their effect size. Even with its promising features, this model has some limitations depending on measurement situations. The first limitation comes from the fact that the effect of item property is fixed. Since this effect is fixed, this model does not allow any individual differences; for example, the open response item format may be easier for some students than other students compared to the multiple choice item format, and the opposite situation is also likely to happen. The original LLTM is, however, designed to estimate the average effect rather than individual effects, and thus the underlying assumption of this model is that the item format effects are same for all students. The second limitation is that the model assumes that a certain number of item properties fully explain item difficulty; it does not specify any residuals in the combination of item properties and their effects. This unrealistic assumption may lead to an inaccurate parameter estimate of item property effect when substantial residuals exist. Since the main purpose of the LLTM is to estimate the item property effect, this limitation is critical. The first limitation is relatively easily to overcome. Rijmen, Tuerlinckx, De Boeck, & Kuppens (2003) addressed the first limitation of the LLTM and introduced the random weight LLTM model. In their model, the weight (i.e., item property effect) is random, and thus the item property effect may vary across students. The second limitation has required technical advances in estimation methods. When the model specifies residuals for item parameters, it becomes a crossed random effect model where both items and persons are random. When software adopting the Bayesian estimation method is introduced, however, estimation in this kind of model ceases to matter; eventually, De Boeck (2008) showed the LLTM with errors and its applications.

3. The Linear Logistic Test Model with Errors

Fischer (1973, 1983) designed the original LLTM to explain item difficulty with respect to underlying cognitive operations or item properties, considering item effects fixed. This implies that item properties can perfectly explain item difficulty, which might not be true. This assumption has been relaxed by incorporating random effects for the items (De Boeck, 2008) later. In the LLTM with random item effects (LLTM-R), item properties (e.g., item format) and an item-specific deviation - which is a random item variation - can thus explain the item parameter. One of the LLTM's advantages is that we can tease out the effects of item properties within this framework. Because item properties often arise from assessment design factors, this model is also useful for evaluating those factors.

In the LLTM-R model, the item parameter (β_i) is expressed by the linear combination of the item format variable (X_i) and its effect (γ), and an error term (ϵ_i). Note that this model estimates the effect of item property (e.g., item format) and the error term rather than the individual item parameter; and thus the individual item parameter can be constructed using the value of the item property, its estimated effect, and the error term as follows:

$$\eta_{pi} = \theta_p - \beta_i \tag{2}$$

$$\beta_i = \gamma X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

η_{pi} indicates the log odds of the response of person p to item i,

θ_p indicates a person ability parameter for person p,

β_i indicates an item parameter for item i,

γ indicates the effect of item format,

X_i indicates the value of item format for item i and

ϵ_i indicates the error term with a normal distribution, $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.

The following figure shows the graphical representation of the LLTM-R.

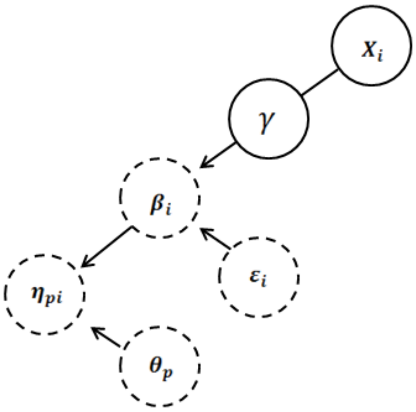

Figure 1. Graphical representation of the LLTM-R

In this representation, the solid and dashed lines indicate the fixed and random effects, respectively.

4. The Hierarchical item response model

According to item format, PISA 2006 science items can be partitioned into three groups (i.e., open response item group; OC, multiple choice item group; MC and complex multiple choice item group; CMC). Since individual items are exclusively nested within these three groups, the hierarchical structure can be seen with respect to the item parameters. The hierarchical IRM is designed to deal with this format of hierarchical structure by incorporating random effects into the item side and specifying within-group item variance. Since this two-level model on the item side gives between-group variance estimates which indicate group effects, it is also useful to evaluate the effect of the items' formats. The two-level hierarchical item response model specifying item groups along with individual items can be formulated as follows:

$$\eta_{pik} = \theta_p - \beta_{ik} \tag{3}$$

Level-1 model

$$\beta_{ik} = \beta_{0k} + \epsilon_{ik},\ \epsilon_{ik} \sim N(0, \sigma_\epsilon^2) \tag{4}$$

Level-2 model

$$\beta_{ok} = \beta_{00} + \tau_{0k},\ \tau_{0k} \sim N(0, \sigma_\tau^2) \tag{5}$$

In the level-1 model, the item parameters nested group (i.e., item foramts) k is explained by the group mean of the item parameter and the item deviation within group. In the level-2 model, the group mean of the item parameter is specified as a linear combination of the grand mean of item difficulty and the group deviation from the grand mean. Figure 2 visualizes this two-level model.

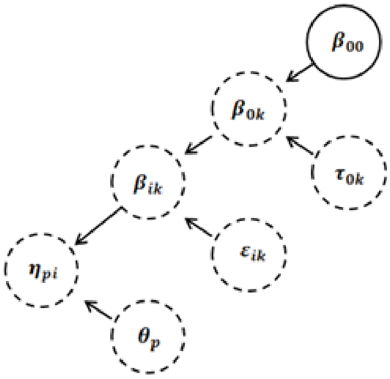

Figure 2. Graphical representation of the 2-level model for the items

In this graphical representation, β_{ik} indicates an item parameter for the item i nested group k, β_{0k} indicates the average item parameter for group k, β_{00} indicates the grand mean of the item parameter, ϵ_{ik} indicates the item deviation within group k, and τ_{0k} indicates the group deviation from the grand mean item difficulty. The dashed lines, thus, show that they are random and the solid line indicates it is fixed effect.

5. Example: PISA 2006 Science Assessment

5.1. PISA 2006 Science Assessment

The Program for International Student Assessment (PISA) is an internationally standardized assessment that was jointly developed by participating countries and was administered to fifteen-year-olds in schools to identify key demographic, social, economic, and educational factors that affect student performance in reading, mathematics, and science. In 2006, fifty-seven counties participated in PISA. Each PISA data collection effort assesses one of the three subject areas in depth, and science literacy was the subject area assessed in depth in PISA 2006 (OECD, 2006a). The PISA 2006 assessment consisted of 140 science items which were partitioned into 7 test clusters. These seven clusters, along with four mathematics and two reading test clusters, were allocated to thirteen test booklets; consequently, each booklet was made up of four test clusters.

The science section of PISA 2006 included multiple choice (MC), complex multiple choice (CMC), and open response (OR) items. The MC items required the selection of a single response from four options, and the CMC items required students to respond to a series of related “Yes/No” questions. The OR items required a relatively extended written or drawn response from a student (OECD, 2006a). Approximately one third of the science items were MC, one third were CMC and one third were OR items. Currently, twenty-three cognitive item texts of the PISA 2006 science assessment are available, so responses to these items will be used in order to answer the research questions. Among the twenty-three items, ten items are MC, five items are CMC, and eight items are in the OR format. Individual items adopt different item formats: open response, multiple choice, and complex multiple choice formats; thus, items can be partitioned into three groups as follows:

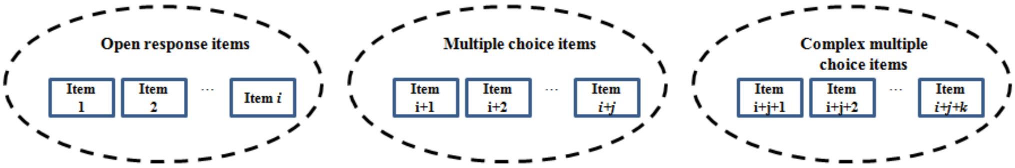

Figure 3. Item groups by item formats

Table 1 is a summary of the item parameter estimates based on Classical Test Theory (CTT). As can be seen in this table, the CTT analysis identifies differences in item difficulty among the different item formats; and, in fact, the item difficulties of the OR items are a little bit more difficult than the MC or CMC items. The item difficulties of the OR items are between 0.19 and 0.66, while it appears that those of the MC and CMC are between 0.34 and 0.87.

Table 1. Classical Test Theory Analysis results

Item format	OR				MC					CMC				
Item number	1	2	8	12	4	5	6	9	14	3	7	10	11	13
Difficulty	0.66	0.19	0.67	0.44	0.85	0.75	0.87	0.85	0.64	0.49	0.75	0.34	0.73	0.81

This result simply shows the tendency of item difficulty from the perspective of CTT, but does not quantify any relationship between item difficulty and item format. In order to investigate how the three item formats affect item difficulties, the current study adopts the random item response model approaches.

5.2. Sample

In this study, Korean samples from the PISA 2006 study (5,611 American students and 5,176 Korean students) were used to examine item format effects. Since the students took only one of thirteen test booklets consisting of four test clusters, they did not respond to all the science items. For the Korean sample, each test booklet was given to about four hundred students, and consequently sixteen hundred students responded to each item. Since only a certain number of item texts are open, the number of items that this study investigates is limited; students who responded to these items are included in the data analyses. The analysis of this study includes 395 students responding to 14 items. Among the fourteen items, four items are in the open response item format, five are multiple choice items, and five are complex multiple choice items.

5.3. Results

In this study, we used the LLTM, the LLTM-R, and the hierarchical IRM. Because the LLTM-R and the hierarchical IRMs allow for random effects for both items' and persons' parameters, it turns out they are the crossed random-effect models; and thus we used WinBUGS 1.4.3 (Spiegelhalter, Thomas, Best, & Lunn, 2003) adopting Bayesian estimation to fit these models to the PISA 2006 science data. The WinBUGS was run with 3 chains for

10000 iterations, and the burn-in period was 3000.

5.3.1 LLTM and LLTM-R

In order to test the item format effects, the current study applied both the LLTM and the LLTM-R models. Since the LLTM and the LLTM-R models are relatively new approaches to investigating item format effects, it is necessary to examine whether item parameter estimates from these two models are comparable with those from other models. Since item parameter estimates are expressed as a linear combination of item format effects (γ) and item format values (X_i), the quality of the estimates for the item format effects depends highly on the quality of the item difficulty estimates. With empirical data, this study examined the quality of the item difficulty estimates by testing the consistency of the estimates across the models. Because both the LLTM and the LLTM-R models are in the Rasch family model, the current study compared item parameter estimates from these two models with those from the Rasch model for this purpose. Table 2 shows item difficulty estimates by item format across the three models, and Figure 4 visualizes these item difficulty estimates by item format.

Table 2. Item difficulty estimates across the three models

Item Format	Item Number	Rasch	LLTM	LLTM-R
OR	1	-0.003	0.026	0.127
	2	2.460	2.525	2.620
	8	-0.058	-0.028	0.069
	12	1.051	1.092	1.192
MC	4	-1.244	-1.261	-1.169
	5	-0.553	-0.555	-0.460
	6	-1.430	-1.447	-1.355
	9	-1.247	-1.262	-1.163
	14	0.090	0.098	0.195
CMC	3	0.816	0.846	0.944
	7	-0.566	-0.561	-0.464
	10	1.573	1.614	1.712
	11	-0.428	-0.414	-0.317
	13	-0.964	-0.956	-0.860

This table clearly shows that the three models provide very similar parameter estimates; and, in fact, there are no substantial differences among the models in terms of item parameter estimates. This result confirms the idea that item parameter estimates from the LLTM and the LLTM-R are stable as in the Rasch model. Figure 4 displays this result. This figure also reveals that, depending on item format, the patterns of item difficulty are somewhat different. OR items tend to be more difficult than MC items, whereas CMC items do not show any clear tendency compared to the other item formats.

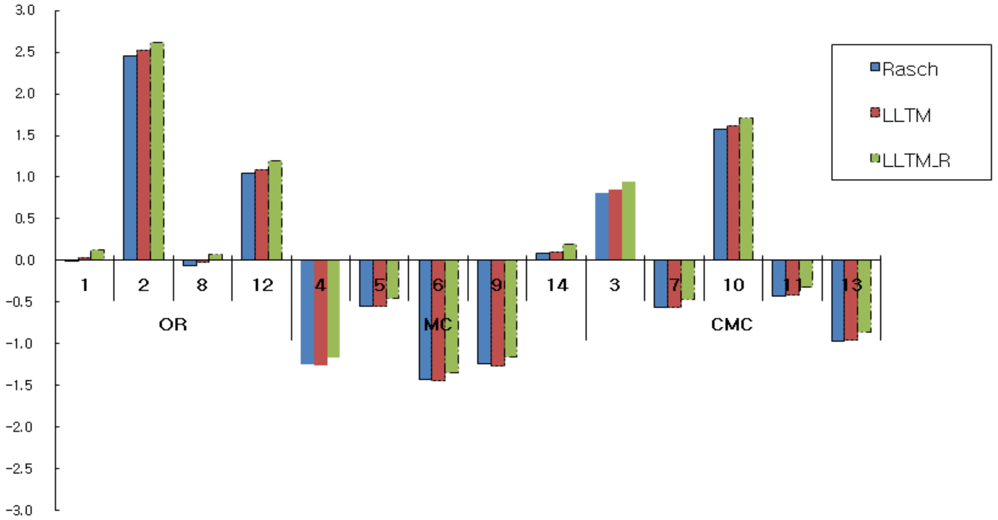

Figure 4. A comparison of item difficulty by item format

This pattern of item difficulty estimates can be consistently observed in the three models. As can be seen in Figure 5, regardless of the model, the average item difficulty is very similar as long as the item format is the same. Average OR item difficulties, for example, are very much same across the models, but they are significantly different from average MC item difficulties.

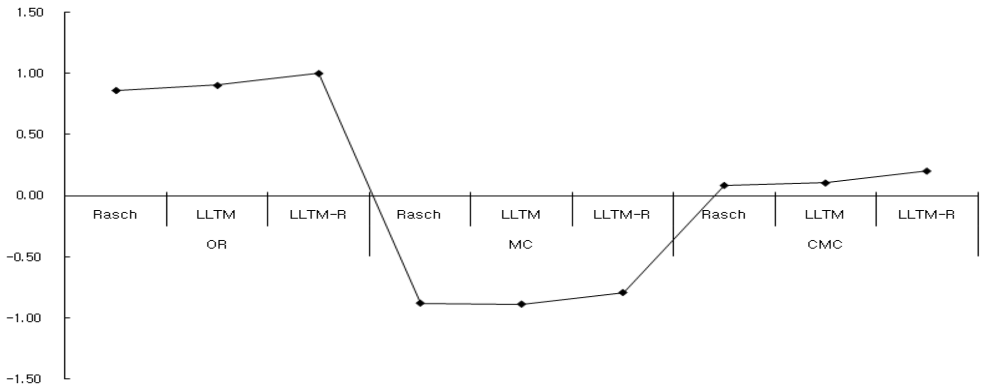

Figure 5. The average item difficulty across the models

This result reveals that there may be a substantial/significant relationship between item difficulty and format. Since the LLTM-R model specifies the item parameter as a linear combination of item properties (X_i) (which are the item formats in this example), their effects on item difficulty (γ), and random errors (ϵ_i), this relationship can be identified by looking at the coefficient γ. In the LLTM-R model, item formats are dummy coded, and the CMC is the reference item format. Table 3 shows the effect of each item format. Since item formats are dummy coded, the value of each cell indicates the difference between the CMC items and others in terms of item format effect on item difficulty. As can be seen in this table, the effects of the OR item format are 0.844 and 0.910 for the LLTM and the LLTM-R, respectively. For the effects of the MC item format, the LLTM and the LLTM-R provide γ coefficients -0.830 and -0.764, respectively. These results indicate that students feel items to be more difficult in the OR item format compared to those in the CMC format. Based the these results, we can see that the differences in item format may cause significant differences in item difficulty; and, in fact, these differences can be 0.844 or 0.910 logits depending upon the model that one applies. In contrast, it appears that students feel more relaxed with MC items compared to CMC items. The change of item format from MC to CMC may cause 0.830 or 0.764 logits in item difficulty. The current results show that these two item formats (i.e., OR and CMC) affect item difficulty parameters in totally different ways in PISA assessments.

Table 3. The effect of item format on item difficulty

	LLTM	LLTM-R
OR	0.844 (0.060)	0.910 (0.108)
MC	-0.830 (0.055)	-0.764 (0.074)

Since the LLTM-R model considers residuals which the LLTM model does not specify, somewhat different estimates from these two models are already expected, but it turns out that this different is not significant as shown in Table 3.

5.3.2 The hierarchical IRM

In the hierarchical IRM analysis, the current study first compared the item parameter estimates across three models (the hierarchical IRM, the LLTM, and the LLTM-R) to show the accuracy of the estimator in the hierarchical IRM. Figure 6 displays the comparison of item parameter estimates among the hierarchical IRM, the LLTM, and the LLTM-R. As shown in this figure, the parameter estimates from the hierarchical IRM are significantly equivalent to those from the LLTM and the LLTM-R.

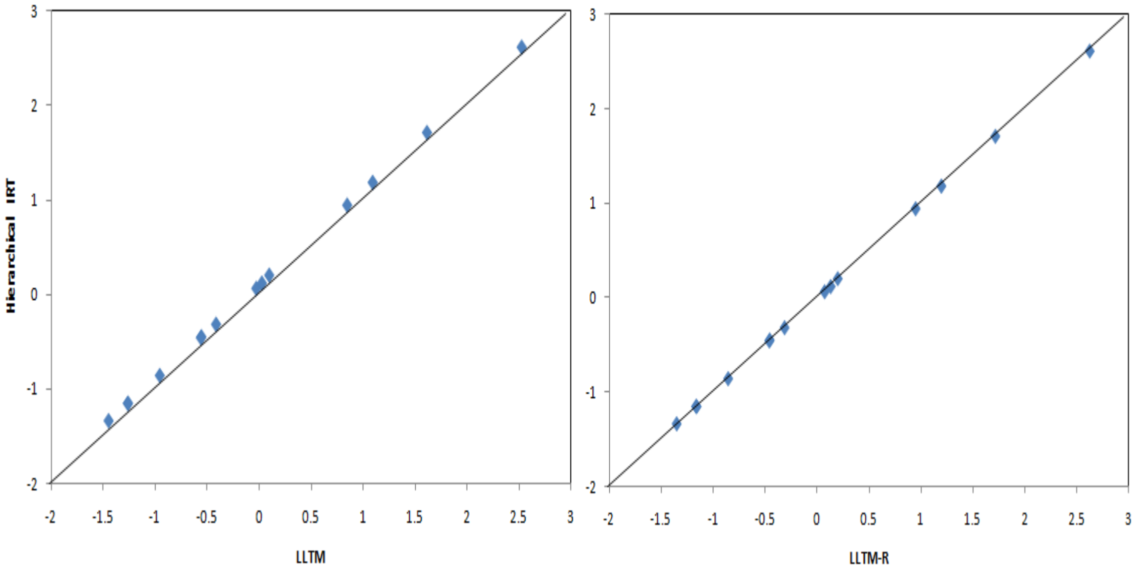

Figure 6. The comparison of item parameter estimates

As addressed above, the hierarchical IRM partitions the item parameter deviation into the within-group deviation (ϵ_{ik}) and the between-group deviation (τ_{0k}). The within-group deviation is the item parameter deviation from the group mean of the item parameters. On the other hand, the between-group deviation is the item group (i.e., item format) deviation from the grand mean of the item parameters, and this deviation can be understood as the group effect on item parameters. In the current study, this between-group deviation thus indicates the item format effect. Table 4 summarizes the item parameter estimates and deviations from the hierarchical IRM analysis.

Table 4. The hierarchical IRM analysis results

Item format	Item difficulty			Item variance	
	Mean	Min	Max	Within-group	Between-group
OR	0.994	0.061	2.616	1.376	1.115
MC	-0.781	-1.342	0.201		
CMC	0.202	-0.863	1.712		
Total	0.077	-1.342	2.616		

This table shows that the within-format variance is 1.376 and the between-format variance is 1.115, respectively. Although the between-format variance is smaller than the within-format variance, this result indicates that the item format substantially affects the item parameters; and, considering the person parameter variance (1.156), in fact, this item format effect is quite large. One can expect, therefore, that student performance in the science assessment could be totally different simply by changing the item format even though the items require the same content in their answers.

6. Discussion and conclusion

The PISA 2006 science assessment is composed of open response (OR), multiple choice (MC), and complex multiple choice (CMC) items. One may expect that the item formats cause differences in item difficulties although these items deal with the same content (Bolger & Kellaghan, 1990; DeMars, 1998, 2000; Garner & Engelhard, 1999; Hellekant, 1994). If so, at the stage of assessment design, the item format should be considered. For the investigation on the item format effects, the current study introduced IRMs with random item effects and applied these models to the PISA 2006 science assessment data for the illustration.

As for IRMs with random item effects, the current study introduced the LLTM-R and the hierarchical IRM, and examined the equivalence of parameter estimates from these models in the PISA 2006 science assessment and compared them with the Rasch and the LLTM models. The empirical study with the PISA 2006 science assesment indicates that the item difficulties are substantially different depending on the item formats; especially, the result from the LLTM-R indicates that the item formats affect the item difficulties in totally different ways.

Since The PISA 2006 science assessment is used for making inter-country comparisons and inferences, the investigation of item difficulty is important. For the PISA assessments, their underlying assumption is that items are fair for different students from different countries. Securing fairness, however, is quite challenging for a number of reasons (e.g., sampling, language, etc.) when conducting international assessments. From a psychometric perspective, this fairness has been tested through Differential Item Functioning (i.e., DIF) analysis to ensure functional equivalence across countries. If DIF analysis shows that some items are more difficult for a certain group of students than for other groups, these items could be said to be unfair. Accordingly, these DIF items can cause a biased outcome for the students, which is a critical issue in international comparison.

At this point, this study contributes to the development of fair assessments because it provides information about the item format effect on item difficulty. If the item format effect causes a significantly different degree of item difficulty, and this item format effect is different among countries, one may be able to eliminate the unfairness caused by the item formats by changing the item format when developing the assessment. Depending on the curricula, instructional methods, and cultures that students experience, a specific item format might work in favor of a specific group of students, indicating DIF. In order to ensure the assessment's quality and provide a useful guideline for designing and revising PISA assessment items, a comparison of the item format effects across countries should be done in the further study.

References

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27,* 165-174. https://doi.org/10.1111/j.1745-3984.1990.tb00740.x

Butter, R., De Boeck, P., & Verhelst, N. D. (2004). An item response model with internal restrictions on item difficulty. *Psychometrika, 63,* 1-17.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73,* 533-559. https://doi.org/10.1007/s11336-008-9092-x

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11,* 279-299. https://doi.org/10.1207/s15324818ame1103_4

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55-77. https://doi.org/10.1207/s15324818ame1301_3

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometirka, 48,* 3-26. https://doi.org/10.1007/BF02314674

Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple choice and constructed response mathematics items. *Applied Measurement in Education, 12,* 29-51. https://doi.org/10.1207/s15324818ame1201_3

Hellekant, J. (1994). Are multiple-choice test unfair to girls? *System, 22,* 349-352. https://doi.org/10.1016/0346-251X(94)90020-5

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp.189–212). New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9_6

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285-306. https://doi.org/10.3102/10769986025003285

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697. https://doi.org/10.1007/BF02296403

Lee, Y, & Wilson, M. (2009). An extension of the MIRID model for polytomous responses and random effects. Paper presented at the annual meeting of American educational research association, San Diego.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12,* 725-737. https://doi.org/10.1177/014662168801200306

Organisation for Economic Co-operation and Development (OECD). (2006a). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2006b). *PISA2006 technical report.* Paris: OECD.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and*

Data Analysis Methods. Thousand Oaks: Sage.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205. https://doi.org/10.1037/1082-989X.8.2.185

Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

Copyright Disclaimer