# Application of multilevel IRT to investigate cross-national skill profiles on TIMSS 2003

**Chanho Park and Daniel M. Bolt**
*University of Wisconsin-Madison, USA*

This article presents a multilevel IRT model developed for group-level diagnosis and applied to study cross-national profiles on the TIMSS 2003 mathematics assessment. Variability in item difficulty (i.e., differential item functioning or DIF) across countries is investigated in relation to item features associated with content and cognitive process categories. Random effects were attached to each feature type at the country level, and their variability studied across countries. The estimated feature effects were shown to provide a basis for examining cross-national differences for individual features as well as cross-feature differences within individual countries, as may be useful for diagnostic purposes. The model was fitted using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS.

## INTRODUCTION

Educational tests are frequently designed to enable comparisons between units at different levels of an educational hierarchy. For example, while many tests are designed to compare individual student performances, others are designed to facilitate comparisons at a higher level (e.g., among schools, districts, states, or countries). Assessments such as the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the National Assessment of Educational Progress (NAEP), for instance, are designed to facilitate comparisons among units above the student level. The best use of the assessments is achieved when statistical methodologies designed for inferences at the appropriate level(s) of comparison are used. The purpose of this study was to propose and investigate a statistical modeling approach based on application of multilevel item response theory (ML-IRT) for group-level diagnosis.

Our real data application of the model used item response data from TIMSS. The aim of the TIMSS assessments is "to improve the teaching and learning of mathematics and science by providing data about students' achievement in relation to different types of curricula, instructional practices, and school environments" (Mullis, Martin, Gonzalez, & Chrostowski, 2004, p. 13). Both the mathematics and science tests are designed to provide comparative information regarding cross-national achievement, in the hope that educators and policy-makers use the information to evaluate the effectiveness of their respective curricula and to provide better education to their students. Cross-national comparisons based on TIMSS can be made not only in overall mathematics and science abilities, but also in specific areas related to particular content and cognitive domains. Indeed, it is through these more specific diagnoses that more meaningful information able to underpin improvements to educational practice might be attained. For example, the mathematics assessment in TIMSS organizes items according to five content domains—number, algebra, measurement, geometry, and data—and four cognitive domains—knowing facts and procedures, using concepts, solving routine problems, and reasoning. Each item is assigned to one content domain category and one cognitive domain category. Currently, TIMSS reports domain scale scores with respect to these content and cognitive domains as well as overall scores (Mullis et al., 2004; Mullis, Martin, & Foy, 2005). It is important to note, however, that because all items are cross-classified into both content and cognitive categories, there is the potential for some confounding of effects when attempting to interpret scores specific to each domain. The goal of this study was to present an item response theory (IRT)-based methodology that would explain differential item functioning (DIF) in relation to the content and cognitive characteristics of the items, and in the process better distinguish these effects by controlling for effects related to one characteristic when examining the effects of another.

IRT has provided a useful scaling methodology for educational measurement, and has also found useful applications to the TIMSS assessment, as well as to similarly designed large-scale assessments (e.g., NAEP, PISA). A common strategy entails the introduction of a separate proficiency for each student with respect to each domain (possibly handled simultaneously within a multidimensional model). Due to the matrix sampling designs used with these assessments, a plausible values methodology (often adding student background variables as covariates) is used to account for the varying amounts of uncertainty of each student's level of attainment on each proficiency. Group-level (e.g., country-level) estimates of proficiency distribution parameters (i.e., mean, standard deviation) can then be derived, often incorporating sampling weights. The current approach to domain score reporting using TIMSS data appears to follow this general approach (Mullis et al., 2005). Various issues related to model estimation with this type of design have been described by von Davier, Sinharay, Oranje, and Beaton (2007) in the context of NAEP.

While this modeling approach is a very natural strategy, others may also prove useful. As noted, one potential challenge with this approach when using TIMSS mathematics items is the tendency for items from different cognitive domains to

belong disproportionately to different content domains. The introduction of distinct IRT scaling models for each domain can thus make it difficult to disentangle the relative contribution of content-domain versus cognitive-domain effects when interpreting each scale. Perhaps a more general limitation is the challenge associated with determining the appropriate number and type of proficiencies to introduce into the model. Highly intercorrelated proficiencies may be better handled statistically in the form of a single unidimensional model (see Haberman & von Davier, 2007, for a discussion of these issues). In addition, there is no guarantee that the multiple proficiencies that may best distinguish students from one another within a country are also those that best distinguish groups (e.g., countries) from one another. Indeed, the TIMSS cognitive domain scores tend to show high intercorrelations at the country level (Mullis et al., 2005).

Other IRT-based procedures for group-level diagnostic assessment have been proposed for TIMSS subscore reporting. Tatsuoka, Corter, and Tatsuoka (2004; see also Chen, Gorin, Thompson, & Tatsuoka, this volume pp. 23–49) applied the rule-space methodology for this purpose. They identified 23 skill attributes that were assumed to fully explain the mathematics item difficulties for Grade 8 students in a revised version of TIMSS administered in 1999. Using these attributes, Tatsuoka and her colleagues presented mathematical content and process skill profiles for a sample of participating countries. For example, they found United States students to be weak in geometry, a content area that "does correlate highly with the attributes measuring higher order mathematical thinking" (2004, p. 920). One potential limitation of this approach is that it appears to be based on an aggregation of student-level diagnoses, which can vary considerably in their reliability across skills.

Other approaches may be used to address these limitations and build alternative frameworks for domain score reporting. In this article, we consider a ML-IRT model for DIF as the basis for score reports. Researchers have observed that IRT models can be viewed as a type of hierarchical generalized linear model (Raudenbush & Bryk, 2002) or, alternatively, as a generalized linear or nonlinear mixed model (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; see also McCulloch & Searle, 2001, for a general description of linear and nonlinear mixed models). By viewing item parameters as fixed effects and person ability parameters as random effects, we can portray traditional IRT models as two-level models or mixed models. This modeling framework accommodates higher grouping levels that, in turn, give rise to multilevel IRT models. For example, Kamata (2001, 2002) demonstrated how the one-parameter logistic (1PL) model or the Rasch model can be represented as a two-level generalized linear model and successfully extended it to a three-level model (see also Cheong & Raudenbush, 2000). ML-IRT has thus been used as a hierarchical extension of traditional IRT models (Adams, Wilson, & Wu, 1997) or as applications of the multilevel paradigm to IRT (Fox, 2003; Fox & Glas, 2001). One advantage of this multilevel representation is that it becomes possible to build in additional levels (e.g., classroom, school, etc.) that might be associated with test performance (De Boeck & Wilson, 2004). Such contextual factors may influence not only the distribution of ability within the population, but also characteristics of the

test items, such as item difficulty. As described below, ML-IRT also makes it possible to model item difficulty with respect to multiple item characteristics simultaneously, thus enabling us to understand group differences on more features than would be achievable through computing domain scores.

One alternative ML-IRT approach to reporting profile scores could effectively use DIF at a group (e.g., country) level as a basis for skill profile reports. As an example of this type of approach, Prowker and Camilli (2007) developed an item difficulty variation (IDV) model as an application of a generalized linear mixed model. This model is characterized by the allowance of random effects for item parameters. Items with substantial variability in difficulty are detected, and the cause of variation can be interpreted using contextual factors, such as how well an item matches a state's curriculum standards.

The focus of the current study is on interpreting DIF in relation to item features. Like Prowker and Camilli (2007), we assume that an item's tendency to display DIF can provide diagnostic information of relevance for score reporting purposes. Unlike the Prowker and Camilli (2007) IDV model, however, our approach seeks to model DIF in relation to item characteristics that are explicitly added to the model to account for difficulty variation across countries and that are assumed to be of value for score reporting purposes.

## A MULTILEVEL ITEM FEATURE MODEL AND EXAMPLE ILLUSTRATION

As noted, the purpose of our study was to develop a new ML-IRT methodology and to illustrate its application in a study of cross-national differences on the TIMSS 2003 Grade 8 mathematics assessment. The current illustration of the methodology is somewhat simplistic, and ignores features of the model that might be added (e.g., an account for sampling weights, school effects) to more accurately reflect the TIMSS assessment and its sampling design. However, the illustration allows us to demonstrate in a general way how average skill profiles associated with different countries could be reported using the methodology. Central to the current application is the coding of items according to characteristics that can help provide an explanation of cross-national differences. We refer to the current model as an item feature model (IFM), and we recognize that there are various other ways of attempting to model country-to-country variability in item difficulty.

The IFM assumes the items can be coded according to what we refer to as *item features*. As noted earlier, the Grade 8 mathematics items on TIMSS are currently categorized according to content and cognitive features. Each item can be classified into exactly one category for each feature. In this model, the TIMSS item content and cognitive categories are studied as potential contributors to DIF across countries. In a multilevel modeling framework, the model is a three-level one in which item responses are nested within students and students are nested within countries. In the current ML-IRT model, ability is assumed to vary both at the student and the country levels; item difficulty is assumed to vary only at the country level, with each

item assumed to have the same difficulty parameter across students from the same country. The objective in fitting the IFM is to investigate features that demonstrate variability across countries.

As mentioned, in the TIMSS item feature categorization, items generally belong to more than one category. Consequently, attempts to study cross-national differences by grouping items (e.g., subscale reporting) may not be appropriate, as the different item characteristics are frequently confounded. This situation makes the currently proposed ML-IRT model a potentially more beneficial way of studying cross-national differences.[1] The next section details the statistical model that is the basis for studying item feature effects across countries.

## MULTILEVEL STRUCTURE OF THE IFM

A multilevel representation of the IFM results in a decomposition of item response variance across three levels, with repeated measures (items) nested within students, and students nested within countries. The statistical representation of the model is as follows. At Level 1:

$$P(X_{ijk} = 1 \mid \theta_{jk}) = \frac{\exp(\theta_{jk} - b_{ik})}{1 + \exp(\theta_{jk} - b_{ik})} \quad ,$$

where $X_{ijk} = 1$ denotes a correct response by student $j$ from country $k$ to item $i$ (partial credit items were recoded so that 0, 1=0 and 2=1),

$\theta_{jk}$ is the ability level of student $j$ in country $k$,

$b_{ik}$ is the difficulty parameter for item $i$ when administered to students in country $k$.

At Level 2:

$$\theta_{jk} = \mu_k + E_{jk},$$

where $\mu_k$ denotes the mean ability level in country $k$, and

$E_{jk}$ is assumed to be normally distributed with a mean of 0 and a variance of $\sigma_k^2$.

Finally, at Level 3:

$$\mu_k = \gamma_0 + U_k,$$

$$b_{ik} = \delta_{i1} + \sum_l w_{kl} q_{il}, \; k = 1, \ldots, K,$$

where $K$ is the number of countries,

$\delta_{i1}$ is the difficulty of item $i$ for country 1 (a reference country),

$q_{il}$ is an indicator variable, indicating whether (content or cognitive) feature $l$ (= 1, ..., L) is associated with item $i$, and

$w_{kl}$ are continuous variables identifying the effect of feature $l$ on the difficulty of items within country $k$; $w_{11}, \ldots, w_{iL} = 0$.

[1] The coding of item features for the 99 released mathematics items for Grade 8 is available from the TIMSS website (http://timss.bc.edu/PDF/T03_RELEASED_M8.pdf).

Note that the item difficulties within each country (except for the reference country) are defined relative to those of the reference country for statistical identification purposes (more detail on identification of the model appears in the next section) and to ensure a comparable interpretation of $\theta$ across countries. The model includes fixed effects associated with the overall ability mean across countries ($\gamma_0$), and item difficulties for the reference country ($\delta_{i1}$). The $w_{kl}$ are (potentially random) effects associated with each attribute. When normalized, these random effects are assumed to be normal, with a mean of zero and an estimated variance of $\tau_l^2$. The $U_k$ are assumed to be normally distributed, with a mean of zero and a variance of $\tau_0^2$.

## MODEL IDENTIFICATION

It is well known that IRT models are over-parameterized and that the estimated model parameters are thus identifiable only up to a linear transformation (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Because item parameters are considered structural parameters and ability parameters incidental parameters (Hambleton & Swaminathan, 1985), it is common practice in IRT to fix the metric of the ability parameters ($\theta$) to resolve the indeterminacy. However, the indeterminacy can also be resolved by assuming the item difficulty parameters have a specified mean and variance. In the current application, we addressed the indeterminacy of the θ metric by assigning the difficulty parameters a mean of zero in a reference country (the United States). Next, to make the $\theta$ metrics for other countries determinate, we assumed the item parameters for items of a particular type (as defined by a reference content feature and a reference cognitive feature) to be invariant across countries. In the current application, these items were defined by the "data" content category and the "using concepts" cognitive category. Consequently, for each country, there were a total of four free deviations estimated among the content categories and three deviations among cognitive categories. It is important to note that we assumed the "data" and "using concepts" items to be invariant only in order to define an initial metric against which all remaining model parameters could be uniquely estimated. As noted below, we later rescaled the solutions by normalizing the effects of the item features within each country (thus allowing the "data" and "using concepts" items to also have varying difficulties across countries).

The reason for defining linking items according to both a content feature and a cognitive feature can be attributed to the rank deficiency of the item feature incidence matrix (the elements $q_{il}$ at Level 3). As noted above, each of the TIMSS Grade 8 mathematics items is categorized with respect to exactly one content feature and one cognitive feature. Thus, the item feature incidence matrix of the IFM can be represented as two submatrices (content and cognitive), where the number of columns of the submatrices is five (number of content features) and four (number of cognitive features). Since each item has only one content and one cognitive feature, one characteristic of each submatrix is that any one column (feature) is perfectly dependent on the other columns (features), which thus leads to rank deficiency in each submatrix. Because of double rank deficiency of the item feature incidence

matrix, item feature effects ($w_{kl}$) at Level 3 of the IFM cannot be estimated without additional constraints, a problem analogous to a multicollinearity problem in multiple linear regression analysis.

We bypassed this problem by fixing the effects of one feature for each of the content and cognitive categories at zero (thereby also allowing the items to serve as "linking items," that is, items linking the $\theta$ metrics across countries). The relative effects of item features can be recovered once the effects are normalized across features. Figure 1 presents an illustration of the normalization procedure for the content feature effects. The shaded part of the left-hand side of the figure is the unnormalized feature effects to be estimated; the other parts of the figure are fixed at zero for model identification purposes. We then normalized these "raw" effects so that all marginal sums of the effects became zeroes, as in the right-hand side of the figure. We applied the same procedure to the cognitive feature effects. This normalization provides one way of dealing with the indeterminacy of the $\theta$ metric across groups. The estimates of $\mu_{k'}$ the mean of student abilities ($\theta$) in country $k$, can then also be adjusted by applying the same normalizing constants to the $\mu_k$ estimates. Because the estimated $\mu_k$'s are, in many ways, of secondary importance compared to the normalized feature effect profiles, we consider their estimates only briefly in the results section.

Figure 1: Normalization procedure for content feature effects

| | Data | Geometry | Measurement | Algebra | Number |
|---|---|---|---|---|---|
| USA | 0 | 0 | 0 | 0 | 0 |
| IDN | 0 | | | | |
| RUS | 0 | | | | |
| JPN | 0 | | | | |
| PHL | 0 | | | | |
| EGY | 0 | | | | |
| IRN | 0 | Unnormalized feature effects | | | |
| ENG | 0 | | | | |
| ITA | 0 | | | | |
| KOR | 0 | | | | |
| MAR | 0 | | | | |
| ROM | 0 | | | | |
| TWN | 0 | | | | |
| SAU | 0 | | | | |
| MYS | 0 | | | | |

Normalization →

| | Data | Geometry | Measurement | Algebra | Number | SUM |
|---|---|---|---|---|---|---|
| USA | | | | | | 0 |
| IDN | | | | | | 0 |
| RUS | | | | | | 0 |
| JPN | | | | | | 0 |
| PHL | | | | | | 0 |
| EGY | | | | | | 0 |
| IRN | | Normalized feature effects | | | | 0 |
| ENG | | | | | | 0 |
| ITA | | | | | | 0 |
| KOR | | | | | | 0 |
| MAR | | | | | | 0 |
| ROM | | | | | | 0 |
| TWN | | | | | | 0 |
| SAU | | | | | | 0 |
| MYS | | | | | | 0 |
| SUM | 0 | 0 | 0 | 0 | 0 | 0 |

**Note:** Full names for the country abbreviations are given in Table 1.

## DATA

Data for this study came from the TIMSS 2003 Grade 8 mathematics administration. Data were collected from 49 countries using a two-stage stratified sampling design. For each participating country, schools were sampled first, followed by random selection of a Grade 8 mathematics classroom within each participating school. All students within the sampled class were administered the test. Items were administered using a matrix sampling design involving a total of 12 possible test booklets, with each student receiving one booklet consisting of two or four mathematics item blocks. In total, 194 items were organized across the booklets. In the current study, we analyzed data from only the 15 most populated countries, and considered item responses only from the seven released blocks (99 items). Table 1 lists the 15 selected countries and their abbreviated names. We then selected from each country a random sample of 1,000 examinees administered any 10 of the 12 booklets, a process that gave a total sample size of 15,000 examinees.

Table 1: Names of the 15 most populated countries and their abbreviations

| | |
|---|---|
| United States | USA |
| Indonesia | IDN |
| Russia | RUS |
| Japan | JPN |
| Philippines | PHL |
| Egypt | EGY |
| Iran | IRN |
| England | ENG |
| Italy | ITA |
| Republic of Korea | KOR |
| Morocco | MAR |
| Romania | ROM |
| Chinese Taipei | TWN |
| Saudi Arabia | SAU |
| Malaysia | MYS |

## MCMC ESTIMATION

We fitted the three-level IFM to the TIMSS dataset using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). This approach requires initial specification of the model and a prior for all model parameters. Using a Metropolis-Hastings algorithm, WinBUGS then attempts to simulate draws of parameter vectors derived from the joint posterior distribution of the model parameters. The success of the algorithm is evaluated by whether the chain converges to a stationary distribution, in which case characteristics of that posterior distribution (e.g., the sample mean for each parameter) can be taken as point estimates of the model parameters. In the current application, we chose the following priors for the model parameters:

$$\gamma_0 \sim N(0, 1),\ \tau_0^2 \sim \text{Inverse Gamma } (1, 1),\ \sigma_k^2 \sim \text{Inverse Gamma } (.5, .5),$$
$$\delta_{i1} \sim N(0, 10),\ w_{kl} \sim N(0, 10);$$

where $\gamma_0$ is the overall ability mean across countries, $\tau_0^2$ is the variance of country means ($\mu_k$), $\sigma_k^2$ is the variance of person abilities ($\theta_{jk}$) within country $k$, $\delta_{i1}$ is the difficulty parameter of item $i$ for country 1 (reference country), and $w_{kl}$ is the random effect of country $k$ for the feature $l$.

In MCMC estimation, several additional issues require consideration during monitoring of the chain's sampling history. WinBUGS, by default, uses an initial 4,000 iterations to "learn" how to generate values from proposal distributions to optimize sampling under Metropolis-Hastings. In the present analysis, we used an additional 1,000 iterations as a "burn-in" period, and simulated and inspected an additional 10,000 iterations for convergence using visual inspection as well as convergence statistics available in CODA (Best, Cowles, & Vines, 1996). We did not incorporate sampling weights in this analysis.

## RESULTS

Our visual inspection of the chain histories and CODA diagnostics supported chain convergence out to 15,000 iterations. In particular, we obtained, by running five chains, the *R*-hat statistics suggested by Gelman and Rubin (1992), the values of which were all 1.0 when rounded to the first decimal, which suggests chain convergence. As a further validity check, we examined if the estimated country means ($\mu_k$) correctly represented the known rank order of overall performance across countries based on normal scoring procedures (see, for example, Mullis et al., 2004). As described previously, we normalized the feature effects and added the corresponding normalizing constant for each country to its estimated $\mu_k$. (This process was unnecessary for the reference country, which already had its feature effects normalized in the initial solution.) We then compared the resulting ability means for each country with the mean of the TIMSS scale scores for each country (Mullis et al., 2004, p. 34). A Pearson product-moment correlation coefficient of 0.99 and a Spearman rank-order correlation coefficient of 0.97 confirmed that we were correctly representing the mean ability levels among countries.

Tables 2 and 3 show the estimated content and cognitive feature effects, along with the posterior standard deviations (PSDs) of these for each country for the first (and primary) MCMC run. In order to compare both across features within country and across countries within feature, we calibrated the normalizing coefficients needed for this purpose at each iteration and added these to the reference country and category effects. Once these were normalized, we could compare the feature effects both within a country as well as between countries and thereby obtain profile information. For example, a look across the rows in Tables 2 and 3 allows us to evaluate the relative difficulty of different feature types for a given country, while a look down the columns allows us to determine the relative difficulty of each feature across countries. As noted, the standard deviations of feature effects across countries were estimated at each posterior simulation, and their means across simulations appear at the bottom of Tables 2 and 3. These indicate the variability of the content and cognitive feature effects across countries. From the tables we can see that the content feature effects show more variability across countries than do the cognitive feature effects. Among the content feature effects, "geometry" has the largest variability and "number" the lowest.

Table 2: Normalized effects of content features and PSDs (in parentheses)

| | Number | Geometry | Measurement | Algebra | Data |
|---|---|---|---|---|---|
| USA | -0.08 (0.03) | 0.43 (0.03) | 0.04 (0.03) | 0.02 (0.03) | -0.40 (0.04) |
| IDN | 0.03 (0.03) | -0.07 (0.03) | 0.17 (0.03) | -0.04 (0.03) | -0.08 (0.04) |
| RUS | 0.04 (0.03) | 0.06 (0.03) | -0.13 (0.03) | -0.20 (0.03) | 0.23 (0.04) |
| JPN | 0.31 (0.03) | -0.34 (0.04) | -0.03 (0.03) | 0.05 (0.03) | 0.01 (0.04) |
| PHL | -0.12 (0.03) | 0.22 (0.04) | 0.04 (0.04) | -0.14 (0.03) | 0.00 (0.04) |
| EGY | -0.10 (0.03) | -0.02 (0.03) | 0.22 (0.03) | -0.03 (0.03) | -0.07 (0.04) |
| IRN | -0.03 (0.03) | -0.25 (0.03) | 0.31 (0.04) | 0.04 (0.03) | -0.06 (0.04) |
| ENG | 0.07 (0.03) | 0.14 (0.03) | -0.17 (0.03) | 0.22 (0.03) | -0.26 (0.04) |
| ITA | 0.01 (0.03) | 0.24 (0.03) | -0.41 (0.03) | 0.10 (0.03) | 0.05 (0.04) |
| KOR | 0.00 (0.03) | -0.26 (0.04) | 0.15 (0.03) | -0.18 (0.03) | 0.29 (0.04) |
| MAR | 0.03 (0.03) | -0.25 (0.03) | 0.12 (0.04) | -0.02 (0.03) | 0.12 (0.05) |
| ROM | 0.07 (0.03) | 0.10 (0.03) | -0.13 (0.03) | -0.16 (0.03) | 0.12 (0.04) |
| TWN | 0.07 (0.03) | 0.03 (0.04) | -0.11 (0.03) | -0.11 (0.03) | 0.12 (0.05) |
| SAU | -0.14 (0.03) | -0.17 (0.04) | 0.11 (0.04) | 0.24 (0.03) | -0.04 (0.05) |
| MYS | -0.18 (0.03) | 0.15 (0.03) | -0.15 (0.03) | 0.22 (0.03) | -0.04 (0.04) |
| *SD* | 0.12 | 0.22 | 0.19 | 0.15 | 0.18 |

Table 3: Normalized effects of cognitive features and PSDs (in parentheses)

| | Using concepts | Knowing facts and procedures | Solving routine problems | Reasoning |
|---|---|---|---|---|
| USA | 0.05 (0.03) | -0.02 (0.02) | 0.02 (0.02) | -0.05 (0.03) |
| IDN | 0.07 (0.03) | -0.09 (0.03) | -0.07 (0.03) | 0.09 (0.03) |
| RUS | 0.00 (0.03) | -0.06 (0.03) | 0.02 (0.02) | 0.03 (0.03) |
| JPN | 0.09 (0.03) | 0.04 (0.03) | -0.13 (0.03) | 0.00 (0.03) |
| PHL | -0.08 (0.03) | -0.08 (0.03) | 0.14 (0.03) | 0.02 (0.04) |
| EGY | 0.09 (0.03) | -0.20 (0.03) | 0.02 (0.03) | 0.09 (0.03) |
| IRN | -0.06 (0.03) | 0.00 (0.03) | 0.03 (0.03) | 0.03 (0.03) |
| ENG | 0.08 (0.03) | 0.16 (0.03) | -0.07 (0.02) | -0.18 (0.03) |
| ITA | 0.03 (0.03) | 0.21 (0.03) | -0.04 (0.02) | -0.20 (0.03) |
| KOR | -0.01 (0.03) | 0.14 (0.03) | 0.02 (0.03) | -0.15 (0.03) |
| MAR | -0.17 (0.03) | 0.02 (0.03) | 0.15 (0.03) | -0.01 (0.04) |
| ROM | 0.02 (0.03) | -0.09 (0.03) | -0.07 (0.03) | 0.14 (0.03) |
| TWN | 0.05 (0.03) | -0.01 (0.03) | -0.19 (0.03) | 0.15 (0.03) |
| SAU | -0.14 (0.03) | 0.00 (0.03) | 0.13 (0.03) | 0.01 (0.04) |
| MYS | -0.03 (0.03) | -0.02 (0.03) | 0.02 (0.02) | 0.03 (0.03) |
| *SD* | 0.08 | 0.11 | 0.10 | 0.11 |

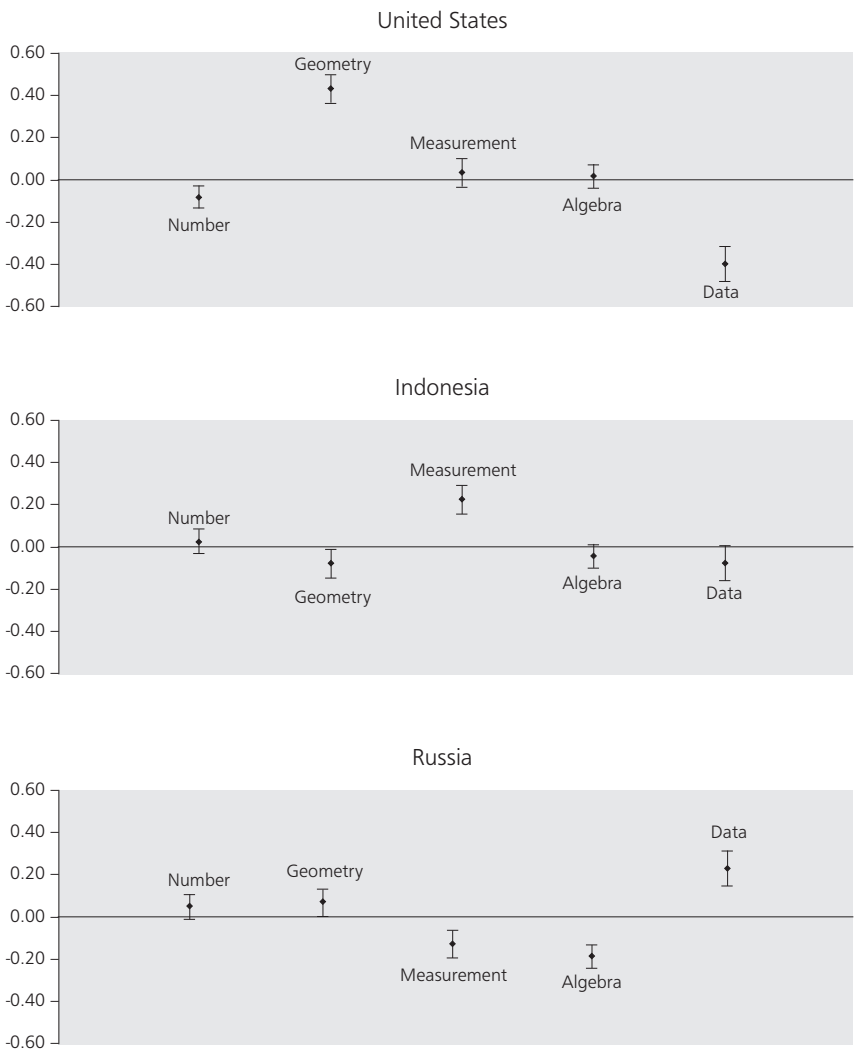
One advantage of the IFM is its ability to define profiles for both countries and features. Figures 2 through 5 demonstrate how our ML-IRT model could ultimately be used to study between-country differences. Figure 2 shows the content feature effects for each of the 15 countries. The most conspicuous coefficient in the graphs is the geometry effect for the United States. We can interpret the large positive value in terms of geometry being a significant contributor to the difficulty of test items for US students relative to students from other countries; that is, geometry items demonstrate disproportionately greater difficulty for US students relative to items associated with other content features. By contrast, for a country such as Japan, geometry items demonstrate disproportionately less difficulty relative to items associated with other content features. This observation aligns with the results reported in the literature. Indeed, Tatsuoka et al. (2004) reported that US students performed poorly on geometry relative to other countries on TIMSS 1999. Mullis et al. (2004) have also reported that US students performed poorly on geometry in TIMSS 2003.

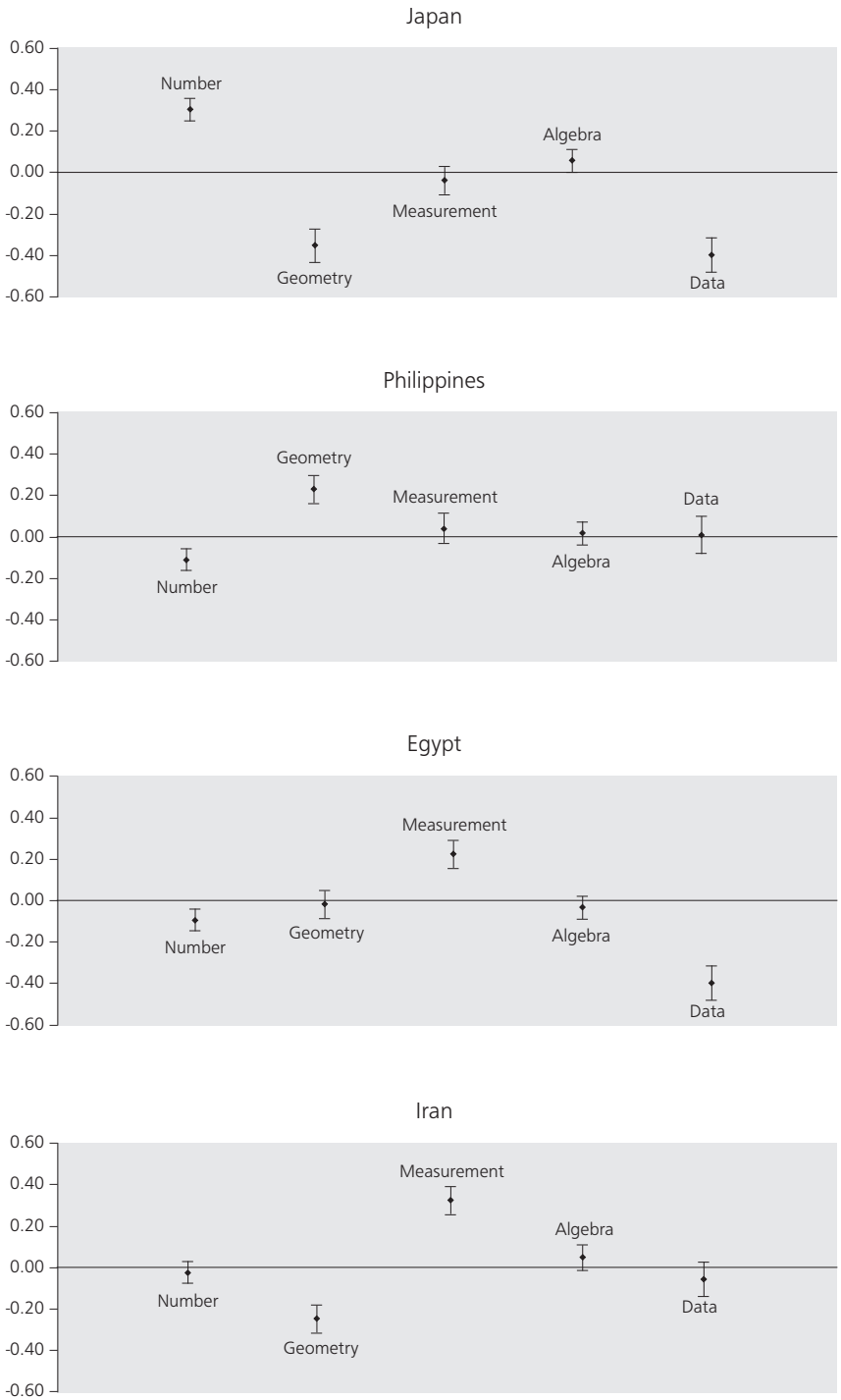
Figure 3 illustrates the normalized cognitive feature effects by country. Relative to the content feature effects, the cognitive feature effects are mostly non-significant contributors to variability in the item difficulties across countries. We can still find some significant contributors, such as the cognitive feature "knowing facts and procedures," which makes items relatively more difficult for students in Italy and makes items relatively easier for students in Egypt. As with the content feature effects,

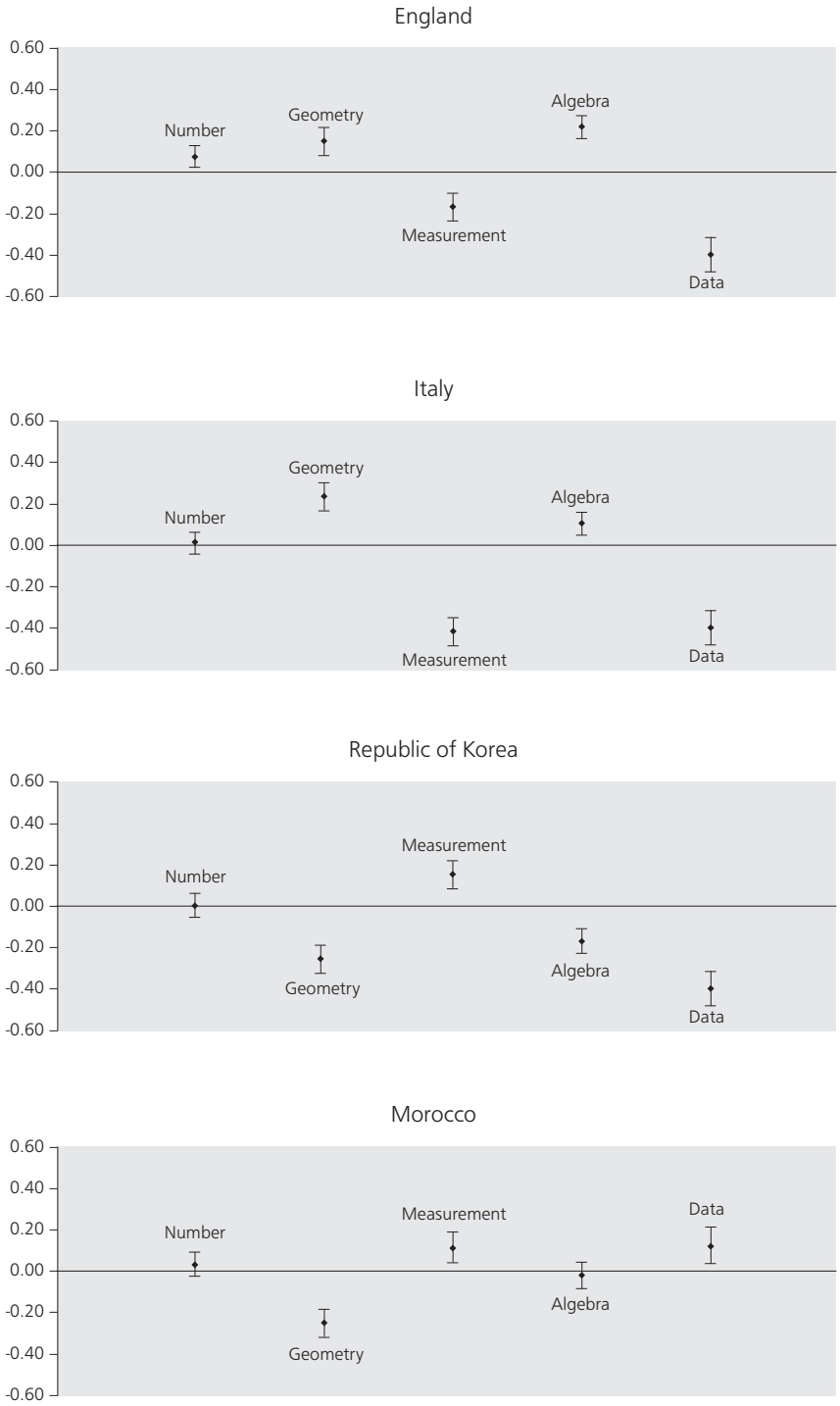
however, it is important to note that the sum to zero constraint requires that these effects always be interpreted in a relative fashion—an equivalent interpretation might emphasize the other cognitive features as relatively easier.

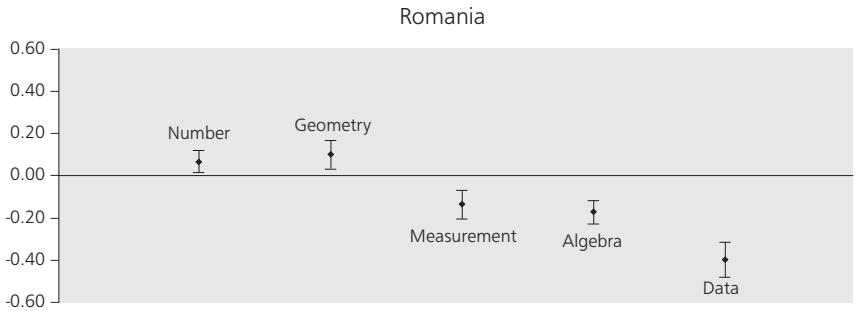
Figures 4 and 5 provide country effects with respect to each of the content and cognitive features. In these graphs, we can compare countries on each feature. For example, the "number" feature makes items relatively harder for Japanese students whereas it makes items relatively easier for Malaysian students. Similar interpretations can be applied for the cognitive features (Figure 5). For example, students in England, Italy, and Korea have relative strengths for "reasoning".

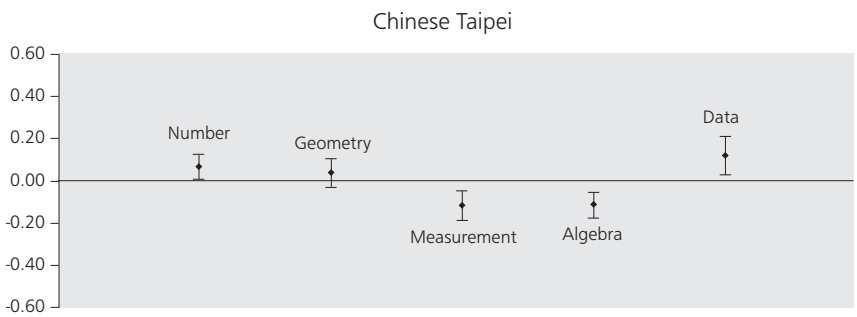
Figure 2: Content feature effects +/– two PSDs, by countries

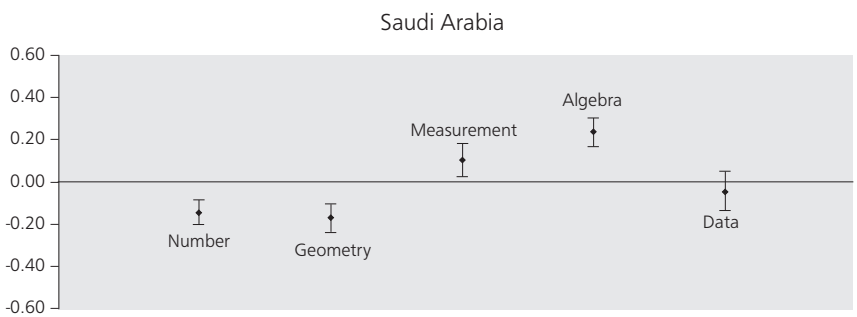



Japan
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data
Philippines
Geometry
Measurement
Data
Number
Algebra
Egypt
Measurement
Number
Geometry
Algebra
Data
Iran
Measurement
Algebra
Number
Data
Geometry


England
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data
Italy
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data
Republic of Korea
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data
Morocco
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data


Romania
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data


Chinese Taipei
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Number
Geometry
Measurement
Algebra
Data


Saudi Arabia
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Measurement
Algebra
Number
Geometry
Data

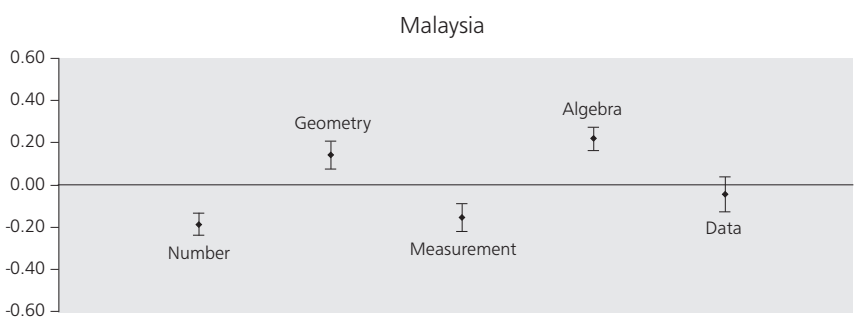
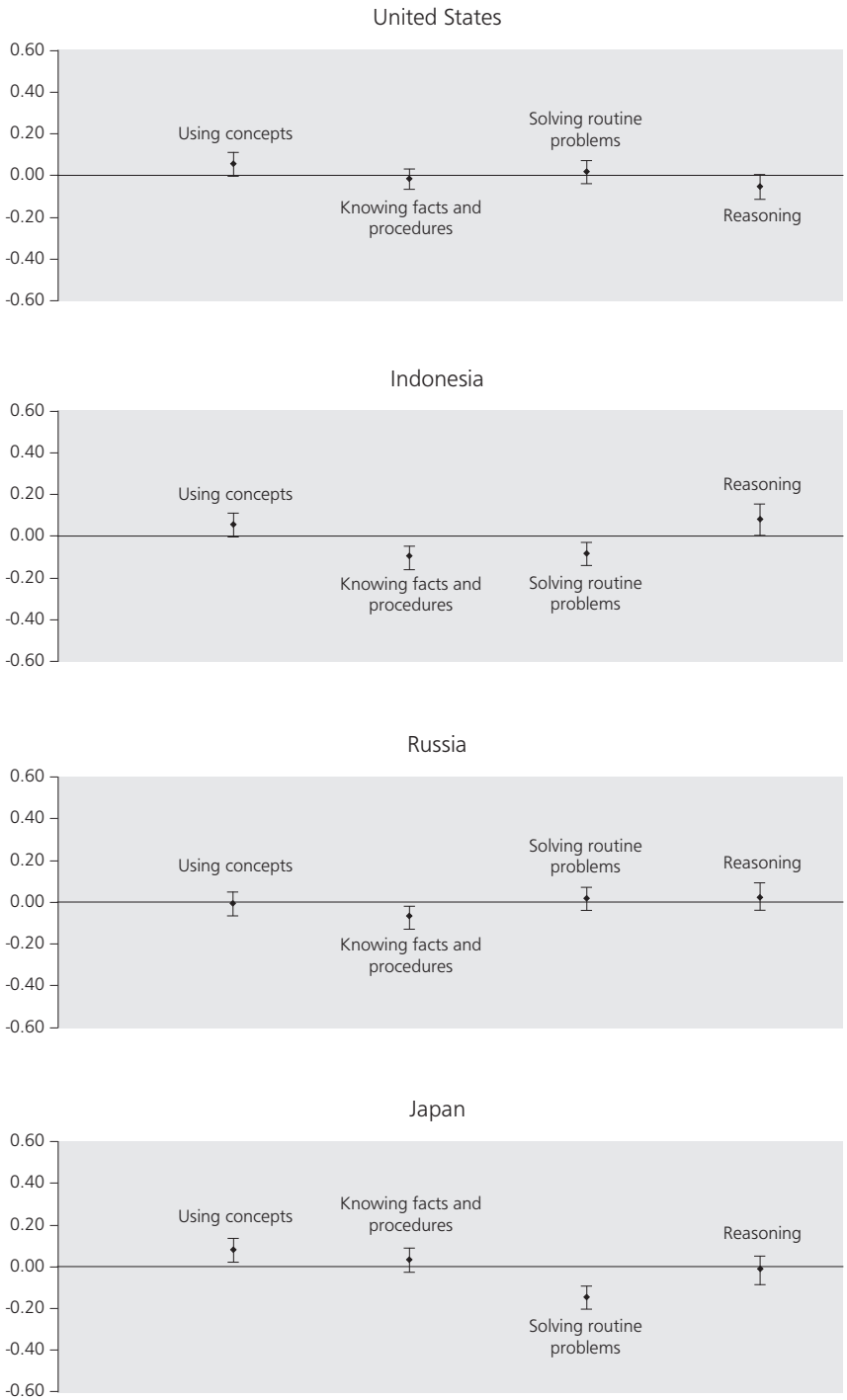
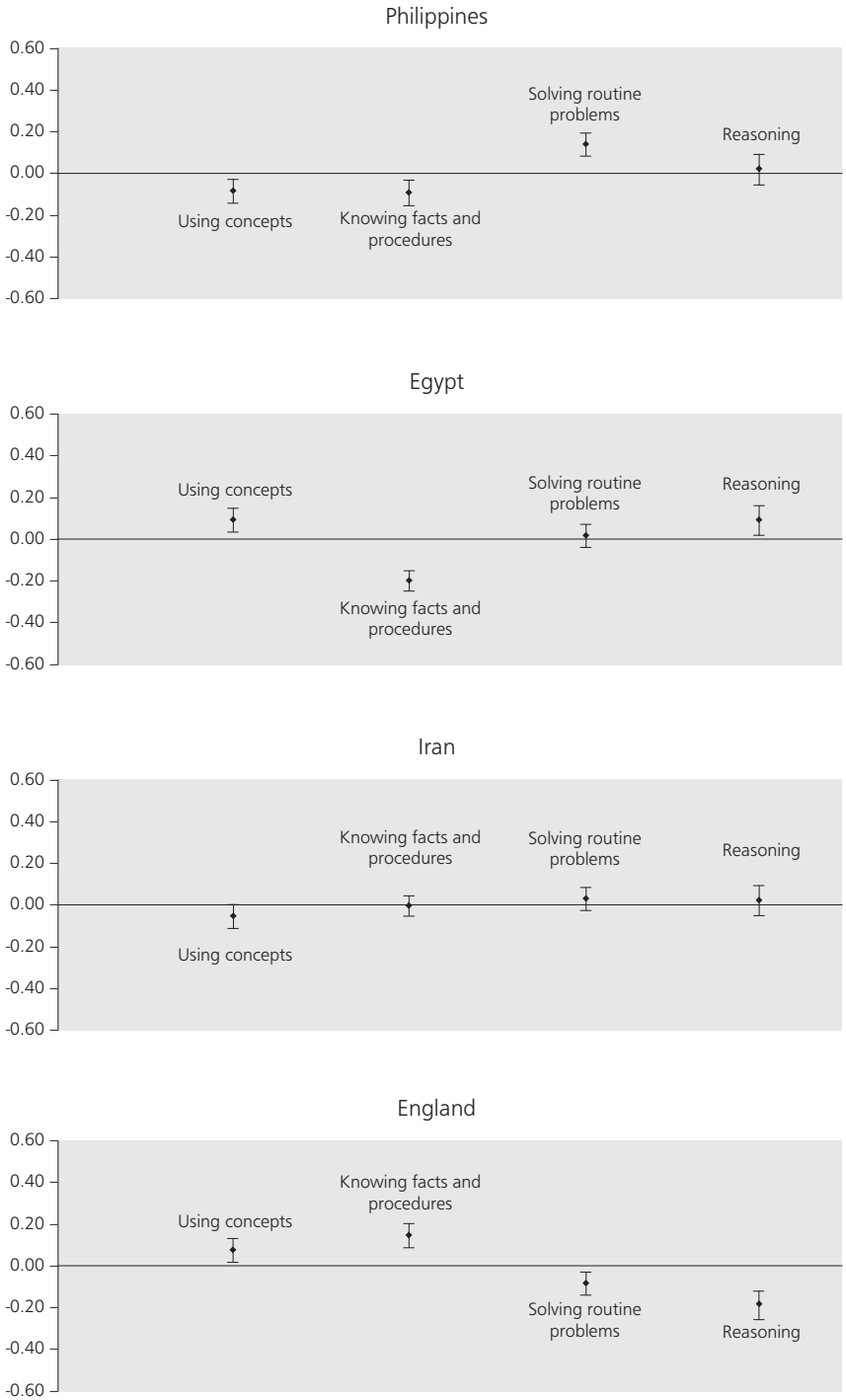
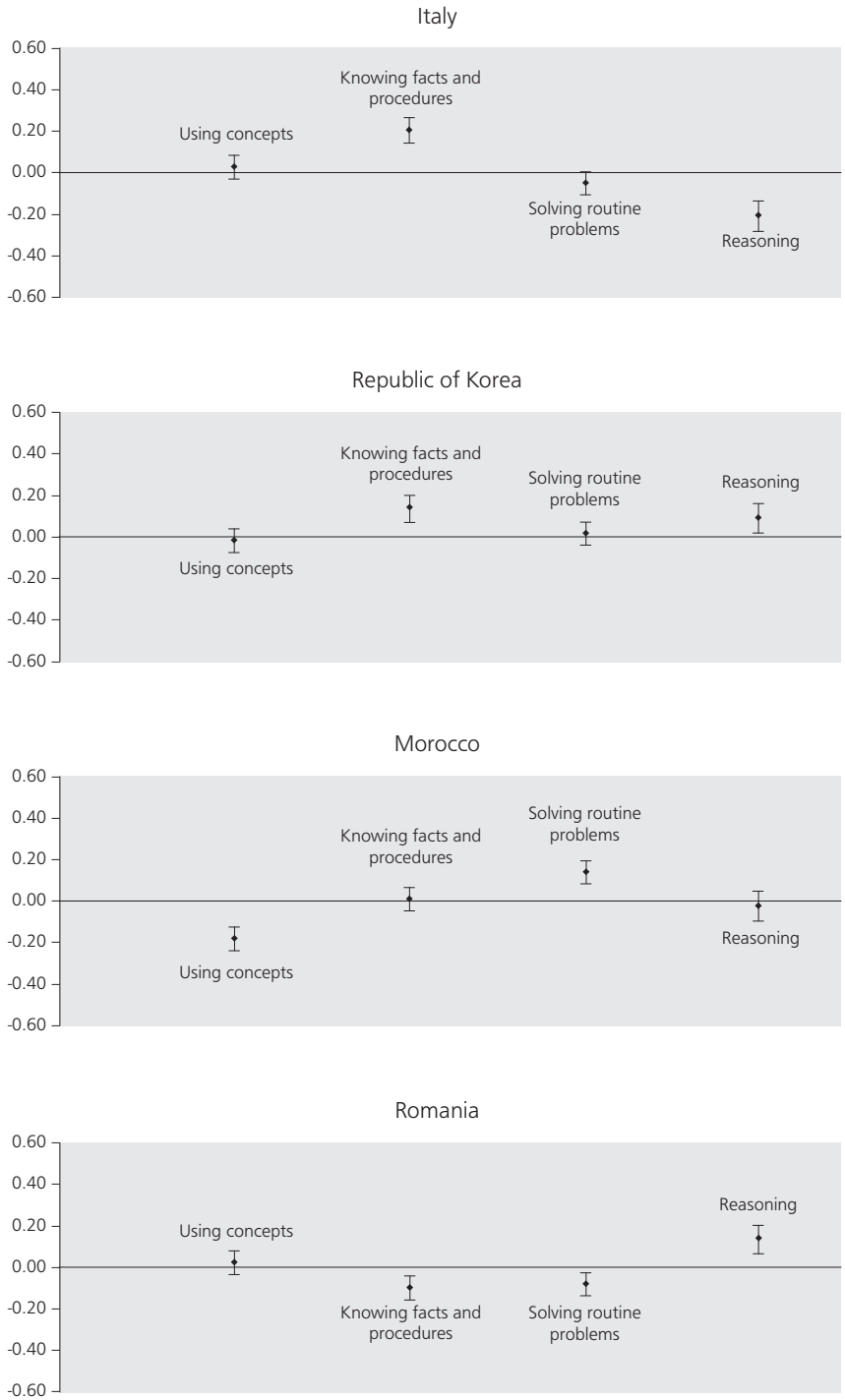

Malaysia
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Geometry
Algebra
Number
Measurement
Data

Figure 3: Cognitive feature effects +/– two PSDs, by countries




Philippines
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Egypt
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Iran
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
England
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning


Italy
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Republic of Korea
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Morocco
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Romania
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning

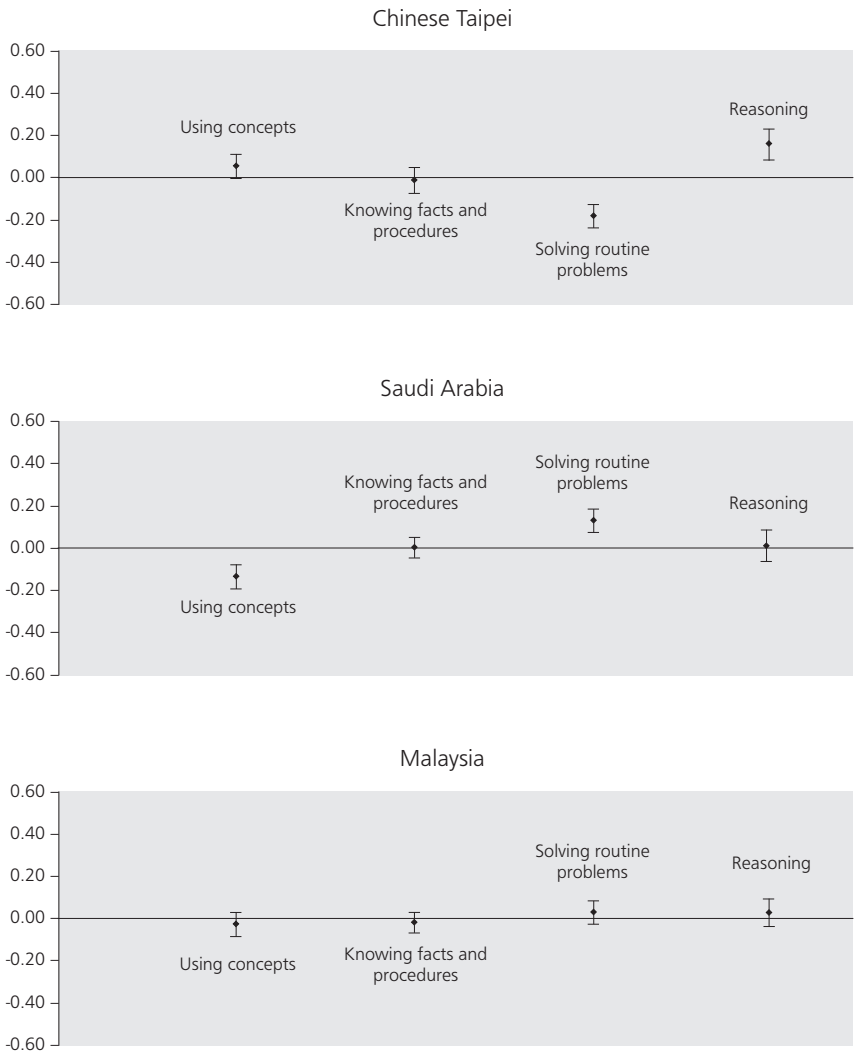
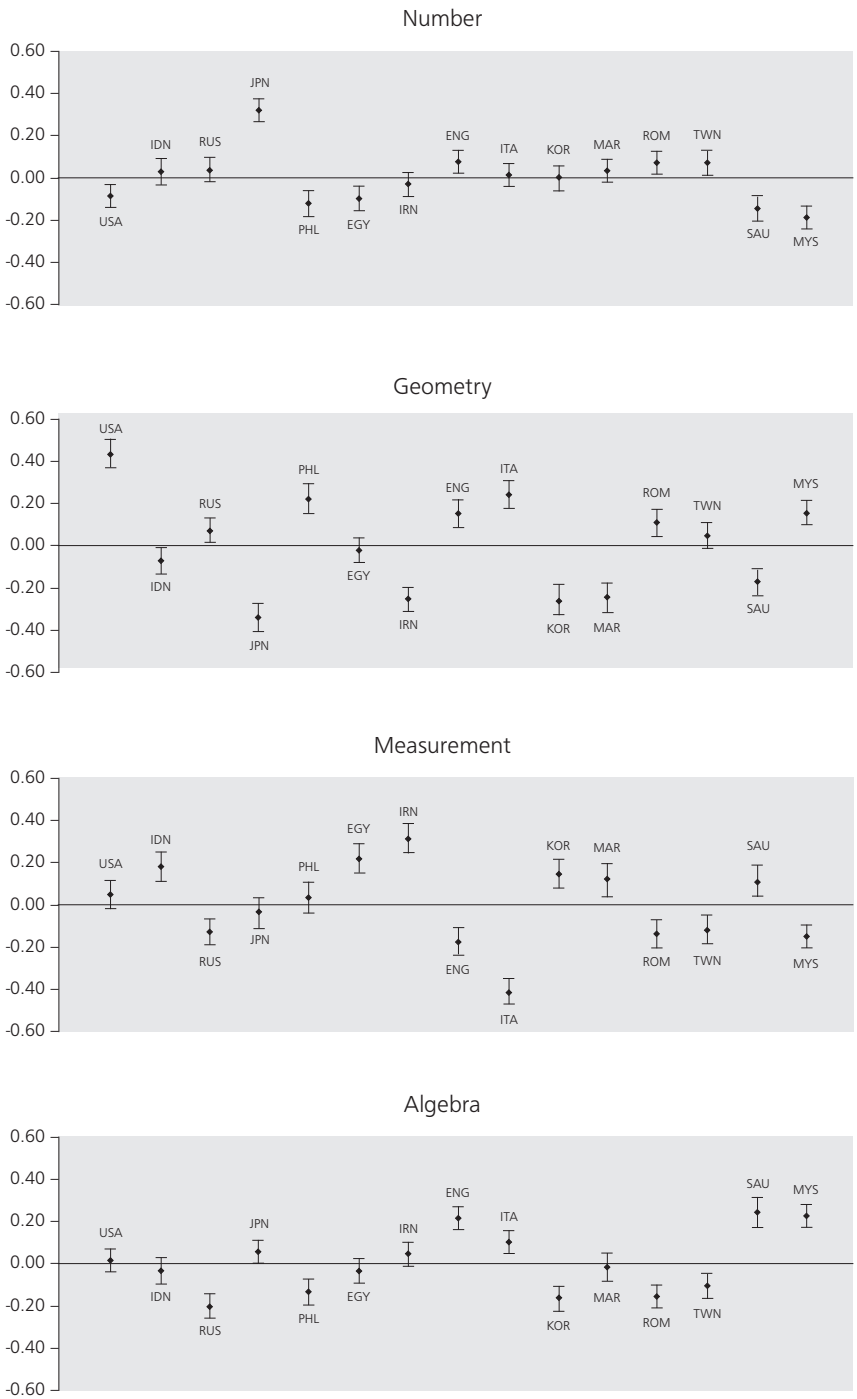

Chinese Taipei
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Saudi Arabia
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning
Malaysia
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Using concepts
Knowing facts and procedures
Solving routine problems
Reasoning

Figure 4: Country effects +/– two PSDs, by content features



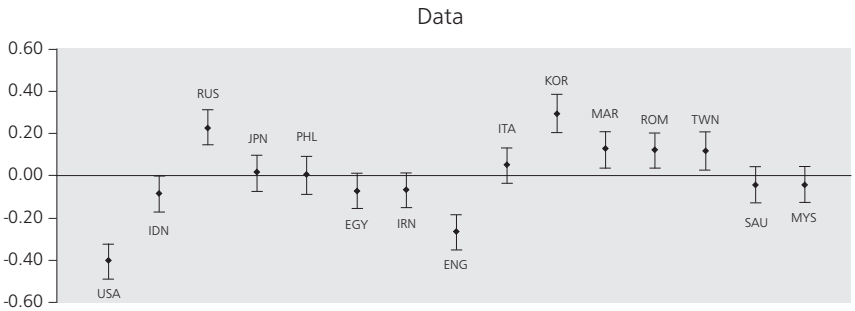
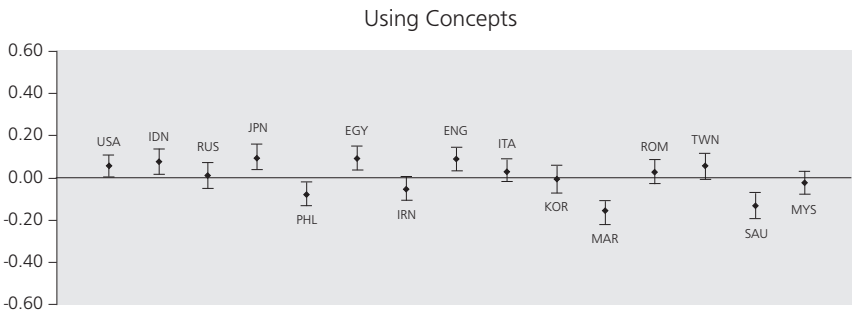
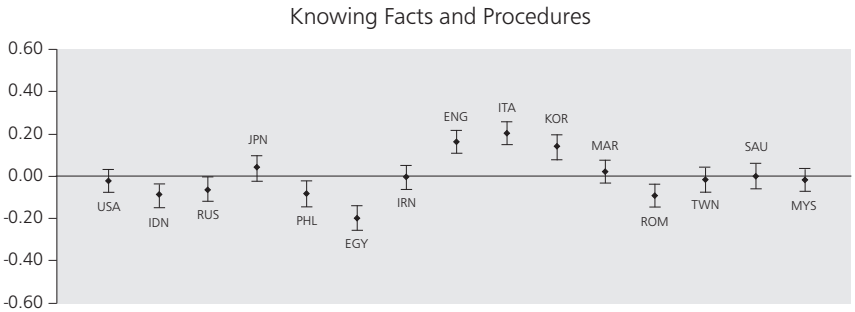
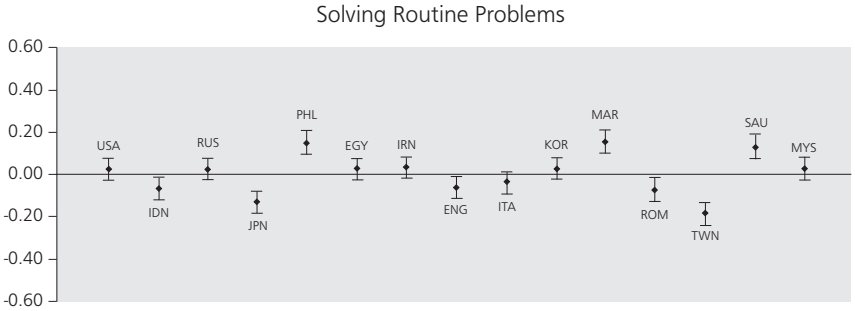
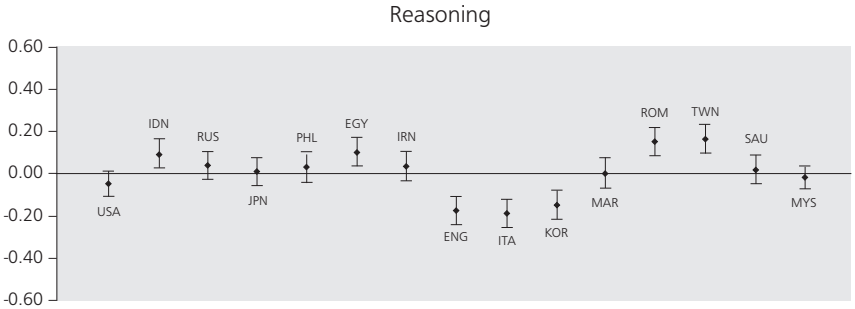



Figure 5: Country effects +/– two PSDs, by cognitive features






When considering Figures 2 to 5, it is important to recognize the effect of the normalization on the results being inspected and the implications this has for interpretation. Because the feature effects for each country are assumed to sum to zero, the figures illustrate only those effects related to the relative difficulty of items; in other words, the effects can be interpreted in comparison to the other feature effects within each country or to the other country effects given to each feature. The other critical component in determining the overall difficulty of the items within countries is the ability distribution within each country. As noted, our analysis estimated a mean ability for each country. It is only by considering this mean ability that we can better understand what the overall difficulties of the items are for a given country. Table 4 shows the adjusted estimates of $\mu$'s ($\hat{\mu}_{Adj}$) and the average TIMSS scale scores for each country.

## DISCUSSION AND CONCLUSION

The purpose of this article was to illustrate a general methodology that could provide the foundation for diagnostic score reports at the country level based on TIMSS. As noted, the approach taken is somewhat simplistic, but it could be extended to incorporate a variety of other features that reflect greater sensitivity to the sampling design used with the assessment. Although, in the current analysis, the results seem to provide comparable results to methods already used with TIMSS, there are several potential advantages of the current methodology relative to that used for current TIMSS reports (Mullis et al., 2005). Because the methodology models how the specific domains simultaneously contribute to DIF at the group (country) level, as opposed to the current TIMSS reports, which are based on the distribution of separate proficiencies at the group level (Mullis et al., 2004, 2005), it is likely that it more effectively separates the relative effects of the cognitive and the content features. The use of DIF for diagnostic score reporting has been demonstrated in previous work (e.g., Prowker & Camilli, 2007), and it seems particularly well suited for assessments like TIMSS, where information is obtained at the group level on a large number of different items (even though relatively few are administered to individual examinees). The potential confounding of cognitive and content feature effects can be examined empirically. Cross-tabulation between content and cognitive features

for the 99 items used in this analysis (Table 5) suggests the potential for dependence between content and cognitive features, although the chi-square statistic in this case was not statistically significant ($\chi^2_{12}$, $p = 0.19$).

Another advantage of the approach is its comparative simplicity in terms of implementation. The relative standing of all countries on specific domains is made in reference to a single latent proficiency, and thus does not require the more extensive linking procedures used when modeling multiple proficiencies either separately or jointly. Indeed, one of the strengths of ML-IRT is that groups as well as individuals are all units of analysis, and thus, in similar vein to concurrent calibration, the linking procedure is incorporated within one general framework (Park, Kang, & Wollack, 2007).

Table 4: Adjusted estimates of ability means (M) and mean ability scale scores on TIMSS 2003 for each country

| | $\hat{\mu}_{Adj}$ | Average TIMSS scale score |
|---|---|---|
| KOR | 0.99 | 589 |
| TWN | 0.93 | 585 |
| JPN | 0.74 | 570 |
| RUS | 0.07 | 508 |
| ENG | -0.01 | 498 |
| MYS | -0.05 | 508 |
| USA | -0.07 | 504 |
| ROM | -0.26 | 475 |
| ITA | -0.29 | 484 |
| EGY | -0.67 | 406 |
| IDN | -0.96 | 411 |
| IRN | -1.04 | 411 |
| MAR | -1.23 | 387 |
| PHL | -1.31 | 378 |
| SAU | -1.60 | 332 |
| Correlation | Pearson | 0.99 |
| | Spearman | 0.97 |

Table 5: Cross-tabulation between content and cognitive features for the 99 items

| Cognitive \ Content | Number | Geometry | Measurement | Algebra | Data | Sum |
|---|---|---|---|---|---|---|
| Using concepts | 7 | 5 | 2 | 5 | 2 | 21 |
| Knowing facts and procedures | 8 | 4 | 7 | 7 | 0 | 26 |
| Solving routine problems | 14 | 3 | 6 | 7 | 4 | 34 |
| Reasoning | 2 | 4 | 2 | 5 | 5 | 18 |
| Sum | 31 | 16 | 17 | 24 | 11 | 99 |

Although we have not examined in this paper how well the specified content and cognitive features account for the DIF observed across countries, it is possible to do this with the current methodology, given that other sources of DIF (besides the item features) are undoubtedly present. One potential extension of the model considered in this article would be to add a residual to each country-specific item difficulty parameter, so as to explicitly account for the presence of other sources of country-level DIF. To get a better sense of the importance of the features in explaining DIF, we calculated simple $R^2$ statistics by regressing the differences of item difficulty parameter estimates between each of the comparison countries and the reference country onto the binary item features incidence matrix. The average $R^2$ statistic across countries was 0.166, indicating substantial other sources of DIF besides the features examined in our analysis. In other words, it would appear that the features currently emphasized in TIMSS score reporting do not go very far in explaining variability that exists between countries. Whether a better set of features exists or whether such variability is simply due to item-specific characteristics remains an important direction for future exploration.

It is necessary to acknowledge a number of limitations of the current analysis, and crucial other directions for future work. As noted, our analysis was not sensitive to the sampling weights present in the TIMSS design. In addition, it ignored other levels (e.g., school) that could be added using the same modeling framework. Also, because the analysis was based on only the 99 released items from the TIMSS assessment and included only the 15 countries with the largest samples, it should be extended to include the full 194 items and more countries. Larger samples may also help in estimating certain parameters of the model, such as the trait variances and country weights, which tend to be more challenging to estimate accurately. Other aspects of the analysis, such as our use of dichotomous scoring for items that were polytomously scored, could be relaxed so that the full range of item scores is considered. Finally, steps can also be taken to improve the estimation of the model. The MCMC run for the analysis presented in this article took more than a day, even on a relatively fast machine.

Although distinct from the Tatsuoka et al. (2004) rule-space approach, the current methodology could also accommodate more specific codings of items, such as those used in rule-space applications to TIMSS. As noted, Tatsuoka et al. (2004) identified 23 content and cognitive skill attributes in the TIMSS Grade 8 mathematics assessment in 1999. Since those attributes were found to explain most of the variation in item difficulties, the same attributes may be applied to account more fully for DIF observed across countries.

This research was supported by a grant to the first author from the American Educational Research Association, which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant #REC-0310268. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

## References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47–76.

Best, N., Cowles, M. K., & Vines, K. (1996). *CODA*: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30*. Cambridge, UK: MRC Biostatistics Unit..

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*, 477–495.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models*. New York, NY: Springer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fox, J.-P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology, 56*, 65–81.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271–288.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively-based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Kamata, A. (2002, April). *Procedure to perform item response analysis by hierarchical generalized linear model*. Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans, LA.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report*. Chestnut Hill, MA: Boston College.

Park, C., Kang, T., & Wollack, J. A. (2007, April). *Application of multilevel IRT to multiple-form linking when common items are drifted*. Paper presented at the 2007 annual meeting of the National Council on Measurement in Education, Chicago, IL.

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement, 44*, 69–87.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185–205.

Spiegelhalter, D. J., Thomas A., Best, N. G., & Lunn, D. (2003). *WinBUGS Version 1.4 user manual*. Cambridge, England: MRC Biostatistics Unit.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*, 901–926.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). North Holland: Elsevier.