DOCUMENT RESUME

ED 467 810 TM 034 352

AUTHOR	McLeod, Lori D.; Lewis, Charles; Thissen, David.
TITLE	A Bayesian Method for the Detection of Item Preknowledge in CAT. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-98-07
PUB DATE	1999-03-00
NOTE	21p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; *Bayesian Statistics; College Entrance Examinations; *Computer Assisted Testing; Estimation (Mathematics); Item Response Theory; Knowledge Level; Law Schools; *Prior Learning; *Test Items
IDENTIFIERS	Law School Admission Test

ABSTRACT

With the increased use of computerized adaptive testing, which allows for continuous testing, new concerns about test security have evolved, one being the assurance that items in an item pool are safeguarded from theft. In this paper, the risk of score inflation and procedures to detect test takers using item preknowledge are explored. When test takers use preknowledge of items, their item responses may deviate from the underlying item response theory (IRT) model, and estimated abilities may be inflated. This deviation may be detected through the use of person-fit indices. A Bayesian posterior log odd ratio index is proposed for detecting the use of item preknowledge. In this approach to person-fit, the estimated probability that each test taker has preknowledge of items is updated after each item response. These probabilities are based on IRT parameters, a model specifying the probability that each item has been memorized, and the test taker's item responses. Simulations based on an operational computerized adaptive test pool were used to demonstrate the risk of item preknowledge to test security and the use of the odds ratio index. An appendix discusses the three classes of models used in the study. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 810

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

A Bayesian Method for the Detection of Item Preknowledge in CAT

Lori D. McLeod, Law School Admission Council
Charles Lewis, Educational Testing Service,
and David Thissen, University of North Carolina at Chapel Hill

Law School Admission Council
Computerized Testing Report 98-07
March 1999

LAW Services

A Publication of the Law School Admission Council

TM034352

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and *LSACD* are trademarks of the Law School Admission Council, Inc.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

ERIC
Full Text Provided by ERIC

Table of Contents

Executive Summary

With the increased use of computerized adaptive testing, which allows for continuous testing, new concerns about test security have evolved, one being the assurance that items in an item pool are safeguarded from theft. As the Law School Admission Council (LSAC) investigates implementing a computerized version of the Law School Admission Test (LSAT), the risk to test security and tools for protecting test items should be explored. The goals of this study include examining test taker success at achieving test score inflation when using item preknowledge and the feasibility of using an odds ratio index as a tool for test security.

This project used simulations based on results from an operational computerized adaptive test (CAT). The design applied a real-world approach to simulate the "item preknowledge" process by incorporating a two-stage process. First, for each condition, the design sent in *n* sources to memorize test items from a 28-item test. These *n* test takers memorized their items perfectly and then combined their lists. (Some overlap was observed among the lists.) The complete list was memorized by another group of test takers, the beneficiaries. Then, the beneficiaries were administered a 28-item test, and if they were administered any of the memorized items, they answered them correctly. (Although we acknowledge that memorizers may not have perfect recall of the item list, the simulation was designed to produce a worst-case scenario for the testing program.)

Simulated test takers were generated at true scores from a discrete uniform distribution at 11 ability, θ, values. The θ values were translated into a number correct true score on a linear 60-item reference test and correspond to the operational test's score range (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 59). Along with varying the proficiency of the sources (50, 55, 59), groups of two, four, and eight sources were simulated for the various memorizing conditions. A control or null condition in which test takers did not have item preknowledge was included. One advantage for using this design is that it depicts a possible reality, especially if recording equipment is used for item theft. Another advantage is that the design maintains the roles of content constraints and the item selection algorithm in the success of using item preknowledge in the CAT environment.

For the memorizing simulees, across all true scores, the mean test score was inflated upward. Therefore, we conclude that by using the source-beneficiary preknowledge strategy, the test takers were successful in attaining higher test scores. The estimates were, of course, more inflated when the test takers had memorized the longer lists gathered by eight sources. Even the lower ability test takers for the eight sources condition had an average estimated test score above 40 (out of a possible 60 score points). Also as expected, information from four sources did not deliver as much test score gain as information from eight-sources. Similarly, item information from two sources did not aid a beneficiary as much as information from four sources. The estimates were more variable at the lower true scores, where the test takers have more room to benefit from the preknowledge, depending on the peculiarities of item selection. The higher ability test takers do not benefit as much from the memorization because they are already the higher scorers.

An odds ratio procedure to detect test takers using item preknowledge was developed and then evaluated. Specifically, three classes of models were introduced for the probability that an item had been memorized. Based on these models, seven Bayesian indices (FLOR1-FLOR7) were developed. Results from the simulated CAT data indicated that these indices had the power to detect item preknowledge. Overall, the best performing index of those studied is FLOR7, because it has the most power to detect those test takers who had preknowledge of more than half of the items on their test. FLOR3 is selected as the second best performing index for these successful test takers. This index has the extra appeal of being simple to compute without a previous simulation.

Abstract

With the increased use of computerized adaptive testing, which allows for continuous testing, new concerns about test security have evolved, one being the assurance that items in an item pool are safeguarded from theft. In this paper, the risk of score inflation and procedures to detect test takers using item preknowledge are explored. When test takers use item preknowledge, their item responses deviate from the underlying IRT model, and estimated abilities may be inflated. This deviation may be detected through the use of person-fit indices. A Bayesian posterior log odds ratio index is proposed for detecting the use of item preknowledge. In

this approach to person-fit, the estimated probability that each test taker has preknowledge of items is updated after each item response. These probabilities are based on the IRT parameters, a model specifying the probability that each item has been memorized, and the test taker's item responses. Simulations based on an operational computerized adaptive test (CAT) pool were used to demonstrate the risk of item preknowledge to test security and the use of the odds ratio index.

Introduction

With the increased use of computerized adaptive testing, which allows for continuous testing, new concerns about test security have evolved, such as how to assure that items in an item pool are safeguarded from theft. In this paper, a new procedure to detect test takers using item preknowledge is explored. When test takers use item preknowledge, their item responses may deviate from the underlying IRT model, and estimated abilities may be inflated. This deviation may be detected through the use of person-fit indices.

A New Approach to Person-Fit

Lewis (1997) proposed a posterior log odds ratio index for detecting the use of prior knowledge in a CAT environment. The concept of odds ratios was extended to describe the increased likelihood (based on item responses) that a response pattern arises from the normal or aberrant models, which is much like the concept behind optimal appropriateness indices developed by Drasgow and Levine (1986). In the posterior log odds ratio approach to person-fit, c represents the dichotomous item preknowledge state (c or $\overline{c}$). If the state is c , then the test taker's response pattern is "nonfitting" and the test taker has memorized at least one of the test items. If the state is $\overline{c}$, then the test taker's response pattern is "fitting," and the test taker has not memorized any of the items and is using underlying proficiency to respond to the test. The probability $p(c)$ that a test taker is using item preknowledge is updated after each item response. These "item preknowledge" probabilities are based on the IRT parameters (assumed known), a model describing the probability that each item has been memorized, and the test taker's item responses (Lewis, 1997). The prior probability of item preknowledge, $p_0(c)$, is a specified value that reflects the expected proportion of test takers believed to be using item preknowledge (e.g. 0.0001). This number may be established using empirical evidence from traditional approaches to detect cheaters, or prior elicitation based on the decision theory literature.

The odds ratio is based on two alternative models for an item response: one that assumes an item preknowledge state (c), and one that assumes a normal ($\overline{c}$) state. The "normal" model is the (usual) 3PL model. The 3PL model produces the probability of an item response for varying values of θ. For a correct response to the i^{th} item, denoted by $u_i = 1$, the value is

$$p(u_i = 1|\overline{c}) = T_{\overline{c}}(u_i = 1|\theta) = g_i + \frac{(1-g_i)}{1+\exp[-Da_i(\theta - b_i)]}. \quad (1)$$

where

- a_i is the slope, or discrimination power of an item, which is proportional to the slope of the curve at its inflection point;
- b_i is the threshold or inflection point of the curve, which is the point on the proficiency scale where the probability of a correct response is 0.5*(1-g_i);
- g_i is the lower asymptote, which is the probability of a correct response for test takers with very low proficiency;
- D is a constant (1.7) chosen to make the scale of the logistic closer to that of a cumulative standard normal; and
- θ is the continuous latent trait.

For an incorrect item response, the value is 1- $T_{\overline{c}}(u_i = 1)$ for each value of θ.

The "item preknowledge" model is a modified 3PL model. For this model, the probability of a correct response to an item is the combination of (1) the probability of answering the item correctly based on the test taker's preknowledge of the item and (2) the probability of answering the item correctly based on the test taker's underlying proficiency in the case that the test taker did not have preknowledge of the specific item. Specifically, if a test taker does have preknowledge of an item (has memorized the item), the item will be answered correctly. If a test taker has not memorized the item, the probability of a correct response is $T_{\bar{c}}(u_i = 1)$. The additional quantity that must be specified is the probability that an item has been memorized.

This paper will focus on three classes of models proposed for the probability $p(m_i)$ that an item i has been memorized. The first class is simply a constant probability for all items in the pool. For example, $p(m_i)$ may be set to 0.75, meaning that each item administered has a 75% chance of being memorized.

It may be assumed that cheating test takers will be more likely to memorize the more difficult test items. Hence, another approach is to model $p(m_i)$ as a function of each item's estimated difficulty, the threshold parameter estimate. Two such functions are

$$p(m_i \mid b_i) = \frac{1}{1 + \exp(-b_i)} \tag{2}$$

and

$$p(m_i \mid b_i) = \frac{1}{1 + \exp(1 - b_i)} . \tag{3}$$

The third class of models for the probability of memorization is a function of the specific item pool and item selection algorithm used to generate the CAT. This class is empirical because it uses the operational administration procedures to compute the probability that a specific item could potentially be memorized. The probability that an item has been memorized is computed using simulations in which some number of simulees memorize their tests. For example, suppose a pair of "source" simulees are sent into a test center to memorize items. If each test is 28 items long and the members of the pair do not receive any of the same items, the pair may take (and memorize) 56 distinct items. In most cases there will be some overlap in the items administered to a pair, and the sources will not see, (and thus memorize,) the maximum possible number of items. Because of content constraints and the use of exposure control algorithms, some items will be administered more often than others. The empirical class of models for the probability that an item has been memorized has the advantage that it incorporates all of the underlying factors in the constraints and the item-selection algorithm to produce a value for the expected vulnerability of an item. Using the empirical approach, a value for the probability that an item has been memorized may be specified for each pool used for a CAT. For example, suppose an item has been exposed to 357 of the 500 simulee pairs administered in a CAT using a specific pool. The probability that the item has been memorized using this model is 357/500, or 0.714.

The models proposed by this study for the probability that an item has been memorized are not an exhaustive set. They represent a subset of models that are independent of a test taker's proficiency. Other models considered include those based on the length of time that an item has been in the item pool and ability-dependent models such as one based on the relative difficulty of an item for a test taker. The ability-dependent models are a more difficult class to evaluate because they rely on the reliability and precision of the estimation technique used to compute test taker abilities.

Combining the probability of a correct response, whether the result of memorizing or of using underlying proficiency, with the probability that an item has been memorized, gives the overall probability of a correct response given that the test taker is using item preknowledge:

$$\begin{aligned} p(u_i = 1|c) &= T_c(u_i = 1) \\ &= p(m_i) + \left(1 - p(m_i)\right)T_{\bar{c}}(u_i = 1) \\ &= p(m_i) + \left(g_i + \frac{(1 - g_i)}{1 + \exp[-Da_i(\theta - b_i)]}\right) \end{aligned} \tag{4}$$

BEST COPY AVAILABLE

$$- p(m_i)\left(g_i + \frac{(1-g_i)}{1+\exp[-Da_i(\theta - b_i)]}\right)$$

$$= p(m_i) + g_i + \frac{(1-g_i)}{1+\exp[-Da_i(\theta - b_i)]}$$

$$- g_i p(m_i) - p(m_i)\frac{(1-g_i)}{1+\exp[-Da_i(\theta - b_i)]}$$

$$= \left(p(m_i) + g_i - g_i p(m_i)\right) + \left(1 - p(m_i)\right)(1-g_i)\frac{1}{1+\exp[-Da_i(\theta - b_i)]}$$

$$= \left(p(m_i) + g_i - g_i p(m_i)\right) + \frac{\left(1 - \left(p(m_i) + g_i - g_i p(m_i)\right)\right)}{1+\exp[-Da_i(\theta - b_i)]}$$

The model may also be written as

$$T_c(u_i = 1) = g_i' + \frac{(1-g_i')}{1+\exp[-Da_i(\theta - b_i)]}, \qquad (5)$$

where g_i' is a modified lower asymptote, $p(m_i) + g_i - g_i(p(m_i))$, and the other parameters are defined previously (in Equation 1). $T_c(u_i = 1)$ is a modified 3PL. With the incorporation of $p(m)$, the model replaces the "guessing" parameter (g) with a new "guessing-plus-item-preknowledge" combination parameter, g_i', that inflates the probability of a correct response. The new g_i' is always greater than or equal to g. For an incorrect item response, the probability $T_c(u_i = 0)$ is 1-$T_c(u_i = 1)$ for each value of θ.

We now have the equations for the probability of a correct response, the probability of an incorrect response, and a prior probability that a test taker is using item preknowledge. Using Bayes' Theorem, these components combine with each item response to give us the posterior probability that a test taker is using item preknowledge. The initial probability that a test taker is using item preknowledge is the prior $p_0(c)$. After each item is administered to a test taker, this probability is updated in a manner that depends on the response to that item. For the first item, we assume that the question of whether test takers cheat is independent of their proficiency. (There are various pressures to cheat on a test at all ranges of proficiency.) So, the joint density of c and θ for the first item response is $p(c, \Theta = \theta) = p_0(c)p_0(\theta)$. The probability that a test taker is using item preknowledge after the first item response is $p_1(c) \propto \int T_c(u_1)p_0(c)p_0(\theta)d\theta$, where $T_c(u_1)$ is $T_c(u=1)^{u_1}\,T_c(u=0)^{1-u_1}$, which is $T_c(u=1)$ or $T_c(u=0)$, depending on the item response u_1. Integrating this with respect to θ for both cases (c and $\bar{c}$) and normalizing, we have the equation for the posterior probability that a test taker is using item preknowledge (given that he or she responded u_1 to the first item) which is not conditional on proficiency level

$$p_1(c) = \frac{\int T_c(u_1)p_0(c)p_0(\theta)d\theta}{\int T_c(u_1)p_0(c)p_0(\theta)d\theta + \int T_{\bar{c}}(u_1)p_0(\bar{c})p_0(\theta)d\theta}. \qquad (6)$$

After the first item, we may no longer assume that c and θ are independent because we have updated the distribution of proficiency using an item response that may or may not have used prior item knowledge. So, after the test taker is administered the next item, we build our new estimate for the probability of item preknowledge using the previous joint density of c and θ,

$$p_1(c,\theta) = \frac{T_c(u_1)p_0(c)p_0(\theta)}{\int T_c(u_1)p_0(c)p_0(\theta)d\theta + \int T_{\bar{c}}(u_1)p_0(\bar{c})p_0(\theta)d\theta}, \tag{7}$$

which is proportional to $T_c(u_1)p_0(c)p_0(\theta)$.
The probability that the test taker is using item preknowledge after the second item is

$$p_2(c) = \frac{\int Tc(u_2)p_1(c,\theta)d\theta}{\int Tc(u_2)p_1(c,\theta)d\theta + \int T\bar{c}(u_2)p_1(\bar{c},\theta)d\theta}, \tag{8}$$

and $p_2(c) \propto \int T_c(u_2)p_1(c,\theta)d\theta$, which is the area of the posterior distribution of θ given item preknowledge. The posterior probability that the test taker is using item preknowledge after *n* items is $p_n(c) \propto \int T_c(u_n)p_{n-1}(c,\theta)d\theta$.

The magnitude of $p_n(c)$ could provide a simple index for identifying test takers with item preknowledge. However, a more useful index is a ratio between the current odds (after *n* items) and the prior odds—the final odds ratio. For numerical convenience, the log (base 10) of the odds ratio is used for analyses. The final log odds ratio index is

$$\log_{10}\left[\frac{p_n(c)/[1-p_n(c)]}{p_0(c)/[1-p_0(c)]}\right]. \tag{9}$$

For example, a final log odds ratio of 0 implies that after the 28 items we do not have any more suspicion that this test taker is using item preknowledge than we had before the 28 items were administered. Therefore, we assume the probability that the test taker is using item preknowledge to be the base rate for item preknowledge use. A final log odds ratio of 1 implies that we are 10 times more suspicious that a test taker is cheating than we were before the 28 items were administered. It follows that with a final log odds ratio of -1 we are 10 times less suspicious that a test taker is cheating than we were before the 28 items were administered. In general, a negative log odds ratio suggests evidence that the test taker is not cheating.

Method

The current project used simulations based on an operational CAT. Simulees were generated at true scores from a discrete uniform distribution at eleven θ values. The θ values were translated into a number correct true score on a linear 60-item reference test and correspond to the operational test's score range (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 59). At each true score there were 10,000 simulees generated for the null group and 10,000 simulees generated for each of the nine memorizing-group conditions. For each simulee, a 28-item response pattern was generated using an operational CAT item pool and weighted deviations item selection criteria (Stocking & Swanson, 1993).

The item pool contained 494 items. Of these, 341 were discrete items and 153 were part of item sets (i.e., associated with one of 22 reading passages or other type of stimuli). Although preknowledge of stimuli is another possible strategy for inflating test scores, this research treats all items as discrete and ignores the stimulus component of preknowledge tactics. For the memorizing conditions, if a simulee was administered one of the memorized items, a correct response was automatically given. The 3PL IRT model with operational item parameter estimates was used to generate a response when one of the remaining items was administered.

Population weights were used in some analyses to allow comparisons that are representative of an operational distribution of proficiencies. See Table 1 for the relationship between true scores, θ values, and population weights. True scores ranged from 7.84 to 59.

TABLE 1

True test scores, θ values, and corresponding simulation population weights representative of an operational administration of the test used in the simulation study

True Score	θ	Weight
10	-3.8394	0.001442
15	-2.1841	0.029116
20	-1.3811	0.100307
25	-0.8118	0.158306
30	-0.3482	0.171876
35	0.0534	0.154741
40	0.4271	0.125484
45	0.8074	0.106487
50	1.2419	0.094023
55	1.8824	0.054866
59	3.5462	0.003353

Null Group Response Patterns

The simulation design for the null group generated 28-item response vectors for 10,000 simulees at each of the 11 true-score values. No item preknowledge was assumed. These data served as a baseline to compare CAT response patterns that fit the IRT model with CAT response patterns based on item preknowledge.

Memorizing Group-Response Patterns

The simulation design for each memorizing group generated 28-item response vectors for 10,000 simulees at each of the 11 true score values. Item preknowledge was assumed. The design used a real-world approach to simulate the item preknowledge process by incorporating a two-stage process. First, for each condition, the design sent in *n* sources to memorize test items. These *n* test takers memorized their items perfectly and then combined their lists. (Some overlap was observed among the lists.) The complete list was memorized by another group of test takers, the beneficiaries. Then the beneficiaries were administered a 28-item test, and if they were administered any of the memorized items, they answered them correctly. (Although we acknowledge that memorizers may not have perfect recall of the item list, the simulation was designed to produce a worst-case scenario for the testing program.) The three highest true scores were used for the sources' proficiency levels. One advantage of using this design is that it is a model of a possible reality, especially if recording equipment is used for item theft. Another advantage is that the design maintains the roles of content constraints and the item selection algorithm in the success of using item preknowledge in the CAT environment.

Design

Table 2 shows the 10 different source-beneficiary conditions for the CAT response patterns. Nine test conditions were formed from three source-true score conditions completely crossed with three number-of-sources conditions. A null case in which response patterns were generated using each simulee=s underlying proficiency was also included. For each cell in Table 2, 10,000 CAT response patterns were generated at each of 11 beneficiary true score values. The total number of simulees for the design was 1,100,000.

TABLE 2

Condition numbers for the research design

Number of Sources	Source True Score Level	Condition Number
0	—	1
2	50	2
4	50	3
8	50	4
2	55	5
4	55	6
8	55	7
2	59	8
4	59	9
8	59	10

Two hundred beneficiary replications were generated for each source true score level condition. Each replication contained 50 simulees at each of the 11 proficiency levels, for a total of 10,000 beneficiary simulees at each true score value (200 X 50= 10,000).

Within a replication, the two-stage process included a source simulation followed by a beneficiary simulation. For example, Condition 3 used four sources at true score of 50. Within a replication, four test takers at true score of 50 were administered a test. Then their item lists were concatenated and used by 550 beneficiary simulees (50 at each of 11 proficiency levels) when they were administered tests. The process was repeated 200 times simulating additional sources and beneficiaries. For this project, source replications were simulated for two, four, and eight sources at source true scores of 50, 55, and 59. Eight sources were selected as the highest number of sources because preliminary analyses showed that the average list length gathered by eight sources was 122-138 items. For this study, we assumed beneficiaries would not be willing to memorize more than 122 to 138 items.

For each simulee, seven final log odds ratios (FLOR1-FLOR7) were calculated based on the responses to the 28 items administered. Each odds ratio was computed using a different model for the probability that an item had been memorized. Table 3 shows the seven models used for the probability that an item had been memorized.

TABLE 3

Models used for the probability that an item has been memorized

Final Log Odds Ratio	Model Class	p(m$_i$)
FLOR1	constant	$p(m_i) = 0.1$
FLOR2	constant	$p(m_i) = 0.5$
FLOR3	difficulty	$p(m_i \mid b_i) = \frac{1}{1+\exp(-b_i)}$
FLOR4	difficulty	$p(m_i \mid b_i) = \frac{1}{1+\exp(1-b_i)}$
FLOR5	empirical	Item relative frequency using 2 sources
FLOR6	empirical	Item relative frequency using 4 sources
FLOR7	empirical	Item relative frequency using 8 sources

BEST COPY AVAILABLE

Calculation of the Empirical Indices

FLOR5 to FLOR7 approaches use preliminary relative item frequency data averaged over pairs, quadruples, or sets of eight sources at true scores of 50, 55, and 59 from a simulation of 10,000 simulees at each of the 11 true scores. FLOR5 is based on pairs of sources. FLOR6 is based on sets of four sources. FLOR7 is based on sets of eight sources. For example, to compute $p(m_i)$ for FLOR6, an item was administered to 71.4 percent of the simulee quadruples at true score 50. This same item was administered to 68 percent of the simulee quadruples at true score 55 and 44.8 percent of the simulee quadruples at true score 59. The probability, $p(m)$, used for the Bayesian index is the average of these three cases, or 0.614. The empirical approach is unique for each item selection algorithm and item pool combination. Different values for $p(m_i)$ are expected if either the item selection algorithm or item pool is altered.

Results

Number of Items Memorized

Table 4 shows the mean and standard deviation of the number of memorized items received under the nine memorizing CAT conditions. By using the source-beneficiary strategy, simulees were successful in receiving items that had been memorized. If the simulees had access to a list of items memorized by two sources, they would expect to receive an average of about six of those items when taking the 28-item test. More items were received if the simulee had a list from more sources. (Note: as the number of sources increases, the longer the list, and the more items a beneficiary has to memorize.) Average list lengths for two, four, and eight sources were approximately 50, 85, and 125 items, respectively. The item pool contained 494 items. The number of items received by a beneficiary did not seem to be influenced by the ability of the source as much as the number of sources.

TABLE 4
Mean and standard deviations (in parentheses) of the number of memorized items received by the memorizing group (Test length was 28 items)

Number of	Sources' True Score		
Sources	50	55	59
2	5.86 (3.78)	5.56 (4.07)	5.48 (4.16)
4	11.36 (5.29)	11.20 (6.17)	10.99 (6.43)
8	18.10 (4.99)	19.61 (6.47)	19.58 (7.04)

Test Score Inflation

Table 4 contains evidence that the source-beneficiary strategy was successful at helping test takers gain preknowledge of items that were later administered to them. It also shows that the ability of the source does not have as large an effect on the number of memorized items received by the beneficiaries as the number of sources. However, the table does not show the success of the strategy at inflating these beneficiaries' test scores. Table 5 presents the mean and standard deviation of the estimated test scores for the test takers. Because the number of sources has a larger effect on success than the source's ability, Table 5 contains the average estimated test score by the number of sources and the beneficiary's true score. When comparing the beneficiaries' estimated test scores with those of the null group, note that the null group's test scores represent the variability and bias produced by the CAT item selection algorithm, the item pool's characteristics, and the estimation method.

TABLE 5
Mean and standard deviation of the estimated test scores by memorizing group and true score level

Number of	True Score Level										
Sources	10	15	20	25	30	35	40	45	50	55	59
None											
Mean	10.18	15.02	20.08	25.10	30.17	35.06	40.04	44.90	49.94	54.98	59.27
SD	1.54	2.58	3.01	3.36	3.61	3.69	3.60	3.27	2.65	1.69	0.79
Two											
Mean	12.74	17.87	23.34	29.14	35.12	40.71	45.49	49.51	53.07	56.37	59.45
SD	6.96	7.01	7.07	7.12	6.85	5.97	4.78	3.59	2.48	1.51	0.70
Four											
Mean	22.03	26.58	31.94	37.43	42.72	47.12	50.47	53.04	55.16	57.32	59.59
SD	16.44	14.85	13.28	11.51	9.26	6.80	4.66	3.13	2.06	1.36	0.63
Eight											
Mean	40.35	43.43	46.67	49.78	52.55	54.48	55.60	56.50	57.35	58.44	59.74
SD	20.25	17.42	14.37	11.08	7.61	4.74	3.01	2.02	1.53	1.17	0.51

The first block in Table 5 shows the mean estimated test score and standard deviation for the noncheating or null group. Note that for fixed true scores, the mean estimated true score was slightly inflated, except for three of the four highest abilities. The estimates were more variable in the middle true scores, but even at these scores the standard deviations were less than four score points. (Range of possible true scores was 7.84 to 59.)

For the memorizing simulees, across all true scores, the mean test score was inflated. Therefore, we conclude that by using the source-beneficiary preknowledge strategy, the test takers were successful in attaining higher test scores. The estimates were more inflated when the test takers had memorized the longer lists gathered by eight sources. Even the lower ability test takers for the eight-sources condition had an average estimated test score above 40 (out of a possible 60 score points). Information from four sources did not deliver as much test-score gain as information from eight sources. Similarly, item information from two sources did not aid a beneficiary as much as information from four sources. The estimates were more variable at the lower true scores, where the takers had more room to benefit from the preknowledge, depending on the peculiarities of item selection. The higher ability test takers did not benefit as much from the memorization because they were already the higher scorers.

Distributional Characteristics of the Bayesian Index

Table 6 shows the means of the seven indices by memorizing condition and test taker true score. Across index and true score, average values for the null condition were lower than for the memorizing conditions. In most cases, there was more difference between the averages for the null and memorizing groups at the lower true scores and more difference between the null and eight-sources group. Only FLOR3, FLOR6 and FLOR7 attained average final log odds ratios above 1, implying 10 times the suspicion after the 28-item test that the test takers are using item preknowledge. However, some reservation should be used when judging Table 6 because it includes all simulees at each condition. Some of these may have achieved more success at test-score inflation than others in the same condition.

TABLE 6

Mean of the final log odds ratios (FLOR1-FLOR7) for the null and memorizing groups by true score level. (Each null cell contains data from 10,000 simulees. Other cells contain data from 30,000 simulees.)

Number of	True Score Level										
Sources	10	15	20	25	30	35	40	45	50	55	59
FLOR1 $p(m_i)$=0.1											
Null	-0.70	-0.34	-0.15	-0.08	-0.04	-0.01	0.02	0.05	0.06	0.08	0.10
Two	-0.24	-0.01	0.09	0.14	0.16	0.16	0.15	0.13	0.11	0.09	0.10
Four	0.05	0.16	0.20	0.22	0.22	0.20	0.17	0.14	0.11	0.10	0.10
Eight	0.11	0.14	0.15	0.15	0.15	0.14	0.13	0.12	0.11	0.10	0.11
FLOR2 $p(m_i)$=0.5											
Null	-5.90	-3.77	-2.38	-1.59	-1.15	-0.86	-0.55	-0.17	0.18	0.44	0.60
Two	-4.27	-2.57	-1.45	-0.69	-0.19	0.14	0.37	0.51	0.56	0.56	0.63
Four	-2.39	-1.19	-0.34	0.24	0.60	0.78	0.81	0.76	0.68	0.60	0.64
Eight	-0.54	0.06	0.45	0.71	0.84	0.85	0.80	0.74	0.67	0.62	0.66
FLOR3 Difficulty											
Null	-3.67	-2.73	-2.10	-1.97	-2.15	-2.33	-2.30	-1.96	-1.36	-0.19	1.16
Two	-2.49	-1.86	-1.48	-1.31	-1.24	-1.13	-0.91	-0.62	-0.23	0.44	1.23
Four	-1.19	-0.79	-0.51	-0.28	-0.10	0.06	0.18	0.30	0.45	0.77	1.28
Eight	0.36	0.55	0.71	0.88	0.98	1.03	1.00	0.98	0.97	1.05	1.33
FLOR4 Shifted Difficulty											
Null	-1.51	-1.07	-0.78	-0.74	-0.83	-0.91	-0.87	-0.68	-0.42	0.08	0.71
Two	-0.59	-0.33	-0.19	-0.11	-0.06	-0.01	0.06	0.12	0.19	0.40	0.76
Four	0.27	0.40	0.48	0.56	0.63	0.65	0.61	0.55	0.51	0.55	0.78
Eight	0.78	0.80	0.83	0.89	0.91	0.89	0.82	0.75	0.68	0.67	0.81
FLOR5 Empirical (2 sources)											
Null	-0.43	-0.35	-0.29	-0.32	-0.39	-0.45	-0.39	-0.21	0.05	0.27	0.32
Two	0.42	0.37	0.33	0.33	0.33	0.33	0.33	0.35	0.35	0.34	0.33
Four	0.96	0.85	0.77	0.76	0.74	0.69	0.61	0.51	0.42	0.36	0.34
Eight	0.84	0.73	0.67	0.66	0.64	0.60	0.54	0.47	0.42	0.37	0.35
FLOR6 Empirical (4 sources)											
Null	-0.95	-0.82	-0.73	-0.81	-1.02	-1.25	-1.22	-0.89	-0.25	0.46	0.63
Two	0.17	0.13	0.07	0.04	-0.01	0.00	0.09	0.28	0.48	0.65	0.65
Four	1.02	0.91	0.84	0.85	0.87	0.86	0.84	0.79	0.75	0.70	0.66
Eight	1.29	1.19	1.14	1.15	1.14	1.09	1.01	0.91	0.81	0.72	0.67
FLOR7 Empirical (8 sources)											
Null	-2.04	-1.83	-1.70	-1.95	-2.52	-3.20	-3.44	-3.02	-1.72	0.31	1.23
Two	-0.73	-0.78	-0.89	-1.11	-1.39	-1.54	-1.37	-0.86	-0.04	0.92	1.26
Four	0.31	0.19	0.07	0.05	0.04	0.10	0.25	0.51	0.84	1.19	1.28
Eight	1.46	1.37	1.35	1.41	1.45	1.47	1.45	1.42	1.39	1.35	1.30

For the indices based on constant models, FLOR1 and FLOR2, the average index value increased as the beneficiaries' true scores increased for the null group. This implies that we are more suspicious of test takers who score high on the test, in general. For the two-, four-, and eight-sources conditions, average FLOR1 values increased as beneficiaries' true scores increased for the lower ability simulees. For the higher ability simulees, the average index values decreased. FLOR2 behaved much like FLOR1 for the four- and eight-sources conditions. Average FLOR1 values ranged from –0.70 to 0.22. Average FLOR2 values ranged from –5.90 to 0.85.

ERIC Full Text Provided by ERIC

FLOR3 and FLOR4, the indices based on the threshold estimates, also had increasing average values across beneficiaries' true scores for the null group. FLOR3 maintained a systematic increase for the memorizing group across true score, except for the eight-sources condition. For this case, the averages dipped for the 40-45 true score range before increasing at 55 and 59. Average values ranged from –3.67 to 1.33 and –1.51 to 0.91 for FLOR3 and FLOR4, respectively.

The empirically based indices, FLOR5, FLOR6, and FLOR7 exhibited more variation in their average values. The average values did not systematically increase as beneficiaries' true scores increased, even for the null condition. FLOR5's average values ranged from -0.45 to 0.96. FLOR6's average values ranged from -1.25 to 1.29. FLOR7 had a larger range at -3.44, for a null condition, to 1.47, for an eight-sources condition.

Table 7 shows the standard deviations of these three indices. The least variable index was FLOR1, and the most variable index was FLOR7. Index values were more variable for those beneficiaries at the lower ability ranges. The variability at these levels may have been due, in part, to the range of success at test-score inflation.

TABLE 7

Standard deviation of the final log odds ratios (FLOR1-FLOR7) for the null and memorizing groups by true score level. (Each null cell contains data from 10,000 simulees. Other cells contain data from 30,000 simulees.)

Number of	True Score Level										
Sources	10	15	20	25	30	35	40	45	50	55	59
FLOR1 $p(m_i)$=0.1											
Null	0.3	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0
Two	0.5	0.4	0.4	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.0
Four	0.5	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.0	0.0
Eight	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0
FLOR2 $p(m_i)$=0.5											
Null	1.2	1.2	0.9	0.8	0.7	0.6	0.6	0.6	0.5	0.4	0.2
Two	2.1	1.7	1.3	1.1	1.0	0.8	0.7	0.6	0.5	0.3	0.1
Four	2.9	2.2	1.7	1.3	1.0	0.8	0.6	0.5	0.4	0.3	0.1
Eight	2.4	1.7	1.2	0.8	0.6	0.5	0.4	0.4	0.3	0.2	0.1
FLOR3 Difficulty											
Null	1.1	1.1	1.0	1.0	1.1	1.1	1.1	1.2	1.2	1.1	0.5
Two	1.5	1.3	1.2	1.2	1.2	1.3	1.3	1.3	1.2	1.0	0.4
Four	1.9	1.7	1.5	1.5	1.4	1.3	1.3	1.2	1.1	0.8	0.4
Eight	1.8	1.6	1.5	1.3	1.2	1.1	1.1	1.0	0.9	0.7	0.3
FLOR4 Shifted Difficulty											
Null	0.7	0.7	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.7	0.3
Two	1.1	1.0	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.6	0.3
Four	1.4	1.3	1.2	1.1	1.1	1.0	0.9	0.8	0.7	0.5	0.3
Eight	1.1	1.0	1.0	0.9	0.8	0.8	0.7	0.6	0.6	0.4	0.2
FLOR5 Empirical (2 sources)											
Null	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.5	0.4	0.2	0.1
Two	0.9	0.8	0.8	0.8	0.7	0.7	0.6	0.4	0.3	0.2	0.1
Four	1.0	0.9	0.9	0.8	0.7	0.6	0.5	0.3	0.2	0.1	0.1
Eight	0.7	0.6	0.6	0.5	0.4	0.3	0.3	0.2	0.2	0.1	0.1
FLOR6 Empirical (4 sources)											
Null	0.6	0.6	0.5	0.6	0.7	0.8	0.9	0.9	0.8	0.4	0.1
Two	1.1	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.6	0.3	0.1
Four	1.2	1.2	1.2	1.1	1.0	0.9	0.8	0.6	0.5	0.3	0.1
Eight	0.9	0.9	0.8	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
FLOR7 Empirical (8 sources)											
Null	1.0	1.0	1.0	1.1	1.4	1.5	1.5	1.5	1.4	1.0	0.3
Two	1.3	1.2	1.2	1.3	1.4	1.6	1.6	1.5	1.2	0.7	0.2
Four	1.4	1.4	1.4	1.4	1.5	1.5	1.4	1.2	0.9	0.5	0.2
Eight	1.2	1.1	1.1	1.1	1.0	0.9	0.8	0.7	0.5	0.4	0.2

BEST COPY AVAILABLE

Table 8 shows the number of simulees in each category by condition and true score. Note that, in general, as the beneficiary true score increased, more simulees received many memorized items. Table 8 supports the suggestion that the number of source<s is more important than sources' true score for a beneficiary's success at receiving memorized items. Table 8 also suggests an interaction between beneficiary true score, number of sources, and the number of memorized items received. A pattern appears across the columns in Table 8.

TABLE 8
Number of simulees by number of memorized items received.

Beneficiary True Score	Number of Mem. Items	Number of Sources								
		Two			Four			Eight		
		Source True Score								
		50	55	59	50	55	59	50	55	59
10	00-07	9255	9510	9502	6340	7043	7097	2924	3401	3519
	08-14	694	456	472	1448	1283	1488	89	177	334
	15-21	51	33	22	2195	1602	1315	5159	2089	1702
	22-28	—	1	4	17	72	100	1828	4333	4445
15	00-07	9093	9388	9360	5986	6626	6874	2468	3012	3176
	08-14	854	580	599	1656	1516	1567	147	203	372
	15-21	53	32	39	2341	1791	1462	5542	2317	1843
	22-28	—	—	2	17	67	97	1843	4468	4609
20	00-07	8790	9210	9236	5019	5969	6374	1777	2409	2751
	08-14	1145	735	714	2237	1868	1861	311	272	398
	15-21	65	55	50	2735	2094	1669	5985	2545	2002
	22-28	—	—	—	9	69	96	1927	4774	4849
25	00-07	8157	8848	8884	3747	5125	5463	983	1741	2072
	08-14	1781	1100	1056	3080	2335	2400	451	412	542
	15-21	62	52	60	3160	2468	2028	6674	2916	2242
	22-28	—	—	—	13	72	109	1892	4931	5144
30	00-07	7108	8274	8393	2205	3728	4269	326	894	1237
	08-14	2820	1651	1517	4086	3245	3042	456	586	720
	15-21	72	75	87	3697	2926	2571	7315	3275	2585
	22-28	—	—	3	12	101	118	1903	5245	5458
35	00-07	5836	7356	7607	983	2292	2818	64	308	483
	08-14	4060	2539	2315	4785	4043	4045	318	536	770
	15-21	104	105	78	4222	3566	2999	7769	3722	2972
	22-28	—	—	—	10	99	138	1849	5434	5775
40	00-07	4566	6085	6471	335	1025	1401	10	67	152
	08-14	5343	3781	3409	5329	4593	4716	229	334	617
	15-21	91	134	120	4328	4280	3742	8008	3967	3353
	22-28	—	—	—	8	102	141	1753	5632	5878
45	00-07	3794	4715	5051	110	372	522	—	6	23
	08-14	6108	5116	4780	5767	4515	4580	221	152	305
	15-21	98	169	169	4116	5013	4726	8135	4137	3356
	22-28	—	—	—	7	100	172	1644	5705	6316
50	00-07	3726	3270	3717	77	105	138	1	—	2
	08-14	6192	6508	6037	6568	4273	3826	386	49	106
	15-21	82	222	246	3347	5513	5872	8076	4244	3217
	22-28	—	—	—	8	109	164	1537	5707	6675
55	00-07	4621	2493	2354	201	24	19	1	—	—
	08-14	5329	7267	7313	7115	4092	3102	569	21	16
	15-21	50	240	330	2677	5779	6688	7931	4376	3103
	22-28	—	—	3	7	105	191	1499	5603	6881
59	00-07	4883	2607	2047	229	33	9	—	—	—
	08-14	5069	7161	7610	7061	4149	3007	601	16	3
	15-21	48	230	341	2701	5721	6801	7884	4416	3086
	22-28	—	2	2	9	97	183	1515	5568	6911

ERIC
Full Text Provided by ERIC

ROC Curve Analysis

Marginal probability ROC curves (Green & Swets, 1966) offer an evaluation of the Bayesian indices. As the points on an ROC curve represent the ratio of false alarms to hits, such curves provide a visual tool for assessing the power of these indices in this simulated CAT environment. Empirical ROC curves were calculated using the simulation data for FLOR1-FLOR7. For each point on the ROC curve, the value on the horizontal axis is the proportion of those from the null group (falsely) detected by an index using a particular cut-off value (false-alarm rate), and the detected proportion of those from the beneficiary group is indicated by the vertical-axis value (hit rate). These are weighted proportions using the population weights given in Table 1.

Figure 1 shows the partial ROC curve for those beneficiaries who received at least 15 memorized items out of the 28 items administered. The FLOR7 index shows the steepest slope, quickly approaches 1.0, and is, therefore, the most powerful index for beneficiaries receiving at least 15 memorized items. For example, for 5% false-alarm rate, over 84% of the memorizing beneficiaries were detected using FLOR7. FLOR6 performs only slightly worse, and the difficulty index holds third place in order by power to detect memorizing beneficiaries.

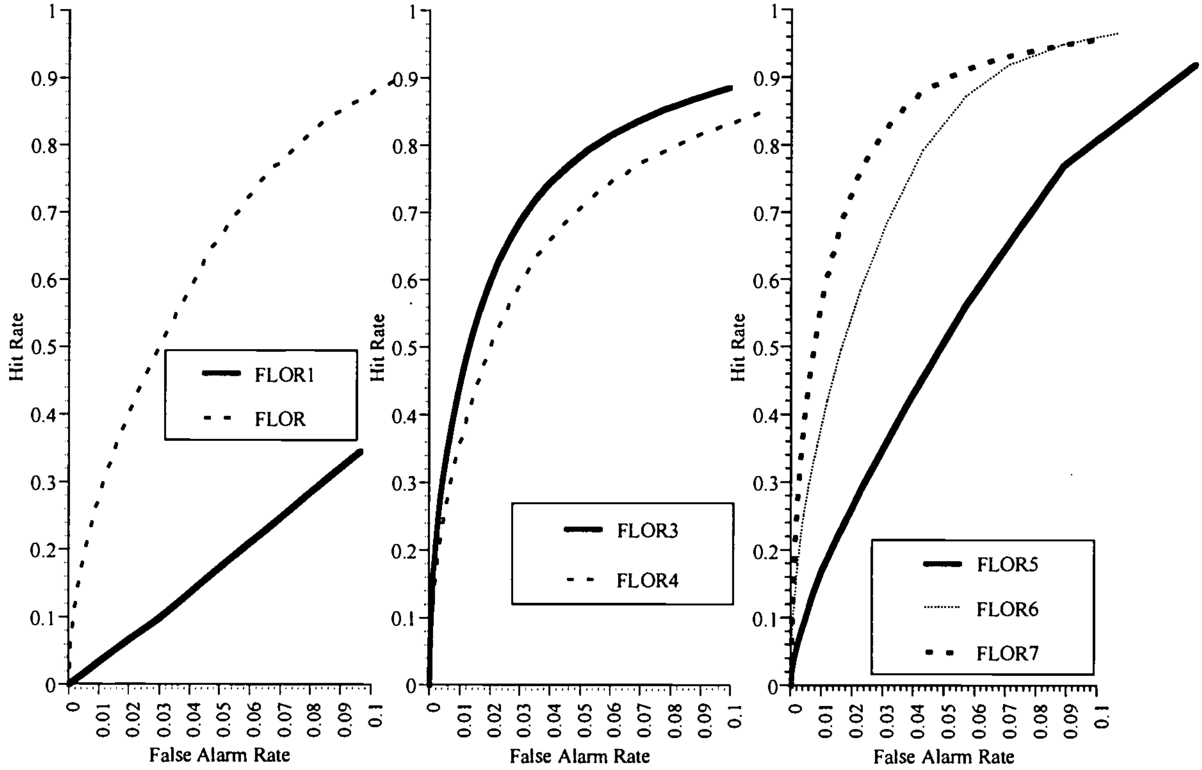

FIGURE 1. *Weighted ROC curve for those simulees that received at least 15 memorized items in the simulation based on 28 items*

BEST COPY AVAILABLE

The distributions for the seven indices for these beneficiaries are displayed in Figure 2 using box plots. The top, bottom, and middle lines through each box correspond to the 75th percentile, 25th percentile, and the 50th percentile (or median) of each distribution, respectively. The end of the top whisker shows the 90th percentile, and the end of the bottom whisker represents the 10th percentile. The dot in each box represents the mean. For each index, the left box plot shows the distribution for the null simulees; these simulees did not receive any memorized items. The right box plot shows the distribution for the simulees who received at least 15 memorized items on the 28-item test. For each index, the null distribution was more variable and always more negative than that for the successful memorizing group. It also appeared that an index that assigned more items higher $p(m_i)$'s, was also more variable. (Refer to the Appendix for the distribution of $p(m_i)$ for the item pool used.)

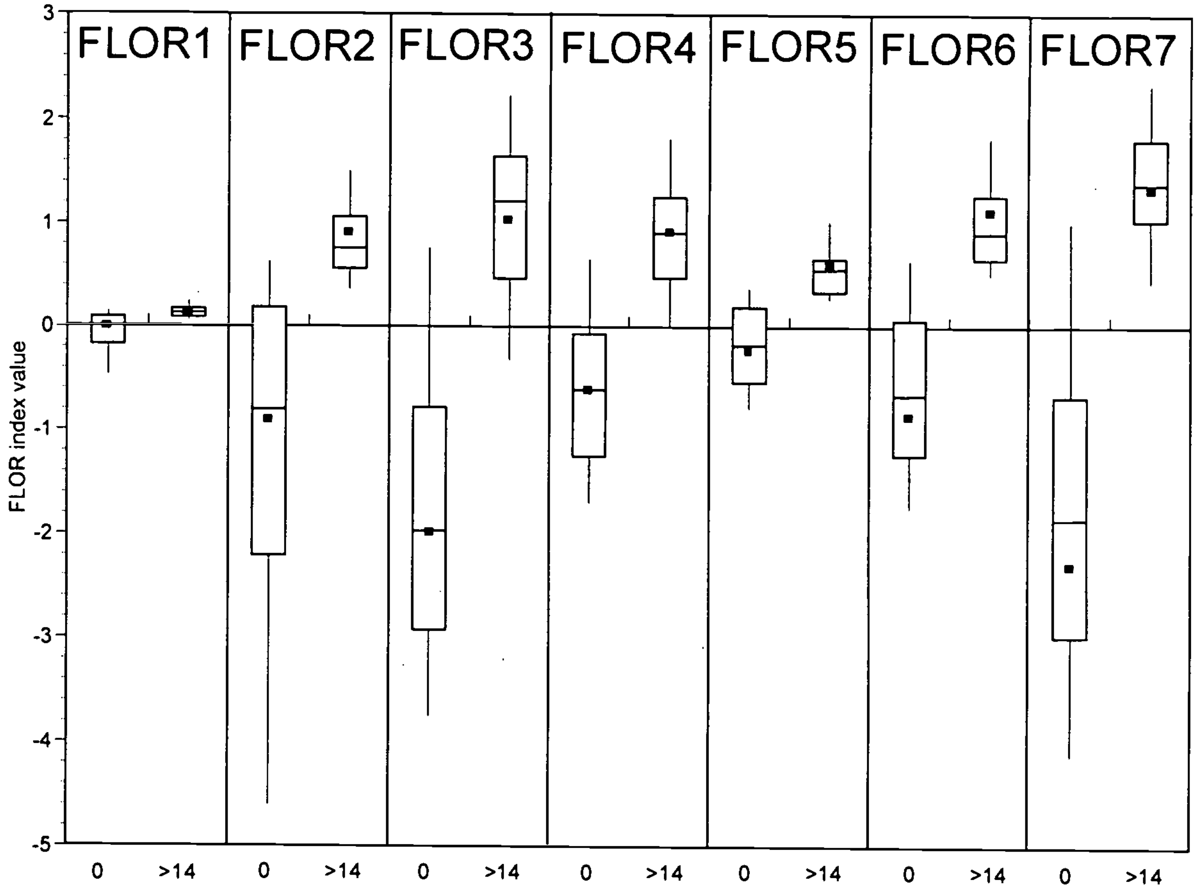

FIGURE 2. *Box plots of FLOR1-FLOR7 values for the simulation by number of memorized items received*

The FLOR7 index showed the largest separation between the two groups. The FLOR3 index showed the next highest amount of separation, and FLOR1 showed the most overlap. The most variable indices were FLOR2, FLOR3, and FLOR7.

Overall, the best performing index of those studied is FLOR7, because it has the most power to detect those beneficiaries who had preknowledge of more than half of the items on their test. FLOR3 is selected as the second best performing index for the successful beneficiaries. It has the extra appeal of being simple to compute without a previous simulation.

Correlation Analyses

Table 9 presents the pairwise weighted correlations among the seven indices for simulation data. The correlations using the null group (*N*=110,000) are in the upper triangle, and the successful memorizing group (*N*=407,900) that received at least 15 memorized items are in the lower triangle. For the null group, the correlations among the indices in the same class (constant, difficulty, or empirical) were higher than among those in different classes. In the memorizing group, the correlations were generally lower within the classes. (FLOR1 and FLOR2 were the exceptions.) For example FLOR7 correlated 0.83 with FLOR5 in the null group and 0.54 in the memorizing group. These lower correlations may reflect the decreased variability for each index in the memorizing group.

TABLE 9
Pairwise weighted correlations among the seven indices using the null group (N=110,000) in the upper triangle and the memorizing group (N=407,920) in the lower triangle. The memorizing group used for this table consists of simulees that received at least 15 memorized items.

	FLOR1	FLOR2	FLOR3	FLOR4	FLOR5	FLOR6	FLOR7
FLOR1	1.000	0.862	0.643	0.721	0.610	0.480	0.274
FLOR2	0.911	1.000	0.640	0.626	0.500	0.426	0.243
FLOR3	0.548	0.742	1.000	0.964	0.683	0.743	0.729
FLOR4	0.787	0.928	0.919	1.000	0.748	0.754	0.684
FLOR5	0.875	0.857	0.454	0.717	1.000	0.962	0.827
FLOR6	0.764	0.849	0.583	0.773	0.947	1.000	0.944
FLOR7	0.353	0.522	0.715	0.639	0.538	0.743	1.000

Discussion

The goals of this study included examining test taker success at test-score inflation when using item preknowledge and the feasibility of using an odds ratio index as a tool for test security. The results of this study show that test takers may be very successful at test-score inflation when using item preknowledge. Furthermore, increasing the number of sources rather than the sources' true score yields more success at test-score inflation. In addition, results for the FLOR7 index show that the combination of the empirical model based on eight sources and 3PL IRT model is useful for modeling behavior that mimics the source-beneficiary strategy. The findings from the simulation study indicate that FLOR7 shows some promise for that application. The correlation analysis results indicate that the empirical models show a moderate amount of agreement for identifying simulees that were using item preknowledge.

Application

The final log odds ratio proposed in this study may be used as an index to detect test takers who use item preknowledge to inflate their test scores. Before doing so, criteria for detection must be established. A straightforward way to set a criterion would be based on the value of the final log odds ratio. A value of 2 means that the probability that a test taker is using item preknowledge is 100 times more than before we knew his or her responses to the test items. A value of -2 means that we are about 100 times less suspicious that a person is using item preknowledge than before we knew his or her responses to the test items. A viable criterion may be 2.0, meaning that anyone receiving a final log odds ratio value of 2.0 or more would be flagged as having item preknowledge.

BEST COPY AVAILABLE

Conclusions and Future Research

Individuals who behave aberrantly on large-scale tests are currently detected using several techniques. These techniques flag response patterns that have discrepancies from a ?fitting≅ response pattern. They do not model the behavior they are designed to detect, but simply look for discrepancies from the behavior that ?fits.≅ In the approach developed in this paper, the ?nonfitting≅ or aberrant behavior is also modeled. A traditional IRT model is used for the fitting patterns and a new model is used for the nonfitting patterns (in our case, patterns that reflect the use of item preknowledge). Then, the concept of odds ratios is extended to describe the increased likelihood (based on the item responses) that a response pattern arises from the old or new models. From the results of the simulations, the new approach shows promise for use as a test-security index in the CAT environment.

The purpose of this research was not to investigate the performance of an index to identify subjects who were using item preknowledge of random items from the item pool, but to identify those using preknowledge of items that sources would be administered given an adaptive test. Therefore, results should be viewed within the specific strategy described.

Another limitation of this study was the source strategy used. Sources gain access to those items received by higher ability test takers. The purpose of this strategy is to gain knowledge of items that would give a higher test score due to the adaptive nature of the test. However, the lower ability subjects may not do well enough on the first items of the test to receive any memorized items. Other strategies will be studied in future research. For example, for another strategy sources may manipulate the type of items they receive by choosing to give incorrect responses to administered items. The strategies may prove more advantageous for lower proficiency users.

Because the use of the Bayesian index ?online≅ during an actual CAT may be expensive, other uses for the index may be more practical. One possible use is that of a quality-control device for the item pool. The index might be used to track the ?freshness≅ or security of a pool using test takers' response patterns. It is hoped that this work will enable testing-program management to more effectively decide how long to leave an item pool in the field.

References

Green, D.M., & Swets, J.A. (1966). *Signal detection theory andpsychophysics*. New York: Wiley.

Drasgow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, *10*, 59-67.

Lewis, C. (June 1997). Personal communication. Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277-292.

Appendix

Three different classes of models were proposed for the probability that an item had been memorized, $p(m_i)$. The first class was simply a constant probability for all items in the pool. FLOR1 used $p(m_i)$ =0.1 and FLOR2 used $p(m_i)$ = 0.5.

The second class was based on the difficulty of the test items. FLOR3 and FLOR4 used logistic functions of each item's estimated difficulty, the threshold parameter estimate. FLOR3 used

$$p(m_i|b_i) = \frac{1}{1+\exp(-b_i)} \quad , \tag{A1}$$

and FLOR4 used

$$p(m_i|b_i) = \frac{1}{1+\exp(1-b_i)} \quad . \tag{A2}$$

The third class of models for the probability of memorization were functions of the specific item pool and item selection algorithm used to generate the CAT. FLOR5 was based on the relative item frequency when teams of size 2 were sent in to memorize items. FLOR6 used teams of size 4 and FLOR7 used teams of size 8. Table A1 presents descriptive statistics (maximum, minimum, mean, and standard deviation) for the item pool used and the five models used for the probability that an item has been memorized for the item pool used for this project. (The item pool contained 494 items.)

TABLE A1
Comparison of item parameters and p(m)'s used for FLOR1-FLOR7 in the CAT simulation.

	Descriptive statistics			
	Max	Min	Mean	Standard Deviation
Item Parameters				
a	1.8	0.2	0.8	0.3
b	2.7	-4.6	0.0	1.2
g	0.5	0	0.1	0.1
p(m)				
FLOR1	0.1	0.1	0.1	—
FLOR2	0.5	0.5	0.5	—
FLOR3	0.9	0.0	0.5	0.2
FLOR4	0.8	0.0	0.3	0.2
FLOR5	0.4	0.0	0.1	0.1
FLOR6	0.7	0.0	0.2	0.2
FLOR7	0.9	0.0	0.3	0.3

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

□ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").