The Context Effects of Multidimensional CAT on the Accuracy of Multidimensional Abilities and the Item Exposure Rates

Yuan H. Li, Prince Georges County (Maryland) Public Schools
William D. Schafer, University of Maryland

Abstract

Under a MIRT CAT's (multidimensional computerized adaptive testing) testing scenario, an ability estimate in one dimension will provide clues for subsequently seeking a solution in other dimensions. This feature may enhance the efficiency of MIRT CAT's item selection as well as its scoring algorithms compared with its counterpart, unidimensional CAT (UCAT). However, when practitioners are planning to employ MIRT CAT on a real testing program, interesting problems present themselves. For the case of simultaneously measuring examinee's Reading and Mathematics abilities, will we administer to examinees Reading items first and Mathematic items next, Mathematics items first and Reading items next, or mixed items (e.g., a Reading item follows by a Mathematics item) ? Will the orders of administering different type of items to examinees make significant difference in terms of ability estimates and item exposure rates ? This sort of context effects in MIRT CAT never occurred in UCAT, but might happen in MIRT CAT. This issue is so critical and should be clarified before a real MIRT CAT program is implemented in place. The current research design intended to assess those context effects.

Presented at the 2004 annual meeting of the
American Educational Research Association, San Diego CA.

The Context Effects of Multidimensional CAT on the Accuracy of Multidimensional Abilities and the Item Exposure Rates

I. Introduction

A. Background of MIRT CAT

The computerized adaptive testing (CAT) involves the selection of the most informative test items from an item pool, so that each individual's latent trait is efficiently estimated with a short test. The present use of CAT heavily relies on unidimensional item response theory (IRT), simply assuming a single ability is necessary to account for examinees test performance on a test (Lord, 1980). This assumption is likely to make CAT produce biased ability estimates for a test dataset which is assumed unidimensional, but actually multidimensional. Further, when an examinee simultaneously takes several content-area (e.g., Reading and Mathematics) tests at a time as is usually required in the application of most academic programs, the cross-information (or correlation) of an examinee's knowledge among various content area can not be efficiently utilized in the process of CAT's item selection and scoring for the unidimensional-based CAT (UCAT).

On the other hand, the CAT based on the rationale of multidimensional IRT (MIRT, Reckase, 1985) might mitigate the problems that UCAT has encountered. Results from Li and Lissitz (2000) suggested that multidimensional IRT (MIRT) models can be applied to not only multidimensional data but also to unidimensional test data as well. It seems apparent that MIRT models are more flexible for fitting test data than unidimensional models. Further, when MIRT is accommodated within the context of computerized adaptive testing (CAT), the result (MIRT CAT) enhances the efficiency of adaptive item selection as well as scoring algorithms (Luecht, 1996, Segall, 1996, 2000). The primary reason for such promising features has been documented by Segall (1996; 2000), who pointed out that "When the dimensions measured by a test or battery are correlated, responses to items measuring one dimension provide clues about the examinee's standing along other dimensions" (Segall, 2000, p53). Such a unique characteristic, that can not be fulfilled in the conventional unidimensional CAT (UCAT), might make MIRT CAT more appealing, such as by increasing the reliability for an examinee's ability estimate(e.g., Luecht, 1996, Segall, 1996).

Results from Luecht (1996) and Segall (1996) MIRT CAT's studies indicated that a shorter MIRT CAT (about 25 % to 40%) could achieve about the same subscore reliability as its longer UCAT counterpart when multidimensional abilities are intercorrelated. Compared with UCAT, Li and Schafer's study showed that MIRT CAT increased the accuracy of ability estimates, especially for the low or high abilities, and reduced the rate of unused items in the item pool. Indeed, cross-information of an examinee's knowledge among various dimensions provides a better mechanism for choosing adaptive items for the examinee, whereas in multiple UCATs, cross-information among content areas is not utilized.

However, when practitioners is planning to employ MIRT CAT on a real testing program, interesting problems present themselves. For the case of simultaneously measuring examinee's Reading and Mathematics abilities, will we administer to examinees Reading items first and Mathematic items next, Mathematics items first and Reading items next, or mixed items (e.g., a Reading item follows by a Mathematics item) ? Will the orders of administering different type of items to examinees make significant difference in terms of ability estimates and item exposure rates ? This sort of context effects in MIRT CAT never occurred in UCAT, but might happen in MIRT CAT. This issue is so critical and should be clarified before a real MIRT CAT program is implemented in place.

B. Practical Problems of MIRT's Data Modeling

Although MIRT or its application in MIRT CAT has promising features over unidimensional IRT from a statistical viewpoint, MIRT has not yet been implemented in real testing programs. One primary reason is that unique solutions for estimated item loadings cannot be attained. The loadings are the coordinates on the axes (or dimensions) that define the space. Hence, rotating the axes will result in a new set of loadings. In addition, the interpretation of each dimension together with the determination of the number of dimensions for a given test dataset could be quite subjective, as is the case in factor analysis. The current MIRT computer software, TESTFACT 4 (Wood, Wilson, Gibbons, Schilling, Muraki & Bock, 2003), was programmed using the full information method. This program was designed to perform exploratory item-factor analysis although it might be used for a specific confirmatory factor analysis know as bi-factor analysis (Gibbons & Hedeker, 1992) which requires a factor pattern that should consist of one main factor plus group factors. This requirement might be limited to be used in some test data, but not suitable to all test data.

The confirmatory item factor-analysis modes in the MIRT context could help resolve these practical problems (McLeod, Swygert& Thissen, 2001). The FACT computer program (Segall, 1998), implementing the full information confirmatory item factor analysis using Markov chain Monte Carlo estimation, might be a suitable method for calibrating interpretable MIRT item parameters. The FACT program is, however, currently undergoing further investigation and is not available to the public. Existing available confirmatory item-factor analysis programs such as NOHARM (Fraser & McDonald, 1988) that models item-covariances rather than the full information approach has therefore become our choice for dealing with MIRT's item calibration. This program specifically allows constraints to be specified for the loading and/or covariances among the latent traits. Introducing the pattern matrices for those parameter matrices (e.g., indicating the values to be fixed at zero or at some other value) seems a flexible approach to seeking interpretable MIRT item parameters.

C. Research Purpose

As illustrated, UCAT might produce biased ability estimates when the condition of an under-fitted model occurs. Further, making use of ability estimates from other dimensions in item selection as well as in the scoring algorithms is precluded, making UCAT less efficient than MIRT CAT. With the compelling advantages of MIRT CAT (Segall, 2000), the intention in this study was to modify the current CAT to facilitate the process of locating Reading and Mathematics true abilities. The reason for choosing these two content areas as the example subject areas to be explored is that they are essential skills for students to be successful in all academic fields and they are often required to be taken at the same period of time in most testing programs. Of course, the methodology employed in these two content measures can be generalized to other content combinations (e.g., Reading and Science).

However, using MIRT CAT, instead of using UCAT, will present a unique and critical issue --- context effect, the accuracy of an examinee's ability estimates might depend on what order of items (e.g., Reading, Mathematic items, or mixed items) is administered to examinees. Further, this context effect might also have impact on the issue of item exposure rate, the ratio of the number of times an item has been administered to the total number of test-takers. The latter issue involves test security and becomes one of key issues of current CAT studies.

II. Methodology

A. Simulated Item Bank and Ability Parameters

We obtained a 765-item pool whose summary of item characteristics is presented in the full-version paper. Five hundred simulees were randomly selected from multivariate normal distribution, MVN ($\underline{0}$, **Φ**), where **Φ** is the population varaince-covarance matrix of the Reading and Math dimensions. The **Φ** is obtained from the NOHARM. The covariance value of both content-area scores was .729 which is the off-diagonal value of **Φ** and the both diagonal values of **Φ** are 1. This sample was used for evaluating the accuracy of MIRT CAT's ability estimates. Similarly, another five thousand simulees were generated and then used for evaluating the exposure rate of the item pool items.

B. Ability Estimates and Item Selection

The Bayesian model approach (Segall, 1996) was chosen for estimating multidimensional abilities. This method will incorporate the prior information of highly correlated Reading and Math measures into the likelihood function so that an ability estimate (e.g., Reading) in one dimension will provide clues for subsequently seeking a solution in other dimensions (e.g., Math). The **Φ** matrix, as mentioned above, was also used for the prior varaince-covaraince while estimating abilities. The initial abilities of all simulated subjects at the beginning of the test were taken as an vector that was randomly drawn from a multivariate normal distribution, MNV($\underline{0}$, **Φ**). The MIRT CAT stopped when the fixed test length was reached.

The maximum DPI (the determinant of the posterior information) criterion (Segall, 2000) together with the shadow-test approach (van der Linden, 2000) was chosen for item-selection method in this study because the maximum DPI criterion selects items to maximize accuracy along all dimensions simultaneously (Segall, 2001), and the shadow-test approach ensures that content balance was achieved. The rationale of DPI and the shadow-test MIRT CAT will be illustrated in the full-version paper.

C. Simulation Conditions

There were 29 constraints imposed to assemble an on-line shadow test. These 29 constraints (listed in full-version paper) corresponded to the 29 objectives of the test specifications that were used for editing the CTBS Reading and Math tests. Hence, the number of items for each objective on the shadow test was constrained as in the original test.

Three simulation conditions are listed in Table 1. The first condition allowed simulees to take the MIRT CAT test in which Reading items were mixed with Mathematics items; whereas, the second condition forced simulees to take Math items first, and Reading items next. The third condition, on the other hand, forced simulees to take Reading items first, and Math items next. The comparisons of results among these three conditions were used to explore which item order would result in the most accurate ability estimates and the most homogeneous item-exposure rates.

Table 1. Research Conditions being Simulated

Conditions	Item Order	Total Test Length	Reading Test Length	Mathematics Test Length	Progress
1	Mixed Items	51	25	26	Finished
2	Math, Reading	51	25	26	None
3	Reading, Math	51	25	26	Finished

D. Evaluation

One hundred replications for each condition were conducted. The BIAS along with the RMSE (root mean squared error) statistics of the ability estimates across these two simulation conditions were computed by the formulas shown below.

$$\mathrm{BIAS}(\theta_i) = \frac{\sum_{i=1}^{r}(\hat{\theta}_i - \theta_i)}{r} \quad \text{and} \qquad (1)$$

$$\mathrm{RMSE}(\theta_i) = \sqrt{\frac{\sum_{i=1}^{r}(\hat{\theta}_i - \theta_i)^2}{r}} \qquad (2)$$

where θ_i is the true ability parameter, $\hat{\theta}_i$ is the corresponding estimated ability parameter, and r is the number of replications, which was 100 in this study.

RMSE is a measure of total error of estimation that consists of the systematic error (BIAS) and random error (SE). These three indexes relate to each other as follows (Rao, 2000):

$$\mathrm{RMSE}(\theta_j)^2 \cong \mathrm{SE}(\theta_j)^2 + \mathrm{BIAS}(\theta_j)^2 \qquad (3)$$

As can be seen from Equation 3, either a large variance (SE^2) or a large BIAS will produce a large RMSE. It is apparent that an estimator will have much practical utility only if it must not only be highly precise (or small SE^2), but also has small BIAS (Rao, 2000). The accuracy of an estimator is inversely proportional to its RMSE so that this RMSE index is the criterion of accuracy for an estimator (Rao, 2000). Accordingly, this index was primarily used to compare the accuracy of ability estimates when they were estimated under various simulation conditions.

In term of comparing item-exposure rates, the following indices (refer to Revuelta & Ponsoda, 1998) will be used to compare the three CAT algorithms: (a) the percentage of items never administered in the population, (b) the standard deviation (SD) of the variable of the item-exposure rate, (c) the minimum and maximum values of this variable. The distribution of the item-exposure rates, grouped in several intervals, will be also computed for each CAT condition.

E. Computer Program

The computer program MIRTCAT was used for running the simulation conditions. The MIRTCAT was coded by the MATLAB matrix language (The MathWorks, 2001), in which the 0-1 linear programming was resolved from the callable library of LINDO API (LINDO Systems, Inc. 2001).

We expect this study will be finished by the end of December 2003.

References

Fraser, C & McDonald, R. P. (1988). NOHARM: Least Squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Gibbons, D. D., & Hedeker, D. R. (1992). Full information item bi-factor analysis. Psychometrika, 57, 423-436.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. Applied Psychological Measurement, 24, 115-138.

LINDO Systems, Inc. (2001). LINDO API: The premier optimization engine. [Computer program]. Chicago Illinois: INDO Systems, Inc.

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

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. Applied Psychological Measurement, 20, 389-404.

McLeond, L.D., Swygert, K. A, & Thissen, D. (2001). Factor analysis for items scored in two categories D. Thissen and H. Wainer (eds.), Test Scoring, 189-216.Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Rao, P. S. (2000). Sampling methodologies with applications. New York: Chapman & Hall/CRC.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35, 311-327.

Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61, 331-354.

Segall, D. O. (1998). IFACT computer program Version 1.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [computer program]. Seaside, CA: Defense Manpower Data Center.

Segall, D. O. (2000). Principles of Multidimensional Adaptive Testing. W. J. van der Linden and C. A. W. Glas (eds.), Computerized Adaptive Testing: Theory and practice, 53-57. Dordrecht, The Netherlands: Kluwer Academic Publishers.

The MathWorks, Inc. (2001). MATLAB (Version 6.1): The language of technical computing [Computer program]. Natick MA: The MathWorks, Inc.

van der Linden, W. J. (2000). Constrained Adaptive Testing with Shadow Tests. W. J. van der Linden and C. A. W. Glas (eds.), Computerized Adaptive Testing: Theory and practice, 27-52. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wood, R., Wilson, D., Gibbons, R, Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT 4: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.