Psychology Science Quarterly, Volume 50, 2008 (3), pp. 403-416

Investigating DIF and extensions using an LLTM approach and also an individual differences approach: an international testing context[1]

YIYU XIE[2] & MARK WILSON

Abstract

This study intends to investigate two ways to generalise *differential item functioning* (DIF) by grouping of items that share a common feature, or an item property as in the Linear Logistic Test Model (LLTM). An item "facet" refers to this type of grouping, and DIF can be expressed in terms of more fundamental parameters that relate to the facet of items. Hence the *differential facet functioning* (DFF) model, a particular version of the LLTM, helps to explain the DIF effects more substantively. Using the mathematics data from the Program for International Student Assessment (PISA) 2003, this study shows that modeling the DFF effect through an interaction of the group-by-facet parameter rather than DIF effect on the individual item level can be handled easily with the NLMIXED procedure of SAS. We found that the results are more interpretable when the bias is interpreted on the facet level rather than the item level. Analogous to the multidimensional DIF model, one natural extension of the DFF model is to make the model multidimensional when DFF facets (i.e., LLTM facets) are considered as dimensions. This extension, multidimensional DFF (MDFF), is also investigated. The MDFF model allows individual differences to be modeled on the dimension that exhibits a DFF effect. However, it is always recommended to check the individual DIF estimates and construct a substantive analysis first before conducting DFF and MDFF analysis.

Key words: DIF, LLTM, differential facet functioning (DFF), multidimensional model, PISA

[1] Note, this work was carried out while the first author was a graduate student at UC Berkeley.

[2] Yiyu Xie; email: yxie05@gmail.com

As international studies of educational achievement have gained increasing popularity, test fairness arises as a vital issue in such studies. Researchers are generally concerned with the question of whether an item is fair for members of certain focal groups compared to members of a reference group. In psychometrics, a statistical analysis to test the fairness on the item level is called a *differential item functioning* (DIF) analysis. An item is said to be fair, or unbiased, if it is equally difficult for persons of the focal and the reference groups who are matched with respect to the underlying dimension that the test is intended to measure. DIF occurs when persons at the same point on the underlying dimension respond differently to an item given his/her group membership, such as gender, ethnicity, etc.

Numerous studies have been focused on the procedures for detecting DIF (for an overview, see Millsap & Everson, 1993). The Mantel-Haenszel procedure modified by Holland and Thayer (1988) is a theoretical milestone in psychometrics. It is a classical approach to detect DIF. Dorans and Kulick (1986) developed a standardized p-difference index, which is also broadly used. There are other approaches in item response modeling, including using loglinear item response models (Mellenbergh, 1982; Kelderman, 1989), logistic regression models (Swaminathan & Rogers, 1990), area measures (Raju, 1988), Wald statistics (Lord, 1980) and likelihood-ratio tests (Thissen, Steinberg & Wainer, 1988). Yet all these studies remain largely technical. They are just various approaches to show how DIF in individual items affects the distribution of the test scores in different groups. William Stout pointed out in his presidential address given at the 67th annual meeting of the Psychometric Society held in Chapel Hill, North Carolina (2002), “one of the subtle ways that DIF has been compartmentalized is the almost total disconnect that has evolved between substantive (content-based) and DIF (statistical) approaches to the understanding and practice of test fairness”. In many situations, no explanation can be given why some substantively sound items show large DIF values statistically whereas some other items expected to be biased from the substantive analysis do not display DIF at all. In an attempt to produce substantively interpretable DIF results, Shealy and Stout (1993a, b) developed a multidimensional method, called SIBTEST, to model DIF. Analogous to DIF which measures the bias at the individual item score level, *differential bundle functioning* (DBF), and *differential test functioning* (DTF), are defined in the multidimensional model to measure the bias at the item bundle score level and the test score level. Wainer, Sireci, and Thissen (1991) also illustrated how to model DIF at the testlet score level.

This study investigates a more general way to model DIF as well, namely *differential facet functioning* (DFF). The term DFF was first introduced by Engelhard (1992). He suggested that “studies of differential facet functioning (DFF) can be conducted by a variety of procedures that are conceptually similar to current approaches for studying differential item functioning" (p. 175). An item facet refers to a group of items that share a common feature, or an item property. Hence DIF can be expressed in terms of more fundamental parameters that relate to the properties of items. Meulders and Xie (2004) presented an approach to explore DFF using a general software package that allows the flexibility of building and estimating a variety of models. They demonstrated on a verbal aggression data set that DFF can be viewed as a parsimonious way to summarize DIF. They also showed one extension of the DFF model that includes random interaction parameters over persons. Denoted as *random-weight differential facet functioning* (RW-DFF), it allows the model to capture the heterogeneity of the interactions between person and item via an interaction parameter.

Most educational surveys report students' performance at the national level for international comparisons in their publicly released documents. The results are usually arranged to show national means on each subject or content area. Under the frameworks of these assessments, items are designed using several domains. For instance, the Trends in International Mathematics and Science Study (TIMSS) 2003 assessment is framed by two domains: content and cognitive. The Program for International Student Assessment (PISA) 2003 framework for mathematics defines three domains, content, process and situation. Each item belongs to one category in each domain. Thus, it may be more appropriate to apply the DFF approach rather than the DIF approach to such data; i.e., the results might be more meaningful if the facets explain, at least in part, the DIF effects.

Model

For dichotomously scored responses, the Rasch model (1960) defines the probability of a response in item *i* for person *n* as

$$P(X_{ni} = 1 \mid \theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \tag{1}$$

where θ and δ are the person proficiency and item difficulty parameters, respectively. It is also common in psychometrics to express Equation 1 in logit form:

$$\log it = \theta_n - \delta_i. \tag{2}$$

The Rasch model explains the variation in the response data through individual person and item parameters.

Fischer (1973) developed a *linear logistic test model* (LLTM) to estimate the effects of item properties instead of individual items. This model is useful in testing hypotheses on the cognitive operations involved in the process of solving the items. The logit expression of the LLTM is:

$$\log it = \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik}, \tag{3}$$

where η_k is the difficulty parameter for item property *k* and Q_{ik} is the indicator weight of item *i* on item property *k*. In the application that follows, Q_{ik} takes the value of 1 if item *i* belongs to item property *k*, and 0 otherwise, but other values are also possible. For identification purpose, the mean of the person parameters is constrained to be 0, so an item intercept is needed for Q_{ik} = 1 for all items when *k* = 0. When the mean of the person parameters is free to be estimated, *k* would take values from 1 to *K* in the above equation. Note $\boldsymbol{Q}$ is a $I \times K$ matrix, "a priori" determined from theory where *I* and *K* are total numbers of items and item properties or facets, respectively.

When DIF occurs, the logit expression in the Rasch model becomes

$$\log it = \theta_n - \delta_i + Z_n \gamma_i , \tag{4}$$

where Z_n is an indicator variable of person n's group membership, and γ_i is the DIF parameter for item i. The additional term in Equation 4 can be viewed as an interaction term between item i and person group membership Z. The interpretation of the DIF parameters depends on the coding scheme of the variable Z. If dummy coding is used, Equation 4 for the reference group is the same as Equation 2, and the logit expression of the focal group is

$$\log it = \theta_n - \delta_i + \gamma_i . \tag{5}$$

If contrast coding is used where the reference group takes the value of -1 and the focal group takes the value of 1, the logit expressions of the two groups become

$$\begin{aligned} \log it &= \theta_n - \delta_i - \gamma_i \text{, and} \\ \log it &= \theta_n - \delta_i + \gamma_i , \end{aligned} \tag{6}$$

respectively.

By analogy with the LLTM treatment, a DFF term can be added to Equation 3 as well:

$$\log it = \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik} + Z_n \sum_{k=1}^{K} \gamma_k Q_{ik} . \tag{7}$$

Here, γ_k is the DFF parameter for item property k and k runs from 1 to K for the DFF term. If contrast coding is used for the indicator variable Z, Equation 7 for the reference and focal groups turns into

$$\begin{aligned} \log it &= \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik} - \sum_{k=1}^{K} \gamma_k Q_{ik} \text{, and} \\ \log it &= \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik} + \sum_{k=1}^{K} \gamma_k Q_{ik} , \end{aligned} \tag{8}$$

for the two groups, respectively.

Data

The data under investigation are from the Program for International Student Assessment (PISA) 2003 international database. The PISA 2003 study has a focus on Mathematics and developed 85 mathematics items along three domains: *content*, *process* and *situation*. There are four categories of mathematical content identified by the Organization for Economic Cooperation and Development (OECD): *space and shape*, *change and relationships*, *quan-*

tity, and *uncertainty*. The process domain defines three cognitive demands imposed by different mathematical problems. They are *reproduction*, *connections*, and *reflection*. The PISA study also classifies the situations represented by the stimulus material for each item. There are four sorts of situations: *personal*, *educational or occupational*, *public*, and *scientific* situations.[3] Table 1 shows the breakdown by each domain of the 85 items.

Data from countries that have comparable mean performance on the overall mathematics scale were chosen to investigate DFF. This is to minimize any overall country effect that may contribute to the difference in performance. Two groups of countries were selected, Canada and Japan, and the United States and Spain. The mean of the mathematics scale of the international comparison is set to 500 with a standard deviation of 100. Table 2 lists the mean scores and standard errors of the four selected countries.

The publicly released data has 84 mathematics items for these countries. For Canada, one item is not available, so there are 83 items in total. Among all items, 13 items were scored polytomously. They were recoded to just 0 and 1 for the current study.[4] The recoding does not substantially affect the purpose of the study much as the investigation is aimed at the facet rather than individual item level, but the estimation process is much easier to carry out. A random sample of 500 students was drawn for each country.

Table 1:
Distribution of items by the domains

Content	Number of items	Process	Number of items	Situation	Number of items
Space and shape	20	Reproduction	26	Personal	18
Change and relationship	22	Connections	40	Educational or occupational	20
Quantity	23	Reflection	19	Public	29
Uncertainty	20			Scientific	18

Table 2:
Mean scores and standard errors of selected countries

Country	Mean (S. E.)	Country	Mean (S. E.)
Canada	532 (1.8)	U. S.	481 (2.9)
Japan	534 (4.0)	Spain	485 (2.4)

[3] For more details about each domain, see *The PISA 2003 Assessment Framework: Mathematics, Reading, Science, and Problem Solving Knowledge and Skills* (OECD, 2003e).

[4] All response categories greater than 1 were recoded to 1. Alternate recodings (such as setting the scores 0 and 1 to be 0, and those above to 1) would also be interesting, but are beyond the scope of this paper.

Results

The two groups of data were analyzed separately using the *SAS* computer program, in particular, its NLMIXED procedure (SAS Institute, 2004). At first, the Rasch model and exploratory DIF models were fit to each group. For an exploratory DIF model the DIF parameters for all items are estimated simultaneously. A main effect term, the country effect, was also added to the model to account for group mean difference in the performance. Note this country effect can also be modeled through the latent regression model of θ with country as the regressor. The results for these two approaches are essentially the same. In both comparisons, contrast coding is used. Canada and the United States are set as reference groups. A design matrix of I item indicator variables is necessary for the Rasch and DIF analyses. In this case, it is an 84×84 matrix with 1s on the diagonal line and 0s for the remainder. In order for the exploratory DIF models to be identified, only I-1 DIF parameters are included in the model. Thus, there are 82 DIF parameters for the Canada-Japan comparison with the last item-by-country interaction omitted, and 83 DIF parameters for the U.S.-Spain comparison with the last item-by-country interaction omitted. The model fit statistics are reported in Table 3 and Table 4.

Table 3 shows that the AIC and BIC values are smaller for the DIF model compared to the Rasch model for the Canada-Japan data. Comparison of the deviance using the likelihood-ratio (LR) test also indicates that the DIF model has a better fit (χ^2=589, df=83, p<0.001). The estimated country effect is -0.338 (SE=0.145), indicating that Japanese students perform better than Canadian students overall. This effect is significant at α=0.05 but not at α=0.01. Table 4 shows that the AIC indicates the DIF model fits the U.S.-Spain data

Table 3:
Model fit statistics for the Canada-Japan comparison

Model	No. of Parameters[5]	Deviance -2*log-likelihood	AIC	BIC
Rasch	85	24430	24600	25017
DIF	168	23841	24177	25001
LLTM	10	27008	27028	27077
DFF	19	26850	26888	26981

Table 4:
Model fit statistics for the U.S.-Spain comparison

Model	No. of Parameters	Deviance -2*log-likelihood	AIC	BIC
Rasch	85	24638	24808	25226
DIF	169	24258	24596	25425
LLTM	10	27352	27372	27421
DFF	19	27294	27332	27425

[5] A parameter for the variance of the person distribution is also estimated in each model.

better whereas the BIC value is smaller for the Rasch model. According to the LR test, the DIF model shows significant improvement in the deviance ($\chi^2=380$, df=84, p<0.001). The estimated country effect is -0.028 (SE=0.149), indicating that Spain has higher performance but this is not statistically significantly higher than the United States. Closer examination of the individual DIF estimates reveals that there are 18 and 14 DIF parameters statistically significantly different from 0 at $\alpha=0.05$ level for the two groups respectively. It is not an easy task to derive sound explanations of why these items exhibit DIF when looking at individual DIF estimates. Alternatively, we can use the results to check if any DFF-like patterns existed for these DIF effects. Table 5 and Table 6 show the breakdown by the domain and the direction of estimated values for the two contexts.

The tables show that for some categories of a domain, the DIF effects are consistently in favor of one country. According to Equation 6, a positive estimate for the DIF parameter means the probability for the reference group to get the item correct is lower. For example, among the 14 DIF items found from the U.S.-Spain comparison, 4 belong to the content of *quantity*, and they are all in favor of Spanish students. Thus, in the next step, the LLTM and DFF models were fit to the two data groups to see if they can help model the patterns.

In the LLTM and DFF analyses, the design matrix is the $\boldsymbol{Q}$ matrix in Equation 3 and Equation 7. Each domain is considered as one facet. As there are 3 facets with multiple categories in each facet in the PISA data, a coding scheme similar to the one for categorical variables in linear regression analysis is used. For each facet with *m* categories, *m*-1 indicator variables are needed. The interpretation of each indicator variable is always in reference to

Table 5:
Distribution of DIF items for the Canada-Japan comparison

Content	+	-	Process	+	-	Situation	+	-
Space and shape	2		Reproduction	3	2	Personal	1	2
Change and relationship	1	2	Connections	2	7	Educational or occupational	3	1
Quantity	2	3	Reflection	1	3	Public	2	5
Uncertainty	1	7				Scientific		4

Table 6:
Distribution of DIF items for the U.S.-Spain comparison

Content	+	-	Process	+	-	Situation	+	-
Space and shape	2		Reproduction	5	1	Personal	3	
Change and relationship	1	2	Connections	5	1	Educational or occupational	1	1
Quantity	4		Reflection	1	1	Public	6	1
Uncertainty	4	1				Scientific	1	1

the base category. In theory, any category can be chosen as a base category. The aim is to find a relationship among the categories so as to make the interpretation easier. For the facet of *content*, *space and shape*, *change and relationship*, *quantity* and *uncertainty* are closely related to geometry, algebra, arithmetic and statistics and probability, respectively, which are four common curricular branches of mathematics. For the facet of *process*, the categories were built to have an association with each other. The *reproduction* process is needed in those items that require the reproduction of practiced knowledge. The *connection* process builds on reproduction to solve problems that still involve or develop beyond the familiar settings. The *reflection* process builds further on the connection process. It requires some insight and creativity from the student. For the facet *situation*, the categories from *personal* to *scientific* show an increasing distance between the student and the situation. The *personal* situations are those that have the most direct impact on the students whereas the *scientific* situations are the most abstract ones. One factor is also called into play in the decision of choosing the base category. Given the information from Table 5 and Table 6, it is better to choose a base category that we would expect to exhibit only a small DFF effect. Thus, different base categories were set for the two data sets. For the Canada-Japan comparison, *quantity*, *reproduction* process and *personal* situations were set as the reference categories for the three facets. For the U.S.-Spain comparison, *change and relationship*, *reflection* process and *scientific* situations were set as the base categories. Thus, combined with an intercept variable, there are a total of 9 columns in the $\boldsymbol{Q}$ matrix.

Table 3 and Table 4 also list the fit statistics of the LLTM and DFF models for the two data groups. Compared with the model fit results from the Rasch model, the LLTM produces larger deviance, AIC and BIC values. The goodness of fit of the LLTM is always lower than the Rasch model, this is because the number of predictors that account for item effects are reduced from 84 to just 9, and the LLTM does not use an error term for fitting the item parameter (Fischer, 1973). Nevertheless, the estimates from the LLTM are worth checking to see how they conform to the intention of the test constructers. Table 7 and Table 8 present the parameter estimates from the LLTM and DFF models for the two groups. Once again, a country effect was added to the DFF model.

According to Equation 3, a positive estimate in the LLTM means the probability of getting the correct answer on the item belonging to the specified category of a facet is lower compared to that of the reference category. Even though the base categories of the facets are different for the two data, the estimates tell essentially the same story. For the content areas, both *space and shape*, and *uncertainty* appear to be more difficult than *quantity*, whereas *change and relationship* is easier. For the Canada-Japan group, *uncertainty* is the most difficult content among the four, while for the U.S.-Spain group, *space and shape* is the most difficult one. As expected, both the *connection* process and the *reflection* process are harder than the *reproduction* process with the *reflection* process being the most difficult. For the *situations* facet, the estimates for the Canada-Japan group show an increasing difficulty from the *personal* to the *scientific* situations, though the *educational/occupational* situation is not statistically different from the *personal* situation. For the U.S.-Spain group, the estimates for the three situations, *educational/occupational*, *public* and *scientific*, are quite close together. Therefore, the classification of the situations does not add to our understanding for the U.S.-Spain data.

Table 7:
Parameter estimates for the Canada-Japan comparison

	LLTM		DFF	
	η	SE(η)	γ	SE(γ)
Intercept	-1.584*	0.057		
Content (Quantity)				
Space and shape	0.259*	0.045	0.157*	0.045
Change and relationship	-0.383*	0.047	0.124*	0.047
Uncertainty	0.374*	0.044	-0.162*	0.044
Process (Reproduction)				
Connection	0.874*	0.038	-0.053	0.038
Reflection	1.548*	0.046	0.135*	0.046
Situation (Personal)				
Educational/Occupational	0.011	0.049	0.223*	0.049
Public	0.194*	0.044	0.063	0.044
Scientific	0.590*	0.054	-0.148*	0.054

* significant at α=0.05 level

Table 8:
Parameter estimates for the U.S.-Spain comparison

	LLTM		DFF	
	η	SE(η)	γ	SE(γ)
Intercept	0.523*	0.056		
Content (Change and relationship)				
Space and shape	0.637*	0.050	0.123*	0.050
Quantity	0.232*	0.046	0.125*	0.046
Uncertainty	0.485*	0.048	0.023	0.048
Process (Reflection)				
Reproduction	-1.670*	0.046	0.130*	0.046
Connection	-0.875*	0.042	0.037	0.042
Situation (Scientific)				
Personal	-0.370*	0.053	-0.105	0.053
Educational/Occupational	-0.040	0.054	-0.173*	0.054
Public	-0.015	0.048	-0.015	0.048

* significant at α=0.05 level

Table 3 shows that the DFF model has better fit than the LLTM for the Canada-Japan group as the DFF model has lower AIC and BIC values, and a statistically significant improvement in the deviance (χ^2=158, df=9, p<0.001). Table 4 shows that the LR test (χ^2=58, df=9, p<0.001) and AIC indicate that the DFF model fits the U.S.-Spain data better whereas the BIC indicates that the DFF is no better than the LLTM. The estimated country effects for the two data sets are -0.129 (SE=0.056) and -0.107 (SE=0.055), respectively. There is no significant country effect at the α=0.01 level for the two data sets. All these results are consistent with those from the comparisons between the Rasch and DIF models.

The DFF estimates presented in Table 7 and Table 8 point out that some DFF terms are statistically significantly different from 0 for each data set. When the results in these Tables are compared with the results from Table 5 and Table 6, the significant DFF parameters reflect the patterns of the significant DIF parameters to some extent. For reporting purposes, it is more meaningful to conclude that Japanese students have higher probabilities to get correct answers on the items that belong to the *space and shape*, *change and relationship* content areas, or those using the *reflection* process, or those that have *educational/occupational* situations in the stimulus materials. On the other hand, items that belong to the content area of *uncertainty*, or those involving the *scientific* situations are in favor of Canadians. Compared to American students, the Spanish have higher probabilities to get correct answers on the items that belong to the *space and shape*, and *quantity* content areas, or those using the *reproduction* process. Items that have the educational/occupational situations in the stimulus materials are in favor of American students. Since contrast coding was used, the effect size of the DFF parameter equals $exp(2\gamma)$, according to Equation 8. The interpretation of the effect size is that, for example, the odds ratio of getting the items requiring the *reflection* process correct for Japanese versus Canadian students after correcting for differences in mean performance between the two countries is 1.3.

The DFF effect is more prominent in the Canada-Japan data, since the DFF estimates and consequently the effect sizes are relatively larger than those in the U.S.-Spain data. In addition, all fit statistics, -2*log-likelihood, AIC and BIC, show that DIF or DFF need to be modeled for the Canada-Japan data. Therefore, further investigation was carried out for this data group only.

Multidimensional DFF

As the PISA mathematics data has three domains and the DFF effects are found in each domain, one natural extension of the DFF model is to make the model multidimensional if each DFF facet is considered as a dimension. This is also analogous to the multidimensional DIF model. The multidimensional DIF model extends the unidimensional DIF model specified in Equation 4 to be

$$\log it = \Theta_n - \delta_{id} + Z_n \gamma_{id}, \tag{9}$$

where $\Theta_n = (\theta_1, \theta_2, \cdots, \theta_D)'$, a vector of proficiency parameters for person *n* on *D* dimensions. The item difficulty parameter δ_{id} and the DIF parameters γ_{id} are all dimension dependent. The DIF term in Equation 9 can be replaced by a DFF term so that the DFF parameters can be dimensional as well. The logit expression of the multidimensional DFF model turns into

$$\log it = \Theta_n - \delta_{id} + Z_n \gamma_{kd}. \tag{10}$$

Note, the items are analyzed at the item level as the multidimensional Rasch model does, whereas the interaction effect is modeled at the facet level to yield a more substantively sound explanation. Figure 1 can be used to illustrate the relation between the multidimensional DIF (MDIF) and multidimensional DFF (MDFF) models. Suppose there are 6 items

Item	D_1	D_2
1	√	
2	√	
3		√
4	√	
5	√	
6		√

Figure 1:
Diagram of a between-item multidimensionality case.

and 2 dimensions. Item 3 and item 6 are the ones that exhibit DIF effects. Since the DIF effects both appear on the second dimension, a DFF parameter can be modeled instead of two individual DIF parameters. This is equivalent to constraining the DIF parameters for item 3 and 6 to be equal. Loading the two items on an additional dimension allows the model to examine the individual differences of the performance on the facet, whereas the unidimensional model only captures the group difference. Adams, Wilson and Wang (1997) distinguished two types of multidimensionalities. When each item in a test measures only one underlying dimension of the person proficiency, this is called between-item multidimensionality. If an item measures more than one dimension, it is called within-item multidimensionality. Figure 1 illustrates an example of between-item multidimensionality.

In the case of the Canada-Japan data, each significant DFF effect from the unidimensional DFF model can be considered as a dimension. However, high dimensionality data are hard to estimate. Here, the *reflection* process is selected to be one underlying dimension to illustrate the MDFF model. Furthermore, one dimension of the overall underlying dimension θ is needed. Thus, it is a within-item multidimensional model. A two-dimensional DFF model was fit to the data with $\Theta = (\theta_{overall}, \theta_{\gamma})$. This time, the means of the two countries are modeled through the multivariate regression of Θ, with country as the regressor. The model is constrained on the item side, so the group means on each dimension are free to be estimated. This multidimensional latent regression analysis was performed using the *ConQuest* software (Wu, Adams & Wilson, 1997). Table 9 and Table 10 present results from the MDFF analysis.

In Table 9, the estimates for the constant are the mean performance of the reference country, which is Canada. The mean performance for Canadians on the second dimension is lower than that on the first one. That is to say, Canadian students' performance on the items requiring the *reflection* process is lower than that on the overall mathematics scale. The estimates for the country regression coefficient show that Japan has higher performance than Canadians on both dimensions. To better interpret the regression coefficient estimates, it is helpful to calculate the effect sizes of the country regression coefficient by dividing the estimates of the regression coefficients by the unconditional standard deviation of the latent dimensions.[6] The unconditional model was also fit to the data without the regressor. Table

[6] The procedure to compute the effect size is described in *ConQuest* manual on multidimensional latent regression.

Table 9:
Regression coefficient estimates from the MDFF

	Dimension	
Regression Variable	**1**	**2**
Constant	0.551	-0.211
Country	0.283	0.203

Table 10:
Variance-covariance matrix from the MDFF

	Dimension	
Dimension	**1**	**2**
1	1.241	0.026
2	0.026	0.197

10 shows that the variances of the two latent dimensions from the conditional model are very different. The variances obtained from the model without the regressor do not differ much from the above results. The effect sizes of the country regression coefficient are 0.254 and 0.454. Therefore, on the overall dimension, the country difference is 25.4% of a standard deviation whereas it is 45.4% on the second dimension. This indicates that the country difference is more prominent on the second dimension. Table 10 also shows that the facet (reflection) effect between the two countries is almost orthogonal to the overall ability dimension, and that the variance of the reflection effects is about 16 percent of the size of the variance of the overall ability.

Conclusion

This study demonstrates the use of the LLTM, DFF and MDFF models in this international assessment data set. The LLTM is useful in testing how the response data conform to the structure of the test design. The DFF model helps to explain the DIF effects more substantively (see discussion of specific effects above). An important next step in this research agenda will be to see how well the findings about country differences above relate to curriculum and other differences between the countries. Modeling the DFF effect through an interaction of the group-by-facet parameter rather than DIF effect on the individual item level can be handled easily with the NLMIXED procedure of *SAS*. However, it is always recommended to check the individual DIF estimates and do a substantive analysis first. The multidimensional extension of the DFF model allows individual difference to be modeled on the dimension that exhibits the DFF effect. Again, relating these findings to curricular and other educational differences between the countries is an important next step. The finding that there are relatively large country differences on the secondary reflection dimension seems particularly interesting here.

There are many possible extensions of the aforementioned models. For example, in the LLTM, parameters can be added to test interactions between the item properties. It is also quite straightforward to model DFF and MDFF for polytomous data. The type of data that suits all these models the most is skills diagnosis assessment data. Tatsuoka (1983, 1990, 1995)'s Rule Space used a skill **Q** matrix defining the mastery and nonmastery of a predetermined list of skills. When the methodology of detecting DIF is combined with the pattern recognition approach from Tatsuoka's **Q** matrix, the DFF model can become a valuable model to compare group performances at the skills-level. Following Stout's (2002) assertion in his presidential address that "formative assessment skills diagnosis is the new test paradigm", we can predict that a blended summative assessment and formative assessment trend can be expected to be a fruitful direction for international assessment as well. It would be informative, to researchers and the general public as well, to acquire results from DFF/MDFF models that address these three important aspects of psychometrics, test fairness, skill diagnosis, and dimensionality.

References

Adams, R. J., *Wilson*, M. R., & *Wang*, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*, 355-368.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 3*, 359-374.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54*, 681-697.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach*. New York: Springer.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institution for Educational Research.

SAS Institute (2004). *SAS OnlineDoc 9.0 for the Web* (software manual). Cary, NC: SAS Institute Inc.

Shealy, R., & Stout, W. (1993a). An item response theory model for test bias and differential item functioning. In P. W. Holland & W. Howard (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shealy, R., & Stout, W. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stout, W. (2002). Psychometrics: from practice to theory and back. 15 Years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika, 67*, 485-518.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*, 361-370.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glazer, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nicholas, S. Chipman & R. Brennen (Eds.), *Cognitively diagnostic assessment* (pp.327-395). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item-response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.147-169). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Wu, M. L., Adams, R. J., & Wilson, M. (1997). *ConQuest: Multi-aspect test software* [computer program]. Camberwell, Vic.: Australian Council for Educational Research.