

N.B. Please do not cite or distribute.

Multilevel IRT for group-level diagnosis

Chanho Park
Daniel M. Bolt

University of Wisconsin-Madison

Paper presented at the annual meeting of the American Educational Research Association,
March 24 – March 28, 2008, New York City, NY

Abstract

In this study we conducted simulation analyses to evaluate the effectiveness of a multilevel item feature model (Park & Bolt, in press) as a basis for group-level diagnosis. The model essentially attempts to explain DIF across groups in relation to item features that can then serve as a basis for group-level score profiles. In order to understand the performance of the model as a function of item features and feature weights, three factors—level of feature confounding, feature effect variability, and the explanatory power of item features—were considered for simulation conditions. The model was fit using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS, and the accuracy of item feature weights' recovery was evaluated using biases and root mean squared errors (RMSEs). This study not only helped better evaluate the model's performance under various conditions, but also sheds light on how to analyze group-level diagnostic assessment data more generally.

Multilevel IRT for group-level diagnosis

Educational policies can benefit from an understanding of the cognitive strategies, knowledge states, or skill profiles that underlie student performances on an exam. Recently there has been growing interest in psychometric models that can lend insight into these more finely-grained aspects of student performances (Junker & Sijtsma, 2001). Most currently available cognitive diagnostic models are student-level models developed to account for individual examinee differences in cognitive strategies or skill profiles, whereas many educational tests (e.g., NAEP, TIMSS, PIRLS, PISA) are designed so as to enable comparisons between units at higher levels of the educational hierarchy (e.g., school, school district, state, or country level). When inferences are made at these higher levels, it is usually undesirable to aggregate results from an individual-level model, as group-level inferences based on aggregated scores may result in erroneous interpretations (Snijders & Bosker, 1999; Raudenbush & Bryk, 2002). The best use of such assessments will be achieved when statistical methodologies designed for inferences at the appropriate level(s) of comparison are used. Recent advances in hierarchical modeling have successfully integrated item response theory (IRT) models to create a multilevel item response theory

(ML-IRT) modeling framework (Adams, Wilson, & Wu, 1997; Kamata, 2001). Park and Bolt (in press) considered an item feature model (IFM) as an extension of ML-IRT model for performing group-level diagnosis. The major goal of this study is to evaluate the IFM as an ML-IRT model for group-level diagnosis.

In the IFM, the item's content and cognitive features are studied as potential contributors to differential item functioning (DIF) across groups. In a multilevel modeling framework, a common framework might entail a three-level model in which item responses are nested within students and students are nested within some higher order grouping units, such as schools. Ability is assumed to vary both at the individual and group levels; item difficulty is assumed to vary only at the group level, with each item assumed to have the same difficulty parameter for examinees from the same group. The objective in fitting the IFM is to investigate features that demonstrate variability across groups.

The IFM can be viewed as an extension of the approach applied to the National Assessment of Educational Progress (NAEP) by Prowker and Camilli (2007). Prowker and Camilli developed an item difficulty variation (IDV) model as an application of a generalized linear mixed model. This model is characterized by allowing 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. Like the IDV model, we assume that an item's tendency to display DIF can provide diagnostic information of relevance for score reporting purposes. Unlike the 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.

Park and Bolt (in press) fitted the IFM to the dataset sampled from the Trends in International Mathematics and Science Study (TIMSS), but did not conduct simulation analyses to evaluate parameter recovery or to study the performance of the model under different conditions. In that study the IFM studied countries (as opposed to schools) as grouping units. Since a primary purpose of the current study is to better understand the model's performance with TIMSS, we will refer to the grouping units as "countries" recognizing the potential of applying the IFM using other types of grouping structures. Simulation analyses are naturally necessary to verify how well the IFM will perform under various conditions and to understand how various factors affect the performance of the IFM as an ML-IRT model. Therefore, the simulation analyses will not only investigate how the IFM is affected by various aspects of the test and test items, but also examine how

successful the IFM is as an ML-IRT model for group-level diagnosis.

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:

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

where $X_{ijk} = 1$ denotes a correct response by student j from country k to item i ,

θ_{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 μ_k denotes the mean ability level in country k , and

E_{jk} is assumed normally distributed with mean 0 and variance σ_k^2 .

Finally at level 3:

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

$$b_{ik} = \delta_{i1} + \sum_l w_{kl} q_{il}, k = 1, \dots, K, \text{ where } K \text{ is the number of countries,}$$

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

q_{il} is an indicator variable indicating whether a given feature l ($l = 1, \dots, L$) 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 ; $w_{11}, \dots, w_{1L} = 0$.

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 and to ensure a comparable interpretation of θ across countries. The model includes fixed effects associated with the overall ability mean across countries (γ_0), and item difficulties for the reference country (δ_{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 estimated variance τ_l^2 . The U_k are assumed normally distributed with mean zero and variance τ_0^2 .

The indeterminacy of the IFM can be resolved by assigning the difficulty parameters a mean of zero in a reference country. Next, to make the θ metrics for other countries determinate, the item parameters for items of a particular type are assumed to be

invariant across countries. These initial solutions are then normalized for interpretation.

Figure 1 shows an illustration of the normalization procedure for 5 item features and 15

countries. See Park and Bolt (in press) for more details on identification of the model.

Insert Figure 1 About Here

MCMC Estimation

The three-level IFM was fit to the TIMSS dataset using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). This approach involves 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 from the joint posterior distribution of the model parameters. The success of the algorithm is evaluated by whether the chain successfully 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, the following priors were chosen for the model parameters:

$$\gamma_0 \sim N(0, 1), \tau_0^2 \sim \text{Inverse Gamma}(1, 1), \sigma_k^2 \sim \text{Inverse Gamma}(.5, .5),$$

$$\delta_{il} \sim N(0, 10), w_{kl} \sim N(0, 10);$$

where γ_0 is the overall ability mean across countries, τ_0^2 is the variance of country means (μ_k), σ_k^2 is the variance of person abilities (θ_{jk}) within country k , δ_{il} is difficulty parameter of item i for country l (reference country), and w_{kl} is the random effect of country k for the feature l .

In MCMC estimation, several additional issues require consideration in monitoring the sampling history of the chain. WinBUGS, by default, will use an initial 4,000 iterations to “learn” how to generate values from proposal distributions to optimize sampling under Metropolis-Hastings. An additional 1000 iterations were thus used as a “burn-in” period, and the subsequent 10,000 iterations were then assumed to represent a sample from the joint posterior, and inspected for convergence using visual inspection as well as convergence statistics available in CODA (Best, Cowles, & Vines, 1996).

Simulation Conditions

The simulation conditions manipulated in this study were as follows. First, different levels of feature confounding were varied. As item features that may contribute

to DIF frequently correlate, as appears to be the case with TIMSS, for example (Martin, Mullis, & Chrostowski, 2004), it is worth studying how different levels of such confounding affect recovery of the feature effect parameters. Second, different levels of variability of feature effects across countries are considered. It may naturally be expected that certain item characteristics will more likely relate to DIF than others. Third, the explanatory power of item features in explaining residual variability of item difficulties for the comparison countries is simulated. Systematic variability of item difficulties for the comparison countries with respect to the reference country is accounted for by the item feature incidence (Q) and item feature effects (W) matrices. To make the simulation data more plausible, additional random variability unrelated to the features should be introduced. The amount of DIF attributable to the features can be captured by R^2 statistics, where the features are used as regressors. Two levels of R^2 were considered (.7 and .3) representing large and small residual variances.

Levels of feature confounding. Three levels of confounding among features were considered (see Table 1). The level of confounding was studied using the Jaccard index of similarity for binary variables, here applied to columns of the Q matrix. Jaccard's index of similarity is a measure of association between two binary features, and is defined as the

ratio of the total number of observations (items in this case) where the feature is present for both items to the total number of items where the features are present for at least one item. The Jaccard index ranges from zero to one, where one indicates that one feature is present whenever the other feature is present, and zero indicates that whenever one feature is present, the other is not present.

Three conditions manipulating the Jaccard index were considered. First, all five features (q1 through q5) were considered as highly distinct, implying minimal confounding of features; that is, the Jaccard index was zero for every pair. This condition simulates the situation where each item is assigned one and only one feature. The second condition introduces a medium confounding between any two of the features (Jaccard index = .33), which is held constant across all pairs. Finally, for a third condition one pair of features has a high level of confounding, while the other features have lower levels of confounding. The upper triangular portions of the Jaccard index matrices for the three conditions are shown in Table 1.

 Insert Table 1 About Here

Variability of feature effects across countries. Five conditions were considered in

manipulating the W matrix. The number of features having large variability (i.e., the standard deviation of effects across countries is larger than .6) was varied from one through five, while all other features had small variability (i.e., the standard deviation of effects across countries is smaller than .2). Since normalized results are presented for interpretation, all of the row sums and column sums add up to zero in these W matrices, and all effects should be interpreted in a relative sense. The five conditions for variability of feature effects are illustrated in Table 2.

 Insert Table 2 About Here

Explanatory power of item features in explaining between-country DIF. As noted, when simulating data, the item difficulties for the countries relative to the reference country are determined by the Q and W matrices, allowing the item difficulties for the reference country to define difficulties against which those of comparison countries are compared. The explanatory power of the Q and W matrices can be manipulated by introducing a residual to the difficulty parameters of items for all comparison countries. By manipulating the variance of this residual, we can control how much of the variance in item difficulty for the comparison countries is explained by the specified features ($R^2 = .7$ vs. R^2

= .3). For example, when the variance of item difficulties for a fixed item across countries is .5, then the residuals of the item difficulties were generated from normal distributions having a mean of zero and variances of .21 and 1.17, producing R^2 values of approximately .7 and .3, respectively.

Data

The numbers of items and countries were fixed at 99 and 15, respectively, which provided the same condition as a previous real data study applied to TIMSS (Park & Bolt, in press). 500 examinees were used for each country. The number of examinees was reduced from the real data study (where 1,000 were sampled per country) due partly to relieve computational burden and also to reflect the matrix sampling design of the real data. Although 1,000 examinees were randomly sampled from each country, the selected examinees answered only a subset of the 99 items, and 500 examinees for each country seem reasonable. The number of item features was fixed at 5. Therefore, the Q matrix had a dimension of 99 by 5, and the W matrix had a dimension of 5 by 15. Using these fixed dimensions, the aforementioned three conditions were manipulated—confounding level in Q matrix, variability of feature weights, and the explanatory power of the features

in explaining variability in the item difficulties. Steps to generate item difficulties for the countries are illustrated in Figure 2. All simulated factors were fully crossed. Overall, therefore, 30 conditions were simulated, and five replications were conducted per combination of factor conditions. Although the number of replications may seem low, each condition results in 75 (15 times 5) unique elements in the W matrix; as a result, there were a large number of feature weight values generated across replications.

Insert Figure 2 About Here

Results

All MCMC analyses were conducted on the simulated datasets using the WinBUGS program developed by Park and Bolt (in press). Visual inspection of the chain histories and R-hat statistics suggested by Gelman and Rubin (1992) supported chain convergence. Recovery of item feature weights was of primary concern, and the recovery was evaluated using biases and root mean squared errors (RMSEs), which were calculated using the following formulas:

$$Bias(w) = \sum \frac{(\hat{w} - w)}{75},$$

$$RMSE(w) = \sqrt{\sum \frac{(\hat{w} - w)^2}{75}}.$$

Table 3 shows the bias of estimated item feature effects across all conditions averaged over replications. The mean bias across all simulation conditions considered in this study is effectively zero. It may thus be concluded that MCMC estimators of item feature effects provide unbiased estimates in various conditions when fitting the IFM. The RMSEs of the item feature effects averaged over replications show how accurately parameters were recovered for the simulation conditions (Table 4). The pattern of RMSE changes is more easily detectable when graphically displayed. Figures 3 and 4 show how RMSEs change as the number of features having large variability increases from one through five for different levels of feature confounding when the R^2 values are .7 (Fig. 3) and .3 (Fig. 4), respectively.

 Insert Tables 3 and 4 About Here

The difference between the larger explanatory power ($R^2 = .7$; Fig. 3) and the smaller explanatory power ($R^2 = .3$; Fig. 4) conditions is most noticeable when the features

are not confounded at all. When the explanatory power is high, the no confounding condition produced substantially lower RMSEs when compared to the other confounding conditions; however, when the explanatory power is low, the no confounding condition produced noticeably higher RMSEs and the increase was shaper as the number of features having large variability increased. Except for the no confounding condition, the pattern of change and the range of values were similar for the two explanatory power conditions. RMSEs tended to increase as the number of features having large variability increased; however, the increase was more noticeable when the item features were not confounded, especially the explanatory power was low ($R^2 = .3$).

Insert Figures 3 and 4 About Here

Among the three simulation factors, the most conspicuous changes were due to different confounding levels. An increased level of confounding substantially increased RMSEs across all variability of feature conditions when the explanatory power was large. When the R^2 value was .3, on the other hand, the effects of medium or high level of feature confounding were not as apparent as when the R^2 value was .7. The no feature confounding condition did not show much advantage over the other confounding conditions

as the number of features having large variability increased when the explanatory power was low. Still, feature confounding level is overall the most significant factor among the simulation factors considered in this study.

In summary, an increased level of confounding substantially deteriorated the recovery of item feature effects as shown in RMSE increases in almost all conditions. Also, RMSEs increased as the more features had large variability, and this pattern was more easily detectable when the features were not confounded, even more so when the explanatory power was low. The level of explanatory power (different R^2 values) did not cause big changes except when the features were not confounded. Among the three simulation factors considered in this study—confounding level in Q matrix, variability of feature effects, and the explanatory power of the features in explaining variability in DIF—confounding level in Q matrix tended to cause the largest changes in RMSEs.

Discussion and conclusion

This study conducted simulation analyses for various test conditions that were expected to impact use of the IFM. The IFM is a model that decomposes the sources of DIF with respect to a priori specified item features. We manipulated various conditions

for the Q and W matrices, both of which account for systematic variation. We also added conditions for different levels of explanatory power in explaining the sources of DIF by adding residual variability to the DIF. As a result, estimates of feature effects were unbiased for all conditions, and RMSEs were within reasonable ranges even for the most unfavorable conditions (small R^2 value, high level of feature confounding, all features having large variability). The best result (i.e., smallest RMSE) was obtained when random noise was small ($R^2 = .7$), when the features were not confounded, and when only one feature had large variability. RMSEs tended to increase as random noise became larger, the confounding level became higher, and more features had large variability. Among the three simulation factors, the most dramatic changes were made by the level of confounding.

The results are promising about the performance of IFM for group-level diagnosis because even in the worst conditions estimates of feature effects were unbiased and the RMSEs did not steeply increase. Among the three simulation factors, the feature confounding level, which caused the most dramatic changes in RMSEs, may be manipulated by the researchers who choose to use IFM as a model to obtain cross-group skill profiles. When simulating the conditions, the high confounding level was simulated

by manipulating the Jaccard index to be .8 for two binary features, which means both features are present 80% of the time either feature is present. In real data analyses, this high level of confounding will lead researchers to suspect that the two features are almost identical, and one of the features may be removed from the analyses. By removing high confounding features, the accuracy of recovering feature effects will increase substantially. It is also worth noting that only a small amount of random noise can dramatically decrease R^2 values because systematic variation is also small, which may be the reason for the similarities between the two explanatory power conditions. In real data analyses, even an R^2 value of .3 may be too high (R^2 value of the TIMSS data was about .17); however, the conclusion from the simulation analyses can be extended to the smaller explanatory power conditions since a small amount of random noise can significantly drop the value of R^2 .

The IFM is a confirmatory method that uses the prespecified Q matrix to diagnose group-level skill profiles. The results of this study will not only apply to the performance of the IFM but also to that of any group-level diagnostic methods based on Q matrix. For example, we may deduce that feature confounding may have a greater impact on the performance of Q matrix-based group-level diagnostic methods. This is also a limitation of the IFM from a methodologist's point of view. Since the performance of the IFM is

largely influenced by the specification of item feature incidences but the specification is conducted by item experts, the performance of the method, as a result, relies largely on item experts.

In education, useful information may often be obtained by comparing groups, and educational policy makers often require evidence based on group-level diagnosis (e.g., Tatsuoka, Corter, and Tatsuoka, 2004). Assessments such as the TIMSS, the NAEP, the Progress in International Reading Literacy Study (PIRLS), or the Programme for International Student Assessment (PISA) are all designed to facilitate comparisons among units above the student level, and are all good sources of information for group-level diagnostic inferences. Further study is needed to better understand the suggested model's performance on the TIMSS data, as well as to shed light on how to analyze general group-level diagnostic assessment data and what to expect from such analyses.

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.
- Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.
- Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472.
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*. 79-93.
- Martin, M.O., Mullis, I.V.S., & Chrostowski, S.J. (Eds.) (2004). *TIMSS 2003 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Park, C., & Bolt, D.M. (in press). Application of Multilevel IRT to Investigate Cross-National Skill Profile Differences on TIMSS 2003. *IEA-ETS Monograph Series*.
- 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.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Spiegelhalter D.J., Thomas A., Best N.G., & Lunn D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge.

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.

Table 1. Upper triangular Jaccard index matrices for three levels of confounding

| 1-1. No confounding | | | | | | |
|---------------------|------|------|------|------|-----|------|
| | q1 | q2 | q3 | q4 | q5 | |
| q1 | 1.00 | .00 | .00 | .00 | .00 | .00 |
| q2 | | 1.00 | .00 | .00 | .00 | .00 |
| q3 | | | 1.00 | .00 | .00 | .00 |
| q4 | | | | 1.00 | .00 | .00 |
| q5 | | | | | | 1.00 |

| 1-2. Medium confounding | | | | | | |
|-------------------------|------|------|------|------|-----|------|
| | q1 | q2 | q3 | q4 | q5 | |
| q1 | 1.00 | .33 | .33 | .33 | .33 | .33 |
| q2 | | 1.00 | .33 | .33 | .33 | .33 |
| q3 | | | 1.00 | .33 | .33 | .33 |
| q4 | | | | 1.00 | .33 | .33 |
| q5 | | | | | | 1.00 |

| 1-3. High confounding for two features with medium confounding for other features | | | | | | |
|---|------|------|------|------|-----|------|
| | q1 | q2 | q3 | q4 | q5 | |
| q1 | 1.00 | .80 | .46 | .46 | .46 | .46 |
| q2 | | 1.00 | .46 | .46 | .46 | .46 |
| q3 | | | 1.00 | .33 | .33 | .33 |
| q4 | | | | 1.00 | .33 | .33 |
| q5 | | | | | | 1.00 |

Table 2. Variability of features in W matrix

2-1. One feature with large variability

| Features Countries | q1 | q2 | q3 | q4 | q5 |
|-----------------------|-------|-------|-------|-------|-------|
| 1 | 0.48 | -0.12 | -0.12 | -0.12 | -0.12 |
| 2 | 0.56 | -0.14 | -0.14 | -0.14 | -0.14 |
| 3 | -0.72 | 0.18 | 0.18 | 0.18 | 0.18 |
| 4 | -0.48 | 0.12 | 0.12 | 0.12 | 0.12 |
| 5 | 0.64 | -0.16 | -0.16 | -0.16 | -0.16 |
| 6 | -0.72 | 0.18 | 0.18 | 0.18 | 0.18 |
| 7 | 0.56 | -0.14 | -0.14 | -0.14 | -0.14 |
| 8 | -0.64 | 0.16 | 0.16 | 0.16 | 0.16 |
| 9 | -0.40 | 0.10 | 0.10 | 0.10 | 0.10 |
| 10 | 0.72 | -0.18 | -0.18 | -0.18 | -0.18 |
| 11 | -0.64 | 0.16 | 0.16 | 0.16 | 0.16 |
| 12 | 0.56 | -0.14 | -0.14 | -0.14 | -0.14 |
| 13 | 0.40 | -0.10 | -0.10 | -0.10 | -0.10 |
| 14 | -0.72 | 0.18 | 0.18 | 0.18 | 0.18 |
| 15 | 0.40 | -0.10 | -0.10 | -0.10 | -0.10 |
| var(w) | 0.37 | 0.02 | 0.02 | 0.02 | 0.02 |
| SD(w) | 0.61 | 0.15 | 0.15 | 0.15 | 0.15 |

2-2. Two features with large variability

| Features | q1 | q2 | q3 | q4 | q5 |
|-----------|-------|-------|-------|-------|-------|
| Countries | | | | | |
| 1 | 0.62 | -0.68 | 0.02 | 0.02 | 0.02 |
| 2 | 0.72 | -0.78 | 0.02 | 0.02 | 0.02 |
| 3 | -0.88 | 0.82 | 0.02 | 0.02 | 0.02 |
| 4 | -0.64 | 0.76 | -0.04 | -0.04 | -0.04 |
| 5 | 0.78 | -0.72 | -0.02 | -0.02 | -0.02 |
| 6 | -0.88 | 0.82 | 0.02 | 0.02 | 0.02 |
| 7 | 0.70 | -0.70 | 0.00 | 0.00 | 0.00 |
| 8 | -0.78 | 0.72 | 0.02 | 0.02 | 0.02 |
| 9 | -0.54 | 0.66 | -0.04 | -0.04 | -0.04 |
| 10 | 0.88 | -0.82 | -0.02 | -0.02 | -0.02 |
| 11 | -0.82 | 0.88 | -0.02 | -0.02 | -0.02 |
| 12 | 0.68 | -0.62 | -0.02 | -0.02 | -0.02 |
| 13 | 0.52 | -0.58 | 0.02 | 0.02 | 0.02 |
| 14 | -0.88 | 0.82 | 0.02 | 0.02 | 0.02 |
| 15 | 0.52 | -0.58 | 0.02 | 0.02 | 0.02 |
| var(w) | 0.58 | 0.58 | 0.00 | 0.00 | 0.00 |
| SD(w) | 0.76 | 0.76 | 0.02 | 0.02 | 0.02 |

2-3. Three features with large variability

| Features | q1 | q2 | q3 | q4 | q5 |
|-----------|-------|-------|-------|-------|-------|
| Countries | | | | | |
| 1 | 0.96 | -0.44 | -0.64 | 0.06 | 0.06 |
| 2 | 0.44 | 0.24 | -0.56 | -0.06 | -0.06 |
| 3 | -0.90 | 0.40 | 0.50 | 0.00 | 0.00 |
| 4 | -0.42 | -0.42 | 0.88 | -0.02 | -0.02 |
| 5 | 0.98 | -0.72 | -0.42 | 0.08 | 0.08 |
| 6 | 0.32 | 0.62 | -0.58 | -0.18 | -0.18 |
| 7 | -0.50 | -0.40 | 0.90 | 0.00 | 0.00 |
| 8 | -0.38 | 0.72 | -0.38 | 0.02 | 0.02 |
| 9 | 0.38 | -0.62 | 0.68 | -0.22 | -0.22 |
| 10 | -0.80 | 0.90 | -0.30 | 0.10 | 0.10 |
| 11 | -0.56 | -0.66 | 0.74 | 0.24 | 0.24 |
| 12 | -0.42 | 0.88 | -0.42 | -0.02 | -0.02 |
| 13 | 0.82 | -0.38 | -0.48 | 0.02 | 0.02 |
| 14 | -0.36 | 0.74 | -0.46 | 0.04 | 0.04 |
| 15 | 0.44 | -0.86 | 0.54 | -0.06 | -0.06 |
| var(w) | 0.42 | 0.43 | 0.37 | 0.01 | 0.01 |
| SD(w) | 0.65 | 0.65 | 0.61 | 0.11 | 0.11 |

2-4. Four features with large variability

| Features | | | | | |
|-----------|-------|-------|-------|-------|-------|
| Countries | q1 | q2 | q3 | q4 | q5 |
| 1 | 0.60 | -0.70 | 0.80 | -0.70 | 0.00 |
| 2 | 0.72 | -0.78 | 0.92 | -0.78 | 0.02 |
| 3 | -0.86 | 0.84 | 0.64 | -0.56 | 0.04 |
| 4 | -0.66 | 0.74 | 0.54 | -0.76 | -0.06 |
| 5 | 0.80 | -0.70 | -0.90 | 0.80 | 0.00 |
| 6 | -0.86 | 0.84 | -0.56 | 0.84 | 0.04 |
| 7 | 0.64 | -0.66 | 0.74 | -0.76 | 0.04 |
| 8 | -0.80 | 0.70 | -0.90 | 0.80 | 0.00 |
| 9 | -0.52 | 0.68 | 0.48 | -0.62 | -0.02 |
| 10 | 0.84 | -0.86 | 0.54 | -0.76 | -0.06 |
| 11 | -0.80 | 0.90 | -0.80 | 0.70 | 0.00 |
| 12 | 0.70 | -0.60 | -0.90 | 0.80 | 0.00 |
| 13 | 0.56 | -0.64 | -0.84 | 0.86 | -0.04 |
| 14 | -0.90 | 0.80 | -0.50 | 0.70 | 0.00 |
| 15 | 0.54 | -0.56 | 0.74 | -0.56 | 0.04 |
| var(w) | 0.57 | 0.59 | 0.58 | 0.58 | 0.00 |
| SD(w) | 0.76 | 0.77 | 0.76 | 0.76 | 0.03 |

2-5. Five features with large variability

| Features | q1 | q2 | q3 | q4 | q5 |
|-----------|-------|-------|-------|-------|-------|
| Countries | | | | | |
| 1 | 0.98 | -0.42 | -0.62 | 0.68 | -0.62 |
| 2 | 0.54 | 0.34 | -0.86 | 0.74 | -0.76 |
| 3 | -0.88 | 0.42 | 0.52 | -0.88 | 0.82 |
| 4 | -0.42 | -0.52 | 0.78 | -0.62 | 0.78 |
| 5 | 0.90 | -0.50 | -0.50 | 0.80 | -0.70 |
| 6 | -0.46 | 0.84 | -0.36 | -0.86 | 0.84 |
| 7 | 0.28 | -0.62 | 0.68 | 0.48 | -0.82 |
| 8 | -0.36 | 0.74 | -0.36 | -0.76 | 0.74 |
| 9 | -0.46 | -0.46 | 0.84 | -0.56 | 0.64 |
| 10 | -0.30 | 0.80 | -0.60 | 0.90 | -0.80 |
| 11 | -0.90 | 0.40 | 0.40 | -0.80 | 0.90 |
| 12 | -0.44 | 0.86 | -0.44 | 0.66 | -0.64 |
| 13 | 0.88 | -0.32 | -0.72 | 0.68 | -0.52 |
| 14 | -0.36 | 0.74 | -0.36 | -0.86 | 0.84 |
| 15 | 0.44 | -0.86 | 0.64 | 0.44 | -0.66 |
| var(w) | 0.42 | 0.40 | 0.38 | 0.57 | 0.60 |
| SD(w) | 0.64 | 0.63 | 0.62 | 0.75 | 0.77 |

Table 4. RMSEs of estimated feature weights

| Number of Features Having Large Variability | R-squared = .70 | | | R-squared = .30 | | |
|--|-------------------|-----------------------|---------------------|-------------------|-----------------------|---------------------|
| | No Confounding | Medium Confounding | High Confounding | No Confounding | Medium Confounding | High Confounding |
| 1 | 0.05 | 0.19 | 0.25 | 0.14 | 0.20 | 0.26 |
| 2 | 0.11 | 0.25 | 0.32 | 0.24 | 0.25 | 0.34 |
| 3 | 0.11 | 0.23 | 0.28 | 0.26 | 0.25 | 0.30 |
| 4 | 0.14 | 0.27 | 0.34 | 0.36 | 0.30 | 0.37 |
| 5 | 0.16 | 0.24 | 0.28 | 0.35 | 0.29 | 0.34 |

Figure 1. Normalization procedure for 5 feature effects and 15 countries

| | Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 | | Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 | SUM | | | | | | | | | | |
|------------------|-----------|---------------------------------|-----------|-----------|-----------|--|-----------|-----------|-----------|-----------|-----------|-------------------------------|---|---|---|---|---|--|--|--|--|---|
| Country 1 (Ref.) | 0 | 0 | 0 | 0 | 0 | | | | | | | 0 | | | | | | | | | | |
| Country 2 | 0 | Unnormalized feature effects | | | | | | | | | | | 0 | | | | | | | | | |
| Country 3 | 0 | | | | | | | | | | | | | | | | 0 | | | | | |
| Country 4 | 0 | | | | | | | | | | | | | | | | 0 | | | | | |
| Country 5 | 0 | | | | | | | | | | | | | | | | 0 | | | | | |
| Country 6 | 0 | | | | | | | | | | | | | | | | 0 | | | | | |
| Country 7 | 0 | | | | | | | | | | | | | | | | 0 | | | | | |
| Country 8 | 0 | | | | | | | | | | | Normalized feature effects | | | | | 0 | | | | | |
| Country 9 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 10 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 11 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 12 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 13 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 14 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| Country 15 | 0 | | | | | | | | | | | | | | | | | | | | | 0 |
| | | | | | | | | | | | SUM | 0 | 0 | 0 | 0 | 0 | 0 | | | | | |

Normalization
→

Figure 2. Steps to generate difficulty parameters for the countries

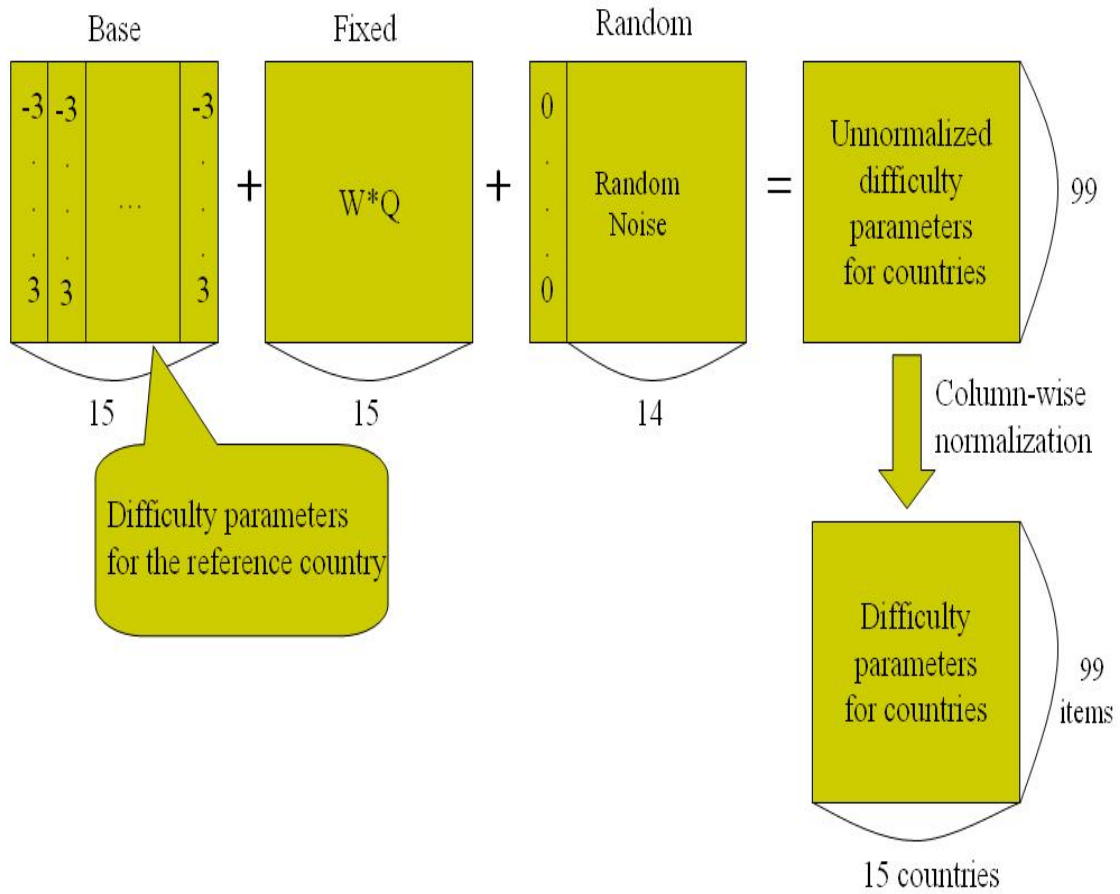


Figure 3. Mean RMSEs of feature weights for different levels of feature confounding as the number of features having large variability changes when $R^2 = .70$

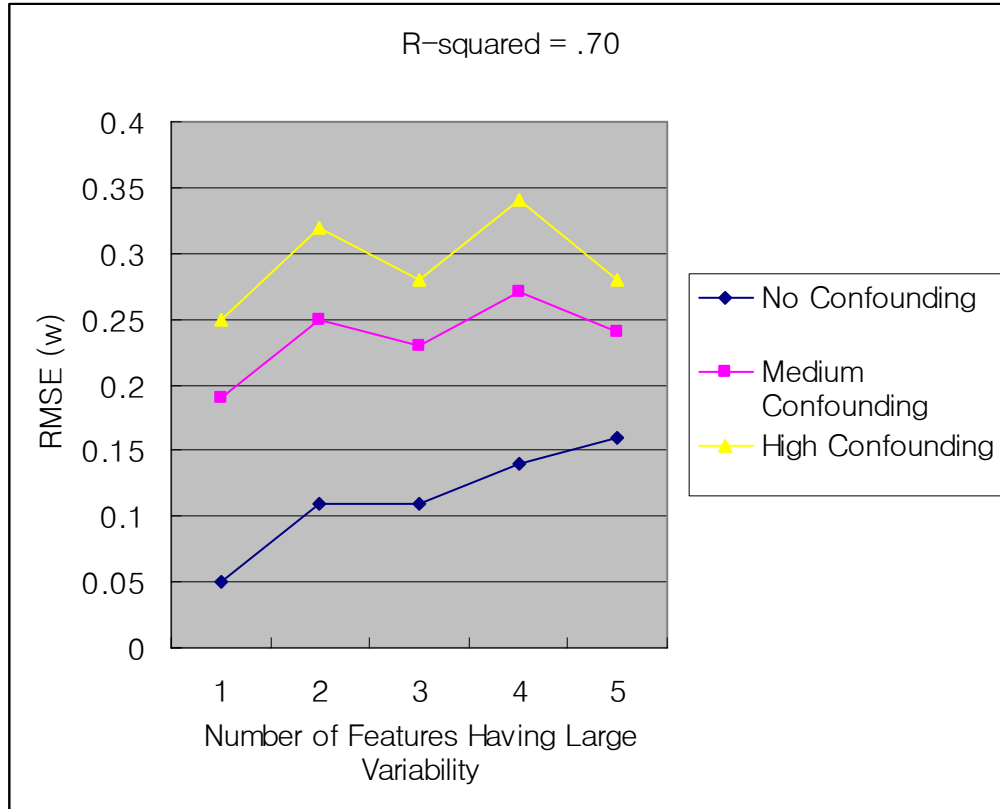


Figure 4. Mean RMSEs of feature weights for different levels of feature confounding as the number of features having large variability changes when $R^2 = .30$

