An Adaptive Exposure Control Algorithm for Computerized Adaptive Testing using a Sharing Item Response Theory Model

Daniel O. Segall
Defense Manpower Data Center
Monterey Bay, CA

Abstract

A new sharing item response theory (SIRT) model is presented which explicitly models the effects of sharing item content between informants and test-takers. This model is used to construct adaptive item selection and scoring rules that provide increased precision and reduced score gains in instances where sharing occurs. The adaptive item selection rules are expressed as functions of the item's exposure rate in addition to other commonly used properties (characterized by difficulty, discrimination, and guessing parameters). Based on the results of simulated item responses, the new item selection and scoring algorithms compare favorably to the Sympson-Hetter exposure control method. The new SIRT approach provides higher reliability and lower score gains in instances where sharing occurs between informants and test-takers.

Introduction

In recent years, computerized adaptive testing (CAT) has grown in popularity because of its many advantages over conventional paper-and-pencil administration. These advantages include, but are not necessarily limited to increased measurement accuracy and shorter test-lengths (Sands, Waters, & McBride, 1997; van der Linden & Glas, 2000; Wainer, 2000). CAT has also grown in popularity through its association with on-demand testing. These exams can be administered at the convenience of the test-taker since they are essentially self -paced and -administered.

Compared to periodic test schedules (where essentially different items are administered on different testing occasions), on-demand testing has a serious shortcoming: test

This paper was presented at the annual meeting of the National Council on Measurement in Education (April, 2003), Chicago, IL. The views expressed are those of the author and not necessarily those of the Department of Defense, or the United States government. Requests for copies should be sent to: Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside, CA 93955-6771. Email: publications@danielsegall.com

security. With on-demand test schedules, the same test items are administered over multiple occasions (spanning weeks, months, or possibly years). This continuous item exposure provides increased opportunities for test compromise. Several item selection algorithms have been proposed to help moderate the effects of compromise. These algorithms (Davey & Parshall, 1995; Stocking, 1993; Stocking & Lewis, 1998; Stocking & Lewis, 2000; Thomasson, 1995), based in large part on the Sympson-Hetter algorithm (Hetter & Sympson, 1997; Sympson & Hetter, 1985) limit the exposure of the pool's most informative items in an attempt to reduce the advantages to test-takers of sharing item content. Although these exposure-control algorithms might help reduce the degradation in measurement precision associated with compromise, evidence suggests that there is still substantial scoring advantages given to examinees who preview item content from one or more friends or informants (Segall, 1995; Segall & Moreno, 1997). Accordingly, some test-developers have gone to extraordinary lengths to help ensure test security, including the frequent replacement of entire item pools.

This paper investigates several issues associated with test compromise in CAT. First, a new sharing item response theory (SIRT) model is derived and evaluated. This model is used to construct new CAT item selection and scoring algorithms that provide increased measurement precision and reduced scoring advantages in the presence of compromise. Second, the performance of this method is evaluated with simulated response data, and is compared to the performance of one of the most commonly used exposure control algorithms, the Sympson-Hetter procedure.

The Sharing Item Response Theory (SIRT) Model

We begin by hypothesizing a specific compromise behavior based on the sharing of item content between informants and test-takers. According to this model, a given test-taker has h informants who have taken the adaptive test before him (where $h = 0, 1, 2, ..., n_h$). The informants disclose r randomly chosen items (out of n items received) to the test-taker. Accordingly, each item in the pool ($i = 1, ..., N$) has two states with regard to preview by the test-taker:

$$v_i = \begin{cases} 0, & \text{if item } i \text{ has not been previewed,} \\ 1, & \text{if item } i \text{ has been previewed.} \end{cases}$$

Then the conditional probability that item i has been previewed given h informants have participated in the disclosure, is stated by

$$p(v_i = 1|h) = 1 - \left(1 - \frac{r}{n} e_i\right)^h \ , \tag{1}$$

where e_i is the exposure rate of item i (defined as the probability of receiving item i).

Next we denote item responses by $u = (u_1, u_2, ..., u_n)$, where a correct response to item i is denoted by $u_i = 1$, and an incorrect response by $u_i = 0$. Assuming that all previewed items are answered correctly by the test-taker, the probability of a correct response to item i conditional on ability θ and number of informants h is given by

$$p_i(u_i = 1|\theta, h) = p(v_i = 1|h) + p_i(u_i = 1|\theta)p(v_i = 0|h) \ , \tag{2}$$

where $p_i(u_i = 1|\theta)$ is the probability of a correct response for item i conditional on θ, modeled by the three-parameter logistic (3PL) function (Birnbaum, 1968):

$$p_i(u_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i\left(\theta - b_i\right)\right]} , \tag{3}$$

and $p(v_i = 0|h) = 1 - p(v_i = 1|h)$.

Following from conditional (on θ and h) independence assumptions among item responses, the joint distribution of parameters and data can be expressed by

$$p(u, \theta, h) = p(\theta)p(h)\prod_{i=1}^{n} p_i(u_i|\theta, h) , \tag{4}$$

where $p(\theta)$ and $p(h)$ are independent prior distributions for θ and h, respectively, and where

$$p_i(u_i|\theta, h) = \left[p_i(u_i = 1|\theta, h)\right]^{u_i}\left[1 - p_i(u_i = 1|\theta, h)\right]^{1-u_i} .$$

Summing (4) over values of h, the joint distribution of θ and u is provided by

$$\begin{aligned} p(u, \theta) &= \sum_{h=0}^{n_h} p(u, \theta, h) \\ &= p(\theta)\sum_{h=0}^{n_h}\left[p(h)\prod_{i=1}^{n} p_i(u_i|\theta, h)\right] . \end{aligned} \tag{5}$$

The posterior distribution of ability θ given data u is provided by

$$p(\theta|u) = p(\theta, u)/p(u) , \tag{6}$$

where $p(u) = \int p(\theta, u)d\theta$.

Central tendency measures of the posterior density (6) (such as the posterior mean) can in principle provide an ability estimate which should be less contaminated by the effects of item disclosure than ability estimates produced by the standard item response theory model. Dispersion measures obtained from (6) (such as the posterior variance) can be used to summarize the level of uncertainty regarding θ in the presence of both measurement-error, and uncertainty due to item disclosure by informants. Numerical approximations to the posterior variance and mean can be computed from (14) and (15).

Note that the joint probability given by (4) implies conditional independence among item disclosure outcomes. According to the model, the probability of item disclosure is conditionally dependent on h (the number of informants). In practice the probability of disclosure is also likely to be dependent on the ability level(s) of the informant(s) as well. The effects of this model simplification will be assessed in the simulation study described in a following section.

Item Selection

The SIRT model can be used to construct an adaptive item selection algorithm which explicitly considers the item's exposure rate and its effect on the reduction of posterior uncertainty. This item selection approach is provided below, along with two enhancements intended to provide improved test-score precision.

Minimum Expected Variance Criterion

The posterior density function given by (6) can be used to construct an item selection algorithm based on a minimum expected variance (MEV) criterion (van der Linden & Pashley, 2000, p. 16, eq. 28). According to this criterion, the next item to be chosen is the one that minimizes the expected posterior variance. The expected posterior variance for item k (where k indexes the item in a pool of previously unadministered items), denoted by $\mathrm{E}\left[\mathrm{Var}(\theta|u,u_k)\right]$, can be expressed by

$$\mathrm{E}\left[\mathrm{Var}(\theta|u,u_k)\right] = \mathrm{Var}(\theta|u,u_k=1)p\left(u_k=1|u\right) + \mathrm{Var}(\theta|u,u_k=0)p\left(u_k=0|u\right) \ , \quad (7)$$

where u contains the responses to previously administered items. The predictive posterior distribution for the response to item k can be calculated from the ratio of two terms:

$$p\left(u_k|u\right) = \frac{p(u,u_k)}{p(u)} \ ,$$

where the numerator is expressed by

$$p(u,u_k) = \int p(\theta,u,u_k)d\theta \ , \quad (8)$$

and denominator by

$$p(u) = \int p(\theta,u)d\theta \ . \quad (9)$$

The integrals on the right sides of (8) and (9) can be approximated by substituting the appropriate item-related terms (and responses) into (5) and then using a one-dimensional numerical integration algorithm. The variance terms $\mathrm{Var}(\theta|u,u_k=1)$ and $\mathrm{Var}(\theta|u,u_k=0)$ in (7) can also be approximated numerically from (5). Computational details are provided by (14) through (16).

Stochastic Minimum Expected Variance Item Selection

The minimum expected variance (MEV) item selection algorithm falls in the class of greedy selection algorithms—at each stage in the item selection process the minimum variance item is always selected, regardless of how close other items are in terms of their variance estimates. Consequently, the usage or administration rates of nearly identical (in terms of their item-response functions) items can vary widely. Two items with nearly identical discrimination, difficulty, and guessing parameters can have very different usage rates. Among items with the same difficulty levels, preference is given to the item with the slightly higher discrimination level (and lower guessing parameter).

To help equalize the administration rates of similar items, the MEV criterion can be modified to include a probabilistic or stochastic component, resulting in the stochastic minimum expected variance (SMEV) criterion. According to this approach, the first item is chosen with probability equal to

$$X_k^{(1)} \propto 1 - \mathrm{E}\left[\mathrm{Var}(\theta|u,u_k)\right] \ , \quad (10)$$

where the subscript k denotes the kth unadministered item in the pool, and the $X_k^{(1)}$ values are normed so that their sum is equal to one. The second and subsequent items ($i = 2, ..., n$) are also chosen stochastically, so that the probability of selection is equal to

$$X_k^{(i)} = \frac{Z_k}{\sum_j Z_j} \,, \tag{11}$$

where

$$\begin{aligned} Z_k &= \frac{\{1 - \mathrm{E}\left[\mathrm{Var}(\theta|u, u_k)\right]\} - \{1 - \mathrm{Var}(\theta|u)\}}{1 - \mathrm{Var}(\theta|u)} \\ &= \frac{\mathrm{Var}(\theta|u) - \mathrm{E}\left[\mathrm{Var}(\theta|u, u_k)\right]}{1 - \mathrm{Var}(\theta|u)} \,. \end{aligned} \tag{12}$$

The value Z_k is equal to the percent of relative increase in explained test-score variance due to the administration of item k, relative to the amount of variance already explained by previously administered items.

Purposely Over-Exposed Items

The goal of exposure control algorithms is to limit the usage of items, most typically the usage of highly discriminating items of moderate difficulty. With the standard item response theory (IRT) model, these frequently administered items are likely to be problematic in instances where examinees share item content. When sharing occurs, responses to highly exposed items are likely to degrade the precision of the final ability estimates.

In contrast, highly exposed items can provide useful information when ability is estimated using the SIRT model. This is especially true if the highly exposed items are difficult highly discriminating items. With these highly exposed difficult items, the probability of a correct response is high through sharing, and low otherwise. Consequently these items can provide useful information regarding the number of informants h that shared item-content with the test-taker. If a large proportion of highly-exposed extremely-difficult items are answered correctly, then it is plausible that $h > 0$ (i.e., the test-taker benefited from the aid of informants). Conversely, if few highly-exposed difficult items are answered correctly, then the plausibility that $h = 0$ is greatly enhanced.

The role of highly exposed difficult items can be examined more formally through a re-examination of components used in the computation of $p(u, \theta)$. From (5), we see that the kernel of this expression is a weighted average of conditional likelihood functions—where each likelihood function is conditional on the number of informants $L(\theta|u, h) = \prod_{i=1}^{n} p_i(u_i|\theta, h)$. In cases where a number of difficult highly-exposed items are answered correctly, the likelihood functions conditional on non-zero informant levels $L(\theta|u, h = 1), ..., L(\theta|u, h = n_h)$ will be large relative to the likelihood function conditional on zero informants: $L(\theta|u, h = 0)$. This is evident from (2). Consequently when highly-exposed difficult items are answered correctly, the posterior density $p(\theta|u)$ will display increased plausibility over the lower ranges of θ, since correct responses can be obtained through sharing (when $h > 0$) as well as through proficiency and guessing. Conversely, when difficult highly-exposed items are answered incorrectly, the likelihood functions conditional on $h > 0$ will be small relative to the likelihood function conditional on $h = 0$. Consequently when highly-exposed difficult

items are answered incorrectly, the shape of the posterior density $p(\theta|u)$ will more closely approximate the posterior density based on the standard IRT model, where the probability of a correct response is expressed as a function of ability θ only.

These observations suggest that the performance of the SIRT procedure, unlike conventional exposure control strategies, might actually be enhanced through the deliberate administration of highly-exposed items to each examinee—given the caveat that these items are difficult and moderate to highly discriminating. According to this strategy, these items, termed *Trojan items*, would be administered early in the adaptive sequence so that responses to these items could influence the choice of subsequent items. If these items are answered correctly, then the SIRT algorithm is likely to choose less exposed items for the remainder of the test. If these difficult highly-exposed items are answered incorrectly, then the algorithm is likely to rely on more heavily exposed (and possibly more informative) subsequent items. The usefulness of Trojan items and their effects on SIRT precision is examined in the simulation study described below.

Simulation Study

A simulation study was conducted to answer several questions related to the SIRT item selection and scoring algorithms. First how well do the SIRT procedures counter the effects of sharing between informants and test-takers? Second, what benefit is there (if any) to the forced administration of Trojan (highly-exposed difficult) items? And third, how does the performance of the SIRT approach compare to another item selection and scoring approach based on the Sympson-Hetter item exposure control algorithm?

Item Pool

The item pool consisted of 300 items where the difficulty parameters b were equally spaced from -1.5 to $+1.5$. The slope a_i and guessing c_i parameters were sampled from independent uniform distributions: $a_i \sim \mathrm{U}[0.5, 1.5]$ and $c_i \sim \mathrm{U}[0, 0.3]$, for $i = 1, ..., 300$.

Informant Distribution

The prior distribution of informants was assumed to be

$$p(h) = \begin{cases} 0.60, & h = 0, \\ 0.20, & h = 1, \\ 0.10, & h = 2, \\ 0.05, & h = 3, \\ 0.03, & h = 4, \\ 0.02, & h = 5. \end{cases} \tag{13}$$

Accordingly, 40% of the population benefited from the disclosure of one or more informants, with 10% of the population previewing items from 3 or more informants. This prior distribution was used in all SIRT item selection and scoring calculations. This distribution was also used to generate the number of informants h for simulation conditions where the number of informants was randomly sampled for each test-taker.

SIRT Calculations

Item selection and scoring calculations were based on the SIRT model. Since there are no closed form solutions for (7), (8), and (9), approximations were obtained using a numerical quadrature approach with $q = 61$ evenly spaced points $\theta_1, ..., \theta_{61}$ in the range ± 3. The general form of the calculations was patterned after an approach based on the standard IRT model, outlined by Bock and Mislevy (1982). The posterior variance for a given set of administered items and associated responses u was approximated by

$$\mathrm{Var}\,(\theta|u) \approx K^{-1} \sum_{j=1}^{q} \left[\theta_j - \hat{\theta}\right]^2 p(\theta_j, u)\ , \tag{14}$$

and the posterior mean by

$$\hat{\theta} \approx K^{-1} \sum_{j=1}^{q} \theta_j p(\theta_j, u)\ , \tag{15}$$

where

$$K = \sum_{j'=1}^{q} p(\theta_{j'}, u)\ ,$$

and where $p(\theta_j, u)$ is given by (5), and $p(\theta)$ [contained in (5)] is a standard normal density function. The marginal probability terms for (8) and (9) can be approximated by

$$p(u) \approx \frac{1}{10} \sum_{j=1}^{q} p(\theta_j, u)\ . \tag{16}$$

To calculate the expected posterior variance associated with the administration of item k given by (7), the posterior variance calculations described by (14) and (15) were augmented by the item-related terms associated with the kth item, including the response to the kth unadministered item, u_k. In a similar manner, the marginal probability calculations given by (16) were also augmented with additional item-related terms and responses to complete the calculation of the expected posterior variance (7) for the administration of the kth item. Once expected posterior variance terms were calculated for each unadministered item contained in the pool, the next item was administered according to the SMEV algorithm described by (10) and (11). A final posterior mean was calculated from the responses to all administered items using (15). This posterior mean was taken as the ability estimate.

Response Generation

For a fixed ability θ and informant level h, the response u_i to item i with parameters a_i, b_i, c_i, e_i was generated by one of two approaches.

Model Based (MB) Response Generation.

According to this approach, a number t was sampled from a uniform distribution and compared to the conditional probability of a correct response:

$$u_i = \begin{cases} 1, & t \leq p_i(u_i = 1|\theta, h) \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

where the conditional probability $p_i(u_i = 1|\theta, h)$ was calculated according to (2), and $r = n$ (all administered items were assumed to be disclosed by the informant). This approach was used to generate item responses for adaptive test sessions used in the estimation of item exposure parameters e_i.

Informant Aided (IA) Response Generation.

According to this approach, h adaptive test administrations were generated, and preview status of item i was calculated: $v_i = 1$ if item i was administered to any of the h informants and $v_k = 0$ if none of the informants received item i. Then

$$u_i = \begin{cases} 1, & \text{if } v_i = 1 \text{ or } t \leq p_i(u_i = 1|\theta) \text{ ,} \\ 0, & \text{otherwise,} \end{cases}$$

where t is a random uniform number, and $p_i(u_i = 1|\theta)$ is the 3PL item response function defined by (3). This approach to response generation was used for all simulation studies.

Exposure Parameter Estimation

Exposure parameters e_i (for $i = 1, ..., N$) were calculated from simulated data by a multi-step iterative process:

(a) Sample θ from a standard normal distribution $p(\theta)$.

(b) Sample h from $p(h)$.

(c) Generate an adaptive testing sequence using the SMEV item selection and SIRT scoring algorithms in conjunction with the model-based (MB) response generation algorithm (17).

(d) Update the frequency of item administration across this and any previously simulated test sessions. Update the estimate of e_i by dividing the total number of times each item in the pool has been administered by the total number of adaptive tests generated so far.

(e) Perform 200 replications of Steps (a) through (d), and after each replication, substitute the updated e_i-values from Step (d) into the "Item selection and response generation" portion of Step (c). The final exposure parameters e_i obtained after the final replication were used in subsequent simulations.

Trojan Items

Except where otherwise noted, each simulated respondent had an opportunity to answer up to 10 Trojan (difficult highly exposed) items. Each of these 10 items (all with parameters $a = 2, b = 3, c = 0.2$) were administered with probability 0.95. Consequently, about 60% of the population would be expected to receive 10 Trojan items, about 32% to receive 9 Trojan items, about 7% to receive 8 Trojan items, and so forth. These Trojan

items were administered at the beginning of the test along with five adaptively selected items. The Trojan items were randomly interspersed among the 5 adaptive items so that their position in the item presentation sequence was not entirely predictable.

Sympson-Hetter Condition

The effects of compromise on test-scores were also simulated using the Sympson-Hetter exposure control algorithm. This algorithm was implemented using maximum information item selection (Lord, 1980, p. 151), where information tables were constructed from 61 equally spaced points in the range ± 3. Items providing the maximum information at the point closest to the maximum *a posteriori* ability estimate were considered for administration. Items passing the stochastic Sympson-Hetter rule based on the exposure control parameter were administered, those items not passing were set aside, and the next most informative item was selected for consideration, and so forth. Exposure control parameters required by the Sympson-Hetter procedure were estimated though a series of simulations using a target ceiling exposure rate of 0.15. Test lengths of 40 items were drawn from the 300-item pool. The final ability estimate was the posterior mean based on the standard IRT model, computed after the administration of the last item.

Results

A series of conditions were examined using the Sympson-Hetter and SIRT approaches. The conditions differed in the generation processes for the examinee ability θ and informant h parameters. Examinee parameters θ were either sampled from a $\mathrm{N}(0,1)$ distribution, or were set equal to fixed values: $-2, -1, -0.5, 0, 0.5, 1, 2$. Informant-level parameters h were either sampled from the distribution $p(h)$ given by (13), or set equal to fixed values: $0, 1, ..., 5$. Except where otherwise noted, all examinees received 40 items selected adaptively by either the Sympson-Hetter or SMEV algorithms. Up to 10 additional Trojan items were administered for the SIRT conditions.

For each condition, 2000 replications (simulated test-taker sessions) were conducted for the Sympson-Hetter procedure, and 500 replications were conducted for the SIRT approach. For each condition, replication outcomes were summarized by three measures: $\mathrm{M}(\hat{\theta})$, the mean of the estimated ability parameters; $\mathrm{SD}(\hat{\theta})$, the standard deviation of the estimated ability parameters, and $\mathrm{PV}(\theta)$, the mean of the examinee-level posterior variance values computed by (14). For conditions where θ was sampled rather than fixed, $\hat{\rho}$ (the squared Pearson product moment correlation between θ and $\hat{\theta}$) was also calculated.

Table 1 provides selected results for two conditions where $\theta \sim \mathrm{N}(0,1)$. When examinees' performance is not effected by sharing among informants ($h = 0$), the Sympson-Hetter and SIRT procedures display similar performance with regards to reliability $\hat{\rho}$, average score $\mathrm{M}(\hat{\theta})$, and dispersion $\mathrm{SD}(\hat{\theta})$. However, when some test-takers do benefit from sharing ($h \sim p(h)$; line 2 of Table 1), the SIRT procedure displays substantially higher precision (0.91 versus 0.77) and no score inflation (-0.04 versus. 0.18).

Table 2 displays some results regarding the effects of Trojan items on test score properties. For total test lengths n_{tot} of 50 items, several combinations of adaptive n_{adp} and Trojan n_{tro} test lengths were examined. Adaptive items were selected according to the SMEV criterion (eqs. 10 and 11). In each condition, up to n_{tro} items were administered, each item administered with probability 0.95 to each simulated respondent. For conditions

Table 1: Simulation Results for Fixed (zero) and Random Informant Levels

	Sympson-Hetter Model[a]				SIRT Model[b]			
Informants	$\hat{\rho}$	$\mathrm{M}(\hat{\theta})$	$\mathrm{SD}(\hat{\theta})$	$\mathrm{PV}(\theta)$	$\hat{\rho}$	$\mathrm{M}(\hat{\theta})$	$\mathrm{SD}(\hat{\theta})$	$\mathrm{PV}(\theta)$
$h = 0$	.94	−.04	.96	.02	.94	.03	.95	.06
$h \sim p(h)$	.77	.18	1.00	.02	.91	−.04	.94	.09

Note. $\theta \sim \mathrm{N}(0, 1)$.
[a]Test length = 40 items.
[b]Test length = 50 items total including up to 10 Trojan items.

Table 2: Comparisons of SIRT Model Outcomes for Alternate Adaptive and Trojan Test Lengths

n_{adp}	n_{tro}	n_{tot}	$\hat{\rho}$	$\mathrm{M}(\hat{\theta})$	$\mathrm{SD}(\hat{\theta})$	$\mathrm{PV}(\theta)$
50	0	50	.87	.01	.92	.11
45	5	50	.91	−.01	.97	.09
40	10	50	.91	−.04	.94	.09
35	15	50	.88	.06	.94	.10

Note. $\theta \sim \mathrm{N}(0, 1)$ and $h \sim p(h)$.

where $n_{\mathrm{tro}} > 0$, the Trojan items were administered towards the beginning of the test with 5 adaptive items randomly interspersed among the Trojan items. The results suggest that given a fixed overall test-length of 50 items, reliability can be increased by the administration of 5 or 10 Trojan items, and that the relative mix of adaptive and Trojan items can effect test-score precision.

Table 3 provides selected results for fixed informant levels h, where $\theta \sim \mathrm{N}(0, 1)$. For conditions where $h > 0$, the Sympson-Hetter procedure displays moderate to large drops in reliability and large increases in mean test scores. For $h = 2$, the Sympson-Hetter procedure displays over one-half SD increase in average score, and about a one SD increase (or higher) for $h \geq 3$. In contrast, the SIRT procedure displays much smaller decreases in reliability and only moderate increases in scores for $h > 0$. For $h \leq 2$, no positive gains in scores are observed, and $\hat{\rho} \geq 0.86$. In addition, only small average gains are observed for higher levels of informants $3 \leq h \leq 4$ (as compared to the Sympson-Hetter procedure). The dispersion of scores $\mathrm{SD}(\hat{\theta})$ appears much more constant across conditions for the SIRT procedure than for the Sympson-Hetter method. For the Sympson-Hetter procedure, $\mathrm{SD}(\hat{\theta})$ increases with h, whereas with the SIRT procedure the $\mathrm{SD}(\hat{\theta})$ decreases slightly for $h > 0$. Note in addition, that the Sympson-Hetter characterizations of posterior uncertainty $\mathrm{PV}(\theta)$ remain relatively constant across informant levels h, and grossly under-estimate the actual uncertainty for high h-levels. In contrast, the SIRT procedure provides $\mathrm{PV}(\theta)$ levels which tend to increase with h-levels, reflecting the increased uncertainty about θ associated with larger numbers of informants.

Table 4 provides selected results for fixed ability levels θ, where $h \sim p(h)$. Compared to the Sympson-Hetter procedure, the SIRT procedure displays lower average scores $\mathrm{M}(\hat{\theta})$

Table 3: Simulation Results for Fixed Informant Levels

	Sympson-Hetter Model[a]				SIRT Model[b]			
h	$\hat{\rho}$	M($\hat{\theta}$)	SD($\hat{\theta}$)	PV(θ)	$\hat{\rho}$	M($\hat{\theta}$)	SD($\hat{\theta}$)	PV(θ)
0	.94	−.04	.96	.02	.94	.03	.95	.06
1	.87	.22	.99	.01	.90	−.14	.89	.12
2	.74	.56	1.09	.02	.86	−.04	.88	.14
3	.60	.95	1.18	.02	.87	.22	.87	.15
4	.45	1.30	1.22	.02	.83	.27	.87	.17
5	.39	1.54	1.25	.01	.79	.46	.89	.17

Note. $\theta \sim \mathrm{N}(0, 1)$.
[a]Test length = 40 items.
[b]Test length = 50 items total including up to 10 Trojan items.

and smaller conditional SD($\hat{\theta}$), suggesting that the SIRT approach produces smaller score-gains from informants and provides increased measurement precision. The columns in Table 4 labeled *Gain* indicate the gain in conditional performance over a group where performance is not effected by informants ($h = 0$). For example, 0.38 reflects an average score-gain among those respondents with fixed $\theta = -2.0$ and sampled $h \sim p(h)$ over those with fixed $\theta = -2.0$ and zero informants ($h = 0$). Note that gains across ability levels for the Sympson-Hetter procedure tended to range between two and four tenths. In contrast for the SIRT procedure, significant positive score gains were only observed for the lowest ability levels, and near-zero gains were observed over most of the ability range ($\theta \geq -0.5$). Also note that the Sympson-Hetter procedure tends to under estimate the posterior uncertainty at each ability level as indicated by the small PV(θ) values juxtaposed against the large conditional SD($\hat{\theta}$)'s.

Discussion

The results of the simulation study suggest that the SIRT model with stochastic minimum expected variance item selection can substantially reduce the negative consequences of sharing item content. Compared to the Sympson-Hetter approach, the SIRT approach produces test scores with substantially higher reliability, and substantially lower inflation in instances where a substantial portion of the test-taking population benefits from item-content provided by one or more informants.

Although no positive score gains were observed for conditions with only one or two informants (Table 3) small to moderate gains were observed for conditions with 3, 4, or 5 informants. These larger gains in average test score might be attributable, in part, to the small prior probability given to these informant levels: $p(3 \leq h \leq 5) = 0.10$. Smaller score gains might have been observed had larger prior probabilities given to these informant levels. However, larger proportions of informants in the upper h-ranges are also likely to degrade measurement precision to a larger degree. For instances where larger percentages of informants occur over the range $3 \leq h \leq 5$, additional simulation studies would be required to examine the effects on score gain and measurement precision.

Table 4: Simulation Results for Fixed Ability Levels

	Sympson-Hetter Model[a]				SIRT Model[b]			
θ	M($\hat{\theta}$)	SD($\hat{\theta}$)	PV(θ)	Gain	M($\hat{\theta}$)	SD($\hat{\theta}$)	PV(θ)	Gain
−2.0	−1.47	.56	.03	.38	−1.62	.49	.11	.28
−1.0	−.74	.40	.01	.23	−.86	.30	.08	.09
−0.5	−.29	.38	.01	.17	−.46	.27	.08	.01
0.0	.22	.42	.01	.23	.00	.27	.08	−.00
0.5	.75	.49	.01	.30	.46	.29	.09	−.03
1.0	1.27	.51	.01	.33	.99	.33	.09	.04
2.0	2.15	.35	.01	.21	1.89	.28	.11	.00

Note. $h \sim p(h)$.
[a]Test length = 40 items.
[b]Test length = 50 items total including up to 10 Trojan items.

One assumption made by the SIRT item-selection and scoring algorithms regards the probability of item-preview conditional on only h, the number of informants that have participated in the disclosure. (See Equation 1.) In practice, the probability of item-preview is likely to be dependent on the ability level(s) θ of the informant(s) as well as the number of informants h. However, favorable precision and score-gain results were observed in spite of this potential model violation. Note that the simulated data were based on informant behavior that violated, to a realistic degree, the conditional independence assumption: Entire informant test-session records were randomly sampled from the population of test-takers, and any items in common between the test-takers and informant(s) were answered correctly by the test-taker. The SIRT item selection and scoring procedures provided satisfactory results, even in spite of the suspected violation to the conditional independence assumption.

This study also demonstrates the surprising benefits of highly-exposed difficult items (termed *Trojan* items). With standard IRT item-selection and scoring procedures, responses to these items would likely provide little or misleading information regarding the respondents ability level. In the context of the SIRT model however, these items help distinguish between those test-takers who have benefited from the help of informants and those who have not. A small number of highly-exposed difficult Trojan items administered to each test-taker can actually increase the precision of estimated ability.

The optimal placement of the Trojan items in the sequence of administered items is likely to be towards the beginning of the test, where the responses to these items can most heavily effect the exposure levels of subsequently selected items. However, the predictable placement of these difficult highly exposed items as the first (say 10) items in the test might lead to deliberate incorrect responding by cheaters, where respondents routinely answer the first 10 items incorrectly (in spite of their disclosure by informants), and then proceed to answer the remaining items correctly based on their ability, and on information provided by informants. In this way, test-takers might be able to more closely mimic the response patterns of honest high-ability test-takers who did not benefit from the aid of informants. However, by interspersing a few adaptively selected items of lower exposure

(based on the SMEV item selection algorithm) among the Trojan items at the beginning of the test, this sort of strategy might be effectively thwarted: The positioning of Tojan items in the sequence of adaptively administered items would be less predictable. This sort of countermeasure was implemented in the simulation study. Good measurement properties of ability estimates were displayed by the SIRT procedure in the simulation study even in spite of the possibly less than optimal placement of the Trojan items.

In at least one sense, the simulation study discussed here might overstate the negative impact of sharing among informants and test-takers, since it was assumed that each informant discloses the contents of all items received (i.e., $r = n$ in Equation 1). Even in instances where informants only disclose a portion of received items (i.e., $r < n$), the SIRT procedure will almost certainly provide improved performance in terms of precision and score gain, as compared to the Sympson-Hetter procedure, or other procedures which do not explicitly model the effects of sharing on responding. However, to determine the specific magnitude of the benefits for lower disclosure levels, additional simulated comparisons would be required.

The SIRT algorithm assumes that the distribution of informant levels $p(h)$ is known. In practical application of the procedure, this is unlikely to be the case. When $p(h)$ is unknown, two approaches are possible. First, a subjective prior could be used. If such an approach is taken, it would be useful to examine the sensitivity of item-selection and scoring results when $p(h)$ is misspecified to varying degrees. This could be accomplished through a series of simulation studies. Alternatively, a methodology to estimate $p(h)$ from empirical data might be derived. Then the estimated distribution could be used in the SIRT item selection and scoring calculations. Here too, the consequences of misspecification (due here to sampling errors) should be studied before the estimated h distribution is used in place of that assumed to be known.

McLeod, Lewis, and Thissen (2003) have suggested the use of a Bayesian index for the detection of examinees with item preknowledge. They suggest additional testing with highly secure items for those test-takers identified by the index as likely cheaters. The approach described in this paper eliminates the first step of identification (of those suspected of cheating) and adaptively selects the level of item-exposure for subsequently administered items based on the expected reduction in posterior variance. It is possible that the SIRT approach presented here is more efficient than the two-step approach suggested by McLeod et al. However, additional research would be required to evaluate the relative benefits of the two approaches.

One important implication of the SIRT model regards item replacement schedules for high-stakes high-volume adaptive tests. According to the sharing item response model, the usefulness of an item does not necessarily diminish over extended periods of exposure. Rather, an item's usefulness (in terms of precision) depends in large part on its *relative* (to other items) exposure, not necessarily on its *absolute* exposure. That is, according to (1), an item's preview propensity depends on its exposure *rate*, which is calculated relative to other items in the pool. According to the SIRT model, an item's functioning is not dependent on the total number of times an item has been used: An item with an exposure rate of 0.33 has the same measurement properties and usefulness whether it has been administered to 100 test-takers, or to 10,000 test-takers. This approach assumes that for a given test-taker, he or she shares item content with a relatively small countable number of informants. The

SIRT approach is not likely to provide reliable scores if item content is banked (which can be thought of as sharing among a large number of informants and test-takers).

In cases where items are banked by a large number of informants and given to a large number of test-takers, then frequent pool replacement might be warranted, and in fact might be the only countermeasure for widespread compromise. However, if banked questions are widely available to test-takers, then this availability should be easily known to the testing organization as well, who ought to be able to stop such practices since the test-questions themselves are likely to be protected under copyright restrictions. Consequently, in many high-stakes high-volume applications of CAT, the SIRT model coupled with vigilance on the part of the testing organization might allow large pools of items to be used over extended periods, without the need for frequent item pool updating and replacement.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

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.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

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

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.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Segall, D. O. (1995, April). *The effects of item compromise on computerized adaptive test scores.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Segall, D. O., & Moreno, K. E. (1997). Current and future challenges. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 257–269). Washington, DC: American Psychological Association.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Tech. Rep. No. 93-2). 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.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston: Kluwer-Nijhoff.

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

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneqpolis, MN.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Boston: Kluwer-Nijhoff.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.