STRATIFIED ITEM SELECTION AND EXPOSURE CONTROL IN
UNIDIMENSIONAL ADAPTIVE TESTING IN THE PRESENCE OF TWO-
DIMENSIONAL DATA

Kevin E. Kalinowski, B.S., M.S.

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2009

APPROVED:

Robin K. Henson, Major Professor
James G. Jones, Minor Professor
Darrell M. Hull, Committee Member
Jon I. Young, Program Coordinator for
Educational Research and Interim
Chair of the Department of Educational
Psychology
Jerry R. Thomas, Dean of the College of
Education
Michael Monticino, Dean of the Robert B.
Toulouse School of Graduate Studies

Kalinowski, Kevin E., Stratified item selection and exposure control in unidimensional adaptive testing in the presence of two-dimensional data. Doctor of Philosophy (Educational Research), August 2009, 47 pp., 7 tables, 14 illustrations, references, 38 titles.

It is not uncommon to use unidimensional item response theory (IRT) models to estimate ability in multidimensional data. Therefore it is important to understand the implications of summarizing multiple dimensions of ability into a single parameter estimate, especially if effects are confounded when applied to computerized adaptive testing (CAT). Previous studies have investigated the effects of different IRT models and ability estimators by manipulating the relationships between item and person parameters. However, in all cases, the maximum information criterion was used as the item selection method. Because maximum information is heavily influenced by the item discrimination parameter, investigating *a*-stratified item selection methods is tenable.

The current Monte Carlo study compared maximum information, *a*-stratification, and *a*-stratification with *b* blocking item selection methods, alone, as well as in combination with the Sympton-Hetter exposure control strategy. The six testing conditions were conditioned on three levels of interdimensional item difficulty correlations and four levels of interdimensional examinee ability correlations. Measures of fidelity, estimation bias, error, and item usage were used to evaluate the effectiveness of the methods. Results showed either stratified item selection strategy is warranted if the goal is to obtain precise estimates of ability when using unidimensional CAT in the presence of two-dimensional data. If the goal also includes limiting bias of the estimate, Sympton-Hetter exposure control should be included. Results also

confirmed that Simpson-Hetter is effective in optimizing item pool usage. Given these results, existing unidimensional CAT implementations might consider employing a stratified item selection routine plus Simpson-Hetter exposure control, rather than recalibrate the item pool under a multidimensional model.

Copyright 2009

by

Kevin E. Kalinowski

Since Lord (1970) established modern methodological foundations for computerized adaptive testing (CAT), the literature on the subject has burgeoned from the psychometric community. Capitalizing on the research, assessment companies have migrated popular paper-and-pencil tests to CAT, such as with Educational Testing Service's Graduate Record Examination® (GRE®; Schaeffer, Reese, Steffen, McKinley, & Mills, 1993). Central to most CAT implementations is item response theory (IRT), which posits that an underlying latent ability (θ) explains correct responses to items of varying difficulty (Wainer & Mislevy, 2000). Furthermore, the unidimensional IRT models used in CAT assume item response patterns can be described for the most part by examinee ranking on the same one-dimensional latent construct on which the items are based (Lord, 1980). That is to say, the item pool is unidimensional on the same scale as examinee ability. Consequentially, departures from unidimensionality in the item pool can pose a threat to the validity of IRT-based tests (Steinberg, Thissen, & Wainer, 2000).

It is not uncommon for unidimensional IRT models to be applied to multidimensional data (Way, Ansley, & Forsyth, 1988). Regardless of the model used, item calibration (estimating item parameters based on examinees' response patterns) is a costly endeavor (Wainer & Mislevy, 2000). Therefore, if a unidimensional model was initially used in item calibration, and it was subsequently found that the items were in fact multidimensional in nature, considerable resources would be wasted calibrating the items under a new model. As an alternative to recalibrating the item pool and attempting to link the old unidimensional ability scale to a new multidimensional ability scale, it would be beneficial if a portion of an operational CAT framework, such as the item

selection routine, could be exchanged for a method that was robust in violations of unidimensionality.

Prior studies have investigated the effects of unidimensional adaptive testing in the presence of multidimensional data. However the maximum information strategy, a common item selection method in CAT implementations (Chang & Ying, 1999), was the only method used to select items to be administered to the examinees. The purpose of this Monte Carlo study was to respond to stated limitations of previous studies (cf. Weiss & Suhadolnik, 1982; Zhao, McMorris, and Chen, 2002) by exploring the effectiveness of various the item selection and exposure control strategies in adaptive testing as a possible means of compensating for two-dimensional data under unidimensional CAT.

Review of the Literature

Item Response Studies

Several Monte Carlo studies have investigated the robustness of the unidimensional IRT model to data of varying types and degrees of multidimensionality. Reckase (1979) used data from a state aptitude test as well as generated data to fit specific factor structures under investigation. He concluded both the one-parameter logistic (1PL) model and the three-parameter logistic (3PL) model adequately estimated parameters for the first principal component when it was large relative to the other factors present. Furthermore, when the first factor accounted for more than twenty percent of the test variance, item calibration parameter estimates were stable. Relying only on simulated data used from a hierarchical factor-analytic model, Drasgow and Parsons (1983) explored the effects of a second-order factor as the general latent trait

controlling the correlations of five first-order common factors. They reported the unidimensional model provided adequate parameter estimation of multidimensional data provided that the dominant latent trait is sufficiently strong.

Using standardized achievement test data as a frame of reference, Ansley and Forsyth (1985) generated two-dimensional data using a noncompensatory multidimensional IRT (MIRT) model. Unlike compensatory models, where a strong ability on one dimension can compensate for a weak ability on another dimension, the weakest ability is the upper limit for the likelihood of answering correctly in noncompensatory models. They concluded the estimated ability ($\hat{\theta}$) was highly related to the average of the ability on the first dimension (θ_1) and second dimension (θ_2). Way et al. (1988) expanded Ansley and Forsyth's study by including data sets based on compensatory as well as noncompensatory models. They replicated prior results for both types of data even though the two models produced two-dimensional data that yielded moderately different unidimensional discrimination and difficulty estimates. Compensatory and noncompensatory MIRT models were also investigated by Ackerman (1989), who found the relationship between $\hat{\theta}$ and the true θ_1 and θ_2 values was stronger than in previous studies. Furthermore, he reported the data became more unidimensional in nature as the relationship between the two ability dimensions became more linear.

Sheng and Wikle (2007) compared the performance of a two-parameter unidimensional IRT model against a multiunidimensional model (also known as the MIRT model with simple structure or the between-item MIRT model). The multiunidimensional model was developed to represent data from an overall ability test

consisting of several subtests, each focusing on a specific ability (e.g., a mathematics test consisting of multiplication, division, and factorization subtests). They found both models worked well when the latent abilities are highly correlated, but the multiunidimensional model was more accurate in ability estimation when the abilities were less correlated.

Adaptive Testing Studies

In addition to IRT Monte Carlo studies, researchers have explored the effects of using unidimensional IRT models in a CAT environment when the item pool is multidimensional. After factor analyzing the response patterns from two forms of the Armed Services Aptitude Battery (ASVAB) Weiss and Suhadolnik (1982) generated multidimensional data based on the results. They found as multidimensionality increased, estimated ability deviated from true ability values, as evident in measures of bias, accuracy, and root mean square error (*RMSE*) values regardless of the degree of dimensionality. However, these effects were more pronounced with multiple factors accounting for nearly equivalent variance in the data rather than with a single strong factor relative to another. Furthermore, they found that these effects were negatively correlated to the number of items administered, with many of the negative effects acceptably diminished by doubling the test length.

Folk and Green (1989) simulated two-dimensional data using a compensatory MIRT model to study the effects of multidimensional between-items tests (i.e., orthogonal ability vectors) and multidimensional within-items tests (i.e., oblique ability vectors). They concluded for between-item samples, estimated ability was more strongly related to either θ_1 or θ_2 as interdimensional ability correlations ($r_{\theta_1\theta_2}$)

decreased. However, for within-item samples, $\hat{\theta}$ was equally related to θ_1 and θ_2 regardless of $r\theta_1\theta_2$.

Using a two-dimensional compensatory IRT model Ackerman (1991) generated data to investigate whether examinees of differing ability would receive parallel adaptive tests with regard to θ_1 and θ_2 compositions. As with results from previous studies, he found estimated abilities were best represented by a linear combination of θ_1 and θ_2 . Furthermore, because item parameters are estimated, Ackerman reported items most likely to be selected by the CAT administration were inevitably predetermined by the unidimensional calibration process, which in turn masked the multidimensional nature of the item pool.

de Ayala (1992) simulated two-dimensional data sets that differed with regards to $r\theta_1\theta_2$, the correlation between the two dimensions of item difficulty (rb_1b_2), and the presence of a confounding condition where as difficulty increased, the items discriminated progressively less along θ_1 and proportionally more along θ_2 (cf., Ackerman, 1989). He reported as rb_1b_2 increased the accuracy of ability estimation, as measured by *RMSE*, increased. In addition, correlations between $\hat{\theta}$ and the θ_1 and θ_2 parameters, as well as their average ($\bar{\theta}$), increased as rb_1b_2 and $r\theta_1\theta_2$ increased. Furthermore, he concluded that for item selection in CAT, the distribution of the difficulty parameter (b), plus the magnitudes of item discrimination parameter (a) and pseudo-guessing parameter (c), are of primary importance.

To simulate items from a standardized achievement test where one dimension of ability is dominant over other dimensions, Zhao et al. (2002) generated two-dimensional data sets such that the discrimination parameters on the first dimension were twice as

large as the second dimension on average (cf. Ansley & Forsyth, 1985). As with previous studies Zhao et al. found the estimated abilities to be approximate averages of the true two-dimensional abilities for both compensatory and noncompensatory MIRT models regardless of how dominant one dimension is over the other. In addition, they found the interdimensional item difficulty correlations had little impact on CAT, but $r_{\theta_1\theta_2}$ did have effects on CAT of varying degree. Furthermore, they reported that the three Bayesian ability estimation methods they tested (Owen's (1975) Bayesian sequential, Samejima's (1969) Bayesian modal, and Bock and Aitken's (1981) Bayes expected a posteriori (EAP) estimators) outperformed maximum likelihood estimation (MLE), but they found it impossible to distinguish between the two better performing, Bayesian modal and Bayes EAP, estimators.

Summary of Existing Studies

Consistent across the studies is the empirical relationship between estimated ability and the true ability parameters. Most of the aforementioned studies found positive correlations of varying degree between $\hat{\theta}$ and the average of the true ability parameters, θ_1 and θ_2 , under a variety of testing conditions, and with a range of effects on the precision of estimation. However, many adaptive testing researchers suggest item selection may also have some impact on results, warranting further investigation (Ackerman, 1991; de Ayala, 1992; Folk & Green, 1989; Weiss & Suhadolnik, 1982; Zhao et al., 2002). In all the adaptive testing simulation studies reviewed, the maximum information (MAXI) strategy was used to select items to be administered to the examinees.

Item Selection Based on Maximum Information

In a 3PL IRT model, the probability of an examinee correctly answering an item given some degree of latent ability (θ) is

$$P(X = 1 | \theta) = c + (1 - c) \frac{1}{1 + e^{-a(\theta - b)}}, \quad (1)$$

where a is the item discrimination parameter, b is the item difficulty parameter, and c is the pseudo-guessing parameter for the item (Lord, 1980). Item information at the examinee's current estimated ability level ($\hat{\theta}$) is subsequently defined as

$$I(\hat{\theta}) = \frac{(1 - c)a^2 e^{a(\hat{\theta} - b)}}{\left[1 + e^{a(\hat{\theta} - b)}\right]^2 \left[1 - c + c(1 + e^{a(\hat{\theta} - b)})\right]} \quad (2)$$

(Chang & Ying, 1999). If m items have already been administered in CAT, an information-based selection rule selects the next item such that I at $\hat{\theta}_m$ is maximized (Chang & Ying, 1996). Hau and Chang (2001) demonstrated for $c = 0$ (i.e., no guessing is modeled) and a fixed a , item information reaches the maximum value, $a^2 / 4$, when the item difficulty equals the examinee's true ability ($b = \theta$). So using the maximum information criteria, if an examinee's true ability is known, the next item selected will efficiently be the one with a difficulty closest to θ and with the largest discrimination parameter. However, Chang and Ying (1999) argued that the expected gain in efficiency may not be attained because $\hat{\theta}$ is used as a continuously updating estimate of θ during CAT. In addition, maximum information risks overexposing highly discriminating items and not using weakly discriminating items, thereby decreasing test security and increasing test cost (Chang & Ying, 1999). As a result, dozens of alternative item

selection methods have been developed, especially for early stages of CAT when a small number of responses may yield poor estimates of ability.

Alternative Selection Strategies

Georgiadou, Triantafillou, and Economides (2007) classified thirty-one item selection strategies appearing in the literature from 1983 to 2005 into five categories. Randomization strategies attempt to limit item exposure rates by randomly selecting items from a subset of the item pool. Conditional selection strategies seek to control the probability of item administration, often by way of item exposure control parameters, which are computed prior to operational CAT through a series of iterative simulations. Stratification strategies attempt to control item exposure by subdividing the item pool into various strata, and then selecting items at their most useful point in testing rather than on maximum information. Combined strategies involve one or more of the above approaches in the attempt to develop a method performing better than each strategy by itself. Finally, multiple stage adaptive test designs control item exposure prior to administration through the tight integration of test development, assembly, selection, administration, scoring, and data management systems. Of the various strategies available, those based on stratification of the item pool, such as *a*-stratification and *a*-stratification with *b*-blocking, and conditional selection, such as the Simpson-Hetter procedure, appear most likely to have practical impact on CAT under violations of unidimensionality. These strategies do not involve item selection decisions based solely on maximum information, and they are relatively easy to implement in an existing adaptive test, either alone or in conjunction with each other, without affecting dimensionality assumptions.

a-Stratification. The *a*-stratified strategy (STRA) is a multistage item selection method that partitions an item pool into K strata according to increasing a values, as illustrated in Figure 1, then partitions the CAT into a similar number of stages such that for each stage n number of items are administered based on the similarity between b and $\hat{\theta}$ rather than on maximizing item information (Chang & Ying, 1999). As a result, weakly discriminating items in lower strata are used earlier in CAT when estimating $\hat{\theta}$ is less precise, and strata of highly discriminating items are used toward the end of the test for more accurate estimations of ability. In their simulation study, Chang and Ying found STRA successfully decreased exposure of highly discriminating items and increased the exposure of items with low a values, while substantially reducing the average number of overlapping items between examinees. Furthermore, the precision of STRA, as measured by mean squared error (MSE) and the bias of ability estimates, was comparable to two selection methods based on the Simpson-Hetter (Simpson & Hetter, 1985) method, one incorporating maximum information and the other using a Bayes algorithm.

Figure 1. Illustration of the *a*-stratification procedure.

a-Stratification with *b*-blocking. Chang, Qian, and Ying (2001) refined the *a*-stratified design to include a blocking factor for the *b* parameter (STRB), as illustrated in Figure 2. The design is to force a balanced distribution of *b* values in each stratum, increasing the likelihood of a match between *b* and $\hat{\theta}$ for each examinee, while maintaining the progressively increasing *a* values across strata as found in STRA. They concluded when the *a* and *b* parameters are perfectly uncorrelated, an unlikely occurrence in operational CAT according to Wingersky and Lord (1984), STRA performed as well as STRB. However, when *a* and *b* are correlated, as substantiated by Chang et al. with evidence of an item bank from a GRE quantitative test ($r_{ab} = .44$), STRB outperformed STRA in reducing item overexposure rates, while increasing item usage and improving precision in ability estimation (i.e., lower *MSE* and *bias*), mainly due to the larger overlap of strata across abilities (cf. Figures 1 and 2). Also, correlations between $\hat{\theta}$ and θ for the two methods were similar on average, but STRB had a stronger relationship for $\theta < -1.95$, suggesting STRB might improve ability estimation for examinees with extreme abilities.

Figure 2. Illustration of the *a*-stratification with *b*-blocking procedure.

Other stratified designs. As an extension to a -stratification with b -blocking, Yi and Chang (2003) included a stratification based on predetermined item content specifications. Chang and van der Linden (2003) proposed a computationally intensive alternative to Yi and Chang's b blocking design using 0-1 linear programming to force balanced distributions of b values across strata. Additionally, Lee, Ip, and Fuh (2008) adapted Chang and Ying's (1999) a -stratification design for multiple dimensions.

Sympson-Hetter. Sympson and Hetter (1985) developed a strategy to control the probability that an item is administered, $P(A)$, by calibrating the probability the item is administered conditioned on it being selected, $P(A|S)$, also known as the item exposure parameter, k . The values for k are determined through a series of iterative CAT simulations where k is adjusted based on observing the probability the item is selected, $P(S)$. After the k values have been determined, the SH strategy is used in operational CAT by comparing a selected item's k to a random number drawn from $U(0,1)$. If the random number is less than or equal to k the item is administered, else it is marked as "used" and a new item is selected. SH is the most common method for item exposure control in CAT (Chang & Ying, 1999; Hau & Chang, 2001), and is popular because it can be applied in addition to most any item selection strategy (Leung, Chang, & Hau, 2002). However, Leung et al. showed although SH used in conjunction with MAXI (MAXI+SH) was slightly more efficient in ability estimation (i.e., increased $r\hat{\theta}$) than STRA or STRA+SH, MAXI+SH consistently had greater number of underutilized and overlapping items for examinees of average ability.

Other conditional selection strategies. To help minimize the clustering of items, Davey and Parshall (1995) extended SH by incorporating conditional item exposure

parameters. Stocking and Lewis (1995, 1998) developed alternative SH-based methods based on multinomial probabilities to attempt to remove dependence on the distribution of abilities. To avoid complexities involved in calibrating the k values required in SH-based strategies and to help balance item content across examinees, van der Linden and Chang (2003) proposed a method involving item selection from optimal shadow tests, which are continuously regenerated during test administration based on the remaining item pool and the current estimate of an examinee's ability.

Purpose of the Current Study

In general, researchers observed positive relationships between estimated ability and a linear combination of the true ability values in two or more dimensions, which strengthened as interdimensional examinee ability correlation strengthened (Ackerman, 1989, 1991; Ansley & Forsyth, 1985; de Ayala, 1992; Folk & Green, 1989; Way et al., 1988; Weiss & Suhadolnik, 1982; Zhao et al., 2002). Additionally, researchers found substantial relationships between interdimensional ability correlations and measures of *bias* and *RMSE*, as well as weaker relationships between these measures and interdimensional item difficulty correlations (de Ayala; Zhao et al.). However, these results were confounded in adaptive testing because in all cases reviewed, maximum information was the only item selection method used. As such, the selection algorithm consistently selected the most discriminating items with difficulties nearest the examinees' estimated ability, primarily due to the dominance of the a^2 term in the numerator of the item information function (see Equation 2). As the selected items tended from a single dimension, the overall test became a test primarily measuring that

dimension, and as such, true ability that corresponded to the over-selected dimension was most related to estimated ability (Folk & Green).

The purpose of this Monte Carlo study was to respond to stated limitations of prior studies investigating the effects of unidimensional adaptive testing in the presence of multidimensional data, by varying the item selection and exposure control strategies implemented in CAT (cf. Weiss & Suhadolnik, 1982; Zhao et al., 2002). Following simulation and assessment guidelines set forth by previous studies, as described below, the current study compared three selection methods, maximum information (MAXI), *a*-stratification (STRA), and *a*-stratification with *b* blocking (STRB). Each method was investigated by itself, as well as in combination with the Simpson-Hetter (SH) exposure control strategy. The six possible item selection / exposure control combinations were conditioned on three levels of interdimensional item difficulty correlations ($rb_1b_2 = 0.0, 0.6, 0.9$) and four levels of interdimensional ability correlations ($r\theta_1\theta_2 = 0.0, 0.3, 0.6, 0.9$). Table 1 lists a summary of the testing conditions.

Table 1
Manipulated Conditions in the Current Study

<i>Item Selection Method</i>	<i>Exposure Control Strategy</i>	<i>Item Difficulty Correlation (rb_1b_2)</i>	<i>Examinee Ability Correlation ($r\theta_1\theta_2$)</i>
MAXI	None	0.0	0.0
STRA	SH	0.6	0.3
STRB		0.9	0.6
			0.9

To evaluate the effectiveness of the six selection/exposure combinations in the presence of twelve correlation conditions, two main areas of analyses were conducted, those involving ability estimation and those involving item usage. To measure the effectiveness of ability estimation, correlations between estimated abilities and true

abilities, also known as fidelity coefficients, were computed. In addition, mean square error (*MSE*), estimation bias, and average standard error of the estimate (*ASE*) were calculated. Item usage was measured by determining the number of items overexposed and underutilized, as well as computing test overlap rates and an overall item pool usage efficiency index. Table 2 shows the measured variables in the current study.

Table 2
Evaluation Criteria in the Current Study

<i>Ability Estimation</i>	<i>Item Usage</i>
Fidelity coefficients	Number of items overexposed
Correlation between $\hat{\theta}$ and θ_1	Number of items underutilized
Correlation between $\hat{\theta}$ and θ_2	Test overlap rate
Correlation between $\hat{\theta}$ and $\bar{\theta}$	Item pool usage efficiency index (χ^2)
Mean square error (<i>MSE</i>)	
Bias	
Average standard error of the estimate (<i>ASE</i>)	

Method

Two-Dimensional Data Generation

Item response model. Ackerman (1989) and Way et al. (1988) found little difference in results between data produced by two-dimensional compensatory and noncompensatory MIRT models. Therefore, Doody-Bogan and Yen's (1983) multidimensional compensatory model, modified for two-dimensions (θ_1, θ_2), was used to generate the item pool,

$$P(X = 1 | \theta_1, \theta_2) = c + \frac{1 - c}{1 + e^{-(a_1(\theta_1 - b_1) + a_2(\theta_2 - b_2))}}, \quad (3)$$

which is an extension of the unidimensional 3PL model presented in Equation 1. All symbols for the two-dimensional model represent the same parameters as for the unidimensional model, except that each examinee has two ability parameters (θ_1 and

θ_2), and each item has two discrimination parameters (a_1 and a_2) and two difficulty parameters (b_1 and b_2).

Item pool. As with de Ayala (1992), difficulty parameters were generated such that for sets of four items, b_1 was fixed at every 0.1 logit between -3.5 and 3.5 , thus producing four items at 71 levels, or 284 total items in the pool. The item's difficulty on the second dimension was computed from b_1 by using a linear transformation of a random sampling from a uniform distribution via Cholesky decomposition (Gentle, 1998) of the desired intercorrelations, 0.0, 0.6, and 0.9. The resulting parameter pairs had actual correlations of -0.040 , 0.626 , and 0.915 , respectively, and are graphically displayed in Figure 3. Furthermore, one set of discrimination parameters (a_1 and a_2) was created by randomly sampling from a uniform distribution in the range of 0.2 to 1.8. As with Ansley and Forsyth (1985) and de Ayala, the pseudo-guessing parameter, c , was set to 0.2 for all items.

Figure 3. Scatterplots for three levels of interdimensional item difficulty correlations used in the current study.

Simulated examinees. True abilities on the first dimension for 1,600 examinees (cf. de Ayala, 1992) were randomly sampled from a standard normal distribution. The

examinees' ability on the second dimension was computed from θ_1 by using a linear transformation of a random sampling from a uniform distribution via Cholesky decomposition of the desired intercorrelations, 0.0, 0.3, 0.6, and 0.9. Actual correlations were 0.029, 0.285, 0.585, and 0.891, respectively, and are graphically displayed in Figure 4.

Figure 4. Scatterplots for four levels of interdimensional examinee ability correlations used in the current study.

Item response matrices. Using the generated item and ability parameters, the two-dimensional compensatory model (Equation 3) was used to create twelve (three levels of rb_1b_2 and four levels of $r_{\theta_1\theta_2}$) examinee (i) by item (j) matrices, with elements p_{ij} , representing the probability of correctly answering item j given abilities, θ_1 and θ_2 .

A random number matrix of equal order (1,600 x 284) was generated with elements r_{ij} from $U(0,1)$. The twelve item response matrices were generated with elements x_{ij} such that

$$x_{ij} = \begin{cases} 1 & \text{if } p_{ij} \geq r_{ij} \\ 0 & \text{if } p_{ij} < r_{ij} \end{cases} . \quad (4)$$

This procedure was repeated four times with unique random number matrices, producing a total of five item response sets for each of the twelve difficulty-ability levels in the study. CAT simulations were performed independently on each set.

CAT Simulation

Software. A computer program was written in C++ to simulate an operational CAT based on a unidimensional 3PL IRT model (see Equation 1). Variables defined in a configuration file controlled the operation of the simulation, including the input and output file names, the number of examinees, the test length, the item selection algorithm used (i.e., MAXI, STRA, or STRB), stratification parameters, and whether to include Simpson-Hetter exposure control. Multiple configuration files were generated, and each replication was run via a command-line batch file.

CAT item pool. One 284-item unidimensional item pool for use with the CAT was generated with characteristics of the first dimension of the two-dimensional pool. Discrimination parameters were randomly sampled from $U(0.2, 1.8)$, four difficulty parameters were fixed at every 0.1 logit between -3.5 and 3.5 , and the pseudo-guessing parameter was set to 0.2 for all items.

Item selection. Three selection strategies were programmed into the simulation, maximum information (MAXI), a -stratification (STRA), and a -stratification with b blocking (STRB). For MAXI, 24 items were administered to each examinee. Items with a difficulty parameter closest to an examinee's estimated ability yielding the most information (see Equation 2) were unconditionally selected for administration.

For both stratified methods, four strata of 71 items each were created based on results obtained from Chang et al. (2001), and Yi and Chang (2003). For the STRA

design, the first stratum consisted of items with the smallest 71 a values, with the remaining strata comprised of increasing values such that the last stratum contained the largest group of a values. During CAT administration, six items were delivered from each stratum based on the similarity between b and $\hat{\theta}$, resulting in a 24-item adaptive test for each examinee.

In the STRB design, the 284-item pool was initially divided into 71 blocks of increasing difficulty parameters based on the procedure implemented by Chang et al. (2001). Each block had four items by design (i.e., four items every 0.1 logit between -3.5 and 3.5), which were sorted by a values. The stratification procedure was performed with one item from each b -block, producing a total of four strata of increasing a values across all 71 b blocks. As with STRA, six items were delivered from each stratum during CAT, resulting in a 24-item adaptive test for each examinee.

Exposure control. Simpson-Hetter (SH) was chosen as the exposure control method because of its popularity in CAT (Chang & Ying, 1999; Hau & Chang, 2001), and its ability to be applied in addition to most any item selection strategy (Leung, Chang, & Hau, 2002). SH was programmed into the simulation using a predetermined table of item exposure parameters, k . The values for k were calibrated through a series of iterative CAT simulations where k was adjusted based on observing the probability the item was selected, $P(S)$. For each iteration, if $P(S)$ was greater than the maximum exposure rate, r , then k was set to $r / P(S)$, otherwise it was set to 1. The calibration simulations continued until the k values stabilized, resulting in the maximum observed $P(A)$ over all items as approximately equal to r . For the current study, r was set to 0.2 based on Chang and Ying (1999), Hau and Chang (2001), and Leung et al. (2002). To

ensure the k values were ideally set for each simulation involving SH, item exposure parameters were recalibrated for each of the three item selection methods under each of the twelve difficulty-ability correlation conditions in the study, resulting in 36 item exposure calibrations. In operational CAT, SH was implemented by comparing a selected item's k to a random number drawn from $U(0,1)$. If the random number was less than or equal to k the item was administered, else it was marked as “used” and a new item was selected from the pool.

Ability estimation. Based on results from Wang and Vispoel (1998) and Zhao et al. (2002), Bock and Aitken's (1981) Bayes expected a posteriori (EAP) ability estimation method was implemented in the simulation,

$$EAP(\theta) = E(\theta | \mathbf{x}) = \frac{\sum_{k=1}^q L(\mathbf{x} | X_k) A(X_k) X_k}{\sum_{k=1}^q L(\mathbf{x} | X_k) A(X_k)}, \quad (5)$$

where $L(\mathbf{x} | X_k)$ is the log-likelihood of the observed response pattern $\mathbf{x} \equiv (x_1, x_2, \dots, x_j)$ to j items given the k^{th} Gauss-Hermite quadrature node X_k , and $A(X_k)$ is the weight of X_k . Gauss-Hermite quadrature is used as a numeric approximation to the area under the conditional ability expectation distribution (Bock & Aitken). Three hundred nodes (q) were pre-computed to balance accuracy of the integral estimate against extensive computing time. The standard normal distribution was used as an informative prior to EAP because true abilities were sampled from $N(0,1)$.

Simulation. An examinee's unidimensional ability estimate ($\hat{\theta}$) was set to 0 at the start of each of the CAT simulations. Updated ability estimates were computed after each response. In all cases, an adaptive testing simulation was terminated when 24

items were administered. A test of 24 items, approximately one-twelfth of the pool, was suggested by Chang and van der Linden (2003) as appropriate for CAT.

Simulation output for each examinee included the final estimate of ability, the standard error of the estimate, and the item numbers of the 24 items administered. Output for each item included the number of times the item was selected and the number of times it was administered. Additionally, the total number of common items for each of the pairs of examinees was computed.

Analysis

Several indices have been established to evaluate CAT results under various testing environments (de Ayala, 1992; Folk & Green, 1989; Wang & Vispoel, 1998; Weiss & Suhadolnik, 1982; Zhao et al., 2002). Each of the following indices was independently calculated on data from one of the five simulation runs, and the five values were averaged to minimize sampling error. All analyses were performed and figures were generated using R 2.8.1.

Fidelity. Fidelity coefficients include Pearson product-moment or Spearman rank-order correlations between $\hat{\theta}$ and θ_1 , θ_2 , and between $\hat{\theta}$ and the average of θ_1 and θ_2 , $\bar{\theta}$. They have been used (cf. de Ayala, 1992) to determine how well the unidimensional CAT estimates of ability capture the true multidimensional nature of the examinees' ability. Because relatively small differences between the two correlation coefficients have been observed in similar studies (cf. de Ayala; Zhao et al., 2002), only Pearson product-moment correlation coefficients between $\hat{\theta}$ and θ_1 , θ_2 , and $\bar{\theta}$ were computed for the 72 testing conditions in the current study. These correlations were directly

compared to ascertain if item selection methods, under variable difficulty and ability relationships, influence the extent to which $\hat{\theta}$ relate to true abilities.

Precision. The mean square error (*MSE*) is a measure of total error of ability estimation, consisting of two parts, systematic error (*bias*) and random sampling error (*SE*; Wang & Vispoel, 1998). These indices are related by $MSE = bias^2 + SE^2$. The three indices are defined by Zhao et al. (2002) as

$$MSE(\hat{\theta}) = \frac{1}{N-1} \sum_{i=1}^N (\hat{\theta}_i - \bar{\theta})^2, \quad (6)$$

$$bias(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \bar{\theta}), \text{ and} \quad (7)$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left(\hat{\theta}_i - \frac{1}{N} \sum_{k=1}^N \hat{\theta}_k \right)^2}, \quad (8)$$

where N is the total number of examinees. In the current study the standard error of the estimate for examinee i , $SE_i(\hat{\theta})$, was obtained directly from Bayes EAP, using a standard normal prior, by

$$SE_i(\hat{\theta}) = \sqrt{\frac{\sum_{k=1}^q L(\mathbf{x} | X_k) A(X_k) (X_k - \hat{\theta})^2}{\sum_{k=1}^q L(\mathbf{x} | X_k) A(X_k)}}, \quad (9)$$

where the parameters were previously defined in Equation 5. For each of the testing conditions *MSE* and *bias* were computed using Equations 6 and 7. The average standard error of the estimate, *ASE*, was calculated as the mean of SE_i across all examinees. Results were tabularized for the 72 testing conditions and directly compared to determine which condition was most precise.

Item overexposure and underutilization. Chang and Ying (1999) defined the exposure rate of an item as the ratio of the number of times the item is administered to the total possible times the item could have been administered during the test (i.e., the total number of examinees). Leung et al. (2002) cautioned high exposure rates could lead to issues with test security and validity, so minimizing overexposure is desirable in CAT. Leung et al. also advised against having large numbers of items with low exposure rates because an underutilized item pool “challenges directly [its] cost-effectiveness and the appropriateness of the item selection method” (p. 381). As with Leung et al. an item in the current study was declared overexposed if its exposure rate was greater than 0.2, and underutilized if its exposure rate was less than 0.02. The total numbers of overexposed and underutilized items for the 72 testing conditions were directly compared to determine which condition produced the least number of overexposed and underutilized items.

Test overlap. Chang and Ying (1999) defined test overlap rate as the number of overlapping items for two randomly selected examinees, averaged across all possible pairwise combinations. Ideally, the number of common items between any two examinees should be minimized. A lower bound was established by Chang and Zhang (as cited by Chang & Ying, 1999) for fixed length adaptive tests with test length, L , and item pool size, P , as L / P . For the current study, the optimal test overlap rate was $24 / 284 = 8.45\%$. To calculate overlap rate for the 72 testing conditions, the total number of common items for each of the pairs of examinees was divided by the total possible (i.e., $LN(N - 1) / 2 = 24 * 1,600(1,600 - 1) / 2 = 30,700,800$). Results were directly compared to determine which condition had an overlap rate closest to optimal.

Pool usage. Chang and Ying (1999) developed a χ^2 statistic to measure the efficiency of the item pool usage,

$$\chi^2 = \sum_{j=1}^P \frac{(er_j - L/P)^2}{L/P}, \quad (10)$$

where er_j is the observed exposure rate for the j^{th} item (i.e., the number of times the item is used per examinee). The desired uniform exposure rate for all items is L/P (Chang and Ying), or $24/284 = 8.45\%$ for current study. χ^2 statistics were computed using Equation 10 for each of the 72 testing conditions, and directly compared to determine which produced the most efficient pool usage.

Results

Each 24-item, 284-examinee simulation took approximately 39 seconds to complete regardless of the item selection method used or the addition of SH exposure control. In addition to the 72 data-generating simulations, approximately 6.8 simulations were required to calibrate the item exposure parameters for each of the 36 SH conditions, resulting in a total of 318 simulations, about 3.4 hours, to complete all simulations necessary for 72 testing conditions. This process was performed for each of the five response sets, for a total of 17.2 hours to collect all of the simulation data.

Ability Estimation

Table 3 summarizes fidelity results. Overall, correlations between estimated ability and true ability increased as interdimensional examinee ability correlation ($r\theta_1\theta_2$) levels increased. Similarly, correlations between estimated and true abilities generally increased as interdimensional item difficulty correlation (rb_1b_2) levels increased. However, the change from one rb_1b_2 level to another was less noticeable than with changes in $r\theta_1\theta_2$ levels.

Also, correlations between $\hat{\theta}$ and $\bar{\theta}$ (Figure 7) were consistently stronger than correlations between $\hat{\theta}$ and θ_1 or θ_2 (Figures 3 and 4). Furthermore, although correlations generally increased as $r\theta_1\theta_2$ and rb_1b_2 levels increased, correlations using average ability were more consistent across testing conditions than results using single dimension abilities (e.g., for no exposure control, the differences (Δ) between $r\theta_1\theta_2 = rb_1b_2 = 0.0$ and $r\theta_1\theta_2 = rb_1b_2 = 0.9$ correlations are, $\Delta r\hat{\theta}\theta_1 = 0.937 - 0.693 = 0.244$, $\Delta r\hat{\theta}\theta_2 = 0.933 - 0.539 = 0.394$, and $\Delta r\hat{\theta}\bar{\theta} = 0.962 - 0.856 = 0.106$).

Correlations between $\hat{\theta}$ and $\bar{\theta}$ under SH were consistently lower than counterparts with no exposure control, with differences diminishing as $r\theta_1\theta_2$ and rb_1b_2 increased. Results using single dimension abilities were mixed, especially at low levels of $r\theta_1\theta_2$ and rb_1b_2 , but differences became negligible as interdimensional correlations increased.

With regards to the item selection strategy used, some differences were detected using single dimension abilities (Figures 3 and 4), but these differences diminished when using average ability (Figure 7). As interdimensional correlations increased, differences became trivial regardless of which measure of true ability was used.

Table 3

Correlations^a Between $\hat{\theta}$ and θ_1 , θ_2 , and $\bar{\theta}$ for Three Item Selection Methods, Conditioned on Four Levels of $r\theta_1\theta_2$ and Three Levels of rb_1b_2 , Grouped by Exposure Control

		$r\theta_1\theta_2$							
		0.0		0.3		0.6		0.9	
rb_1b_2	Selection Method	$\hat{r}\hat{\theta}_1$	$\hat{r}\hat{\theta}_2$	$\hat{r}\hat{\theta}_1$	$\hat{r}\hat{\theta}_2$	$\hat{r}\hat{\theta}_1$	$\hat{r}\hat{\theta}_2$	$\hat{r}\hat{\theta}_1$	$\hat{r}\hat{\theta}_2$
		$(\hat{r}\hat{\theta})$		$(\hat{r}\hat{\theta})$		$(\hat{r}\hat{\theta})$		$(\hat{r}\hat{\theta})$	
No Exposure Control									
0.0	MAXI	0.693	0.539	0.754	0.652	0.825	0.771	0.892	0.880
		(0.856)		(0.877)		(0.896)		(0.911)	
	STRA	0.647	0.590	0.729	0.681	0.808	0.784	0.888	0.877
0.6	STRB	0.638	0.599	0.721	0.688	0.811	0.788	0.889	0.877
		(0.861)		(0.880)		(0.895)		(0.907)	
		(0.862)		(0.879)		(0.898)		(0.908)	
0.9	MAXI	0.657	0.609	0.741	0.696	0.829	0.804	0.905	0.897
		(0.882)		(0.897)		(0.917)		(0.927)	
	STRA	0.688	0.570	0.759	0.672	0.837	0.791	0.906	0.891
	STRB	0.670	0.585	0.743	0.689	0.824	0.801	0.902	0.894
		(0.874)		(0.893)		(0.912)		(0.924)	
		(0.874)		(0.893)		(0.912)		(0.924)	
0.9	MAXI	0.685	0.665	0.769	0.756	0.857	0.855	0.941	0.938
		(0.941)		(0.952)		(0.961)		(0.966)	
	STRA	0.681	0.654	0.767	0.740	0.855	0.844	0.937	0.930
	STRB	0.678	0.656	0.763	0.748	0.852	0.848	0.937	0.933
		(0.930)		(0.941)		(0.954)		(0.960)	
		(0.929)		(0.944)		(0.955)		(0.962)	
SH Exposure Control									
0.0	MAXI	0.636	0.542	0.711	0.640	0.791	0.756	0.868	0.862
		(0.820)		(0.843)		(0.869)		(0.890)	
	STRA	0.618	0.581	0.699	0.678	0.786	0.778	0.872	0.862
0.6	STRB	0.618	0.584	0.707	0.678	0.782	0.777	0.873	0.867
		(0.835)		(0.860)		(0.878)		(0.892)	
		(0.838)		(0.865)		(0.875)		(0.894)	
0.9	MAXI	0.604	0.622	0.701	0.696	0.798	0.797	0.885	0.886
		(0.854)		(0.873)		(0.896)		(0.910)	
	STRA	0.656	0.588	0.742	0.682	0.823	0.792	0.899	0.889
	STRB	0.649	0.596	0.732	0.677	0.819	0.795	0.896	0.891
		(0.866)		(0.889)		(0.907)		(0.920)	
		(0.867)		(0.880)		(0.906)		(0.919)	
0.9	MAXI	0.647	0.668	0.740	0.755	0.840	0.849	0.930	0.931
		(0.917)		(0.934)		(0.949)		(0.957)	
	STRA	0.661	0.664	0.754	0.748	0.847	0.844	0.933	0.931
	STRB	0.658	0.665	0.753	0.748	0.849	0.846	0.934	0.932
		(0.923)		(0.938)		(0.950)		(0.958)	
		(0.922)		(0.938)		(0.952)		(0.959)	

^a Pearson product-moment coefficients.

Figure 5. Correlations between $\hat{\theta}$ and θ_1 for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 6. Correlations between $\hat{\theta}$ and θ_2 , for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 7. Correlations between $\hat{\theta}$ and $\bar{\theta}$ for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Table 4 summarizes precision results. The bias between estimated ability and $\bar{\theta}$ was relatively consistent across interdimensional examinee ability correlation levels as seen in Figure 8. However, as rb_1b_2 levels increased, *bias* in general tended towards 0 (reference grey lines set at *bias* = 0 in Figure 8), with best results obtained from stratified item selection strategies. In fact, both stratified methods under SH consistently had least *bias* regardless of any interdimensional correlation levels.

Overall, average standard error (ASE), as estimated by Bayes EAP during CAT, was relatively similar under all testing conditions (e.g., the range of all ASEs was between 0.235 and 0.311). Without SH exposure control, MAXI consistently had the least average standard error (see Figure 9). However, under SH differences between the three item selection strategies were negligible.

Mean square error (MSE), as illustrated in Figure 10, was consistent across levels of $r\theta_1\theta_2$, with results tending towards 0 as rb_1b_2 increased. Noticeable differences between all methods occurred when $rb_1b_2 = 0$, with SH consistently producing higher MSE than item selection strategies operating without exposure control, but generally, stratified methods tended to have lower MSE than MAXI.

Item Usage

Figure 11 illustrates the effectiveness of SH exposure control for one testing condition, which was typical for all conditions. Without SH, items with large discrimination values were administered without restriction, resulting in exposure rate spikes. STRA, which stratified on item discrimination, somewhat limited item exposure. STRB, which first blocked on item difficulty, and then stratified on discrimination, performed better at limiting exposure. However, a few items with relatively large discrimination values had high exposure rates regardless of item selection method. When SH exposure was included, MAXI+SH, STRA+SH, and STRB+SH limited exposure rates to around the maximum exposure rate, $r = 0.2$, set a priori.

Table 4

Bias, Average Standard Error (ASE), and Mean Standard Error (MSE) Between $\hat{\theta}$ and $\bar{\theta}$ for Four Levels of Interdimensional Ability Correlations and Three Levels of Interdimensional Item Difficulty Correlations

		$r\theta_1\theta_2$							
		0.0		0.3		0.6		0.9	
rb_1b_2	Selection Method	<i>bias</i> (MSE)	<i>ASE</i>	<i>bias</i> (MSE)	<i>ASE</i>	<i>bias</i> (MSE)	<i>ASE</i>	<i>bias</i> (MSE)	<i>ASE</i>
<i>No Exposure Control</i>									
0.0	MAXI	-0.321 (0.549)	0.250	-0.321 (0.582)	0.253	-0.356 (0.592)	0.255	-0.354 (0.587)	0.255
	STRA	0.153 (0.389)	0.297	0.153 (0.412)	0.298	0.145 (0.430)	0.299	0.140 (0.427)	0.300
	STRB	0.102 (0.437)	0.293	0.102 (0.454)	0.295	0.089 (0.463)	0.297	0.094 (0.475)	0.298
0.6	MAXI	-0.126 (0.137)	0.235	-0.126 (0.143)	0.236	-0.134 (0.142)	0.238	-0.145 (0.145)	0.239
	STRA	0.006 (0.177)	0.294	0.006 (0.176)	0.294	-0.003 (0.163)	0.294	-0.002 (0.163)	0.294
	STRB	-0.045 (0.172)	0.288	-0.045 (0.174)	0.290	-0.048 (0.164)	0.291	-0.049 (0.159)	0.291
0.9	MAXI	0.021 (0.077)	0.236	0.021 (0.073)	0.235	0.006 (0.071)	0.235	0.008 (0.071)	0.235
	STRA	-0.014 (0.088)	0.289	-0.014 (0.089)	0.289	-0.013 (0.083)	0.289	-0.016 (0.082)	0.289
	STRB	-0.043 (0.086)	0.287	-0.043 (0.083)	0.288	-0.049 (0.079)	0.288	-0.049 (0.077)	0.289
<i>SH Exposure Control</i>									
0.0	MAXI	-0.321 (0.893)	0.295	-0.354 (0.888)	0.294	-0.350 (0.837)	0.294	-0.324 (0.795)	0.293
	STRA	-0.014 (0.556)	0.300	-0.043 (0.579)	0.300	-0.015 (0.571)	0.300	0.005 (0.585)	0.299
	STRB	0.002 (0.527)	0.301	-0.001 (0.524)	0.302	0.010 (0.568)	0.300	0.022 (0.552)	0.301
0.6	MAXI	-0.150 (0.269)	0.311	-0.164 (0.283)	0.307	-0.167 (0.265)	0.303	-0.166 (0.255)	0.300
	STRA	-0.022 (0.174)	0.295	-0.038 (0.171)	0.295	-0.031 (0.165)	0.295	-0.042 (0.160)	0.294
	STRB	-0.020 (0.173)	0.295	-0.010 (0.183)	0.294	-0.006 (0.166)	0.293	-0.008 (0.161)	0.293
0.9	MAXI	-0.050 (0.091)	0.311	-0.060 (0.088)	0.303	-0.061 (0.085)	0.299	-0.066 (0.081)	0.296
	STRA	-0.016 (0.094)	0.296	-0.028 (0.091)	0.295	-0.026 (0.089)	0.294	-0.024 (0.083)	0.292
	STRB	-0.023 (0.094)	0.304	-0.029 (0.090)	0.301	-0.028 (0.084)	0.293	-0.029 (0.081)	0.291

Figure 8. Bias between $\hat{\theta}$ and $\bar{\theta}$ for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 9. Average standard error (ASE) of the EAP estimate for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 10. Mean standard error (MSE) between $\hat{\theta}$ and $\bar{\theta}$ for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 11. Item exposure rates of 284 items for three item selection methods, used alone and with SH exposure control, for $r\theta_1\theta_2 = 0.6$, $rb_1b_2 = 0.6$.

Table 5 summarizes item overexposure and underutilization results. Few patterns emerged from the results of overexposed items, as seen in Figure 12. When $rb_1b_2 = 0$, MAXI clearly had more overexposed items than the two stratified methods, regardless of the presence or absence of SH exposure control. For the other two levels of interdimensional item difficulty correlations, the results were nebulous, but often the least overexposed conditions occurred with both stratified item selection strategies, and no exposure control.

Table 5
Number of Items Overexposed (OVER) and Underutilized (UNDER) for Four Levels of Interdimensional Ability Correlations and Three Levels of Interdimensional Item Difficulty Correlations

rb_1b_2	Selection Method	$r\theta_1\theta_2$							
		0.0		0.3		0.6		0.9	
		OVER	UNDER	OVER	UNDER	OVER	UNDER	OVER	UNDER
<i>No Exposure Control</i>									
0.0	MAXI	48.4	190.8	47.2	183.0	47.4	181.0	47.6	177.6
	STRA	39.0	152.2	36.6	145.8	38.2	136.4	37.4	131.4
	STRB	36.6	138.2	34.8	130.6	35.0	123.0	33.4	119.2
0.6	MAXI	44.0	213.6	44.2	208.6	46.6	205.8	46.6	202.2
	STRA	53.6	162.8	52.6	157.0	50.6	152.6	47.2	143.4
	STRB	57.2	157.0	52.8	151.4	47.8	148.0	44.0	142.2
0.9	MAXI	50.4	204.8	52.2	199.2	54.2	193.4	54.6	190.0
	STRA	47.4	160.2	45.6	152.6	42.8	145.4	40.4	141.0
	STRB	47.6	159.4	45.6	149.4	45.8	143.2	42.8	138.6
<i>SH Exposure Control</i>									
0.0	MAXI	50.2	141.8	51.6	138.0	52.0	137.4	49.0	132.2
	STRA	35.8	120.4	28.0	114.0	29.4	108.4	31.4	104.8
	STRB	36.8	118.8	35.6	113.8	33.4	103.8	32.2	100.6
0.6	MAXI	57.8	148.4	51.6	145.4	54.2	143.8	56.8	142.6
	STRA	66.4	153.2	63.4	147.4	57.2	141.0	47.6	135.6
	STRB	67.6	149.8	52.0	143.0	47.4	140.4	53.2	143.4
0.9	MAXI	57.0	149.2	56.4	150.0	53.0	144.2	50.6	142.2
	STRA	77.4	152.8	75.6	144.6	58.2	131.8	46.8	126.8
	STRB	63.4	146.2	60.6	138.2	61.6	135.4	57.6	128.4

Figure 12. Number of items overexposed for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Figure 13. Number of items underutilized for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Numbers of underutilized items were consistent across the four levels of $r\theta_1\theta_2$, and mostly regular across the three levels of rb_1b_2 (see Figure 13). In each of the twelve conditions, MAXI had more underutilized items than the stratified strategies when no exposure control was used. However, under SH, differences between the three item selection methods diminished.

Table 6 and Figure 14 illustrate test overlap results. Without SH exposure control, MAXI consistently had higher overlap rates than stratified strategies, with the greatest differences occurring at $rb_1b_2 = 0.6$. The inclusion of SH exposure control unconditionally limited overlap rates to approximately 0.2.

Table 6
Overlap Rate for Three Item Selection Methods, Conditioned on Four Levels of $r\theta_1\theta_2$ and Three Levels of rb_1b_2 , Grouped by Exposure Control

Selection		$r\theta_1\theta_2$			
rb_1b_2	Method	0.0	0.3	0.6	0.9
<i>No Exposure Control</i>					
0.0	MAXI	0.409	0.393	0.375	0.365
	STRA	0.351	0.333	0.315	0.310
	STRB	0.331	0.317	0.300	0.292
0.6	MAXI	0.524	0.505	0.480	0.469
	STRA	0.300	0.291	0.283	0.279
	STRB	0.269	0.260	0.252	0.249
0.9	MAXI	0.439	0.419	0.395	0.382
	STRA	0.331	0.318	0.294	0.283
	STRB	0.322	0.305	0.284	0.276
<i>SH Exposure Control</i>					
0.0	MAXI	0.186	0.184	0.183	0.182
	STRA	0.168	0.163	0.162	0.161
	STRB	0.172	0.171	0.161	0.162
0.6	MAXI	0.192	0.188	0.188	0.188
	STRA	0.213	0.207	0.197	0.187
	STRB	0.199	0.184	0.184	0.190
0.9	MAXI	0.192	0.201	0.188	0.186
	STRA	0.224	0.218	0.187	0.179
	STRB	0.195	0.187	0.191	0.187

Figure 14. Overlap rate for three item selection methods, conditioned on four levels of $r\theta_1\theta_2$ and three levels of rb_1b_2 , grouped by exposure control.

Table 7 summarizes pool usage results as measured by Chang and Ying's (1999) χ^2 statistic. The Pearson product-moment correlation between overlap rates (Table 6) and χ^2 was 1.0. Therefore, the same trends described for overlap rates apply for χ^2 results (i.e., without SH, MAXI had the highest values, and under SH, the three selection methods were comparable).

Table 7

Efficiency of Item Pool Usage (χ^2) for Four Levels of Interdimensional Ability Correlations and Three Levels of Interdimensional Item Difficulty Correlations

rb_1b_2	Selection Method	$r\theta_1\theta_2$			
		0.0	0.3	0.6	0.9
<i>No Exposure Control</i>					
0.0	MAXI	92.18	87.81	82.47	79.94
	STRA	75.72	70.74	65.72	64.04
	STRB	70.06	66.06	61.30	59.04
0.6	MAXI	124.81	119.48	112.41	109.15
	STRA	61.14	58.60	56.57	55.23
	STRB	52.47	49.85	47.76	46.79
0.9	MAXI	100.71	95.00	88.29	84.63
	STRA	70.10	66.41	59.69	56.63
	STRB	67.67	62.73	56.87	54.50
<i>SH Exposure Control</i>					
0.0	MAXI	29.01	28.52	28.22	27.94
	STRA	23.83	22.44	22.19	21.87
	STRB	24.96	24.63	21.86	22.24
0.6	MAXI	30.76	29.43	29.49	29.63
	STRA	36.54	34.99	32.23	29.15
	STRB	32.79	28.39	28.38	30.27
0.9	MAXI	30.54	33.23	29.46	29.09
	STRA	39.74	37.89	29.31	26.87
	STRB	31.67	29.34	30.37	29.26

Discussion

Regarding ability estimation, current results confirm previous conclusions (cf. Way et al., 1988; Zhao et al., 2002) that estimated ability ($\hat{\theta}$) best represents the average of the abilities in the first and second dimensions ($\bar{\theta}$; see Table 3). These results were minimally impacted by interdimensional examinee ability correlation and interdimensional item difficulty correlation, and stable with respect to the item selection strategy utilized in the current study (MAXI, STRA, and STRB), either with SH exposure control or used alone, as seen in Figure 7. Furthermore, as with previous studies (e.g., de Ayala, 1992; Zhao et al.) correlations between estimated ability and true abilities

(including their average) increased as both interdimensional ability and difficulty correlations increased.

The bias between the estimated unidimensional ability and $\bar{\theta}$ was most reduced when either stratified item selection strategy was used, especially when the interdimensional item difficulty correlation was low. Similar results were found for mean square error, suggesting that either *a*-stratified or *a*-stratified with *b*-blocking item selection could be used to compensate for the effects of two-dimensional data under unidimensional CAT. Furthermore, *bias* was effectively eliminated across all $r\theta_1\theta_2$ and rb_1b_2 conditions when Simpson-Hetter exposure control was included with either stratified selection method.

Unlike the conclusions drawn from de Ayala (1992) and Zhao et al. (2002), results from the current study showed the greatest impact to *bias* and *MSE* came from changes in rb_1b_2 rather than changes in interdimensional examinee ability correlations. However, the apparent contradiction seems to have more to do with research design than in a direct comparison of results. de Ayala was primarily interested in recovering examinee ability, and as such, his *bias* and *RMSE* results were best illustrated through graphs across a wide range of ability scores (e.g., his theta scales ranged from -5 to +5). In the current study, abilities were sampled from a standard normal distribution, and results were averaged over all abilities to point estimates. Additionally, de Ayala was interested in how *bias* and *RMSE* changed as $\hat{\theta}$ was considered an estimate of θ_1 , θ_2 , or $\bar{\theta}$, therefore his graphs required wide scales for *bias* (-3.5 to +3.5) and *RMSE* (0 to +5, equivalent to an *MSE* scale of 0 to +25) to capture the large fluctuations of

comparisons made with respect to θ_1 and θ_2 . As a comparison, the largest average value for *MSE* in the current study was 0.893, for MAXI+SH when $r\theta_1\theta_2 = rb_1b_2 = 0.0$.

Because of results from prior research, such as de Ayala's (1992), *bias* and *MSE* in the current study were only measured with respect to $\bar{\theta}$. Therefore, much smaller differences were analyzed in the current study than in de Ayala's study. Furthermore, because of the method used to generate two-dimensional data with specific correlational properties, de Ayala and Zhao et al. (2002) were not able to test for the condition where b_1 and b_2 were perfectly uncorrelated. In fact, the lowest interdimensional item difficulty correlation used by Zhao et al. was 0.10. As seen in Figures 6 and 8, the greatest differences in *bias* and *RMSE* in the current study were detected when $rb_1b_2 = 0.0$. It is possible that had previous researchers used a method allowing them to test perfectly uncorrelated interdimensional item difficulties, greater changes might have been detected.

Regarding item usage, no studies were found in the literature investigating the effects of unidimensional adaptive testing in the presence of multidimensional data. However, several studies (cf. Chang & Ying, 1999; Chang et al., 2001; Hau & Chang, 2001; Leung et al., 2002) compared various item selection and exposure control strategies in a strictly unidimensional CAT framework. As originally demonstrated by Leung et al. with the *a*-stratified design, the inclusion of Simpson-Hetter exposure control in the current study clearly limited item exposure to the maximum set a priori, $r = 0.2$, regardless of the item selection method used (see Figure 11). However, the results displayed in Table 5 and Figure 12 do not necessarily reflect this conclusion. Following the design used by Leung et al., the criterion used in the current study to label an item

as overexposed was set to an exposure rate greater than 0.2, the same value as the maximum exposure rate, r . Although SH exposure control limits exposure rates to r by design (Stocking & Lewis, 1998), it does not guarantee exposure rates are strictly less than r , as seen in the lower three graphs in Figure 11. Because the criterion for overexposure in the current study was equal to r , many items were counted as overexposed, when in fact they were only nominally greater than r , but well below their non-SH exposure rates. If, for example, the criterion was set modestly higher, say to an exposure rate greater than 0.3, the average number of items classified as overexposed would have been 24.0 with no exposure control, and 0.9 with SH, which better represents the results and better corresponds to results found in the literature.

The current study confirms previous conclusions (cf. Chang & Ying, 1999; Leung et al., 2002) that stratified item selection methods reduce the number of underutilized items over the maximum information strategy. However, the magnitudes of these reductions were not realized in the current study. Conversely, although a reduction of overlap rates and χ^2 values resulted when item selection strategies included SH exposure control, the reduction was actually greater than in previous research (cf. Leung et al.). The apparent discrepancies to prior research are most likely a result of the item pools used in the studies. Like Chang and Ying, Leung et al. used a 400-item pool consisting of difficulty values sampled from a standard normal distribution, and fixed discrimination values such that the four 100-item strata had a values set to 0.5, 1.0, 1.5, and 2.0, respectively. The item pool in the current study similarly had b values sampled from $N(0,1)$, but the a values were randomly sampled from $U(0.2, 1.8)$. As seen in Figure 11, only a few items in the pool had extremely large discrimination

values, and consequently were highly exposed, which contrasted to prior studies where 25% of the pool had equally high discriminating items.

Item overexposure results notwithstanding, when item underutilization results are considered in conjunction with overlap rates (Table 6 and Figure 14) and χ^2 values (Table 7), item usage, in general, was most impacted across all $r\theta_1\theta_2$ and rb_1b_2 conditions when Simpson-Hetter exposure control was included, regardless of the item selection method used.

Limitations

There are several limitations on the generalization of the results of the current study. First, only two-dimensional data were explored. It is possible to obtain different conclusions with three or more factors of varying dominance (cf. Reckase, 1979), as well as with higher order factor models (cf. Drasgow & Parsons, 1983). Second, it was assumed that correlations between difficulty and discrimination parameters were zero. Wingersky and Lord (1984) have stated a positive correlation often exists between the two parameters in operational item pools, and Chang et al. (2001) gave one such example with the GRE quantitative test. Given the stratification methods used in STRA and STRB, it is possible that ability estimation would be impacted because parameter distributions for each stratum would no longer represent the entire pool. Third, only certain item selection methods based on the *a*-stratification framework were studied. Several more stratified strategies exist, such as *a*-stratification with freezing (see Georgiadou et al., 2007). Furthermore, other types of item selection and exposure control strategies have yet to be investigated, such as randomization, conditional, and other combined strategies. Fourth, a fixed length test of 24 items was used in the

current study. Although at approximately one-twelfth of the item pool, a 24-item test is appropriate as suggested by Chang and van der Linden (2003), previous studies have shown increased test length to positively impact ability estimation (Ackerman, 1991; Weiss & Suhadolnik, 1982) as well as item usage (Leung et al., 2002). Additionally, the effect of other stopping criteria, such as a threshold on the standard error of the estimate, is unknown. Fifth, the pseudo-guessing parameter was fixed at $c = 0.2$ (cf. Zhao et al., 2002). As suggested by Chang and Ying (1999), the effect of c in adaptive testing using stratified item selection methods is relatively unknown. Finally, a unidimensional CAT was used in the current study. Although operational CAT implementations have grown considerably (Wainer, 2000), substantial progress has recently been made in both multidimensional CAT theory and computer processing capability. It is well within the capability of researchers to develop and administer MIRT-based CAT, such as those methods proposed by Lee et al. (2008). Further simulation studies in each of these areas might be pursued.

Conclusions

Way et al. (1988) stated “unidimensional IRT models are often applied to multidimensional item response data” (p. 242). Therefore it is important to understand the implications of summarizing two or more dimensions of ability into a single estimate ($\hat{\theta}$), especially when effects are confounded when applied to an adaptive testing implementation using a fraction of the item pool. Previous simulation studies have investigated the effects of different MIRT models (compensatory vs. noncompensatory, first-order vs. second-order, two-factor vs. multiple-factor), as well as various ability estimators (maximum likelihood vs. various Bayesian methods), by manipulating the

relationships between item parameters (a , b , c) and person parameters (θ). However, in all cases, the maximum information (MAXI) criterion was used as the method of selecting the next item in the adaptive testing session. Because MAXI is heavily influenced by the item discrimination parameter (a), investigating item selection methods that are conditioned on discrimination (i.e., those based on the a -stratified framework) seems warranted.

The current Monte Carlo confirmed that estimated ability best represents the average of the abilities in the first and second dimensions for all interdimensional ability and difficulty correlation conditions tested. The study also showed that if the goal of unidimensional CAT, in the presence of two-dimensional data, is to obtain precise estimates of ability, either a -stratified or a -stratified with b -blocking item selection strategies should be used. Additionally, if the goal includes limiting bias of the estimate, Simpson-Hetter exposure control should be included. If the goal is to optimize item pool usage, the current study also shows that Simpson-Hetter is effective, regardless of the item selection tested. Given these results, existing unidimensional CAT implementations might consider employing an a -stratified item selection routine plus Simpson-Hetter exposure control, rather than recalibrate the item pool under a multidimensional model and attempt to link the old unidimensional ability scale to the new multidimensional ability scale.

REFERENCES

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.
- Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, 15, 13-24.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
- Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-444.
- Chang, H.-H., Qian, J., & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a-stratified computerized adaptive testing. *Applied Psychological Measurement*, 27, 262-274.
- Chang, H.-H., & Ying, Z. (1996). A global informative approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Chang, H.-H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the

- annual meeting of the American Educational Research Association, San Francisco, CA.
- de Ayala, R. J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement*, 52, 513-528.
- Doody-Bogan, E., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, QC.
- Dragow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
- Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-389.
- Gentle, J. E. (1998). *Numerical linear algebra for applications in statistics*. New York: Springer.
- Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8). Retrieved February 17, 2009 from <http://jtla.org/>
- Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, 38, 249-266.

- Lee, Y.-H., Ip, E. H., & Fuh, C.-D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement, 68*, 215-232.
- Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Simpson-Hetter algorithm. *Applied Psychological Measurement, 26*, 376-392.
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. (pp. 139-183). New York: Harper and Row.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*, 100-114.
- Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L. & Mills, C. N. (1993). *Field test of a computer-based GRE General Test* (Research Rep. No. RR-93-07). Princeton, NJ: Educational Testing Service.

- Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67, 899-919.
- Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 185-230). Mahwah, NJ: Lawrence Erlbaum Associates.
- Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Rep. No. RR-95-25). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57-75.
- Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
- van der Linden, W. J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, 27, 107-120.
- Wainer, H. (2000). CATs: Whither and whence. *Psicológica*, 21, 121-133.
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

- Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 109-135.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12, 239-252.
- Weiss, D. J., & Suhadolnik, D. (1982). Robustness of adaptive testing to multidimensionality. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8, 347-364.
- Yi, Q., & Chang, H.-H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56, 359-378.
- Zhao, J. C., McMorris, R. F., & Chen, R. (2002, April). *The robustness of the unidimensional 3PL IRT model when applied to two-dimensional data in computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.