Detecting Item Memorization in the CAT Environment

Lori Davis McLeod, Law School Admission Council

Charles Lewis, Educational Testing Service

The purpose of appropriateness/person-fit indices is to identify response patterns for which a given item response theory model is inappropriate for an examinee even though that model is appropriate for a group. This study was concerned with those cases in which examinees had prior knowledge of items from an item bank used to generate a computerized adaptive test (CAT) and used the memorized information to inflate their test scores. The objective was to evaluate procedures that could identify these individuals by examining the application of person-fit indices in the CAT environment. The l_z and ECI4$_z$ indices were selected for comparison. Using information from these indices, a new method was developed. All three indices showed little power to detect the use of memorization. Some possibilities for altering a test when the model becomes inappropriate for an examinee are also discussed. *Index terms: aberrancy detection, appropriateness measurement, item memorization, item response theory, person fit.*

Appropriateness/person-fit indices based on item response theory (IRT) are designed to detect response patterns that indicate that a given IRT model is inappropriate for an examinee even though the model might be appropriate for a group of examinees. The model may be inappropriate for a person for a number of reasons. First, examinees might answer items at random, not using their underlying abilities because of a "warm-up" effect (Wang & Wingersky, 1992). Second, examinees might skip an item on a paper-and-pencil test without skipping the corresponding item on the answer sheet or vice-versa. The IRT model assumes that the examinees answered the remaining items based on their underlying abilities. Therefore, these examinees will have spuriously low ability estimates (Levine & Rubin, 1979). A spuriously low score may also be produced when examinees turn easy items into difficult items. These examinees create difficulty in the items that was not in the test design and thus answer incorrectly (Hoffman, 1978). Finally, an IRT model is inappropriate for examinees who copy some (but not all) of the answers from a neighbor's test. In this situation, the ability being measured is not the examinee's but a combination of the neighbor's ability, the accuracy of the examinee's copying, and the examinee's ability. The score attained might then be spuriously high. In the case of test preview or memorization of some of the test items, the model is also inappropriate and may result in inflated test scores.

In computerized adaptive testing (CAT), the issue of memorizing items is a major concern. If the item bank is small and does not include many difficult items, the examinee with prior knowledge of some of the more difficult items might have an advantage. Due to the adaptive nature of CAT, memorizing those items most frequently exposed overall would not inflate scores as much as memorizing the more difficult items. Because CAT uses an examinee's performance on test items to select the next item, knowing answers to the more difficult items should help an examinee route into more memorized difficult items. Answering the more difficult items correctly will seriously inflate test scores. This class of deviations from the assumptions of the IRT model, which arises

Applied Psychological Measurement, Vol. 23 No. 2, June 1999, 147–160

from a lack of security of some items, might be able to be detected by person-fit indices (PFIs). If memorization is perfect, then the examinee's responses are all correct and no PFI will determine that the score has been obtained by memorization. However, in most situations, examinees will not have perfect success in memorization, and an index might be useful for gauging a test's security.

For example, a high ability examinee might take a CAT exam and memorize specific items. Later, these items might be shared with future examinees who use this information to route into similar memorized items. If successful, these examinees will inflate their test scores. These individuals include those with access to some of the more difficult items or those items most frequently exposed to the top scorers. This study was designed to evaluate several PFIs for detecting examinees who used memorized information to inflate their test scores.

IRT-Based PFIs

Many IRT-based PFIs have been developed, e.g., l_o (Levine & Rubin, 1979), l_z (Drasgow, Levine, & Williams, 1985), and ECI1_z, ECI2_z, ECI4_z, and ECI6_z (Tatsuoka, 1984; see Meijer, 1996, for a review). Because prior research suggested that l_z and ECI4_z were the most useful of these indices, they were used in this study.

The l_z index. l_z is a standardized function of the maximum of the likelihood function (l_o), and is defined as

$$l_z = \frac{\ln\left[L(\hat{\theta})\right] - \mathrm{E}\left\{\ln\left[L(\hat{\theta})\right]\right\}}{\sqrt{\mathrm{Var}\left\{\ln\left[L(\hat{\theta})\right]\right\}}}\ , \tag{1}$$

where the logarithm of the likelihood is

$$\ln\left[L(\hat{\theta})\right] = \sum_{i=1}^{n}\left\{u_i \ln\left[P_i(\hat{\theta})\right] + (1-u_i)\ln\left[1-P_i(\hat{\theta})\right]\right\}\ , \tag{2}$$

the expected value is

$$\mathrm{E}\left\{\ln\left[L(\hat{\theta})\right]\right\} = \sum_{i=1}^{n}\left\{P_i(\hat{\theta})\ln\left[P_i(\hat{\theta})\right] + \left[1-P_i(\hat{\theta})\right]\ln\left[1-P_i(\hat{\theta})\right]\right\}\ , \tag{3}$$

and the variance is

$$\mathrm{Var}\left\{\ln\left[L(\hat{\theta})\right]\right\} = \sum_{i=1}^{n}\left\{P_i(\hat{\theta})\left[1-P_i(\hat{\theta})\right]\left\{\ln\left[P_i(\hat{\theta})\right] \Big/ \left[1-P_i(\hat{\theta})\right]\right\}^2\right\}\ , \tag{4}$$

where

i indexes the item ($i = 1, \ldots, n$),
θ is the continuous latent trait,
u is a response to an item in the test (1 = correct, 0 = incorrect),
$P(\hat{\theta})$ is the probability of a correct item response for a given θ based on the model, and
$\hat{\theta}$ is the maximum likelihood estimate of θ.

Large negative values of l_z indicate misfit or unlikely response patterns. Large positive values indicate overfit. For this study, the negative of l_z ($-l_z$) was used to simplify comparison with other indices, i.e., negative values became positive and vice-versa.

If the examinee responds according to the IRT model, l_z has a sampling distribution that is asymptotically normal with a mean of 0 and standard deviation (SD) of 1. In this case, the standardization allows examination of person fit for examinees tested on different items and with different θ estimates. It lessens the degree to which person fit will be confounded with θ (Drasgow et al., 1985). If the examinee does not respond according to the IRT model, l_z might not have a similar distribution.

There is justifiable concern about the use of l_z. Some research has shown that l_z is most efficient at detecting non-model-fitting response patterns when the test has items of varied difficulty and small lower asymptote parameters (Reise & Due, 1991). However, Drasgow & Levine (1986) noted that l_z performed satisfactorily with detection rates approximately 65% of the optimal for response patterns that indicate cheating behavior. Nering (1996) found l_z superior to ECI4_z, especially when item responses were misfitting within the first five items administered.

The ECI4$_z$ index. ECI4_z detects response patterns when item responses do not agree with the modeled probability of answering them correctly. This index is based on the relationship between modeled performance and observed performance given the items' difficulties. This statistic is calculated as

$$\text{ECI4}_z = \frac{\sum_{i=1}^{n} \left\{ \left[P_i(\hat{\theta}) - u_i \right] \left[P_i(\hat{\theta}) - \overline{P}(\hat{\theta}) \right] \right\}}{\sqrt{\sum \left\{ \left\{ P_i(\hat{\theta}) \left[1 - P_i(\hat{\theta}) \right] \right\} \left[P_i(\hat{\theta}) - \overline{P}(\hat{\theta}) \right]^2 \right\}}} \tag{5}$$

where $\overline{P}(\hat{\theta})$ is defined as the examinee's mean $P(\hat{\theta})$ for the set of items administered.

ECI4_z, like l_z, is standardized with a mean of 0 and SD of 1 under the null hypothesis. Unlikely misfitting patterns cause ECI4_z to become large. Large positive values may result from prior knowledge of some difficult items without knowledge of easier items. Conversely, large negative values may result from an examinee correctly answering more easy items and fewer difficult items than would be expected on the basis of the IRT model.

A New PFI

This index, Z_c, was designed to detect response patterns that may result when an examinee has memorized some of the items. It is an extension of ECI4_z. Z_c separates test items into three categories: easy, medium, and difficult. These categories are based on threshold/difficulty (b) estimates. The highest one-third of the b values were classified as difficult items; the lowest third of the bs were the easy items, and the middle third were medium difficulty items. The residual performance is computed for each item as the difference between the probability of correctly answering the item and the scored (1-0) response:

$$Z_c = \frac{\overline{\text{Easy}\left[P(\hat{\theta}) - u \right]} - \overline{\text{Difficult}\left[P(\hat{\theta}) - u \right]}}{\sqrt{\left\{ \sum_{\text{Easy}} \left\{ P(\hat{\theta}) \left[1 - P(\hat{\theta}) \right] \right\} \Big/ n^2_{\text{Easy}} \right\} + \left\{ \left\{ \sum_{\text{Difficult}} \left\{ P(\hat{\theta}) \left[1 - P(\hat{\theta}) \right] \right\} \right\} \Big/ n^2_{\text{Difficult}} \right\}}}, \tag{6}$$

where

$\overline{\text{Easy}[P(\hat{\theta}) - u]}$ is the mean residual for the easy items administered,

$\overline{\text{Difficult}[P(\hat{\theta}) - u]}$ is the mean residual for the difficult items administered, and

n_{Easy} and $n_{\text{Difficult}}$ are, respectively, the number of easy and difficult items administered.

Z_c is thus a function of the average of the residuals for the easy items minus the average of the residuals for the difficult items. If Z_c is positive, then the examinee did not answer easy items correctly but answered difficult items correctly, indicating a misfit with the underlying model.

One drawback of the CAT environment when using this index is the need for two ranges of items. If an examinee does not receive at least one easy item and one difficult item, the index cannot be calculated. For a better estimate, several items from each category are necessary. There are several possible solutions to this problem. One solution is to design a CAT to administer at least one item from each category. A second solution is to classify all the items in the bank as easy or difficult, eliminating the middle category. A third solution is to design a CAT algorithm that administers at least one easy item to each examinee. This would allow computation of Z_c for those examinees who have a greater chance of inflating their test scores, namely those receiving many difficult items. Administering at least one difficult item to every examinee might not be necessary because there is little threat of score inflation if an examinee receives only easy items. Therefore, not being able to compute Z_c in these "received no difficult items" cases may not be a concern, and a value of Z_c = 0 could be assigned.

Analytical Comparisons of the Indices

Because these indices were derived using slightly different philosophies, they reflect misfit using different weighting systems. Previous studies have found a large negative correlation between l_z and ECI4_z (e.g., Birenbaum, 1985; Harnisch & Tatsuoka, 1983). The differences may be investigated using the structure of their numerators (the denominators are only for standardization). When the numerator of $-l_z$ in Equation 1 is written as

$$\sum_{i=1}^{n} \left\{ \left[P_i(\hat{\theta}) - u_i \right] \ln \left\{ P_i(\hat{\theta}) \Big/ \left[1 - P_i(\hat{\theta}) \right] \right\} \right\} , \tag{7}$$

it closely resembles the numerator in ECI4_z (Equation 5). For ECI4_z, the second term in the numerator sum weights each item's residual performance by its relative difficulty compared to the average item difficulty, $\overline{P}(\hat{\theta})$, for all items administered to an examinee. This index adjusts for the overall difficulty in the items administered; the sum of these residuals is 0. More weight is given to residuals for items that are farther from the average difficulty. Weights for each item set are linearly related to $\overline{P}(\hat{\theta})$. In the $-l_z$ computation, the second term in the numerator is the weight for each item's residual performance. This weight is a constant function—the logit of the probability of a correct response—and is not dependent on the set of items administered. It does not have the restriction that the weights sum to 0. Therefore, the weights for $-l_z$ may all be the same sign. Overall, both indices are influenced by $\hat{\theta}$. A graphical display of the standardized weights for an easy item set and a more difficult item set using $-l_z$ and ECI4_z is shown in Figure 1.

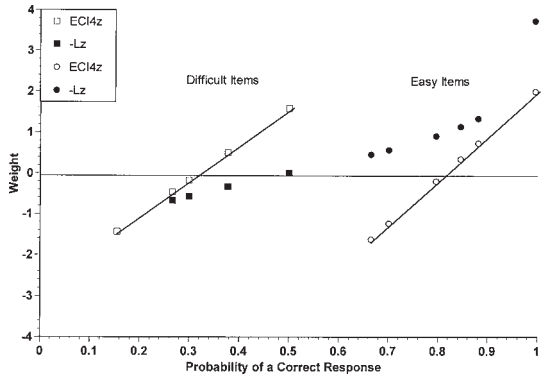
Z_c does not weight the residuals in terms of the probability of a correct response. This index uses the relative difficulty as compared with all other items in the bank to assign the relative weights. Therefore, unlike $-l_z$ and ECI4_z, the relative weights are not a function of the probability of a correct response and are only indirectly influenced by $\hat{\theta}$.

Method

Data Simulation

Two simulations were based on an operational Graduate Record Examination Quantitative (GRE-Q) CAT bank. In the first simulation, the 50 most frequently exposed items (in a bank of 348)

Figure 1
Standardized Weights for ECI4$_z$ and $-l_z$ Computed for Two Item Sets



for the upper 5% of the scorers were assumed to be memorized. Item parameters from the 3-parameter logistic IRT model (3PLM) for these items are listed in Table 1. Although these items were administered to the high scorers, they were not the most difficult items. Two of the memorized items had b values below the average b ($-.02$) and the median b (.11) for the CAT bank used. The average b for the memorized items was 1.26, the average discrimination (a) was 1.27, and the average lower asymptote (c) was .13. For the entire bank, the average a was .91 and the average c was .13. When a simulee was administered one of the 50 memorized items, a correct response was automatically given in the CAT simulation. The simulees were generated to be successful at memorization to produce a worst case for test security and, therefore, a good case for comparing these indices. The 3PLM with operational item parameter estimates was used to generate a response when one of the 298 remaining items was administered. In the second simulation, the null case, none of the items was assumed to be memorized. Therefore, the IRT model was used to generate all item responses.

Each simulation generated 1,650 response patterns using the operational GRE-Q CAT algorithm (Stocking & Swanson, 1993). The θs used were from a discrete uniform distribution containing 150 simulees at 11 θ values selected to correspond to the operational test's score range. Table 2 shows the relationship between estimated number-correct (ENC) values, θs, and population weights. ENCs refer to a reference test that was used for score reporting. The population weights represent an estimated distribution of the θs for an operational administration of the test and were used in some analyses to permit comparisons more representative of an operational distribution.

Twenty-eight items from a bank of 348 were administered to each simulee. For each simulee, the three indices were calculated based on the responses to the first 10 items administered and the responses to all 28 items.

Calculation of Z_c

For the Z_c calculations, the bank of 348 GRE-Q items was divided into three categories by the value of b. The one-third with the highest bs were classified as difficult items. The lowest one-third, those with the smallest bs, were designated as the easy items.

In the 10-item calculation, Z_c was not computed for 365 of the simulees in the memorized simulation: 95 did not receive at least one easy item and 270 did not receive at least one difficult item. In the 28-item calculation, the number of examinees for whom Z_c was not computed decreased to 317 (60 and 257, respectively). For the null case simulation (without memorization), the number

Table 1
IRT Item Parameters for Memorized Items

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	1.630	−.908	.139	26	1.840	1.259	.326
2	1.027	−.456	.070	27	1.158	1.265	.057
3	.910	.485	.341	28	1.457	1.266	.280
4	1.729	.791	.242	29	1.285	1.291	.081
5	1.578	.851	.134	30	1.065	1.299	.065
6	.433	.855	0.000	31	1.037	1.357	.082
7	1.777	.871	.150	32	1.235	1.374	.097
8	1.011	.902	.031	33	1.248	1.424	.124
9	1.274	.978	.203	34	1.413	1.442	.125
10	1.276	1.019	.204	35	1.199	1.446	.071
11	1.298	1.033	.055	36	1.469	1.452	.230
12	1.144	1.072	.123	37	.895	1.481	.112
13	.754	1.083	.039	38	1.125	1.490	.044
14	1.357	1.094	.201	39	.753	1.496	.047
15	1.508	1.112	.182	40	1.233	1.636	.162
16	1.199	1.172	.293	41	1.007	1.666	.130
17	.701	1.182	.170	42	.799	1.776	.094
18	1.840	1.194	.169	43	1.119	1.817	.092
19	1.096	1.201	.161	44	1.579	1.820	.147
20	1.528	1.204	.086	45	1.777	1.830	.094
21	1.766	1.209	.256	46	.726	1.878	.061
22	1.306	1.211	.169	47	1.130	1.971	.044
23	1.489	1.233	.052	48	1.409	2.073	.065
24	1.408	1.238	.165	49	1.567	2.091	.112
25	1.627	1.245	.123	50	1.169	2.182	.157

of cases for which Z_c could not be computed increased. In the 10-item calculation, Z_c was not computed for 565 simulees (80 did not receive an easy item and 485 did not receive a difficult item) out of 1,650. This number decreased to 481 (54 and 427, respectively) in the 28-item calculation. A Z_c value of 0 was assigned for those simulees for which Z_c could not be computed.

Table 2
ENC, θ, and Corresponding Population Weights

ENC	θ	Weight
10	−3.8394	.001442
15	−2.1841	.029116
20	−1.3811	.100307
25	−.8118	.158306
30	−.3482	.171876
35	.0534	.154741
40	.4271	.125484
45	.8074	.106487
50	1.2419	.094023
55	1.8824	.054866
59	3.5462	.003353

Problems in θ Estimation

Initial results showed that after 10 items, 1,015 of the 1,319 simulees in the Memorized group had attained perfect scores and $\hat{\theta} = \infty$. Because of this problem, there is no PFI that will distinguish those simulees receiving perfect scores based on θ from those that received perfect scores by memorization. Therefore, the 10-item calculation was not investigated further.

For the 28-item computation, there were 260 simulees with $\hat{\theta} = \infty$ in the Memorized group and 73 in the Null group. A value of 0 was assigned for these simulees for all three indices to signify perfect or null fit. Reise (1995) used a similar rule for investigating l_z. For those simulees with $\hat{\theta} = -\infty$, $P(\hat{\theta})$ was set to the value of the lower asymptote for each item unless the lower asymptote was 0. In the latter case, $P(\hat{\theta} = -6)$ was used. This rule was applied for 44 simulees, all from the Null group. Eight of these had answered all items incorrectly.

Results

Memorization Success Rates

Evidence of two categories of memorization was found in the first simulation: 1,319 simulees that received 16 or more memorized items and 326 that received 4 or fewer. Five simulees received between 5 and 15 memorized items. 41% (679) of the 1,650 simulees received a memorized item as the first item. Of the simulees that received a memorized item as their first item, 96% were given a memorized item for Item 2; over 50% of these 679 simulees were administered memorized items for each of the first 10 items. Thus, even with the exposure control methods used in the CAT item selection algorithm, these examinees received many of the memorized items. If the memorized items had been selected randomly from the bank, this pattern would have been very unlikely. However, these memorized items were those most frequently exposed to the highest 5% of examinees and were, therefore, some of the more difficult items.

Only those simulees that received 16 or more memorized items were included in the Memorized group for the comparison studies. This group is the larger sample of memorizers and contains those simulees that received many items. Also, only 80 of these were assigned a value of 0 for Z_c because it could not be computed. Table 3 shows average test score inflation for the Memorized and Null groups. The Memorized group's average test score using the population weights given in Table 2 was 58.4. Approximately 43% of those simulees with a low ENC of 10 received 16 or more memorized items. These simulees averaged 57.8 (out of 60 possible points) for their final test score by using memorized information. These examinees should be detected by PFIs.

Detection Rates

Cutoff values for PFIs, e.g., 1.65 for a nominal one-tailed $\alpha = .05$ error rate, have been suggested based on a normal distribution (Reise & Due, 1991). Several cut values were investigated here. Cut 1 was the empirical one-tailed $\alpha = .05$ error rate cut point. Cut 2 was 1.65. Cut 3 was 2.58, corresponding to nominal $\alpha < .005$. Cut 4 was a more conservative cutoff value of 3.3, corresponding to nominal $\alpha < .0005$; this value reflects concerns for the seriousness of a Type 1 error in the context of the intended application. Several characteristics of the $-l_z$, ECI4_z, and Z_c indices became apparent, as shown in Figure 2, which compares the indices for the Null and Memorized groups.

Except for those simulees at the highest ENCs, the Memorized group had a mean $-l_z$ value (Figure 2a) that was consistently higher than expected under the null hypothesis, as shown by a positive value that indicated misfit due to prior knowledge. The $-l_z$ index maintained a difference in the mean value between the Memorized and Null groups for simulees at lower ENCs. However, the average $-l_z$ value for the simulees with many items memorized remained well below any of

Table 3
Mean and SD of Test Score Inflation
(Estimated Minus True Test Score)

	Null			Memorized		
ENC	Mean	SD	*n*	Mean	SD	*n*
10	−.5	1.4	150	47.8	.6	64
15	.2	2.0	150	42.8	.7	80
20	−.1	2.9	150	37.9	.8	90
25	.2	3.1	150	33.0	.7	110
30	.0	3.3	150	28.0	.8	121
35	.2	3.1	150	23.0	.7	120
40	−.4	3.6	150	18.2	.8	135
45	.3	2.9	150	13.4	.8	149
50	−.1	2.0	150	8.6	.9	150
55	.0	1.4	150	4.2	.8	150
59	.3	.8	150	.9	.4	150

the cut points. Also, at the higher ENCs, the Null mean approached the Memorized mean; a very low percentage of the memorizers would be detected under these circumstances.

Figure 2b shows that results for $ECI4_z$ were similar to those of $-l_z$. For the Null group, the average values for $ECI4_z$ approached the Memorized group at higher ENCs. When the Memorized group answered all 28 items, their ENC estimates were so inflated and their memorization so successful that $ECI4_z$ did not detect them well. For example, 131 of the 150 Memorizers at ENC = 59 had response patterns of all 1s and $\hat{\theta} = \infty$. These were assigned a value of 0 for the three indices. Many of the Memorizers successfully inflated their scores and were much more difficult to detect using these indices at higher ENCs. They behaved much like the higher θ simulees in the Null group.

Figure 2c shows the results for Z_c. The mean for the Memorized group did not exceed any of the cut points. The group means for those simulees with many memorized items and no memorized items maintained some distance across the lower and middle ENC categories. At higher ENCs, mean Z_c for the Memorized group was essentially the same as for the Null group.

Table 4 compares the proportion detected, maximum, minimum, average, and SDs for the three indices. Population weights from Table 2 were used to compute these values. Z_c showed the largest difference between the Memorized and Null groups. Each index had success in not detecting many simulees from the Null group. The SDs were approximately .2, except for Z_c's larger SD when computed for the Memorized group. None of the indices appeared to be well-standardized for the Null case, with means less than 0 and SDs less than 1. These findings support those of Reise (1995). He found that l_z's null distribution had a reduced variance when θ was estimated. For the empirical cut point (Cut 1), $ECI4_z$ detected more of the Memorized group (15.9%) than the other indices. However, Z_c out-performed the other two indices at all of the other cut points.

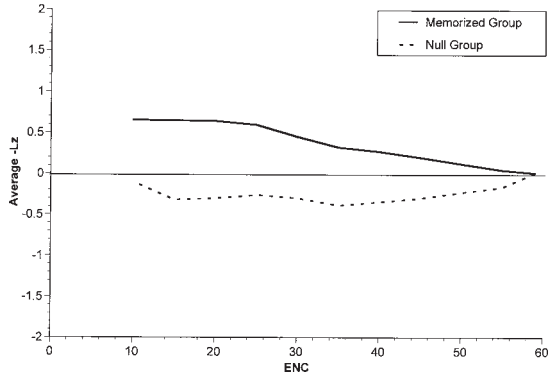
Z_c had the most extreme values, with a maximum value of 11.47 associated with a case in which the simulee received 26 memorized items out of the 28 items administered. This simulee had an ENC of 25 and an estimated test score of 58.9. The simulee answered 27 of the 28 items correctly. The $ECI4_z$ value for this simulee was 1.17; the $-l_z$ value was 1.27. Overall, none of the indices showed great power at detecting the use of memorization.

Correlation Analysis

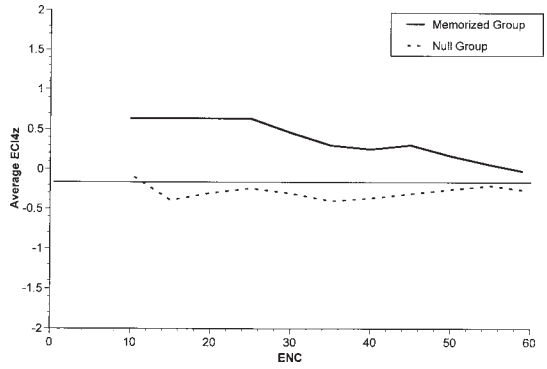
Correlations among $-l_z$, $ECI4_z$, and Z_c (Table 5) confirmed that the three indices behaved differently from each other in the CAT environment. Z_c's low correlations with $-l_z$ and $ECI4_z$ were

Figure 2
Person-Fit Index Value by ENC for 28 Items for Memorized and Null Groups

a. Average $-l_z$



b. Average $ECI4_z$



c. Average Z_c

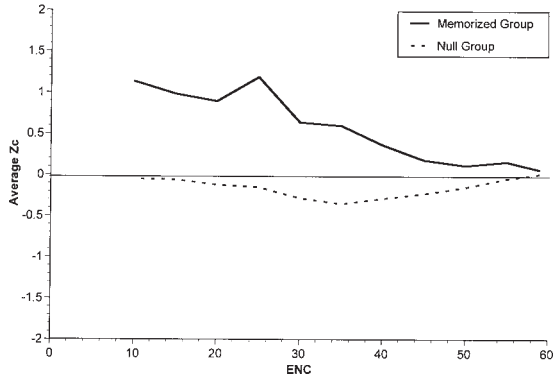


Table 4
Proportion of Simulees Identified as Misfitting Using Four Cut Points, and Descriptive Statistics for $-l_z$, ECI4$_z$, and Z_c in the Null and Memorized Groups

	$-l_z$		ECI4$_z$		Z_c	
Value	Null	Memorized	Null	Memorized	Null	Memorized
Cut 1	.050	.087	.050	.159	.050	.091
Cut 2	.014	.019	.017	.054	.013	.084
Cut 3	.002	0.000	.002	0.000	.001	.084
Cut 4	.0011	0.0000	.0014	0.0000	.0001	.0830
Max	4.10	2.55	4.18	2.40	6.12	11.47
Min	−2.77	−.46	−2.88	−1.57	−1.92	−.62
Mean	−.30	.34	−.31	.36	−.22	.54
SD	.2	.2	.3	.2	.2	.5

consistent with the fact that it gathered information about misfit using a different weighting system. The correlation between $-l_z$ and ECI4$_z$ (.967) was high for the Null group as shown by the upper triangle in Table 5. The correlation was relatively high for the Memorized group (.703), but not as high as reported in previous studies (e.g., Birenbaum, 1985; Harnisch & Tatsuoka, 1983).

Table 5
Weighted Correlations Among the Three Indices for the Null Group (N = 1,650; Upper Triangle) and the Memorized Group (N = 1,319; Lower Triangle)

Index	$-l_z$	ECI4$_z$	Z_c
$-l_z$	—	.967	.524
ECI4$_z$	.703	—	.483
Z_c	.560	.315	—

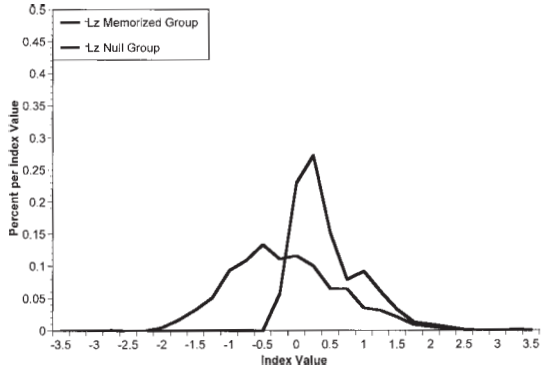
Frequency Distributions

Figure 3 shows the empirical relative frequency distributions for the three indices. To obtain these distributions, the discrete uniform θ distribution was weighted by population weights to depict a sample more representative of an operational θ distribution. For each index, the relative proportion at each index value is plotted for the Memorized and Null groups. For an index to show good discrimination there should be relatively little or no overlap between the Memorized and Null groups.

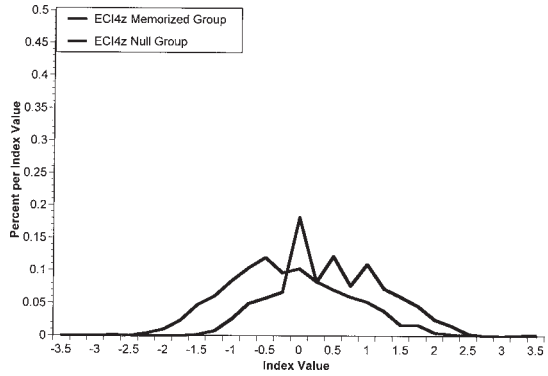
In Figure 3a, the distributions for $-l_z$ almost completely overlap. The Memorized group shows a slight positive shift, but most of its area is near 0. In Figure 3b, the ECI4$_z$ shows more variability for the Memorized group, but still considerable overlap in the tails between the two distributions. Less of the area of the distribution for the Memorized group is above 0 when compared to $-l_z$'s distributions. Z_c (Figure 3c) was slightly better, with a shorter positive tail for the Null group and more area in the positive tail for the Memorized group. For all three distributions, much of the area was near or at 0 for the Memorized group. None of the distributions were well standardized for the Null group. Some of the area at 0 represents the simulees that attained response patterns of all 1s and were assigned a value of 0. Z_c had more area at 0 for those that did not receive at least one easy/difficult item and were assigned a value of 0 for Z_c. Z_c had more area at the 3.5 or above value than for any other index. The values in this area ranged from 3.5 to 11.47 and were relatively evenly spread.

Figure 3
Distribution of Person-Fit Indices for Memorized and Null Groups for 28 Items

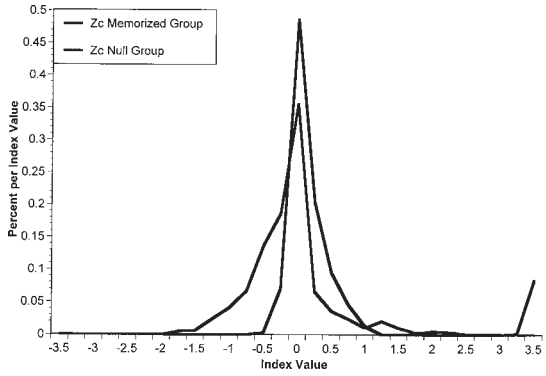
a. $-l_z$



b. $ECI4_z$



c. Z_c



ROC Curves

Marginal probability ROC curves (Green & Swets, 1966) offer an additional evaluation of PFIs as discussed by Hulin, Drasgow, & Parsons (1983). The points on an ROC curve represent the ratio of false alarms to hits. Empirical ROC curves were calculated for $-l_z$, ECI4_z, and Z_c. For each point on the ROC curve, the value on the horizontal axis is the proportion of those from the Null group "detected" by an index using a particular cutoff value (false-alarm rate) and the value on the vertical axis is the detected proportion in the Memorized group (hit rate). These proportions were weighted using the population weights given in Table 2.

Figure 4a contains the ROC curve for 28 items. This figure shows that approximately 90% of simulees in the Memorized group were detected when the false-alarm rate was 42% using $-l_z$. The hit rates for the other two indices at this false-alarm rate were approximately 75% for ECI4_z and 65% for Z_c. At a 16% false-alarm rate, Z_c had slightly more power to detect, with a hit rate of 40%. At this point, $-l_z$ had the least power of the three indices. (An index operating only by chance would produce a curve on the diagonal.)

Figure 4b shows the lower left-hand corner of Figure 4a. This graph has been magnified to show the ROC curves for false-alarm rates up to 10%. These are more representative of rates useful for operational decision making. For example, in Figure 4b, for a 3% false-alarm rate, over 8% of all simulees in the Memorized group were detected using Z_c, whereas $-l_z$ detected 4% and ECI4_z detected 10%. Z_c had more power to detect when the false-alarm rate was less than 2.5% and ECI4_z had more power between 2.5% and 10% false-alarm rates.

Discussion

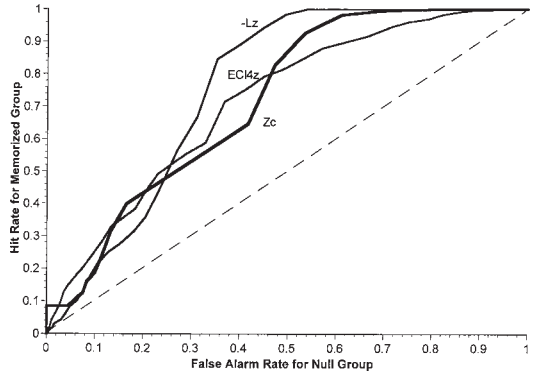
In agreement with Nering (1997), it was found that $-l_z$ and ECI4_z were not distributed as expected within the context of CAT. Because of the item selection algorithm implemented in CAT, these results also support those of Reise & Due (1991) who found it very problematic to detect misfitting response patterns for tests with limited ranges of item difficulty. Although the results are specific to the test and item bank selected, the distributions of the indices in this study were nonnormal and, overall, showed little power to detect memorizers in CAT.

As the technology for detecting memorizers improves, the question of what to do when an examinee is suspected of memorizing becomes important. The first priority should be to salvage the test administration. One strategy is to continue testing in the CAT environment using highly secure items with known characteristics. These may be items that have been calibrated in a selected and secure field test sample. Items in this category will have very low exposure rates and will be very expensive to produce.

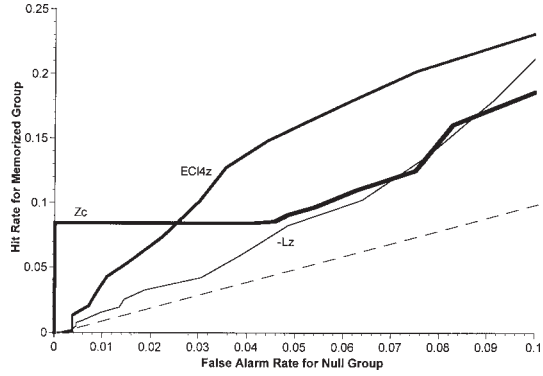
Another, less expensive, strategy is to administer a few easy items that the examinee will probably answer correctly. For example, if the CAT administration has estimated the examinee's θ at 1.6, the CAT may give a "suspect" of prior knowledge a few items at the -1.0 θ range. If the examinee is truly knowledgeable, he or she should consistently answer these items correctly. If the examinee has attained his or her score merely through prior item knowledge or memorization, some of these items may prove difficult. For this strategy to work, however, the memorizer must not have a genuinely high θ level. If the memorizer does, the easy items given would be answered correctly. It is assumed that high θ memorizers have little to gain by memorizing and thus are not those most likely to do so. After an analysis of the examinee, the CAT could continue to be administered if the responses indicated that the examinee was responding consistently.

A more expensive strategy is to stop the CAT mode of testing and continue testing using a secure linear form. This test can be administered by computer so the examinee is unaware of the change. Possible disadvantages of this strategy are increased test time and a less accurate θ estimate.

Figure 4
ROC Curves for Z_c, ECI4$_z$, and $-l_z$ for 28 Items
a. False-Alarm Rate 0.0 to 1.0



b. False-Alarm Rate 0.0. to .10



Future Research

With the use of on-line calibration in the CAT environment, the next step may be to consider item fit. In addition to detecting examinees as misfitting, items may also be detected as misfitting. Items that have been memorized, for example, will no longer fit the IRT model and will perform aberrantly. Once detected, an item may be replaced to maintain test security.

Along with the use of a PFI to monitor the use of examinee memorization to attain higher scores comes the need for differential analyses of these indices. Do they give different results depending on gender or ethnic classification? Are certain misfitting patterns due to cultural differences? If these differences are present, how may they be used to improve testing and thus make testing more fair?

References

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45,* 523–534.

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized

indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

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

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. Hambleton, (Ed.) *Applications of item response theory.* Vancouver BC: Educational Research Institute of British Columbia.

Hoffman, B. (1978). *Tyranny of testing.* Westport CT: Greenwood.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood IL: Dow Jones-Irwin.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Meijer, R. R. (Ed.). (1996). Person-fit research: Theory and applications [Special issue]. *Applied Measurement in Education, 9*(1), 9–18.

Nering, M. L. (1996). *The effects of person misfit in computerized adaptive testing.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115–127.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19,* 213–229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217–226.

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

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Wang, M., & Wingersky, M. (1992, April). *Incorporating post-administration item response revision into a CAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Acknowledgments

The authors express their appreciation to David Thissen, two anonymous reviewers, and the editor for their insightful comments on earlier versions of this manuscript. The authors also thank Martha Stocking for her help in generating the computer simulations. This work was completed while the first author was an ETS Harold Gulliksen Psychometric Fellow at the University of North Carolina at Chapel Hill.

Author's Address

Send requests for reprints or further information to Lori D. McLeod, Law School Admission Council, 661 Penn Street, Newtown PA, 18940-0040, U.S.A. Email: lmcleod@lsac.org.