Graduate Management Admission Council®

A Hybrid Simulation Procedure for the Development of CATs

Steven W. Nydick

and

David J. Weiss

University of Minnesota

Presented at the Item and Pool Development Paper Session, June 3, 2009

Abstract

Frequently, because they tend to be large, CAT item banks are calibrated using concurrent-calibration methods, which estimate IRT parameters from an incomplete data matrix including a set of linking items (e.g., Kim & Cohen, 1998). This paper proposes and evaluates the performance of a hybrid simulation procedure for use in developing CATs that employs these sparse, concurrent-linking matrices. The hybrid procedure estimates θ for each examinee with the item parameters estimated from the sparse linking matrix in conjunction with the set of item responses for each examinee. Then, the θ estimate for each examinee is used with monte-carlo simulation methods to impute the examinee's missing data, resulting in a complete response vector for each examinee—part real item responses and part imputed simulated data. A post-hoc simulation is then implemented with the hybrid response matrix. Results suggest that meaningful hybrid simulations can be performed with sparse data matrices involving up to almost 80% missing/imputed data.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Steven Nydick, N657 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.
Email: nydic001@umn.edu or djweiss@umn.edu

A Hybrid Simulation Procedure for the Development of CATs

Practical Challenges in CAT Development

Even though the immediate scoring and the reduced numbers of items to measure an examinee facilitates time conscious test implementers and examinees, many of the CAT challenges proposed in the early 1980s (see Weiss, 1982, 1985 for theoretical overviews and Mills & Stocking, 1996 for an overview of practical challenges in CAT development) have yet to be sufficiently unraveled.

Test security (including item exposure controls) has been a main research focus over the past two decades (e.g., Sympson & Hetter, 1985; Stocking & Lewis, 1995; Chang & Ying, 1999; Leucht, 2003; and Georgiadou, Triantafillou, & Economides, 2007). Moreover, Wainer, 2000b, found that 50% of item-examinee interactions only use about 14% of the item bank. Because, in high-stakes testing, examinees might have motivation to cheat (academic, financial, or otherwise), and because a perpetual stream of examinees take tests from the same item bank, the only way of preventing item theft from contaminating test validity is to construct a large, and continuously changing, bank of items. Some researchers (e.g., Wise & Kingsbury, 2000; Stocking, 1994) even advocate large, continuously changing, and continuously rotating banks of items; their reason for multiple banks, instead of one large item bank, was the ease with which thieves could electronically steal items and the exorbitant cost of prevention (Colton, 1998 as cited in Wise & Kingsbury, 2000). Stocking (1994), found that doubling the number of banks reduced test overlap to a much greater extent than doubling the number of items in each bank.

Furthermore, as mentioned by Weiss (1982), in order for a test to measure examinees equiprecisely, there must be numerous items across a wide range of θ. If a test developer wants to accurately measure examinees with high ability, he or she needs to create enough difficult items to differentiate between people with exceptionally high ability and people with only moderately high ability. Stocking (1994) advocated an item bank of at least six times the length of a conventional test (or 12 times the length of an adaptive test if the adaptive test length was one-half the length of a conventional test, though she did add that this assessment was conservative). This means that, in an adaptive test of 30 items, the item bank would need to include at least 360 items; moreover, if a researcher wanted a test to use multiple, rotating item banks, then the total number of items could easily reach 1,000 or more. Add that to the, ideally, constantly changing bank of items, the expense of calibrating each item (Guo, 2007), and other necessary precautions (including recalibration of items and continuous checks on item security; e.g., McLeod, Lewis, & Thissen, 2003) to prevent item parameter drift (Wise & Kingsbury, 2000), and it becomes apparent that a considerable amount of cost and labor are necesssary to maintain an adequate CAT program.

Linking and Concurrent Calibration

Developing a CAT for implementation requires all of the items to be on the same scale in order to make comparative judgments about examinees' scores. However, because security and measurement concerns influence the large number of items in a CAT bank, and because the bank must continuously update items and update the calibration of existing items, no examinee could answer all the items in the bank(s) from which the CAT is based. Either factors external to the trait (e.g., fatigue) would influence his/her score, security issues would prevent the examiner from wanting to present all of the items to one examinee (Makransky, 2008), examinees could

not physically take all of the items because of illness (e.g., Bjorner, et al. 2007; this applies to CATs designed for medical diagnostics), or the relative cost would make it unnecessarily expensive to present more than a few items to each examinee (e.g., Wainer, 2000a).

Thus, test devlopers needed a means of linking items onto the same scale. Several methods transform item parameters to the scale of already existing item parameters (using a sequence of anchor items, or items taken by all examinees employed in test calibration). Some of the traditional methods consist of linearly transforming the parameter values by the means and standard deviations of the θ or b estimates (Vale, 1986), or minimizing the true score estimates of the anchor items across groups (Stocking & Lord, 1983). However, often the examiner wants a metric consisting of all of the information about items and groups, instead of arbitrailly transforming the scale of some items to the scale of other items or to an already existing scale. This method of "concurrent calibration" was used in LOGIST (Wingersky, Barton, & Lord, 1982) and is currently implemented in XCALIBRE (Assement Systems Corporation, 1995). In these programs, non-administered items are coded as "not reached," estimating parameters on a subset of people (the people who take the item) and estimating θ on a subset of items (the items given to each person); instead of linking new items to existing items, this method creates a "full-metric" scale that transcends a subgroup.

A final consideration of concurently calibrating a set of items is the number of linking items to include—the number of items that every calibration examinee would take—in order to link all of the items onto a common scale; more linking items would create a longer test, while fewer linking items might not provide ample control for measurement error in test calibration. Wingersky and Lord (1984), suggested that as few as two items could calibrate a test with sufficiently accurate b estimates, while other researchers have recommended more, especially if the groups are not equivalent (Ree & Jensen, 1980; Vale, 1986). Once the linking items are chosen, and the tests given, the researcher ultimately ends up with only a sparse data matrix of examinee responses (only responses to items that the researcher has given to each examinee, which is, of course, incomplete), which can result in item parameter estimates for each item and θ estimates for each person.

Simulation Methods

In order to implement a CAT, a researcher has to make decisions about CAT initiation (how to start the test), continuation (how to choose items), termination (how to end the test), and scoring. Though a CAT can reduce the number of items 50% to over 90% (Weiss & Gibbons, 2007; Gibbons, et al., 2008) with little loss in measurement accuracy, researchers still need to estimate CAT performance prior to implementation, to make the best decisions to maximize test effectiveness. Furthermore, justification of a CAT instead of a conventional test requires estimates of the accuracy/efficiency tradeoff. For researchers to implement CAT constraints such as item exposure, they need to know the relative appearance of each item in the tests. Two general methods have been applied to test aspects of CAT performance prior to implementation of a live CAT: monte-carlo (M-C) simulations and post-hoc (P-H) simulations.

Harwell, Stone, Hsu, and Kirisci (1996) extensively explicated the rationale and application of M-C methods to solve IRT problems, including CAT. A typical dichotomous M-C CAT simulation (e.g., Kim & Plake, 1993) has four parts:

1. Generate simulee θs according to a specified distribution.

2. Generate item parameters based on specified distributions or use item parameters based on an already calibrated item bank.
3. Generate a matrix of item responses based on a specified IRT model. For each individual item/simulee combination, the item parameters and θ are used in the IRT model to generate a probability (P) of endorsing a specific item response. A uniform random number is generated and compared to the model-generated probabilities to create a simulee's scored (e.g., 0-1) response.
4. Implement a CAT on the matrix of item responses, using each simulee's model-generated responses to each item, with a prespecified set of CAT options.

A P-H simulation differs from a M-C simulation, in that the matrix of item responses in a P-H simulation are actual responses from a conventional test of a certain length. The researcher usually starts out with a long test and then tries to determine how short an adaptive test could be with the item bank that made up the conventional test. To determine potential test reduction, the researcher assumes that the examinees would respond in the same way to each item regardless of whether that item was presented amidst many other items at varying difficulties or few items around the same difficulty. However, in a P-H the researcher has less control over aspects of test performance, such as the degree of misfit, the parameter distributions, and the knowledge of the truth behind the responses. In a M-C simulation, the researcher would be able to specify everything in advance. When actually executing a CAT, the test implementer would not have control over these test aspects, so P-H simulation would allow researchers a means of predicting how a particular CAT would perform in practice.

A typical P-H CAT simulation (e.g., Weiss & Gibbons, 2007; Gibbons et al. 2008; Waller & Reise, 1989) also has four steps:

1. Administer a conventional test, typically consisting of an entire item bank, to a group of examinees.
2. Score the responses on the test as a persons × items matrix of integers representing the examinee's response choice (e.g., 0 or 1 for dichotomously scored items).
3. Calibrate the item bank parameters by concurrent full-metric calibration using the matrix of responses and a parameter estimation program (e.g., XCALIBRE, Assessment Systems Corporation, 1995).
4. Implement a CAT on the matrix of item responses (with a prespecified set of CAT options), using the examinee's actual response to each item but the item parameter estimates and θ estimates to choose the sequence of items.

In a variation of P-H simulation, item parameters might be estimated in advance from a prior data set (which might be a sparse data set resulting from a linking design). These parameters might them be used to implement the P-H simulation based a smaller sample of examinees, all of whom would have taken all of the items in the item bank (e.g., Weiss & Gibbons, 2007; Gibbons et al., 2008)

The primary difference between a M-C simulation and a P-H simulation is that a M-C simulation starts with a prespecified distribution of simulees and asks "what would happen if examinees randomly selected from this *distribution* took a CAT on these items" and thus generates *hypothetical* responses based on *hypothetical* people. A P-H simulation starts with

actual responses to *actual* items and asks "what would happen if *these examinees* took a CAT based on these items?"

Hybrid Simulation

Though M-C simulation is frequently used to initiate research into CAT effectiveness (e.g., Kim & Plake, 1993; Stocking, 1996; Weiss & McBride, 1984), ultimately CAT implementers would need to predict item and test characteristics based on real responses to real items. If every person had responded to every item, then Weiss & Gibbons (2007; Gibbons, et al., 2008, demonstrated that P-H simulations well predict the outcome of a live CAT. However, bank size and security constraints (as outlined above) usually prevent the test developer from having a full matrix of responses to every item. Thus, a necessary precondition of performing a P-H simulation, namely that of every item having the possibility of appearing on every test, could not occur. Consequently, a hybrid simulation is proposed in order to allow sparse data matrices to be used in estimating CAT properties through P-H simulation methods. The hybrid simulation would be used to evaluate the performance of a CAT on a sparse data matrix, prior to implementing a live CAT.

A hybrid simulation starts with an *incomplete matrix* of *actual responses* to *actual items* and asks "what would happen if *these examinees* COULD take a CAT on these items," or in other words, "what would happen if *every examinee* COULD take *every item*," and thus generates *hypothetical responses* based on *real people* to *items that they did not take*. A researcher would have a sparse data matrix either if calibrating a large bank of items necessary for a CAT using a linking design or if estimating the effectiveness of a CAT on a new group of examinees without requiring them to take all the items in the item bank. In neither case could an examinee respond to all of the items without external factors such as fatigue or boredom factoring in his or her score. However, because all of the item parameters and person parameters can be accurately estimated using concurrent-calibration methods, all of the necessary information (parameters and model) will be available for the simulation of missing responses.

A hybrid simulation combines useful aspects of a M-C simulation and a P-H simulation in an attempt to solve a heretofore unresolved practical challenge in estimating the usefulness (i.e., efficiency and accuracy) of a CAT. It permits implementing a P-H simulation from an incomplete data matrix. A hybrid simulation has two possible options. One option can be implemented if the parameters of the items to be used in a CAT have not previously been estimated, while the second can be implemented if assessing the effectiveness of a CAT with a new group of examinees. The first option—a hybrid simulation with item parameters estimated from the sparse data matrix (HPP)—has eight steps:

1. Divide the item bank into prespecified subsets of items along with a prespecified group of linking items.
2. Divide the examinees such that each examinee only takes one of the item subsets along with all of the linking items.
3. Administer only the selected items to the selected examinees.
4. Score the items to create a matrix of scored item responses (e.g., 1s and 0s) and place a missing data value (e.g., a "-") in the intersection of items that the examinee did not take.

5. Calibrate the item parameters with a concurrent-calibration method that will work properly with a sparse data matrix and will link the item parameters onto the common scale defined by the linking items (e.g., XCALIBRE; see Vale, 1986, p. 338, for a description of the concurrent-calibration method implemented in XCALIBRE).
6. Estimate the θ of each examinee based on the item parameter estimates and the matrix of responses, ignoring items that have not been administered (i.e., treating them as "not reached," rather than "skipped.").
7. Use the θ estimates from Step 6 and the item parameter estimates from Step 5 to impute the missing data by simulating responses to items that examinees did not take, using the same IRT model that was used to estimate item parameters.
8. Implement a P-H CAT simulation on the now complete hybrid matrix of item responses in the same way as above.

The second method—a hybrid simulation based on a sparse data matrix, but using item parameters previously estimated on another sample (HFP)—also has eight steps:

1. Use items and item parameters from an already calibrated item bank.

2-5 are the same as Steps 1-4 of HPP above.

6-8 are the same as Steps 6-8 of HPP above with the exception that the *already calibrated* item parameters are used to estimate the θ of each examinee and to simulate responses (i.e., Step 5 of the HPP simulation is skipped).

An advantage of the second approach is that its sample size requirements are much less than that of the HPP approach, which requires large samples to estimate item parameters.

HPP could be used when test developers create a new test as a CAT; thus, the item bank would be too large for any examinee to answer all the items in a bank, and item parameters would have to also be estimated with a concurrent-calibration procedure. HFP could occur when test developers try to turn an existing series of tests, measuring the same trait or ability, into a CAT, for which they should already have the item parameters from previous administrations. This circumstance could also occur if an item bank were calibrated on one group of people, and if a researcher were interested in how this item bank performed, as a CAT, on a separate group of people without giving anyone in this second group the entire test. In general, for both scenarios, interest lies in the amount of information needed to estimate statistical properties of a CAT under practical constraints with P-H simulation.

Method

Simulees, Items, and Programs

The M-C/hybrid simulation study used a full-test length of 620 items and a sample size of 1,000 simulees randomly drawn from a standard normal distribution, where $\theta \sim N(0,1)$. Responses were simulated, for each simulee, from the logistic approximation to the normal ogive IRT model,

$$P[1 \mid \theta_j, a_i, b_i, c_i] = c_i + (1 - c_i)\frac{\exp[1.7a_i(\theta_j - b_i)]}{1 + \exp[1.7a_i(\theta_j - b_i)]}. \quad (1)$$

where θ_j is the latent person parameter for examinee *j*,

a_i is the slope/discrimination parameter of item *i*,

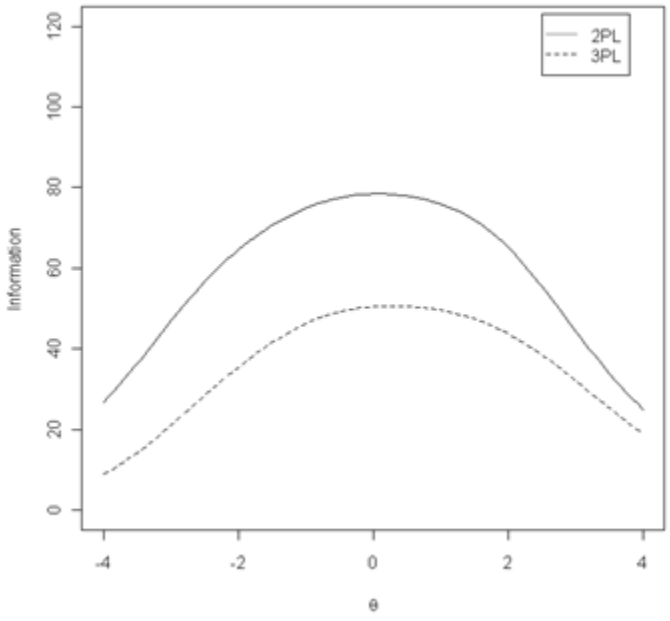
b_i is the location/difficulty parameter of item *i* and

c_i is the lower asymptote parameter of item *i* (sometimes called the pseudo-guessing parameter).

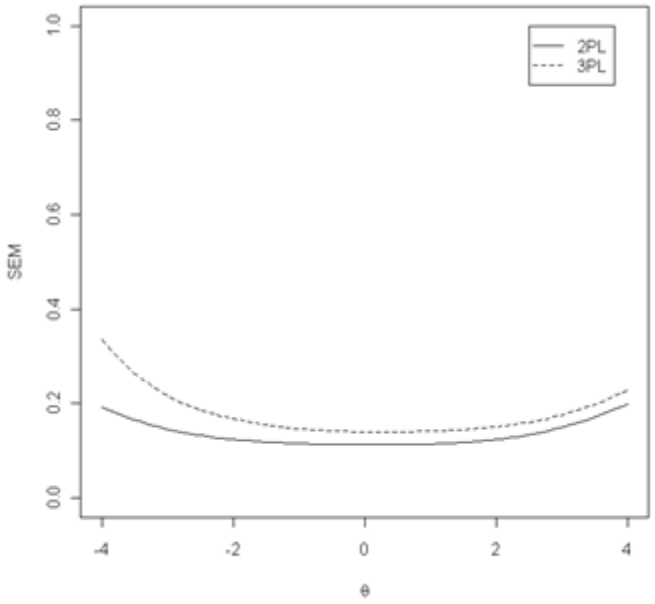
A vector of discrimination parameters was generated from a lognormal distribution, with $a \sim \ln(-.15, .25)$, which is approximately equivalent to a mean of .88 and a standard deviation of .23 in the IRT metric, and a vector of difficulty parameters was generated from a uniform distribution, with $b \sim \text{unif}(-3.0, 3.0)$; for the three-parameter model (3PL), the vector of pseudo-guessing parameters remained a constant $c = .20$; for the two-parameter model (2PL), the vector of pseudo-guessing parameters was set to $c = 0$, effectively eliminating the third parameter from the model. The parameter distributions were designed to approximate an ideal item bank for CAT, with a reasonable size item bank of fairly high information across a wide range of θ (see Weiss, 1982, p. 478, for a rationale of equiprecise CAT measurement using a rectangular distribution of difficulties). Though each replication generated new parameters, and thus new item banks, Figure 1 provides example bank information function (BIF) and bank standard error of measurement (SEM) functions from two 2PL and 3PL item bank [See Hambleton, Swaminathan, & Rogers (1991, pp. 91–99) for equations for test information functions (BIFs in this context) and the related SEMs.)].

Figure 1. Bank Information and SEM Functions for Example 620-Item 2PL and 3PL Item Banks

a. Bank Information Functions



b. SEM Functions



For the hybrid based on M-C simulated data, four programs performed the requisite sequence of data generation and estimation. A program written in R (R Development Core Team, 2008) generated simulees/items, simulated a full matrix of responses to those items, randomly divided simulees into specified groups, and analyzed the final output. XCALIBRE (Assessment Systems Corporation, 1995) estimated the matrix of item parameters and vector of $\hat{\theta}s$ from the R output matrix of item responses. The options for XCALIBRE were as follows: for all conditions, a floating prior was used; for all conditions, starting priors were $a_i \sim N(.80, .20)$ and $b_i \sim N(0.00,100.0)$; for all 3PL conditions, the starting prior for was $c_i \sim N(.20, .025)$. SCOREALL (Assessment Systems Corporation, 1998) estimated the vector of $\hat{\theta}s$ given the R output matrix and XCALIBRE parameters. Finally, POSTSIM (Assessment Systems Corporation, 2007) took the final R output matrix with XCALIBRE parameters and perfomed a P-H CAT simulation. XCALIBRE estimated item parameters for the HPP matrix by concurrent full-metric linking; instead of calibrating parameters by linking items to one group, XCALIBRE created a meta-scale by means of the anchor items, and linked all of the other items to that scale (see Gierl & Ackerman, 1996; Vale, 1986).

The real-data hybrid (R-D/hybrid) study was based on a personality questionnaire of 615 dichotomous items completed by 204 people. Because the 3PL offers no advantage over the 2PL in model fit to personality data (e.g., Chernyshenko et al., 2001), the data were assumed to fit a 2PL for estimation purposes; furthermore, item parameter estimates, based on the 2PL, were already provided for the response data. Three of the four programs mentioned above performed the sequence of data generation and estimation. A program written in R randomly selected items to delete listwise, deleted responses to those items, simulated responses according to the 2PL, and analyzed the final output. SCOREALL estimated the vector of $\hat{\theta}s$ given the R output matrix with deleted items and previously estimated item parameters. POSTSIM took the final R output matrix with provided item parameters and performed a P-H CAT simulation. The main differences between the R-D/hybrid and the M-C/hybrid were that the R-D/hybrid had real

responses to a full set of items with already estimated parameters, while the M-C/hybrid simulated θs, item parameters, and item responses.

Procedure

M-C/hybrid study. The M-C/hybrid procedure consisted of four parts: initial generation, estimation, final generation, and P-H simulation. During initial generation, an IRT model was selected: either a 2PL or a 3PL. A grouping of two, four, five, or ten item/examinee blocks was selected; the numbers of blocks was chosen so that the effects of differing levels of missingness could be assessed and so that the numbers of blocks divided evenly into the number of items (which resulted in no simulee in one group responding to the same item as a simulee in another group, except for the 20 linking items). For each item/examinee block, 620 items and 1,000 examinees were generated, 20 of those items were randomly selected as anchor items, and the remaining items and the 1,000 simulees were divided equally into each of the specified number of groups; a simulee took the 20 anchor items along with items corresponding to its group. Thus, 0-1 responses were simulated for the full matrix of items by simulees, and responses were deleted for items not in a simulee's group

The final output of initial generation consisted of two matrices: (1) a matrix of full responses for each simulee to every item, and (2) a partial (sparse or incomplete) matrix of responses to the 20 anchor items and the items in a simulee's group, with all other responses deleted. Table 1 provides details of the number of items and percentage of total items each simulee took based on its assigned group. Note that for each condition, every simulee took the same number and the same subset of items. The linking items were used in estimating parameters from the HPP matrix, and were also used to estimate θ (but not the parameters) for HFP. The HFP parameters were taken from FFP, which was a standard M-C simulation, with all 1,000 simulees taking all 620 items; FFP was needed in order to compare how well the hybrid simulations performed relative to a standard simulation. Furthermore, even though the linking items were not used in item parameter estimation for HFP, because simulees in HPP and HFP took the same number of items and the same items, this would control for extraneous item effects.

Table 1. Conditions of the M-C/Hybrid Study

	Group Condition			
Condition	2 Groups	4 Groups	5 Groups	10 Groups
Number of items each simulee took[a]	320	170	140	80
Percentage of items each simulee took[a]	.516	.274	.226	.129
Percentage of missing data for each simulee[a]	.484	.726	.774	.871
IRT model 1	2PL	2PL	2PL	2PL
IRT model 2	3PL	3PL	3PL	3PL
Base condition	FFP	FFP	FFP	FFP
Hybrid condition I	HFP	HFP	HFP	HFP
Hybrid condition II	HPP	HPP	HPP	HPP
Termination criterion 1 (fixed)	40 Items	40 Items	40 Items	40 Items
Termination criterion 2 (variable)	SEM ≤ .20	SEM ≤ .20	SEM ≤ .20	SEM ≤ .20

[a]Including the 20 anchor items, in the hybrid conditions (not the base condition).

During estimation, three sets of parameter/θ combinations were created: (1) XCALIBRE was used to estimate the item parameter matrix and $\boldsymbol{\theta}$ vector from the full matrix of responses; (2) XCALIBRE was used to estimate the parameter matrix and $\boldsymbol{\theta}$ vector from the partial matrix of responses, using the linking items in concurrent calibration; or (3) the parameters estimated from the full matrix were used (with SCOREALL) to estimate the $\boldsymbol{\theta}$ vector using the partial matrix of responses. Each of these conditions corresponds to three theoretical scenarios:

1. The full matrix of responses is available (from which a P-H simulation can proceed directly)
2. Only a partial matrix of responses is available (both items and persons need to be calibrated with this partial matrix).
3. The item parameter estimates are available and only θ needs to be estimated for use in imputing missing data in the hybrid simulation.

The final step in data generation was imputing responses in the partial matrix of responses with the item and person parameter estimates from Scenarios 2 and 3 above in place of the missing data. For Scenario 2, the item parameters were estimated from the partial matrix, θ was estimated by maximum likelihood estimation (MLE; Harwell, Baker, & Zwarts, 1988) from the partial matrix item parameters and the partial matrix item responses, and simulated responses to unadministered items in the partial matrix were based on monte-carlo data generation. For scenario 3, the item parameters were estimated from the full matrix, θ was estimated by MLE from the full matrix item parameters and the partial matrix item responses, and responses to the unadministered items were generated in the same manner Thus, ultimately, there were three matrix/parameter pairs: (1) a matrix of full responses with parameters estimated from the full matrix (FFP), (2) a hybrid matrix of actual and imputed responses with parameters estimated from the partial matrix and responses imputed based on the partial matrix (HPP), and (3) a hybrid matrix of actual and imputed responses with parameters estimated from the full matrix and responses imputed based on the partial matrix responses and the full matrix parameters (HFP). For each of these matrix types, there were three datasets consisting of 2, 4, 5 and 10 groups (Table 1) to investigate the effects of varying proportions of imputed data (see Table 1 for a complete description of conditions).

During P-H simulation, a CAT was simulated with POSTSIM for each of the matrix/parameter pairs, selecting items based on Fisher information, and estimating θ using MLE. Starting with an initial θ estimate ($\hat{\theta}$) of 0.0, if a simulee responded in the keyed direction, $\hat{\theta}$ was set to 3.0 until a non-mixed response pattern was achieved; if a simulee responded in the non-keyed direction, $\hat{\theta}$ was set to -3.0 until a non-mixed response pattern was achieved. Finally, the P-H CAT was terminated after both a fixed criterion (40 items) and a variable criterion (SEM $\leq$ .20). Thus, there were eight conditions: (2PL or 3PL model) $\times$ (2 groups, 4 groups, 5 groups, or 10 groups), with three matrix/parameter pairs for each condition (FFP, HFP, and HPP), and two termination criteria for each condition (fixed and variable); each condition was replicated four times and all analyses were performed in R and averaged over replications.

R-D/hybrid study. The real-data hybrid procedure consisted of four parts: deletion, θ estimation, final generation, and P-H simulation. During deletion, items were selected for which item responses were deleted; the differing levels of missingness were nested, and the percentages

of missingness were the same as in the M-C/hybrid study (48.4%, 73.6%, 77.4%, and 87.1% missing data). Nested missingness means that for one replication, 48.4% of the items were selected as missing data, and R deleted responses to those items; following that deletion, an additional 73.6 – 48.4 = 25.2% of the total items was chosen such that all of those items came from the remaining 51.6% of items remaining once the 48.4% of items was deleted; this pattern of using all of the items deleted in the previous level of missingness and adding items to that bank was continued for the final two levels of missingness. Furthermore, as listwise deletion was used, for each of the items selected as missing data, all of the responses were deleted to those items. Following deletion, SCOREALL estimated θ from the partial matrix of responses and previously estimated item parameters.

Note that, in the R-D/hybrid procedure, only two types of matrices were used—one that consisted of a full matrix of real responses by 204 people to 615 items (similar to FFP above, with real responses instead of generated responses), and one which consisted of partially real responses to varying percentages of items, and partially estimated responses from already provided item parameter estimates (similar to HFP above). There was no real-data hybrid matrix corresponding to HPP. The final two steps were the same as the M-C/hybrid procedure; however, both 25 and 40 items were implemented as fixed termination criterion, along with SEM $\leq$ .25 as the variable termination criterion.

Analysis

For both the M-C/hybrid and the P-C/hybrid studies, the POSTSIM $\hat{\theta}s$ from the full matrix were compared with CAT θ estimates from the full matrix and both hybrid matrices using correlations and root mean square indices with a fixed termination criterion of 40 items (for both types of studies), a fixed termination criterion of 25 items (only for the R-D/hybrid), and a variable termination criterion of SEM $\leq$ .20. A variable termination criterion did not preclude final $\hat{\theta}s$ having an SEM exceeding .20; however, if an SEM fell below .20, the CAT ceased, and if an SEM exceeded .20 at the end of the test, the simulee had exhausted the entire item bank of all 620 items, which only rarely happened.

All correlations were Pearson product-moment correlations. RMSE was defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\theta}_{jkC} - \hat{\theta}_{jkF})^2}{N}} \qquad (4)$$

where j is the simulee index, k is the specific procedure used (FFP, HPP, or HFP), C is CAT, F is the full bank, and N is the total number of simulees (i.e., 1,000). The RMSE, thus, is the square root of the average squared deviation between the $\hat{\theta}$ estimated from all of the items in the full matrix and the CAT $\hat{\theta}$; because the full matrix is the best estimate of θ if the simulee completed the full item bank, the differences in the two $\hat{\theta}$ estimates essentially measure how deviant the CAT $\hat{\theta}$ was from the best estimate of θ (i.e., the amount of accuracy lost in order to gain the efficiency from a sparse matrix).

RMSD was defined as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\theta}_{jk} - \hat{\theta}_{jm})^2}{N}} \quad (5)$$

where j is the simulee index, k and m are the procedures used (FFP, HPP, or HFP), and N = 1,000. Defined in this way, the RMSD is the square root of the average squared deviation between a specific $\hat{\theta}$ and a $\hat{\theta}$ from a different procedure.

The model predicted SEM difference was defined as

$$d(\text{SEM}) = \frac{\sum_{i=1}^{N} (SEM_{j1\text{F}} - SEM_{jk\text{C}})}{N} \quad (2)$$

where i is the simulee index, k is the procedure used (FFP, HPP, or HFP), C is the SEM from the CAT, F is the SEM from the full bank, and N = 1,000; furthermore, a 1 in place of k in the first SEM indicates that all of the CAT SEMs were compared to the FFP SEM. Because the SEM from the FFP full bank will always be smaller than the CAT SEMs (due to the larger number of items), this number was always negative; furthermore, the larger this number is in absolute value, the larger the average increase in SEM is when implementing a specific hybrid CAT instead of having the simulee take all the items in the bank.

The full bank SEM is the expected SEM (based on expected information) for differing levels of θ using the full bank of items, while the model-predicted SEM is based on observed information, instead of expected information, and solely includes the CAT items given. In this case, for the full bank SEM, the researcher would estimate θ following the test, find the location on X axis of the bank SEM function (see Figure 1b) matching the $\hat{\theta}$, and determine the SEM from the full test information function. For observed SEM, the researcher would calculate the negative second derivative of the log-likelihood for all of the given items, then use the actual response in the second derivative (a 1 if the response was in the keyed direction, a 0 otherwise)—this is the observed information for each item. Finally, the researcher would sum all of the observed information values and take the reciprocal square-root of that sum. The difference between observed and expected information is that observed information is based on the 0-1 responses examinees gave to each item while expected information is based solely on the information function at $\hat{\theta}$.

Results

Standard Deviation Over Replications

Each of the statistics computed had small replication standard deviations for most conditions. For the fixed termination criterion, the replication standard deviation of the correlations ranged from 2.89×10^{-4} to 1.01×10^{-2} (with most in the 10^{-3} to 10^{-4} range), the replication standard deviation of the model-predicted full test SEM ranged from 2.50×10^{-4} (for FFP, 2PL, 2 groups) to 4.68×10^{-3} (for HPP, 3PL, 5 groups), the replication standard deviation of the model-predicted CAT SEM ranged from 2.22×10^{-4} (for FFP, 2PL, 2 groups) to 1.37×10^{-2} (for HPP, 3PL, 5 groups), and the replication standard deviation of the RMSE/RMSD ranged from 1.78×10^{-3} to 3.09×10^{-2}. For the variable termination criterion, the replication standard deviation of the

correlations ranged from 1.26×10^{-4} to 7.19×10^{-3} and the replication standard deviation of the RMSE/RMSD ranged from 1.67×10^{-3} to 1.55×10^{-2}; when examining the mean number of items taken in a variable-length CAT, the replication standard deviation was under 3.25 for all 2PL conditions and near 4.5 for all of the remaining conditions except the 3PL with 5 groups, where the replication standard deviation was near 6 for all conditions except HPP, which had a replication standard deviation of 8.13. For the median number of items taken in a variable-length CAT, the replication standard deviation was similar to the respective replication standard deviations for the mean number of items, though it was slightly higher for most conditions (reaching a maximum of 8.64 with HPP, 3PL, 5 groups).

M-C/Hybrid Simulation

The first three columns of Table 2 are the base comparisons (or FFP) for the fixed termination criterion. They represent a standard M-C CAT simulation. As would be expected, all of the correlations were high, with the 2PL the highest and a slight reduction when simulating data with the 3PL; furthermore, the RMSE rarely exceeded .20, and the average SEM difference [d(SEM)] was between –.0916 and .0954 for the 2PL and from –.114 to –.118 for the 3PL.

The middle three columns of Table 2 compare the FFP condition with the HFP condition for the fixed termination criterion; in these results, the imputed values did not include the extra error from parameter estimation, as the item parameters estimated from the complete full matrix were used to impute the missing responses in the partial matrix. The final three columns of Table 2, for HPP, compare the full-matrix θ estimates with those from the complete hybrid simulation for the fixed-length CAT; in this data set, the imputed values included error from parameter estimation with the partial matrix before imputation. Table 2 shows that the RMSE increased substantially in both conditions as the amount of missing data increased. HPP RMSEs ranged from .188 to .311 for the 2PL, and from .244 to .410 for the 3PL. HFP RMSEs ranged from .180 to .281 for the 2PL, and .241 to .371 for the 3PL; correlations for HFP decreased as the percentage of missing data increased, with a slightly larger increase for the 3PL. However, in the case of 87% missing data, every simulee would only have taken only 80 of the 620 items with the remainder of the item responses imputed (see Table 1), from which a CAT of 40 items was administered. Combined with the difficulty of estimating parameters for the 3PL, correlations of .940 and .934 with the FFP $\hat{\theta}$ for HFP and HPP, respectively, reflect positively on the practical implications of a hybrid CAT. Furthermore, the negligible difference between the two correlations of .0057 resulted from a substantially reduced number of simulees per item in order to estimate the parameters. Correlations for all other conditions remained above .95 (versus .98 for the FFP baseline condition); thus, the reductions in the correlations emerged at the limits of estimating the 3PL.

Table 2. Correlations, RMSE, and SEM Comparisons of Full Matrix $\hat{\theta}_F$ Estimates With Three $\hat{\theta}_C$ Estimates for a Fixed-Length 40-Item CAT

Percent of	FFP			HFP			HPP		
Missing Data	*r*	RMSE	*d*(SEM)	*r*	RMSE	*d*(SEM)	*r*	RMSE	*d*(SEM)
2PL Model									
48.4%	.989	.154	-.0916	.984	.180	-.0917	.984	.188	-.0910
72.6%	.989	.155	-.0947	.976	.227	-.0948	.975	.234	-.0934
77.4%	.989	.156	-.0933	.973	.239	-.0935	.973	.248	-.0946
87.1%	.988	.158	-.0953	.963	.281	-.0955	.959	.311	-.110
3PL Model									
48.4%	.983	.198	-.115	.974	.241	-.116	.974	.244	-.118
72.6%	.982	.202	-.117	.964	.285	-.118	.960	.309	-.126
77.4%	.983	.19	-.115	.958	.305	-.116	.955	.330	-.131
87.1%	.983	.200	-.117	.940	.371	-.119	.934	.410	-.147

On average, the correlations of HFP $\hat{\theta}_C$ with FFP $\hat{\theta}_F$ were only .0024 higher than the correlations of HPP $\hat{\theta}_C$ with FFP $\hat{\theta}_F$; moreover, on average the RMSE of HFP CAT $\hat{\theta}_C$ with FFP $\hat{\theta}_F$ were only .0183 lower than the correlations of HPP $\hat{\theta}_C$ with FFP $\hat{\theta}_F$. However, the average *d*(SEM) of HFP $\hat{\theta}_C$ with FFP $\hat{\theta}_F$ (–.0917 for 2PL with 48.7% missing data, to –.119 for the 3PL with 87.1% missing data) was much closer, for most conditions, to the base comparison (–.0916 for 2PL with 48.7% missing data, to –.117 for the 3PL with 87.1% missing data) than the average *d*(SEM) of HPP $\hat{\theta}_C$ with FFP $\hat{\theta}_F$ (–.0910 for the 2PL with 48.7% missing data, to –.147 for the 3PL with 87.1% missing data).

Table 3 presents the same comparisons as Table 2, but for the variable termination condition. Note, in comparison with Table 2, that the correlations between $\hat{\theta}_F$ and $\hat{\theta}_C$ remained consistently high for both models, actually increasing slightly when simulating responses with the 3PL; furthermore, the RMSE consistently decreased to some extent with the 3PL for FFP (for example, with 2 groups, the RMSE decreased from .178 for the 2PL to .168 for the 3PL) and remained around the same value for HPP (.209 with the 2PL to .217 with the 3PL for 48.4% missing data and HFP (.201 with the 2PL to .209 with the 3PL for 48.4% missing data). These results were likely due to the lower information in the 3PL bank, so that under the variable termination criterion, more items were required (approximately 25–30 items, on average) to reach a fixed SEM.

Similar to the fixed test length condition, with a variable test length the correlations and RMSE when imputing responses based on the full matrix item parameters (HFP) improved very slightly over the correlations and RMSE when imputing responses based on the partial matrix item parameters (HPP). The average correlation of HPP $\hat{\theta}_C$ with FPP $\hat{\theta}_F$ was .966, while the average correlation of HFP $\hat{\theta}_C$ with FPP $\hat{\theta}_F$ was .967, an improvement of only .001; furthermore, the average RMSE of HPP $\hat{\theta}_C$ with FPP $\hat{\theta}_F$ was .274, while the average RMSE of HFP $\hat{\theta}_C$ with FPP $\hat{\theta}_F$ was .262, a difference of only .012. However, as observed in Table 3, for all conditions, the average number of items given by the HFP CAT (42.7) was closer to the

average number of items in the FFP CAT (42.2) than to the average number of items in the HPP CAT (46.5), with the largest difference occurring in the 3PL where the difference increased as the percent of missing data increased.

Table 3. Correlations, and RMSE Comparisons of Full Matrix $\hat{\theta}_F$ Estimates With Three $\hat{\theta}_C$ Estimates for a Variable-Length CAT (Termination SEM = .20)

	FFP			HFP			HPP		
Percent of Missing Data	r	RMSE	Avg. Items	r	RMSE	Avg. Items	r	RMSE	Avg. Items
2PL Model									
48.4%	.985	.177	29.4	.981	.201	29.5	.98	.209	29.3
72.6%	.986	.174	31.3	.972	.242	31.3	.972	.248	30.8
77.4%	.985	.177	30.5	.969	.256	30.6	.969	.262	31.2
87.1%	.985	.177	31.8	.96	.293	31.9	.957	.317	34.5
3PL Model									
48.4%	.987	.168	52.7	.98	.209	53	.979	.217	54.3
72.6%	.988	.166	55.7	.969	.261	56.2	.968	.272	60.0
77.4%	.988	.166	53	.963	.285	54.2	.962	.295	62.1
87.1%	.987	.17	53.4	.946	.35	54.5	.942	.376	68.7

Tables 4 and 5 compare the three $\hat{\theta}_C$ estimates for the fixed termination criterion (Table 4) and the variable termination criterion (Table 5). As alluded to above and shown in Table 4, the three CAT procedures had similar SEMs for 48.4% and 72.6% missing data in the 2PL; however, with a large percentage of missing data (.774 or .871) with the 3PL, the SEM from the HPP CAT procedure diverged from both the HFP CAT SEM and the FFP CAT SEM. For example, the average SEM for the FFP CAT, with the 2PL and .484 missing data, was .175, close to both the respective HFP CAT SEM (.175) and the HPP CAT SEM (.174); however, by increasing missingness to .871, using the 2PL, the average SEM for the FFP CAT was .181, the same as the respective HFP CAT SEM, but slightly further from the respective HPP CAT SEM (.186). Furthermore, with the 3PL, the average FFP CAT SEM was .224, slightly closer to the average HFP CAT SEM of .226 than the average HPP CAT SEM of .239.

Moreover, even with a lower SEM (with a fixed termination criterion), and using fewer items (with a variable termination; see Table 5), the HFP CAT still slightly outperformed the HPP CAT in terms of correlations and RMSD with the full matrix CAT. The average correlation of HFP $\hat{\theta}_C$ with FFP $\hat{\theta}_C$ under the fixed termination criterion was .959, while the average correlation of HPP $\hat{\theta}_C$ with FFP $\hat{\theta}_C$ under the fixed termination criterion was .954; though small, this difference was observed for every condition (see Table 4). The average correlation of HFP $\hat{\theta}_C$ with FFP $\hat{\theta}_C$ under the variable termination criterion was .960, while the average correlation of HPP $\hat{\theta}_C$ with FFP $\hat{\theta}_C$ under the variable termination criterion was .954 (see Table 5). The RMSDs showed little differences across conditions (see Table 5).

The median was examined for the variable-length condition to determine whether the mean numbers of items were affected by a few outliers; the median entries in Table 5 were averaged

across replications. Comparing the median number of items in Table 5 with the mean number of items in Table 3, the mean number of items was always slightly larger (evidencing a positive skew), but the difference was small (the difference between a median of 66.25 to a mean of 68.7 for HPP 10 groups was the largest difference).

Table 4. Correlations, RMSD, and SEM Comparisons of Three $\hat{\theta}_C$ Estimates for a Fixed-Length CAT, and Average CAT SEM

Percent of	FFP & HFP		FFP & HPP		HFP & HPP		Average SEM		
Missing Data	*r*	RMSD	*r*	RMSD	*r*	RMSD	FFP	HFP	HPP
2PL Model									
48.4%	.984	.186	.982	.199	.983	.188	.175	.175	.174
72.6%	.970	.25	.968	.265	.974	.241	.180	.180	.179
77.4%	.965	.272	.964	.285	.972	.249	.178	.178	.179
87.1%	.954	.317	.948	.345	.969	.271	.181	.181	.186
3PL Model									
48.4%	.971	.257	.968	.270	.970	.262	.222	.224	.226
72.6%	.954	.324	.948	.351	.960	.3080	.228	.228	.237
77.4%	.948	.341	.941	.374	.956	.323	.223	.224	.239
87.1%	.927	.412	.918	.454	.953	.348	.224	.226	.254

Table 5. Correlations and RMSD Comparisons of Three $\hat{\theta}_C$ Estimates for a Variable-Length CAT, and Median Number of Items

Percent of	FFP & HFP		FFP & HPP		HFP & HPP		Median Number of Items		
Missing Data	*r*	RMSD	*r*	RMSD	*r*	RMSD	FFP	HFP	HPP
2PL Model									
48.4%	.979	.201	.977	.222	.978	.216	28.5	28.75	28.75
72.6%	.965	.275	.963	.285	.968	.265	30.75	30.75	30.25
77.4%	.959	.296	.957	.309	.966	.277	29.75	30.25	30
87.1%	.948	.336	.944	.358	.964	.288	31	31	33.5
3PL Model									
48.4%	.978	.218	.975	.237	.978	.222	51	51.25	52.75
72.6%	.963	.287	.960	.302	.971	.256	54.5	54.75	60
77.4%	.955	.314	.952	.331	.968	.274	52	52	60
87.1%	.936	.382	.930	.411	.965	.292	51.75	52	66.25

Table 6 compares the $\hat{\theta}_F$ of each condition. Because both the fixed and variable termination criteria did not affect the $\hat{\theta}_F$, Table 6 is identical for both conditions. Interestingly, the full bank SEM was relatively similar for all conditions except the 3PL with .871 missing data, in which it diverged slightly for HPP (.112) from both (HFP) (.108) and FFP (.107). However, the comparability of all but the latter suggests that the main benefit of using HFP over HPP (and, thus, employing already calibrated parameters from which to estimate θ, simulate responses, and perform a P-H simulation, instead of estimating both item and person parameters from the partial

matrix of responses) lies, mostly, in the CAT procedure. Moreover, the slight benefit in accuracy (both in terms of correlation and RMSE) of HFP over HPP stayed relatively the same through Tables 2–6, regardless of whether the comparison was between CAT vs. full bank, CAT vs. CAT, or full bank vs. full bank.

Table 6. Correlations, RMSE, and SEM Comparisons of Three $\hat{\theta}_F$ Estimates and Full Bank Average SEM

Percent of	FFP & HFP		FFP & HPP		HFP & HPP		Average SEM		
Missing Data	*r*	RMSE	*r*	RMSE	*r*	RMSE	FFP	HFP	HPP
2PL Model									
48.4%	.995	.097	.995	.103	.997	.082	.083	.083	.083
72.6%	.988	.158	.986	.169	.994	.116	.085	.085	.085
77.4%	.985	.176	.984	.186	.993	.121	.084	.084	.084
87.1%	.975	.231	.971	.255	.991	.142	.085	.085	.086
3PL Model									
48.4%	.992	.129	.992	.133	.994	.111	.108	.108	.108
72.6%	.981	.202	.980	.209	.991	.140	.111	.112	.112
77.4%	.976	.228	.974	.240	.990	.152	.108	.109	.110
87.1%	.958	.205	.955	.330	.988	.174	.107	.108	.112

R-D/Hybrid Simulation

Table 7 is similar to Table 3 in layout; however, the HPP condition was removed because the item parameters were provided before the simulation. Also, because FFP had no missing data, given a particular termination criterion, and because the matrix of full responses did not change, all of the FFP comparisons with itself (correlation between FFP $\hat{\theta}_F$ and FFP $\hat{\theta}_C$, mean and median number of items, RMSE, etc.) remained identical for the four levels of missingness. The first four rows of Table 7 provide results for a fixed termination criterion of 25 items, while the last four rows present results for a fixed termination criterion of 40 items. As the number of CAT items increased, all of the statistics improved. The correlation of FFP $\hat{\theta}_F$ with FFP $\hat{\theta}_C$ increased from .937 with 25 items to .953 with 40 items, the RMSE decreased from .485 to .421, and the difference in SEM decreased in absolute magnitude from .196 to .142. This trend held within conditions for the hybrid simulation ,as well. FFP $\hat{\theta}_F$ correlated with HFP $\hat{\theta}_C$ .958 with 25 items and 48.4% missing data, and FFP $\hat{\theta}_F$ correlated with HFP $\hat{\theta}_C$ .971 with 40 items and 48.4% missing data. Similarly the RMSE of FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_F$ decreased from .392 to .327, and the difference in SEM decreased in absolute magnitude from .196 to .141.

Table 7. Correlations, RMSE, and SEM Comparisons of Full Matrix $\hat{\theta}_F$ Estimates With Two $\hat{\theta}_C$ Estimates for Fixed-Length CATs Based on Real Data

Percent of	FFP			HFP		
Missing Data	*r*	RMSE	*d*(SEM)	*r*	RMSE	*d*(SEM)
25 Items						
48.4%	.937	.485	-.196	.958	.392	-.196
72.6%	.937	.485	-.196	.959	.408	-.208
77.4%	.937	.485	-.196	.957	.397	-.196
87.1%	.937	.485	-.196	.954	.420	-.197
40 Items						
48.4%	.953	.421	-.142	.971	.327	-.141
72.6%	.953	.421	-.142	.969	.350	-.153
77.4%	.953	.421	-.142	.968	.348	-.142
87.1%	.953	.421	-.142	.963	.379	-.143

Comparing the performance of the HFP CAT relative to the FFP CAT (with the FFP full matrix as the level of comparison), the correlation increased for the HFP CAT from .937 (comparing FFP $\hat{\theta}_F$ with FFP $\hat{\theta}_C$ with a fixed termination criterion of 25 items) to from .954 (comparing FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ for 87.1% missing data) to .958 (comparing FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ for 72.6% missing data). For either termination criterion, the level of missingness did not appear to make much of a difference for any of the statistics (correlations of .963 comparing FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ for 40 items and 87.1% missing data to .971 comparing FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ for 40 items and 48.4% missing data, RMSE of .392 to .420 for 25 items and .327 to .379 for 40 items, and SEM difference in absolute magnitude of around .197 for 25 items and around .142 for 40 items). However, the HFP CAT consistently performed better than the FFP CAT regardless of condition [higher correlation, lower RMSE, and lower *d*(SEM) in absolute magnitude].

As shown in Table 8, for the variable termination criterion, the correlations of FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ were still somewhat higher than the correlations of FFP $\hat{\theta}_F$ with FFP $\hat{\theta}_C$ (.973 for 48.4% missing data, .970 for 72.6% and 77.4% missing data, and .965 for 87.1% missing data for FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$, compared to a correlation of .957 for FFP $\hat{\theta}_F$ with FFP $\hat{\theta}_C$). Moreover, the RMSE of FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_C$ was .093 (for 48.4% missing data) to .036 (for 87.1% missing data) lower than the RMSE of FFP $\hat{\theta}_F$ with FFP $\hat{\theta}_C$. However, the increase in correlation and decrease in RMSE translated into at most a one-item improvement of the HFP CAT over the FFP CAT (48.4 average items for FFP CAT versus 47.3 average items for HFP CAT with 48.4% missing data). Except for the 48.4% missing data condition (which was only approximately one item lower than the FFP CAT), the rest of the item averages were similar comparing FFP to HFP.

Table 8. Correlations, RMSE, and Average Number of Items Comparisons of Full Matrix $\hat{\theta}_F$ Estimates With Two $\hat{\theta}_C$ Estimates for a Variable-Length CAT (Termination SEM = .25) Based on Real Responses

Percent of Missing Data	FFP			HFP		
	r	RMSE	Avg. Items	*r*	RMSE	Avg. Items
48.4%	.957	.402	48.4	.973	.309	47.3
72.6%	.957	.402	48.4	.970	.341	48.1
77.4%	.957	.402	48.4	.970	.331	47.7
87.1%	.957	.402	48.4	.965	.366	48.9

Comparing the fixed termination CATs of FFP with HFP (see Table 9), the correlations remained above .90. With 48.4% missing data, the correlation between FFP $\hat{\theta}_C$ and HFP $\hat{\theta}_C$ was .940 for a 25-item CAT and .959 for a for a 40-items CAT; with 87.1% missing data, the correlation between FFP $\hat{\theta}_C$ and HFP $\hat{\theta}_C$ was .908 for a 25-item CAT and .928 for a 40-item CAT. The RMSDs were approximately .500 for every condition, except 48.4% missing data with a 40-item CAT (where the RMSD was .393); these RMSDs would be high except for the fact that with 87.1% data missing completely at random, most of the HFP CAT responses were computer generated, while the FFP CAT responses had error from potential misfit. Furthermore, the average SEM of .310 for the FFP $\hat{\theta}_C$ with 25 items was approximately equal to .309 (with 48.4% missing data) to .322 (with 72.6% missing data) for the HFP $\hat{\theta}_C$, while the average SEM of .356 for the FFP $\hat{\theta}_C$ with 40 items was between .255 to .266 for the HFP $\hat{\theta}_C$.

Table 9. Correlations, RMSD, and SEM Comparisons of Two $\hat{\theta}_C$ Estimates for Fixed-Length CATs, and Average SEM

Percent of Missing Data	FFP & HFP		Average SEM	
	r	RMSD	FFP	HFP
25 Items				
48.4	.940	.473	.310	.309
72.6	.910	.593	.310	.322
77.4	.913	.572	.310	.310
87.1	.908	.599	.310	.311
40 Items				
48.4	.959	.393	.256	.255
72.6	.931	.515	.256	.266
77.4	.934	.499	.256	.256
87.1	.928	.528	.256	.257

As shown in Table 10, comparing the two variable termination CAT estimates, the correlations ranged from a high of .961 (comparing FFP $\hat{\theta}_C$ with an HFP $\hat{\theta}_C$ re-simulating 48.4%

of the data) to a low of .928 (comparing FFP $\hat{\theta}_C$ with an HFP $\hat{\theta}_C$ re-simulating 87.1% of the data); furthermore, the RMSD ranged from .377 for the former condition to .525 for the latter condition above. The correlations and RMSDs for the respective conditions were similar to the statistics of the 40-item fixed termination criterion. Also, the median number of items (35) for the HFP CAT was consistently one item lower than the median number of items for the FFP CAT (36).

Table 10. Correlations and RMSD Comparisons of Two $\hat{\theta}_C$ Estimates for a Variable-Length CAT, Median Number of Items, and Correlations, RMSE, and SEM Comparisons of Two $\hat{\theta}_F$ Estimates and Full Bank Average SEM

Percent of Missing Data	FFP $\hat{\theta}_C$ & HFP $\hat{\theta}_C$		Median No. of Items		FFP $\hat{\theta}_F$ & HFP $\hat{\theta}_F$		Average SEM	
	r	RMSD	FFP $\hat{\theta}_C$	HFP $\hat{\theta}_C$	*r*	RMSE	FFP $\hat{\theta}_F$	HFP $\hat{\theta}_F$
48.4%	.961	.377	36	35	.994	.141	.114	.114
72.6%	.934	.498	36	35	.986	.229	.114	.115
77.4%	.937	.483	36	35	.985	.238	.114	.115
87.1%	.928	.525	36	35.25	.979	.278	.114	.115

Finally, comparing the FFP $\hat{\theta}_F$ with HFP $\hat{\theta}_F$, the correlations were consistently high (.979 for 87.1% missing data to .994 for 48.4% missing data), the RMSE was consistently low (.141 for 48.4% missing data to .278 for 87.1% missing data), and the average full test SEMs were approximately the same (.114) for all conditions. The latter results suggest that randomly removing 48.4% to 87.1% of the data did not substantially affect the precision of the θ estimates.

Discussion and Conclusions

This study addressed a methodological problem in CAT implementation—namely, if researchers cannot administer every item in a CAT item bank, how can they predict the performance of a CAT with a real item bank? Beyond measuring the RMSE, bias, SEM, and average number of items administered under a specified configuration of CAT options, other procedural concerns (see Chang & Ying, 1999, van der Linden & Chang, 2003, van der Linden & Velkamp, 2004 for references on item exposure controls) depend on knowing the relative frequency of items (based, solely, on the item selection algorithm), and this necessitates every item having the possibility of occurring in every test. Thus, some method of imputing responses to non-administered questions was needed; the approach examined here was to use parameter and θ estimates from a sparse data matrix and simulate missing responses from an assumed IRT model.

Simulating responses adds a potential source of error—namely that of incorrectly estimating θ, and consequently, inaccurately specifying the probability of response. In addition, estimating item parameters from a sparse matrix adds another potential source of error—the loss in precision by having fewer responses to estimate item parameters. These sources of error would be expected to increase the RMSEs and SEMs of the θ estimates. That is what was found in the hybrid simulations based on monte-carlo data; the FFP CAT, with the least amount of estimation error had a lower RMSE and SEM than the HFP, with only the added estimation error of θ estimates based on a smaller number of items, which had a lower RMSE and SEM than the HPP, which included the estimation error of incorrect θ estimates plus estimation error from parameter

estimation using a sparse data matrix. Furthermore, as the amount of missing data increased, the RMSEs and SEMs also became larger.

Because an increase in RMSE and SEM was inevitable, it was necessary to determine if the tradeoff in the ability to fully estimate θ from a CAT was worth the added imprecision. Remarkably, the results from Tables 2 and 3 suggest that if item parameters are available from a preliminary calibration, estimating the gain in efficiency by using a CAT over a standard test would lose almost no information with a partial matrix of responses – even with up to 80% missing data. The main loss of information would reside in the estimated accuracy; however, in Tables 2 and 3, the correlations remained high and the SEM remained low comparing either hybrid matrix with the full matrix, except at the limits of estimating the model.

The results of this study suggest that, for every condition, although the two hybrid matrices were similar in the accuracy of their θ estimates, the HFP, which had parameters estimated from the full matrix (or, in a real application of the procedure, had parameters already provided), did so with efficiency close to that of the full-matrix CAT. Furthermore, even with 87% missing data, using HPP, in which only 100 simulees took each of the items (in addition to the 20 anchor items that every simulee took), correlations of θ estimates still ranged in the middle .90s.

Because the correlations were high, it was necessary to determine the difference between the simulation methods. In this study, if the original responses perfectly fit the model, the hybrid statistics were biased to underpredict performance; this is clear when examining the full parameter CAT (the base condition) versus any of the other two conditions. The base condition generally had higher correlations and lower RMSEs. However, with the conditions using real responses to real data, the hybrid statistics were biased to overpredict performance, and this was more drastic for the lower number of groups. Thus, it appeared as though, in the monte-carlo based hybrid stimulation, there was only one source of error; in the real-data hybrid simulation there were two sources of error—the original source of error (incorrectly estimating θ with fewer responses to fewer items), and a source of error whereby new responses (once θ was estimated) perfectly fit the model. With fewer groups, the latter source of error was predominant and most estimation statistics were drastically overperforming; however, as the number of groups increased and as missing data increased, the two sources of error started to balance each other, and the performance estimates started to approach actual performance levels.

Limitations

Since this report is the first implementation of the hybrid simulation procedure, the θ distribution was kept simple, though, in the future, researchers should replicate this study with different distributional assumptions. Furthermore, an ideal CAT item bank was created, with relatively high discriminations and a wide range of item difficulties. A researcher seeking to administer an adaptive test might not have the resources to develop an item bank similar to the one used in this study; with a more peaked item bank or with lower discriminations, the correlations between θ estimates from CATs using a hybrid matrix and the θ estimates from all items in the full matrix might be lower. Mills and Stocking (1996), addressed a wide range of CAT administration issues, including item bank characteristics, item ordering and grouping (e.g., CATs based on testlets; Wainer & Kiely, 1987), and item selection initiation and termination, each of which would have to be examined within the context of a hybrid simulation paradigm. However, though more research is needed, the preliminary results show the hybrid simulation method appears to effectively solve a heretofore unresolved problem in CAT.

References

Assessment Systems Corporation. (2007). POSTSIM 3 [Computer Software]. St. Paul, Minnesota: Author.

Assessment Systems Corporation. (1998). SCOREALL for Windows (Version 2.00) [Computer Software]. St. Paul, Minnesota: Author.

Assessment Systems Corporation. (1995). XCALIBRE for Windows (Version 1.10c) [Computer Software]. St. Paul, Minnesota: Author.

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research , 16*, 95-108.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Chang, H.-H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement , 23*, 211-222.

Colton, G. D. (1998). Exam security and high-tech cheating. *The Bar Examine , 67* (3), 13-35.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment , 5* (8), 1-39.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services , 59* (4), 361-368.

Gierl, J. M., & Ackerman, T. (1996). XCALIBRE — marginal maximum-likelihood estimation program, Windows version 1.10. *Applied Psychological Measurement , 20*, 303-307.

Guo, F. (2007). CAT security: A practitioners perspective. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved April 13, 2009 from www.psych.umn.edu/psylabs/CATCentral/

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Washington DC: Sage Publications.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243-271.

Harwell, M. R., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement , 22*, 131-143.

Kim, S. H., & Plake, S. B. (1993). Monte carlo simulation comparison of two-stage testing and computerized adaptive testing. *Paper presented at the annual meeting of the National Council on Measurement in Education.* Atlanta, GA.

Leucht, R. M. (2003, April). Exposure control using adaptive multi-stage item bundles. *Paper presented at the annual meeting of the National Council on Measurement in Education.* Chicago, IL.

Makransky, A. G. (2008). Computer adaptive testing: An introduction to the practial considerations related to developing a computer adaptive test for occupational testing. In K. Benny, *Metodologiske Indblik Og Udsyn* (pp. 197-223). Copenhagen: Psykologisk Institut.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement , 27*, 121-137.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement In Education , 9*, 287-304.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ree, M. J., & Jensen, H. E. (1980). *Item characteristic curve parameters: Effects of sample size on linear equating.* Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* (ETS Research Report No. 93–2). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365-389.

Stocking, M. L., Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (ETS Research Report No. 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement , 7*, 201-210.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *In Procedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Centre.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement , 10*, 333-344.

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.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics , 29*, 273-291.

Wainer, H. (2000a). CATs: Whither and whence. *Psicológica, 21*, 121-133.

Wainer, H. (2000b). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics , 25*, 203-224.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement , 24*, 185-201.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive testing: An illustration with the absorbtion scale. *Journal of Personality and Social Psychology , 57,* 1051-1058.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement , 6,* 473-492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774-789.

Weiss, D. J. & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Weiss. D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

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.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica , 21,* 135-155.