A Sharing Item Response Theory Model for Computerized Adaptive Testing

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.

1. 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, 2000a). 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 the opportunity for sharing test content among examinees is limited by the small number of testing occasions), on-demand testing has a serious shortcoming: test security. With on-demand test schedules, the same test items are administered on many 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.

Accordingly, high-stakes testing programs have sought ways to increase the security of their CAT exams. Wainer (2000b) has pointed out that “in order for an item pool’s security to increase linearly the size of that pool must increase exponentially” (p. 212), indicating that increasing an item pool’s size is not a practical or cost-effective approach. Rather than funneling all items into a single large pool, the ASVAB program has used an alternate more cost-effective way to increase the security benefit of additional test items (Segall & Moreno, 1997). This approach involves the construction of a number of moderate-sized item pools that are randomly assigned to examinees at the start of testing, in much the same way conventional test forms are randomly assigned or spiraled. Developers of the CAT Graduate Record Exam have also gone to extraordinary lengths to help ensure test security. In some instances GRE item pools are replaced or updated after only one or two weeks of operational use (Mills, 1999, p. 131).

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.

2. The Sharing Item Response Theory (SIRT) Model

We begin by hypothesizing a specific compromise behavior based on the sharing of item content between test-takers who act as informants and test-takers who are potential beneficiaries. According to this model, a given test-taker has \( h \) informants who have taken the adaptive test before him (where \( h = 0, 1, 2, ..., n_h \)). Let \( r \) denote the number of randomly chosen items (out of \( n \) items received) that have been disclosed by the informant, and memorized in sufficient detail by the test-taker to result in correct responses. Accordingly, each item in the pool \( (i = 1, ..., N) \) has two states with regard to preview by the test-taker:

\[
v_i = \begin{cases} 
0, & \text{if item } i \text{ has not been previewed (in sufficient detail to result in a correct response)}, \\
1, & \text{if item } i \text{ has been previewed (in sufficient detail to result in a correct response)}. 
\end{cases}
\]

\(^1\)Chang and Zhang (2002) studied CAT test security issues and proposed terminology for particular types of sharing. They propose the terminology item pooling for the type of sharing between informants and potential beneficiaries modeled here.
Then the conditional probability that item $i$ has been previewed given $h$ informants have participated in the disclosure, is stated by

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

(1)

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

The rationale for (1) proceeds as follows. Note first that the term $r/n \times e_i$ represents the probability of disclosure for item $i$ for a single randomly sampled informant, indicating that item disclosure probabilities are directly proportional to their exposure rates. Consequently, the term $(1 - r/n \times e_i)^h$ provides the probability that item $i$ has not been disclosed by any of the $h$ informants. Then the complementary probability that item $i$ has been disclosed by one or more informants (i.e., previewed) is given by (1).

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 $u_i = 0$. The probability of a correct response to item $i$ conditional on ability $\theta$ and number of informants $h$ is given by

$$p_i(u_i = 1|\theta, h) = p_i(u_i = 1, v_i = 1|\theta, h) + p_i(u_i = 1, v_i = 0|\theta, h)$$

$$= p_i(u_i = 1|v_i = 1, \theta, h)p_i(v_i = 1|\theta, h) + p_i(u_i = 1|v_i = 0, \theta, h)p_i(v_i = 0|\theta, h)$$

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

(2)

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

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

(3)

and $p(v_i = 0|h) = 1 - p(v_i = 1|h)$. (McLeod, Lewis, and Thissen, 2003, and Segall, 2002, provide other related examples of compromise response models.)

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

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

(4)

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

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

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

$$p(u, \theta) = \sum_{h=0}^{n} p(u, \theta, h)$$

$$= p(\theta) \sum_{h=0}^{n} \left[ p(h) \prod_{i=1}^{n} p_i(u_i|\theta, h) \right].$$

(5)

The posterior distribution of ability $\theta$ given data $u$ is provided by

$$p(\theta|u) = p(\theta, u)/p(u),$$

(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 \( \theta \) in the presence of both measurement-error, and uncertainty due to item disclosure by informants. Numerical approximations to the posterior variance and mean are presented in Section 4.

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 Section 4.

3. Item Selection

This section describes an adaptive item selection algorithm that explicitly considers the item’s exposure rate and its effect on the reduction of posterior uncertainty. This algorithm is based on a general IRT item-selection approach adapted for use with the SIRT model. The algorithm, along with two enhancements intended to provide improved test-score precision are described below.

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 items previously unadministered to the examinee), denoted by \( E[\text{Var}(\theta|u, u_k)] \), can be expressed by

\[
E[\text{Var}(\theta|u, u_k)] = \text{Var}(\theta|u, u_k = 1)p(u_k = 1|u) + \text{Var}(\theta|u, u_k = 0)p(u_k = 0|u),
\]

where \( u \) contains the responses to previously answered items, and \( u_k \) denotes the yet-to-be observed response to item \( k \). The predictive posterior distribution for the response to item \( k \) can be calculated from the ratio of two terms:

\[
p(u_k|u) = \frac{p(u, u_k)}{p(u)},
\]

where the numerator is expressed by

\[
p(u, u_k) = \int p(\theta, u, u_k) d\theta,
\]

and denominator by

\[
p(u) = \int p(\theta, u) d\theta.
\]
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 \( \operatorname{Var}(\theta|u, u_k = 1) \) and \( \operatorname{Var}(\theta|u, u_k = 0) \) 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

\[
M_k^{(1)} \propto \sigma^2 - \operatorname{E}[\operatorname{Var}(\theta|u_k)]
\]

where the subscript \( k \) indexes the item in a pool of items previously unadministered to the examinee, \( \sigma^2 \) denotes the prior variance of \( \theta \), and where the \( M_k^{(1)} \) 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

\[
M_k^{(i)} = \frac{Z_k}{\sum_j Z_j},
\]

where

\[
Z_k = \frac{\{\sigma^2 - \operatorname{E}[\operatorname{Var}(\theta|u, u_k)]\} - \{\sigma^2 - \operatorname{Var}(\theta|u)\}}{\sigma^2 - \operatorname{Var}(\theta|u)}
\]

where

\[
Z_k = \frac{\operatorname{Var}(\theta|u) - \operatorname{E}[\operatorname{Var}(\theta|u, u_k)]}{\sigma^2 - \operatorname{Var}(\theta|u)}.
\]

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 with difficulty-parameters falling at or near the mean of the \( \theta \) distribution. 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 $L(\theta|u, h = 1), \ldots, L(\theta|u, h = n_h)$ 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 $\theta$, 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 $\theta$ 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.

4. 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 (in terms of increased precision) 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 Symposon-Hetter item exposure control algorithm?

**Item Pool**

Item pools consisted of items with difficulty parameters $b$ equally spaced from $-1.5$ to $+1.5$. The slope $a_i$ and guessing $c_i$ parameters were sampled from independent uniform distributions: $a_i \sim U[0.5, 1.5]$ and $c_i \sim U[0, 0.3]$. Except where otherwise noted, the item pool consisted of 300 items.

**Informant Distribution**

The prior distribution of informants was assumed to be

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

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. Except where otherwise noted, the prior distribution specified by (13) was used in SIRT item selection and scoring calculations, and 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. Although this level of compromise is likely to be more severe than that expected in many operational testing programs, it was chosen here to highlight differences in performance among alternative item-selection and scoring approaches. Issues regarding routine prior specification are outlined in the Discussion section.

**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 $\pm 3$. 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

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

and the posterior mean by

$$\hat{\theta} \approx K^{-1} \sum_{j=1}^{q} \theta_j p(\theta_j, u), \quad (15)$$
where
\[ K = \sum_{j'=1}^{q} p(\theta_{j'}, u) , \]
and where \( p(\theta, u) \) 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
\[ p(u) \approx \frac{3}{q-1} \sum_{j=1}^{q} p(\theta_{j}, u) . \]  

(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 \( k \)th item, including the response to the \( k \)th 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 \( k \)th 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 \( \theta \) 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:
\[ u_i = \begin{cases} 1, & t \leq p_i(u_i = 1 | \theta, h) \\ 0, & \text{otherwise,} \end{cases} \]  

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

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 $\theta$ 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.

Sympon-Hetter Condition

The effects of compromise on test-scores were also simulated using the Sympon-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 $\pm 3$. Items providing the maximum information at the point closest to the maximum $a posteriori$ ability estimate were considered for administration. Items passing the stochastic Sympon-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 Sympon-Hetter procedure were estimated through a series of simulations using a target ceiling exposure rate of 0.15. (This target exposure rate falls within the range of values used by operational testing programs.) 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.
Table 1: Simulation Results for Fixed (zero) and Random Informant Levels

<table>
<thead>
<tr>
<th>Informants</th>
<th>Sympson-Hetter Model</th>
<th>SIRT Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\rho}$</td>
<td>$M(\theta)$</td>
</tr>
<tr>
<td>$h = 0$</td>
<td>.94</td>
<td>−.04</td>
</tr>
<tr>
<td>$h \sim p(h)$</td>
<td>.77</td>
<td>.18</td>
</tr>
</tbody>
</table>

Note. $\theta \sim N(0,1)$ and Pool Size = 300 items.

$^a$Test length = 40 items.

$^b$Test length = 50 items total including up to 10 Trojan items.

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 $\theta$ and informant $h$ parameters. Examinee parameters $\theta$ were either sampled from a $N(0,1)$ distribution, or were set equal to fixed values: $−2, −1, −0.5, 0, 0.5, 1, 2$. Informant-level parameters $h$ were either set equal to fixed values: $0, 1, ..., 5$, or unless otherwise noted, sampled from the distribution $p(h)$ given by (13). Except where otherwise noted, all examinees received 40 items selected adaptively by either the Sympson-Hetter or SMEV algorithms from a pool of 300 items. 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: $M(\hat{\theta})$, the mean of the estimated ability parameters; $SD(\hat{\theta})$, the standard deviation of the estimated ability parameters, and $PV(\theta)$, the mean of the examinee-level posterior variance values computed by (14). For conditions where $\theta$ was sampled rather than fixed, $\hat{\rho}$ (the squared Pearson product moment correlation between $\theta$ and $\hat{\theta}$) was also calculated.

Table 1 provides selected results for two conditions where $\theta \sim N(0,1)$. When examinees’ performance is not affected by sharing among informants ($h = 0$), the Sympson-Hetter and SIRT procedures display similar performance with regards to reliability $\hat{\rho}$, average score $M(\hat{\theta})$, and dispersion $SD(\hat{\theta})$. 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 $n_{\text{tot}}$ of 50 items, several combinations of adaptive $n_{\text{adp}}$ and Trojan $n_{\text{tro}}$ test lengths were examined. Adaptive items were selected according to the SMEV criterion (eqs. 10 and 11). In each condition, up to $n_{\text{tro}}$ items were administered, each item administered with probability 0.95 to each simulated respondent. For conditions where $n_{\text{tro}} > 0$, 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 item pool size of 300 items and an overall test-length of 50 items, reliability can be increased from 0.87 (with the administration of zero Trojan items) to 0.91 (by the administration of 5 or 10 Trojan items).
Table 2: Comparisons of SIRT Model Outcomes for Alternate Adaptive and Trojan Test Lengths

<table>
<thead>
<tr>
<th>$n_{\text{adp}}$</th>
<th>$n_{\text{tro}}$</th>
<th>$n_{\text{tot}}$</th>
<th>$\hat{\rho}$</th>
<th>M($\hat{\theta}$)</th>
<th>SD($\hat{\theta}$)</th>
<th>PV($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>50</td>
<td>.87</td>
<td>.01</td>
<td>.92</td>
<td>.11</td>
</tr>
<tr>
<td>45</td>
<td>5</td>
<td>50</td>
<td>.91</td>
<td>-.01</td>
<td>.97</td>
<td>.09</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>50</td>
<td>.91</td>
<td>-.04</td>
<td>.94</td>
<td>.09</td>
</tr>
<tr>
<td>35</td>
<td>15</td>
<td>50</td>
<td>.88</td>
<td>.06</td>
<td>.94</td>
<td>.10</td>
</tr>
</tbody>
</table>

Note. $\theta \sim N(0, 1)$, $h \sim p(h)$, and Pool Size = 300 items.

Figure 1 displays the simulated reliability values for SIRT conditions with different item pool sizes. These values resulted from simulated SIRT tests in which zero Trojan items were administered. The 34 conditions represent pool-sizes spanning the range from 300 to 795 items. For each condition, test length was 40 items, $\theta \sim N(0, 1)$, $h \sim p(h)$, and the SIRT reliability was calculated as the squared correlation between true and estimated ability from 500 simulated respondents. Expected reliability values (for a given pool-size) were obtained from the least-squares linear regression of reliability on log pool-size. When no Trojan items are administered, a pool size of about 500 items is necessary to achieve a reliability of 0.91. (See Figure 1.) However, when a small number of Trojan items is administered, this same reliability level can be achieved from a pool-size of 300 items (Table 2). Consequently, the administration of 5 to 10 Trojan items increased the reliability of test scores by about the same amount as a 200-item (or 67%) increase in pool size.

Table 3 provides selected results for fixed informant levels $h$, where $\theta \sim N(0, 1)$. For conditions where $h > 0$, the Sympon-Hetter procedure displays moderate to large drops in reliability and large increases in mean test scores. For $h = 2$, the Sympon-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 Sympon-Hetter procedure). The dispersion of scores SD($\hat{\theta}$) appears much more constant across conditions for the SIRT procedure than for the Sympon-Hetter method. For the Sympon-Hetter procedure, SD($\hat{\theta}$) increases with $h$, whereas with the SIRT procedure the SD($\hat{\theta}$) decreases slightly for $h > 0$. This decrease is consistent with the performance of Bayesian estimators: As likelihood functions become flatter, Bayesian estimators (such as the posterior mode) exhibit greater bias towards the mean of the prior, and consequently exhibit a smaller variance of the estimated parameter $\hat{\theta}$ values as well. Note in addition, that the Sympon-Hetter characterizations of posterior uncertainty 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 PV($\theta$) levels which tend to increase with $h$-levels, reflecting the increased uncertainty about $\theta$ associated with larger numbers of informants.

Table 4 provides selected results for fixed ability levels $\theta$, where $h \sim p(h)$. Compared to the Sympon-Hetter procedure (which displays positively biased ability estimates, i.e., M($\hat{\theta}$) $-$ $\theta$ > 0), the SIRT procedure displays lower average scores M($\hat{\theta}$) and smaller condi-
Figure 1. SIRT reliability for tests with zero Trojan items expressed as a function of item-pool size.

Table 3: Simulation Results for Fixed Informant Levels

<table>
<thead>
<tr>
<th>h</th>
<th>( \hat{\rho} )</th>
<th>M(( \hat{\theta} ))</th>
<th>SD(( \hat{\theta} ))</th>
<th>PV(( \hat{\theta} ))</th>
<th>( \hat{\rho} )</th>
<th>M(( \hat{\theta} ))</th>
<th>SD(( \hat{\theta} ))</th>
<th>PV(( \hat{\theta} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.94</td>
<td>-.04</td>
<td>.96</td>
<td>.02</td>
<td>.94</td>
<td>.03</td>
<td>.95</td>
<td>.06</td>
</tr>
<tr>
<td>1</td>
<td>.87</td>
<td>.22</td>
<td>.99</td>
<td>.01</td>
<td>.93</td>
<td>-.14</td>
<td>.89</td>
<td>.12</td>
</tr>
<tr>
<td>2</td>
<td>.74</td>
<td>.56</td>
<td>1.09</td>
<td>.02</td>
<td>.86</td>
<td>-.04</td>
<td>.88</td>
<td>.14</td>
</tr>
<tr>
<td>3</td>
<td>.60</td>
<td>.95</td>
<td>1.18</td>
<td>.02</td>
<td>.87</td>
<td>.22</td>
<td>.87</td>
<td>.15</td>
</tr>
<tr>
<td>4</td>
<td>.45</td>
<td>1.30</td>
<td>1.22</td>
<td>.02</td>
<td>.83</td>
<td>.27</td>
<td>.87</td>
<td>.17</td>
</tr>
<tr>
<td>5</td>
<td>.39</td>
<td>1.54</td>
<td>1.25</td>
<td>.01</td>
<td>.79</td>
<td>.46</td>
<td>.89</td>
<td>.17</td>
</tr>
</tbody>
</table>

Note. \( \theta \sim N(0, 1) \) and Pool Size = 300 items.

\(^a\)Test length = 40 items.

\(^b\)Test length = 50 items total including up to 10 Trojan items.
Table 4: Simulation Results for Fixed Ability Levels

<table>
<thead>
<tr>
<th>θ</th>
<th>Symson-Hetter Modela</th>
<th>SIRT Modelb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M(θ)</td>
<td>SD(θ)</td>
</tr>
<tr>
<td>−2.0</td>
<td>−1.47</td>
<td>.56</td>
</tr>
<tr>
<td>−1.0</td>
<td>−.74</td>
<td>.40</td>
</tr>
<tr>
<td>−0.5</td>
<td>−.29</td>
<td>.38</td>
</tr>
<tr>
<td>0.0</td>
<td>.22</td>
<td>.42</td>
</tr>
<tr>
<td>0.5</td>
<td>.75</td>
<td>.49</td>
</tr>
<tr>
<td>1.0</td>
<td>1.27</td>
<td>.51</td>
</tr>
<tr>
<td>2.0</td>
<td>2.15</td>
<td>.35</td>
</tr>
</tbody>
</table>

Note. h ∼ p(h) and Pool Size = 300 items.

a Test length = 40 items.
b Test length = 50 items total including up to 10 Trojan items.

tional SD(θ). These results indicate 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 affected by informants (h = 0). For example, 0.38 reflects an average score-gain among those respondents with fixed θ = −2.0 and sampled h ∼ p(h) over those with fixed θ = −2.0 and zero informants (h = 0). Note that gains across ability levels for the Symson-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 (θ ≥ −0.5). Also note that the Symson-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(θ)’s.

Table 5 provides an analysis of SIRT model outcomes for 18 conditions examining six different informant distributions. Each informant distribution is characterized by an \( h_{\text{MAX}} \) value (Table 5, column 2), where

\[
p(h) = \begin{cases} 
2/3, & h = 0, \\
1/3, & h = h_{\text{MAX}}, 
\end{cases} \tag{18}
\]

and where \( h_{\text{MAX}} = 1, 3, 5, 10, 15, 20 \). For each informant distribution, two-thirds of the test-taker population had zero informants. For the distribution with \( h_{\text{MAX}} = 1 \) (Conditions 1, 7, and 13), the remaining 1/3 of the population had exactly 1 informant; for \( h_{\text{MAX}} = 3 \) (Conditions 2, 8 and 14), the remaining 1/3 of the population had exactly 3 informants, and so forth. In each analysis (18) was used as the prior for SIRT item selection and scoring calculations. Conditions 1–6 (Table 5, top) display outcomes where \( h \) was sampled from (18) for each test-taker. As the numbers of informants increase, the reliability values decrease. Note however, that average scores remain relatively constant (around zero—the prior mean) even for conditions characterized by large numbers of informants. Also note that the SD(θ)
Table 5: SIRT Model Outcomes for Alternate Informant Distributions

<table>
<thead>
<tr>
<th>Condition</th>
<th>$h_{\text{MAX}}$</th>
<th>$\hat{\rho}$</th>
<th>$M(\theta)$</th>
<th>$SD(\theta)$</th>
<th>$PV(\theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h \sim p(0, h_{\text{MAX}})$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>.93</td>
<td>-.04</td>
<td>.95</td>
<td>.07</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.91</td>
<td>-.05</td>
<td>.97</td>
<td>.09</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>.88</td>
<td>-.05</td>
<td>.92</td>
<td>.11</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>.80</td>
<td>.00</td>
<td>.89</td>
<td>.19</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>.71</td>
<td>.07</td>
<td>.89</td>
<td>.27</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>.61</td>
<td>-.03</td>
<td>.81</td>
<td>.31</td>
</tr>
<tr>
<td>$h = h_{\text{MAX}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>.91</td>
<td>.05</td>
<td>.90</td>
<td>.08</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>.83</td>
<td>-.01</td>
<td>.88</td>
<td>.14</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>.80</td>
<td>.09</td>
<td>.85</td>
<td>.19</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>.55</td>
<td>.05</td>
<td>.77</td>
<td>.39</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>.39</td>
<td>.13</td>
<td>.68</td>
<td>.60</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>.21</td>
<td>.10</td>
<td>.53</td>
<td>.77</td>
</tr>
<tr>
<td>$h = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>.94</td>
<td>.03</td>
<td>.96</td>
<td>.06</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>.93</td>
<td>-.08</td>
<td>.94</td>
<td>.06</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>.95</td>
<td>-.06</td>
<td>1.01</td>
<td>.06</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>.93</td>
<td>-.02</td>
<td>.97</td>
<td>.06</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>.94</td>
<td>-.09</td>
<td>.96</td>
<td>.06</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
<td>.93</td>
<td>-.04</td>
<td>.96</td>
<td>.06</td>
</tr>
</tbody>
</table>

Note. $\theta \sim N(0, 1)$ and Pool Size = 300 items.

diminishes and $PV(\theta)$ increases with increasing numbers of informants. These trends associated with larger numbers of informants are consistent with larger posterior uncertainty, and with an increasing influence of the normal-based $\theta$-prior on ability estimates. Similar, but more pronounced trends are observed for Conditions 7–12 (Table 5, middle) which examine test-score properties of test-takers who benefit from a fixed number of informants (i.e., $h = h_{\text{MAX}}$). Conditions 13–18 (Table 5, bottom) display results for test-takers who benefit from zero informants ($h = 0$) when the SIRT item selection and scoring is based on the six informant-priors which assume otherwise (i.e., eq. 18). Reliability indices $\hat{\rho}$ for these non-cheaters remain consistently high, and $M(\theta)$, $SD(\theta)$, and $PV(\theta)$ remain relatively constant over conditions. These results indicate that non-cheaters are accurately measured, and do not appear to be disadvantaged by the SIRT item-selection and scoring algorithms even under conditions which assume a substantial portion of the population benefits from large numbers of informants.
5. 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 Sympon-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 been 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.

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) \( \theta \) 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 respondent’s 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. Results suggest that the administration of 5 to 10 Trojan items can increase the reliability of test scores by about the same amount as a 67% increase in pool size. Faced with a choice of either administering a few Trojan items or increasing item pool size as a means of achieving higher precision levels, many test-developers might opt for the former because of its significant potential savings in item pool development costs. Note however that in some cases, testing time and associated resources might be in short supply. In these cases the administration extra Trojan items (even though relatively small in number) might also entail some additional costs.

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 purposely 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 Trojan 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). Clearly there are limits to the numbers of items informants and test-takers can memorize and process. However 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. To determine the specific magnitude of the benefits associated with lower disclosure levels, additional simulated comparisons would be required.

The SIRT model assumes that the disclosure probability $r/n$ is a constant across informants and administered items. In practice, these values are likely to vary across informants, with some informants disclosing more items than others. And in some instances, the disclosure probability is also likely to vary across administered items, with probabilities influenced by such factors as: (a) serial item administration position, (b) item difficulty, and (c) item content (length or wordiness of the stem and response alternatives, existence of graphical or tabular stimuli, etc.). Before the SIRT model is used routinely, it would be useful to examine the robustness of the model to violations of this assumption. The effects of mis-specified $r$ values can be studied by simulating response data where $r$ is mis-specified to varying degrees, and where the single-valued $r$ SIRT model is applied to data where $r$ happens to vary across informants. The effects of model violations can also be examined from live data. These data can be used to compare the measurement properties (reliability, validity, mean trends over time) of test scores produced by the SIRT model with those resulting from commonly used exposure control algorithms.

The evaluation of the SIRT algorithm (Section 4) assumed that the distribution of informant levels $p(h)$ was 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 mis-specified 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 mis-specification (due here to sampling errors) should be studied before the estimated $h$
distribution is used as the prior distribution.

Although application of the SIRT model has been studied in the context of fixed-length adaptive tests, there may also be improvements in measurement efficiency when the model is applied to variable-length testing algorithms. For standard item-selection and scoring (based for example on the Sympon-Hetter algorithm), the simulation results suggest that the characterizations of posterior uncertainty $PV(\theta)$ remain relatively constant across informants levels, and grossly under-estimate the actual uncertainty regarding $\theta$ for conditions involving many informants (Table 3). This result suggests that in cases of non-zero informant levels, standard CAT variable-length tests (which incorporate the posterior variance statistic into their stopping-rule) will terminate too soon, and will not achieve the desired precision level. In contrast, the SIRT procedure provides $PV(\theta)$ levels which tend to increase with the number of informants, more accurately reflecting the increased uncertainty about $\theta$ associated with larger numbers of informants. This raises the intriguing possibility that the SIRT model might be used to test all test-takers to the same level of precision, regardless of the test-taker’s level of item-preview. According to the SIRT algorithm, those test-takers benefiting from one or more informants would generally (because of their larger $PV(\theta)$ statistic values) be administered additional low-exposure items. However, additional simulation studies specifically modeling variable-length testing algorithms would be necessary to verify this expected outcome.

The evaluation of the SIRT model (Section 4) assumes that the item parameters ($a$, $b$, and $c$) are known. When data for the calibration of new items are collected online (i.e., new items are interspersed among operational items), it is conceivable that compromise can effect the responses to these items and the parameter estimates derived from these responses. These biases in the parameter-estimates of new items can lead to systematic score-scale drift when these items and their parameter-estimates are used operationally. There are however at least two data collection designs that might help reduce the influence of compromise on parameter estimates of new items. First, item-response data might be collected over a compressed time-period, thus limiting the opportunity for sharing among examinees. Alternatively, exposure rates of new items might be kept sufficiently low by administering to any given examinee a small randomly selected subset of new items drawn from a much larger pool. It is also conceivable that a new item parameter estimation approach based on the SIRT model could reduce or eliminate the effects of compromise on item parameter estimates. Such an approach would use the SIRT model (equation 5) instead of the standard IRT model to characterize examinee response probabilities. Additional study would be required to judge the performance of these alternative approaches.

The ascending-$a$ approach (Chang & Ying, 1999) has been suggested as a method to reduce the effects of compromise on CAT scores. With little or no sacrifice to measurement precision, this approach increases the usage rates of low discriminating items, thus flattening overall item exposure-rates, and in particular lowering the administration rates of some highly-exposed highly-discriminating items. The SIRT model accomplishes the same goal (reduces the effects of compromise on test scores) by adapting the exposure-levels of items to the sharing propensity of examinees. Examinees whose responses are consistent with sharing will receive a larger number of less exposed items, which are presumably less discriminating than more highly exposed items. So, like the ascending-$a$ approach, it is likely that the
SIRT procedure also utilizes more low-discriminating items than standard descending-\( a \) approaches. Although the SIRT item-selection algorithm might be modified to explicitly select low \( a \)-value items towards the beginning of the test (as prescribed by the ascending-\( a \) approach), this modification is unlikely to provide much additional benefit with regard to compromise protection, and would likely lead to lower test-score precision. The trade-off between exposure and discrimination levels is already incorporated into the SIRT item selection algorithm, which maximizes test-score precision in the context of item sharing. Consequently, the SIRT model is likely to provide greater precision and protection against sharing than provided by ascending-\( a \) approaches which treat all examinees as equally likely to have shared item content. Further study providing a direct comparison between the two approaches is necessary however to confirm this outcome. Additional comparisons between the SIRT model and other item selection algorithms that effect distributions of item usage rates, such as the global information approach (Chang & Ying, 1996), would also provide useful results.

McLeod et al. (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. (Note that this implication of the SIRT model contradicts commonplace item retirement schedules, which assign a specific life-span to each item. According to these retirement schedules, an item is replaced after say 20,000 administrations, or after 3 months of use.) According to the SIRT approach, an item’s preview propensity (Equation 1) 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 it 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.

Results indicate that on average, SIRT score inflation does not occur even for test-takers who receive item content from large numbers of informants. (See Table 5.) However, test-score precision is affected by the number of informants: As the number of informants increases, the test-score precision decreases. In cases where the number of informants is extremely large (which is functionally equivalent to an organized effort to steal the contents of an item pool), the test-score reliability among those who have benefitted from the aid of informants tends towards zero. Even though on average test-takers do not benefit from the aid of informants when tested according to the SIRT approach, some test-takers (especially low-ability test-takers) might still attempt to gain advantage by enlisting the aid of informants. By doing so, test-takers could effectively add a lottery component to their
test-score in an attempt to capitalize on chance measurement errors.

A procedure developed by Segall (2002) can be used to assess the level of sharing among informants and test-takers. In cases where a large number of test-takers have pre-viewed large sections of the item pool, frequent pool replacement might be warranted, and in fact might be the only countermeasure for widespread compromise. In other instances, the testing organization might be able to stop such sharing 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


