Detecting Inappropriate Test Scores with Optimal and Practical Appropriateness Indices

Fritz Drasgow, Michael V. Levine, and Mary E. McLaughlin
University of Illinois

Several statistics have been proposed as quantitative indices of the appropriateness of a test score as a measure of ability. Two criteria have been used to evaluate such indices in previous research. The first criterion, standardization, refers to the extent to which the conditional distributions of an index, given ability, are invariant across ability levels. The second criterion, relative power, refers to indices' relative effectiveness for detecting inappropriate test scores. In this paper the effectiveness of nine appropriateness indices is determined in an absolute sense by comparing them to *optimal indices;* an optimal index is the most powerful index for a particular form of aberrance that can be computed from item responses. Three indices were found to provide nearly optimal rates of detection of very low ability response patterns modified to simulate cheating, as well as very high ability response patterns modified to simulate spuriously low responding. Optimal indices had detection rates from 50% to 200% higher than any other index when average ability response vectors were manipulated to appear spuriously high and spuriously low.

Some examinees' scores on a multiple-choice test may fail to provide valid measures of the trait measured by the test. Examinees' scores may be *spuriously high* if they copy answers from more talented neighbors or if they have been given the answers to some questions. Examinees' scores can be *spuriously low* due to alignment errors (answering, say, the 10th item in the space provided for the 9th item, answering the 11th item in the space provided for the 10th item, etc.), language difficulties, atypical education, and unusually creative interpretations of normally easy items.

Detecting inappropriate test scores is very important in many situations. In academic admissions testing, spuriously high scores can lead to the enrollment of unqualified individuals in undergraduate, graduate, and professional programs. Such individuals would be expected to have higher rates of academic probation and dismissal than the qualified students. Furthermore, in selective academic programs, admitting a student on the basis of a spuriously high test score would typically cause a more deserving student to be denied admission. Individuals with spuriously low scores would be unlikely to be admitted to selective academic programs. While this may not harm the institution, the cost to the individual examinees can be high: Artificially low test scores may damage their self-image and thwart some of their career aspirations.

Inappropriate test scores can lead to serious problems in employment testing. For example, spuriously high test scores can result in the selection of individuals who are unqualified for jobs and training programs.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 1, March 1987, pp. 59–79

0146-6216/87/010059-21$2.30

Terminating unsuccessful employees is difficult and can be very expensive if contested in the courts. Individuals who fail to complete a management or vocational training program conducted by an organization (e.g., a military training school) may cost the organization many thousands of dollars.

Spuriously low scores can also cause serious difficulties in employment testing. The costs to individuals are similar to those in educational testing, but there can be a difference in the institutional point of view. For example, spuriously low scores on military aptitude tests (perhaps due to "deliberate failure") during a period of national mobilization might cause manpower shortages.

The goal of *appropriateness measurement* is to identify inappropriate test scores. In recent years, several methods for identifying inappropriate scores have been devised. In all approaches, response patterns are characterized in a way that permits a quantitative assessment of the degree to which an observed response vector is atypical. This quantitative measure is then used to classify response patterns into normal and aberrant categories.

In a series of studies, it has been found that simulated spuriously high response patterns and simulated spuriously low response patterns can be detected by appropriateness measurement. High detection rates have been obtained despite model misspecification, errors in item parameter estimates, and the inclusion of inappropriate response patterns in the test norming sample (Levine & Drasgow, 1982). Very high detection rates have been obtained when response patterns of low-ability examinees have been modified to simulate cheating and when response patterns of high-ability examinees have been modified to simulate spuriously low responding (Drasgow, Levine, & Williams, 1985).

It is relatively easy to propose new appropriateness indices; in the next section the 10 indices evaluated in this paper are presented. Evaluation of the relative merits of the various indices has been very difficult in previous research. Cliff's (1979) description of a related problem cogently summarizes the situation here as well: "Now the trouble is that the formulas multiply not just like rabbits, or even guppies, but rather like amoebae: by both fusion and conjugation, and there seemed to be no general principle to use in selecting from among them" (p. 388). Harnisch and Tatsuoka (1983), for example, correlated 14 different indices in order to see which pairs were more and less related, but this approach has limited value in determining which index is *best*. Moreover, this approach does not show which indices, if any, are good enough for operational use.

In the past, two criteria for evaluating appropriateness indices have been used. The first criterion is *standardization*. It refers to the extent to which the conditional distributions (given particular values of the latent trait) of an index are invariant across levels of the latent trait. There is little confounding between ability and measured appropriateness for a well-standardized index. Well-standardized indices have two attractive features. First, high rates of detection of aberrant response patterns by well-standardized indices in experimental studies of index effectiveness cannot be due merely to differences in ability distributions or number-correct distributions across normal (i.e., non-aberrant) and aberrant samples. In contrast, high detection rates obtained by poorly standardized indices may be due largely to differences in ability distributions. (This point is illustrated in a later section of this paper and in Table 2.) The second consequence of the independence of ability and measured appropriateness for a well-standardized index is that it is easy to use in practice because index scores for individuals with different standings on the latent trait can be compared directly. In contrast, scores on poorly standardized indices can be interpreted only in relation to their conditional distributions and, consequently, a single cutting score for classification into aberrant and normal groups is not possible. Because computation of the conditional distributions of an appropriateness index is usually time-consuming, the practical usefulness of a poorly standardized index is very limited.

The second criterion used to evaluate appropriateness indices is *relative power:* Given a particular rate of misclassification of normal response patterns as aberrant (Type I error rate), which of the indices

under consideration has the highest rate of correctly classifying aberrant response patterns as aberrant? If a well-standardized index has acceptable power, then it can be used in operational settings. Unfortunately, no unequivocal conclusion about the detectability of aberrance is possible if none of the indices under consideration has adequate power, because some other index that was not included in the comparisons might have acceptable power. In addition, even if an index is found to have adequate power for operational use, it is not known whether another index could be devised that is substantially superior to all known indices.

Levine and Drasgow (1984) recently addressed the difficult problem of evaluating the power of an appropriateness index. They introduced a general method for ascertaining the maximum rate of detection of a specified form of aberrance that can be achieved by any index. By means of a new numerical algorithm, Levine and Drasgow were able to apply the Neyman-Pearson Lemma to obtain an appropriateness index that is optimal in the sense that no other index computed from the item responses can achieve a higher detection rate (at each error rate) of the given form of aberrance. Drasgow and Levine (1986) demonstrated this approach by determining the absolute effectiveness of two versions of the standardized ℓ_0 index for detecting two particular types of aberrance. The absolute effectiveness of an index is determined by comparing its detection rate with the detection rate of the corresponding optimal index. This approach was used here to evaluate nine different appropriateness indices for their capacity to detect spuriously high and low response patterns on a long unidimensional power test, namely the Verbal section of the Scholastic Aptitude Test (SAT-V).

The Appropriateness Indices

Optimal Indices

Suppose a simple null hypothesis is to be tested against a simple alternative hypothesis. If the probability of a Type I error is α, then a *most powerful test* is one that minimizes the probability of a Type II error among the set of tests with a Type I error rate of α. The Neyman-Pearson Lemma asserts that maximum power is achieved by a likelihood ratio test. More specifically, let $L_N(\mathbf{x})$ and $L_A(\mathbf{x})$ denote the likelihoods of the data $\mathbf{x}$ under the null and alternative hypotheses, respectively. Then the Neyman-Pearson Lemma states that of all tests with a Type I error rate of α, none is more powerful than a test obtained from the likelihood ratio $L_A(\mathbf{x})/L_N(\mathbf{x})$.

Levine and Drasgow (1984) showed how the Neyman-Pearson Lemma, in the context of appropriateness measurement, could be used to construct most powerful tests and, consequently, optimal appropriateness indices. Suppose that local independence holds, $\mathbb{u} = (u_1, \ldots, u_n)$, and $P_i(u_i|\theta)$ is the probability of response u_i to item i by an examinee of ability θ under the null hypothesis that the response pattern is normal. Then the likelihood of a response vector $\mathbb{u}$ by an examinee of ability θ is

$$P_{\text{normal}}(\mathbb{u}|\theta) = \prod_{i=1}^{n} P_i(u_i|\theta) \quad . \tag{1}$$

If the ability density is $f(\theta)$, then using elementary probability

$$P_{\text{normal}}(\mathbb{u}) = \int P_{\text{normal}}(\mathbb{u}|\theta) f(\theta) d\theta \quad . \tag{2}$$

To apply the Neyman-Pearson Lemma it is necessary to compute $P_{\text{aberrant}}(\mathbb{u})$. This quantity can be obtained by carrying the conditioning-integrating argument one step further. For concreteness, suppose that the type of aberrance under consideration consists of m randomly selected items being modified by the spuriously low treatment. Let S_k denote a set indicating the kth way of selecting m of n items, of the $\binom{n}{m}$ ways possible; let $P_{\text{aberrant}}(\mathbb{u}|\theta, S_k)$ denote the probability of response pattern $\mathbb{u}$ for an examinee with

ability θ, when the items in S_k are subjected to the spuriously low treatment; and let $P(S_k)$ denote the probability of S_k, such that $P(S_k) = 1/\binom{n}{m}$. Then

$$P_{\text{aberrant}}(\mathbf{u}|\theta) = \sum_k P_{\text{aberrant}}(\mathbf{u}|\theta,S_k)P(S_k) \quad , \tag{3}$$

so that

$$P_{\text{aberrant}}(\mathbf{u}) = \int [\Sigma P_{\text{aberrant}}(\mathbf{u}|\theta,S_k)P(S_k)]f(\theta)d\theta \quad . \tag{4}$$

By taking advantage of the symmetry in $P_{\text{aberrant}}(\mathbf{u}|\theta,S_k)$, it is possible to obtain an efficient numerical algorithm for computing $P_{\text{aberrant}}(\mathbf{u}|\theta)$. A numerical 2uadrature formula can be used to evaluate the right-hand side of Equation 4 with an acceptable amount of computation. Details about these calculations and a theoretical treatment of the general problem are provided by Levine and Drasgow (1984).

Thus it is possible to compute the likelihood ratio

$$\text{LR} = P_{\text{aberrant}}(\mathbf{u})/P_{\text{normal}}(\mathbf{u}) \tag{5}$$

and test the simple null hypothesis that a response pattern is normal against the simple alternative hypothesis that the response pattern is aberrant. Due to the Neyman-Pearson Lemma, the LR statistic provides a most powerful test; consequently, when it is used as an appropriateness index, LR is as powerful as any index that can be computed from the item responses.

It should be noted that optimal indices are not intended for use in practical settings at present (the form of aberrance must be completely specified to use the Levine and Drasgow algorithm). Instead, optimal indices (1) provide benchmarks for evaluating practical indices, and (2) yield the highest detection rates that can be obtained from the item responses.

Practical Indices

The practical appropriateness indices examined are described below. These indices were selected for a variety of reasons, including effective performance in earlier studies, ease of computation, and wide use in field settings. In addition, two other indices were included because they had an appealing heuristic motivation, based on the curvature of the likelihood function, that had not been investigated previously.

One of the most important goals of the present research was to determine the extent to which the practical indices are less than optimal. From the Neyman-Pearson Lemma, it is known that the practical indices cannot achieve higher detection rates than the optimal indices. Prior to the present study, little else has been revealed concerning the power of many of the most important practical appropriateness indices. For example, the power of the indices proposed by Tatsuoka (1984), Sato (1975), and Wright (1977) was unknown.

Standardized ℓ_0. Let z_3 denote the standardized ℓ_0 index studied by Drasgow et al. (1985). It may be computed by the formula

$$z_3 = \frac{\ell_0 - M(\hat{\theta})}{[S(\hat{\theta})]^{1/2}} \quad . \tag{6}$$

In this formula, ℓ_0 is the logarithm of the three-parameter logistic likelihood function evaluated at the maximum likelihood estimate $\hat{\theta}$ of θ:

$$\ell_0 = \sum_{i=1}^{n} [u_i \log P_i(\hat{\theta}) + (1-u_i) \log Q_i(\hat{\theta})] \quad , \tag{7}$$

where u_i is the dichotomously scored (1 = correct, 0 = incorrect) item response for item i ($i = 1, 2, \ldots, n$),

$$Q_i(\theta) = 1 - P_i(\theta),$$
$$P_i(\theta) = \hat{c}_i + \frac{1 - \hat{c}_i}{1 + \exp[-D\hat{a}_i(\theta - \hat{b}_i)]}, \quad (8)$$
$$D = 1.702, \text{ and}$$

$\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are item parameter estimates.

The conditional expectation of ℓ_0, given $\theta = \hat{\theta}$, is

$$M(\hat{\theta}) = \sum_{i=1}^{n}[P_i(\hat{\theta}) \log P_i(\hat{\theta}) + Q_i(\hat{\theta}) \log Q_i(\hat{\theta})] \quad (9)$$

and its conditional variance is

$$S(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta})Q_i(\hat{\theta})\{\log[P_i(\hat{\theta})/Q_i(\hat{\theta})]\}^2 . \quad (10)$$

Justifications of these formulas can be found in Drasgow et al. (1985).

Fit statistics. Two fit statistics for the three-parameter logistic model were suggested by Rudner (1983) as generalizations of the Rasch model fit statistics used by Wright and his colleagues. The first is the mean squared standardized residual

$$\text{F1} = \frac{1}{n}\sum_{i=1}^{n}\frac{[u_i - P_i(\hat{\theta})]^2}{[P_i(\hat{\theta})Q_i(\hat{\theta})]} . \quad (11)$$

The other fit statistic is

$$\text{F2} = \frac{\sum_{i=1}^{n}[u_i - P_i(\hat{\theta})]^2}{\sum_{i=1}^{n}P_i(\hat{\theta})Q_i(\hat{\theta})} , \quad (12)$$

which Rudner found to be quite effective in some cases (see Rudner, 1983, p. 214 and p. 216, where W3 denotes an expression proportional to F2).

Likelihood function curvature statistics. Two indices that provide measures of the "flatness" of the likelihood function were also evaluated. These indices are based upon the notion that inappropriate responses will flatten the likelihood function near its maximum because no single value of θ will allow the item response model to provide a good fit to the response vector. Therefore, the likelihood function will not have a sharp maximum; instead it will be relatively flat.

The first curvature statistic is the normalized jackknife variance estimate. In order to compute this index, let $\hat{\theta}$ denote the three-parameter logistic maximum likelihood estimate of θ based on all n test items, and let $\hat{\theta}_{(j)}$ denote the estimate based on the $n - 1$ items remaining when item j is excluded. The *pseudo-values* (see, e.g., Mosteller & Tukey, 1968) are

$$\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}_{(j)}^* \qquad (j = 1,2, \ldots, n) . \quad (13)$$

The jackknife estimate of θ is then

$$\hat{\theta}^* = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j^* \quad (14)$$

and the jackknife estimate of its variance is

$$\text{Var}(\hat{\theta}^*) = \frac{\sum(\hat{\theta}_j^*)^2 - \frac{1}{n}(\sum\hat{\theta}_j^*)^2}{n(n-1)} . \quad (15)$$

The jackknife variance estimate is not a standardized appropriateness index; there is more Fisher information about θ in some ranges than in others, hence $\mathrm{Var}(\hat{\theta}^*)$ is expected to depend upon θ. Lord's (1980) formula for the information of the three-parameter logistic maximum likelihood estimate of θ,

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i(\theta)']^2}{P_i(\theta)Q_i(\theta)} \quad , \tag{16}$$

can be used to reduce this problem. Because the reciprocal of $I(\theta)$ is the asymptotic variance of $\hat{\theta}$, the jackknife estimate of variance can be approximately normalized by evaluating the information function at $\hat{\theta}$ and computing

$$\text{JK} = \mathrm{Var}(\hat{\theta}^*)I(\hat{\theta}) \quad . \tag{17}$$

It is possible to arrange the calculations for computing JK very efficiently. During pretesting it was found that one Newton-Raphson iteration was adequate to move from $\hat{\theta}$ to $\hat{\theta}_{(j)}$. Then, because the first and second derivatives of the log likelihood functions for the whole test are sums over n items, the first and second derivatives of the log likelihood functions for the tests containing $n - 1$ items can be obtained by single subtractions of already computed quantities. Consequently, all the pseudo-values, $\hat{\theta}^*$, and JK can be obtained with fewer arithmetic calculations than are required in a single Newton-Raphson iteration in the calculation of $\hat{\theta}$.

The second index based on the likelihood function's curvature compares the expected and observed curvatures. If the likelihood function is flatter for aberrant response patterns than for normal response patterns, then the observed information, defined as the negative of the second derivative of the log likelihood function at $\hat{\theta}$ given the response vector $\mathbf{u}$ (see Efron & Hinkley, 1978, p. 457), would be expected to be less than the information $I(\hat{\theta})$ given in Equation 16, which (given $\hat{\theta}$) does not depend upon $\mathbf{u}$. Thus, this index is the ratio of the observed and expected information

$$\text{O/E} = \frac{-\left.\dfrac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\hat{\theta}}}{I(\hat{\theta})} \quad , \tag{18}$$

where ℓ is the log likelihood

$$\ell = \sum_{i=1}^{n} [u_i \log P_i(\theta) + (1-u_i) \log Q_i(\theta)] \quad . \tag{19}$$

Item-option variance. Consider the subset of N_{ik} examinees in the test norming sample who selected option k to item i. It is easy to compute the mean number-correct score $\bar{X}_{ik}$ for these examinees. In this way it is possible to identify options to item i that are typically selected by high-ability examinees (e.g., the correct option) and options that are typically selected by lower-ability examinees. For spuriously high and low response patterns, inconsistencies in $\bar{X}_{ik}$ are expected: Sometimes options with low $\bar{X}_{ik}$ should be selected and sometimes options with high $\bar{X}_{ik}$ should be selected. For this reason the item-option variance

$$\text{IOV} = \mathrm{Var}(\bar{X}_{ik}) \tag{20}$$

was evaluated as a measure of appropriateness. Notice that this index is very easy to compute.

Caution indices. Three "caution indices" were also examined. The first is Sato's (1975) caution index S (see also Tatsuoka & Linn, 1983, but replace $y_{.j}$ with $P_{.j}$ for a simpler version of their Equation 1). It was included in this study because it is easy to compute and is widely used in Japan. To compute S, suppose that the n test items are ordered from easiest to most difficult on the basis of proportion correct ($\hat{p}_i$) in the test norming sample. Let

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i \tag{21}$$

be the mean proportion correct and suppose that an examinee answers k items correctly. If $\hat{\mathbf{p}}$ is a vector containing the $\hat{p}_i$ and $\mathbf{g}$ is a perfect Guttman response pattern with 1s as its first k elements and 0s for the next $n-k$ elements, then

$$S = 1 - \frac{\mathrm{Cov}(\mathbf{u},\hat{\mathbf{p}})}{\mathrm{Cov}(\mathbf{g},\hat{\mathbf{p}})} = 1 - \frac{\sum_{i=1}^{n} u_i(\hat{p}_i - \bar{p})}{\sum_{i=1}^{k}(\hat{p}_i - \bar{p})} \quad . \tag{22}$$

Note that the summation in the denominator of the rightmost expression is from 1 to k (i.e., over the k items with the smallest $\hat{p}_i$ values), not 1 to n.

Two indices that are related to Sato's caution index are the second and fourth standardized extended caution indices T2 and T4 presented by Tatsuoka (1984, p. 104). These two indices (of the four studied by Tatsuoka) were included because Harnisch and Tatsuoka (1983) found that they are not related (linearly or curvilinearly) to true score and, therefore, $\hat{\theta}$.

T2 and T4 can be computed relatively easily. Let $\hat{\theta}_j$ denote the three-parameter logistic maximum likelihood estimate of θ for the jth person in the test norming sample of N examinees, and let $P_{ij}(\hat{\theta}_j)$ be the probability of a correct response to item i by this person computed from Equation 8. Then define

$$G_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij}(\hat{\theta}_j) \tag{23}$$

and

$$\bar{G} = \frac{1}{n}\sum_{i=1}^{n} G_i \quad . \tag{24}$$

To compute T2 and T4 for an examinee in the normal sample or an aberrant sample, let

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i(\hat{\theta}) \quad . \tag{25}$$

Then

$$T2 = \frac{\sum[P_i(\hat{\theta}) - u_i][G_i - \bar{G}]}{\left[\sum P_i(\hat{\theta})Q_i(\hat{\theta})(G_i - \bar{G})^2\right]^{1/2}} \tag{26}$$

and

$$T4 = \frac{\sum[P_i(\hat{\theta}) - u_i][P_i(\hat{\theta}) - \bar{P}]}{\left\{\sum P_i(\hat{\theta})Q_i(\hat{\theta})[P_i(\hat{\theta}) - \bar{P}]^2\right\}^{1/2}} \quad . \tag{27}$$

It should be noted that Equations 22, 26, and 27 are generalizations of the original caution indices to the situation where item parameters are estimated in a test norming sample.

Standardization

The Problem

Measured appropriateness can be confounded with ability. Figure 2 in Drasgow et al. (1985), for example, shows a strong, nearly linear relation between estimated ability and an unstandardized index. A score of, say, -50 on this index at one ability level indicates a good fit of the model to a response vector, but the same index score at other ability levels indicates a very poor fit. Consequently, an observed difference between the distributions of index scores for normal and aberrant response vectors is not

unequivocal evidence of index effectiveness. Instead, it may simply reflect differences in ability or number-correct distributions. This problem does not occur if an appropriateness index is well-standardized, that is, if the conditional distributions (given θ) of the index are (approximately) equal across possible values of θ for non-aberrant examinees.

In practical applications of appropriateness measurement, it would be convenient if a single cutting score could be used to classify response patterns as aberrant or normal. If the conditional distributions of an index are not identical, then a score on a practical appropriateness index must be interpreted with respect to the associated conditional distribution. Consequently, it is not possible to use a single cutting score to classify response patterns as aberrant, nor is it possible to directly compare index scores of examinees with different abilities.

Little degradation would be expected in the performance of a well-standardized index if the ability distribution were to change abruptly. Such a change might be expected, for example, with the Armed Services Vocational Aptitude Battery examinee population in a period of national mobilization.

ROC Curves

If an index is properly standardized, its distribution will be nearly the same in subpopulations of normal examinees who differ in ability. Hence the index could *not* be used to distinguish between such groups. A standard, very general method for studying the extent to which some statistic can differentiate between groups is the receiver operating characteristic (ROC) curve. A ROC curve can be used to study index standardization in order to determine whether the index distinguishes between groups of normal examinees who differ in ability.

A ROC curve is obtained by specifying a cutting score t for an index and then computing

$x(t)$ = proportion of group X (say, normal, low-ability examinees) response patterns with index values less than t (assuming that small index values indicate aberrance);

$y(t)$ = proportion of group Y (say, normal, high-ability examinees) response patterns with index values less than t.

A ROC curve consists of the points $[x(t),y(t)]$ obtained for various values of t. The proportion $x(t)$ is called the false alarm rate and $y(t)$ is called the hit rate. A detailed example of the construction of a ROC curve is given by Hulin, Drasgow, and Parsons (1983, pp. 131–134).

An appropriateness index is well-standardized across two ability levels if the ROC curve computed for samples of normal examinees with these two abilities lies along the diagonal line $y = x$. Thus, diagonal ROC curves indicate positive results *in the context of index standardization*. In contrast, a ROC well above the diagonal should be obtained for a powerful appropriateness index when a sample of aberrant response patterns is compared to a sample of normal response patterns.

Method

Polychotomous item responses (five-option multiple-choice items with omitting allowed) were simulated using the histograms constructed by Levine and Drasgow (1983). They used the three-parameter logistic model to estimate the abilities of 49,470 examinees from the 85-item April 1975 administration of the SAT-V. Then the examinees were sorted into 25 groups on the basis of estimated ability. The 4th, 8th, ..., 96th percentile points of the normal (0,1) distribution were used as cutting scores when sorting examinees. The proportions of examinees choosing each option (treating items that were skipped or not reached as a single response category) were computed for each of the 25 ability groups. Probabilities of option choice were then computed by linear interpolation between category medians at the 2nd, 6th, ..., 98th percentile points from the normal (0,1) distribution.

Five samples of normal response patterns were generated by first sampling 3,000 numbers (θs) from the normal (0,1) distribution truncated to the $[-2.05,2.05]$ interval. (It was necessary to truncate the θ distribution because interpolation below the 2nd percentile or above the 98th percentile was not possible with the histograms.) Then low (-2.05 to -1.50), moderately low ($-.70$ to $-.55$), average ($-.05$ to .05), moderately high (.55 to .70), and high (1.49 to 2.05) θ samples of $N = 200$ each were formed.

Polychotomous response vectors were then generated for each θ value. For each item, the associated histogram was used to compute the conditional (given θ) probabilities of the six possible responses (treating non-response as the sixth response). A number was sampled from the uniform distribution on the unit interval and a simulated response was obtained by determining where the random number was located in the cumulative distribution corresponding to the conditional probabilities. Finally, each of the nine practical appropriateness indices was computed for each response vector in each sample. Then ROC curves were computed for each of the $\binom{5}{2} = 10$ possible pairs of samples and each of the nine appropriateness indices.

Results

Figures 1 through 3 present the results for the low-average, average-high, and low-high comparisons. The results for the other seven comparisons were consistent with the trends seen in these three figures; consequently they will not be presented. Furthermore, only the lower left quarter of the unit square is plotted because it is unlikely that anyone would set a cutting score that yielded a false alarm rate of more than 50%.

In Figure 1 it is evident that IOV and S are poorly standardized. This result is not surprising because no explicit steps were taken to standardize these indices. The standardizations of the Z_3, F1, F2, JK, and O/E indices seem reasonably good across low θ and average θ groups. The standardizations of T2 and T4 seem somewhat less adequate, although T2 is well-standardized for false alarm rates of less than .20.

The pattern of results in Figure 2 is somewhat different from the pattern in Figure 1. IOV is poorly standardized in both figures, and Z_3, F2, JK, and O/E are again well-standardized. But F1 is much less well-standardized in Figure 2. In contrast, the results for S and T4 have improved considerably. The standardization of T2 was better in Figure 1.

Finally, Figure 3 presents the results comparing the low θ normals to the high θ normals. The pattern of results indicates that this comparison is the most severe test of standardization. Note that at low misclassification rates, only Z_3, F2, and JK have ROC curves near the diagonal. The standardizations of IOV, F1, and S are all poor. T4 seems somewhat better standardized than T2.

Conclusions

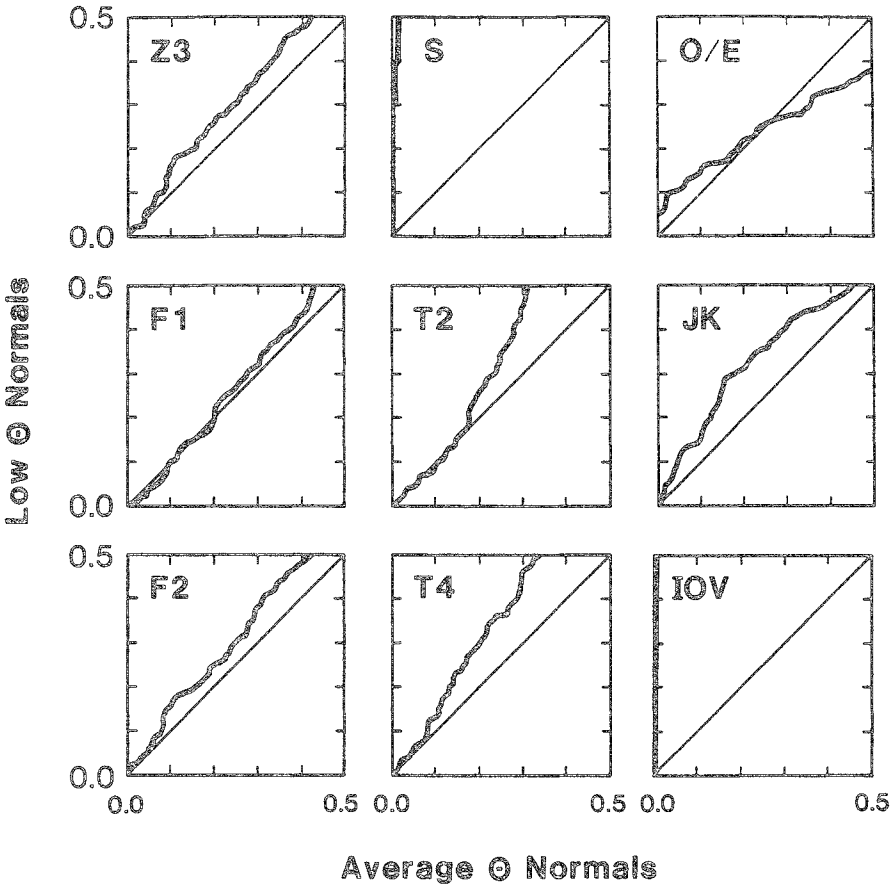
IOV, S, and F1 are poorly standardized. Therefore, the detection rates reported below cannot be interpreted directly for these three indices. For example, normal low ability examinees tend to have IOV scores that are larger than normal high ability examinees. Consequently, IOV will appear to detect aberrance effectively when members of the aberrant sample have lower abilities than members of the normal sample. Thus, high hit rates for IOV are meaningful only when the aberrant sample has higher ability than the normal sample.

Power

The Problem

Do any of the well-standardized practical appropriateness indices have adequate power for detecting

Figure 1
ROC Curves Obtained From 200 Normal Low θ Response Vectors and 200 Normal Average θ Response Vectors

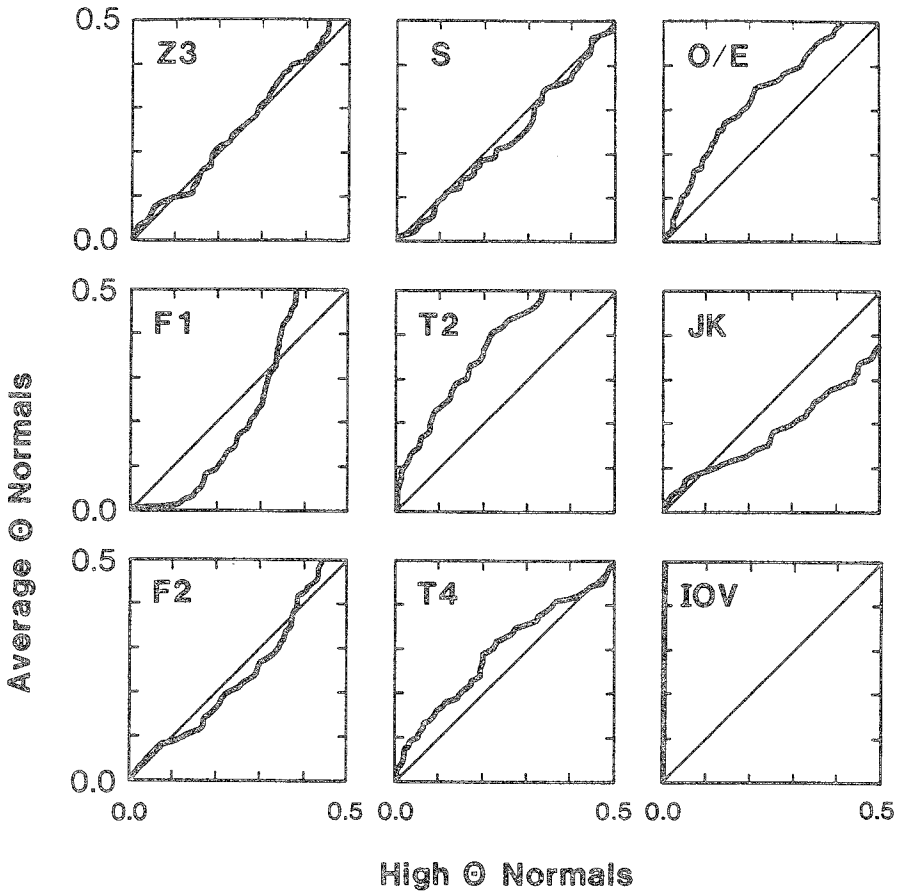


some form of aberrance? Are any nearly as powerful as the index that is optimal for the given form of aberrance? Although results for the poorly standardized indices are reported in Tables 3 through 8 for the sake of completeness, it is to be emphasized that an index must be well-standardized if hit rates are to be interpreted directly. Ability is confounded with measured appropriateness for poorly standardized indices; hence a high hit rate may simply reflect a difference in the ability distributions of the normal and aberrant samples.

Method

Datasets. A test norming sample of 3,000 response vectors was created by sampling 3,000 numbers (θs) from the normal (0,1) distribution truncated to the $[-2.05, 2.05]$ interval. A normal sample of 4,000 response vectors was also generated in this way. Two thousand aberrant response vectors were created in each of 12 conditions. The 12 conditions resulted from varying three factors: the type of aberrance

Figure 2
ROC Curves Obtained From 200 Normal Average θ Response Vectors and 200 Normal High θ Response Vectors



(spuriously high, spuriously low); the severity of aberrance (mild, moderate); and the distribution from which simulated abilities were sampled.

Six of the aberrant samples contained spuriously high response vectors and the remaining six samples contained spuriously low response vectors. Spuriously high response patterns were created by first generating normal response vectors (polychotomously scored) and then replacing a given percentage k of simulated responses (randomly sampled without replacement) with correct responses. Spuriously low response patterns were also created by first generating normal response vectors. Then a fixed percentage of items were randomly selected without replacement and the responses to these items were replaced with random responses (i.e., a response was replaced by option A with probability .2, by option B with probability .2, ..., and by option E with probability .2). Mildly aberrant response patterns were generated by using $k = 15\%$. Moderately aberrant response patterns were created using $k = 30\%$.

The third variable manipulated was the ability level of the aberrant sample. Abilities for the spuriously high samples were sampled from three parts of the normal (0,1) distribution truncated to $[-2.05, 2.05]$:

Figure 3
ROC Curves Obtained From 200 Normal Low θ Response Vectors and 200 Normal High θ Response Vectors

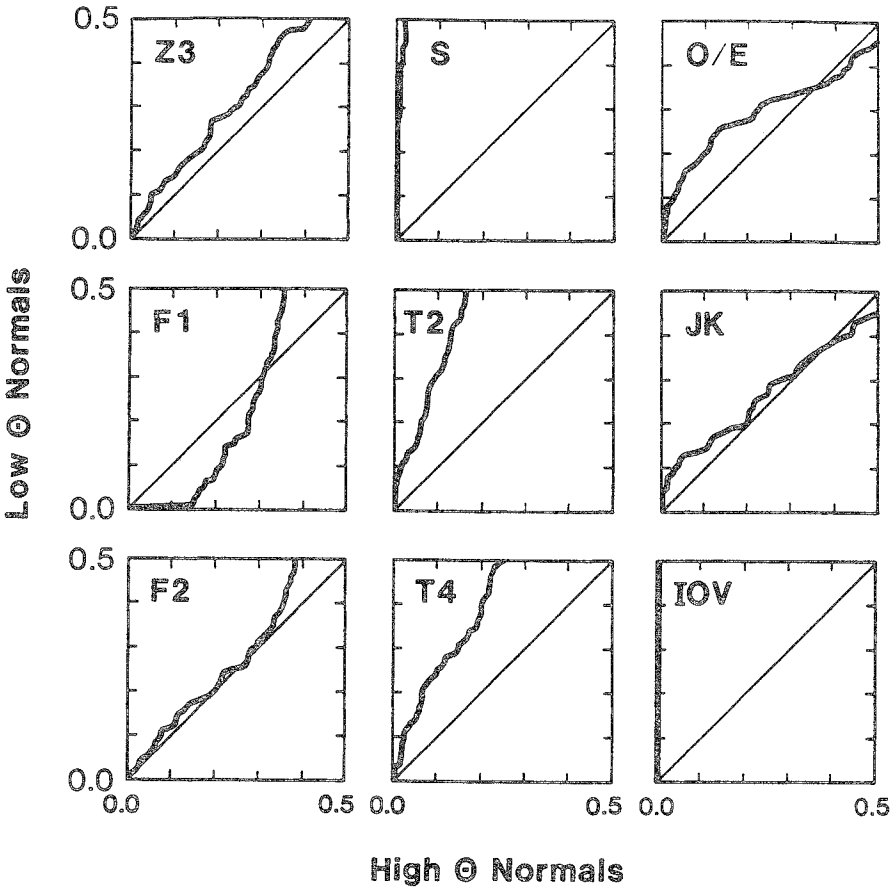


Table 1
Ability Distributions Used to Generate Aberrant Samples

Percent of Aberrant Responses	Type of Aberrance	
	Spuriously High	Spuriously Low
15%	N[-2.05,-1.34]	N(1.41,2.05]
15%	N(-1.34,-0.52]	N(0.35,1.41]
15%	N(-0.52,-0.05]	N(-0.05,0.36]
30%	N[-2.05,-1.34]	N(1.41,2.05]
30%	N(-1.34,-0.52]	N(0.35,1.41]
30%	N(-0.52,-0.05]	N(-0.05,0.36]

Note: N(a,b] is used to denote the standard normal distribution truncated to the interval (a,b].

Table 2
Proportion of Aberrant Response Patterns Generated from 0-9% Ability Range Detected by Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
Normal Group = 200 High Ability Normals										
001	60	38	00	12	62	59	18	100	18	13
005	60	44	00	16	76	63	39	100	19	14
01	63	47	00	34	78	72	40	100	21	15
02	64	56	00	42	82	75	60	100	26	21
03	66	61	00	53	86	83	65	100	30	22
04	68	67	00	54	87	84	69	100	40	26
05	68	73	00	57	91	84	69	100	40	30
07	70	77	01	65	93	89	79	100	46	30
10	71	79	07	74	96	93	82	100	54	35
Normal Group = 200 Low Ability Normals										
001	16	26	25	25	00	14	21	00	00	00
005	50	31	33	27	06	38	26	05	00	00
01	67	44	34	36	08	44	30	10	01	03
02	72	48	46	42	11	49	32	13	03	05
03	80	50	48	46	11	50	37	16	04	09
04	85	52	49	54	20	53	45	18	06	12
05	85	61	54	57	24	59	48	23	09	17
07	88	67	63	61	30	64	53	29	11	19
10	91	72	69	67	35	76	62	33	18	23

very low (0th through 9th percentiles), low (10th through 30th percentiles), and low average (31st through 48th percentiles). In all cases percentile points were determined after the truncation to $[-2.05,2.05]$. These intervals were used because it is more important to detect spuriously high response patterns for low-ability examinees than for high-ability examinees. Similarly, it is more important to detect spuriously low responses by high-ability examinees. Consequently, abilities were sampled from three above-average ability strata for the spuriously low samples: very high (93rd percentile and above), high (65th through 92nd percentiles), and high average (49th through 64th percentiles). The ability percentiles used here correspond to the percentiles forming the United States Air Force Qualifying Test (AFQT) categories. Table 1 summarizes the 12 samples of aberrant response vectors. Each of these 24,000 ($12 \times 2{,}000$) response vectors was independently generated.

Analysis. All of the item and test statistics required to compute the practical appropriateness indices were computed using the test norming sample. These quantities were computed as the first step in the analysis and then used in all subsequent analyses. LOGIST (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters and a FORTRAN program was written to compute the other quantities required.

The nine practical appropriateness indices were then computed for the 4,000 response vectors in the normal sample. The item and test statistics estimated from the test norming sample were used in these calculations. This procedure simulated the process by which practical appropriateness indices would be

Table 3
Proportion of Aberrant Response Patterns
Generated from 0-9% Ability Range Detected by
Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously High										
001	30	26	00	12	10	13	13	13	00	00
005	43	40	00	27	31	25	21	29	01	02
01	50	46	00	34	45	37	30	42	04	06
02	59	54	08	44	59	50	41	53	11	12
03	64	60	22	51	67	56	49	63	16	17
04	67	64	32	55	72	62	54	70	20	21
05	70	69	40	60	78	66	59	75	24	23
07	73	74	52	69	83	73	65	82	31	30
10	77	80	63	76	89	81	73	89	42	39
30% Spuriously High										
001	85	78	00	63	22	75	69	20	01	07
005	91	87	01	81	51	86	80	37	09	17
01	93	90	11	85	65	91	87	50	22	33
02	95	93	44	90	79	95	92	60	43	49
03	95	95	69	93	85	96	94	70	53	56
04	96	96	80	94	88	97	95	76	57	65
05	97	96	86	95	92	97	96	80	62	65
07	97	97	92	97	94	98	97	86	72	72
10	98	98	95	98	96	98	98	91	80	78

computed in many applications. Four optimal indices were also computed for the normal sample: 15% spuriously high, 30% spuriously high, 15% spuriously low, and 30% spuriously low. The ability density f used in Equations 2 and 4 was the normal (0,1) density truncated to the interval $[-2.05,2.05]$. The histograms used to generate the data were also used to compute the optimal indices; that is, polychotomous option characteristic curves were *not* estimated. (In order for an optimal index to be truly optimal for the corresponding form of aberrance, it is necessary to use the true option characteristic curves.)

The nine practical appropriateness indices were computed for each of the 12 aberrant samples. In addition, the 15% spuriously high optimal index was computed for the three samples with this form of aberrance, the 30% spuriously high optimal index was computed for the three samples with this form of aberrance, and so forth.

Note that the ability density used in Equations 2 and 4—a normal (0,1) truncated to $[-2.05,2.05]$—does not match the ability density of any aberrant sample. The proper interpretation of the optimal index is the following: It is the optimal index for the specified form of aberrance, say 15% spuriously high, in a population in which the ability density is normal (0,1) truncated to $[-2.05,2.05]$ for both the normal and aberrant populations and in which a response vector is either normal or 15% spuriously high. The normal group does in fact have this ability distribution. By restricting the abilities of the aberrant group

Table 4
Proportion of Aberrant Response Patterns Generated from 10-30% Ability Range Detected by Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously High										
001	23	14	00	06	01	13	11	00	00	00
005	37	23	00	16	05	22	17	02	01	02
01	45	30	00	21	10	33	25	05	03	05
02	55	38	05	31	19	44	36	09	08	11
03	60	45	15	38	25	49	43	13	12	15
04	63	49	22	43	30	53	47	17	15	19
05	66	53	28	47	38	57	51	21	18	21
07	70	59	41	56	46	64	58	30	26	27
10	75	65	52	63	58	71	66	40	35	35
30% Spuriously High										
001	76	56	00	45	04	61	60	01	08	17
005	85	71	04	67	15	72	72	04	22	29
01	89	75	11	73	27	81	79	08	35	43
02	92	82	34	81	40	87	86	13	52	58
03	93	86	57	86	49	90	90	18	61	64
04	94	88	68	88	56	92	92	22	65	69
05	95	90	75	90	64	93	93	26	70	72
07	96	92	83	93	71	94	95	35	77	77
10	97	94	88	95	80	96	96	45	84	84

to a subinterval of $[-2.05,2.05]$, the power in a particular subpopulation is determined for the index that is optimal for the population as a whole.

Evaluation criteria. The main criteria for evaluating the appropriateness indices are the proportions of aberrant response patterns that are correctly identified as aberrant when various proportions of normal response patterns are misclassified as aberrant. These proportions are presented for all 12 aberrance conditions. This shows what types of aberrant response patterns have acceptably high detection rates using optimal methods and using practical methods. The characteristics of response patterns that cannot be detected are revealed by examining the 12 aberrance conditions separately.

Results

Problems caused by poor standardization. Before presenting the results for the 12 aberrant samples, some problems caused by poorly standardized appropriateness indices are illustrated. Table 2 presents ROC curve points for the 15% spuriously high aberrant sample when the "normal sample" consists of (1) the 200 response vectors with the highest θ values from the normal sample of $N = 4{,}000$ previously described, and (2) the 200 response vectors with the lowest θ values.

Table 5
Proportion of Aberrant Response Patterns Generated from 31-48% Ability Range Detected by Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously High										
001	13	07	00	04	00	09	08	00	01	02
005	26	13	00	12	00	15	14	00	05	05
01	34	18	01	15	01	23	20	00	08	10
02	46	24	06	23	03	32	29	00	16	17
03	51	31	13	29	05	37	35	00	21	22
04	55	34	19	33	07	42	39	01	25	26
05	58	38	25	37	12	45	44	01	29	28
07	64	44	33	45	17	51	50	02	36	34
10	70	52	42	53	26	58	57	05	43	41
30% Spuriously High										
001	59	31	01	31	00	30	38	00	11	20
005	72	45	08	47	03	41	49	00	26	31
01	78	51	15	53	07	51	57	00	38	44
02	84	59	29	63	14	59	67	00	53	58
03	87	65	44	69	19	64	72	01	60	63
04	89	68	50	72	23	68	76	01	64	67
05	91	72	56	75	30	72	79	02	68	70
07	93	77	64	81	39	76	82	04	74	75
10	95	82	71	85	49	81	87	07	79	80

In Table 2, the IOV index seems extremely effective when the normal group consists of high-ability normals: It correctly identifies every single aberrant response vector without a single misclassification of a normal. The S index appears to be an excellent index, although not as powerful as IOV. In contrast, F1 seems to be a very poor index. These results are almost completely contradicted when the normal sample consists of low-ability normals. At a 1% false alarm rate, the detection rate of the IOV index is 10% when the normal group consists of low-ability response patterns; it was 100% when the normals were high-ability. The comparable rates for S are 78% and 8%. The results for F1 are in the opposite direction: The detection rate is 0% when the normals have high ability but 34% when normals have low ability.

The differences in detection rates for F1, S, and IOV result from their poor standardization. The high detection rates for these indices are caused by their confounding of ability and measured appropriateness. In contrast, the well-standardized z_3 has detection rates of 47% and 44% at a 1% misclassification rate. F2 also has similar detection rates: 34% and 36%. T2 is not standardized as well as T4, but note that the detection rates for T2 are higher than the rates for T4. Finally, O/E and JK have moderately dissimilar detection rates across the two sets of normals.

Detection of aberrant response patterns. The results for the 15% and 30% spuriously high samples for the low ability range (0th through 9th percentiles) are shown in Table 3. Here the normal group

Table 6
Proportion of Aberrant Response Patterns Generated from 49-64% Ability Range Detected by Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously Low										
001	29	06	00	03	00	04	04	00	00	01
005	43	12	01	08	00	08	07	00	01	02
01	47	16	03	11	00	14	11	00	03	06
02	56	22	09	17	02	20	17	01	09	12
03	61	27	17	21	03	24	21	02	12	17
04	63	30	24	25	05	28	26	04	15	20
05	67	35	29	29	08	32	29	06	18	23
07	71	40	37	37	13	38	35	10	23	29
10	76	49	46	44	20	46	42	17	32	37
30% Spuriously Low										
001	56	19	00	09	00	09	12	01	00	01
005	75	29	00	20	02	14	20	07	02	05
01	79	35	01	26	06	23	28	14	07	14
02	86	44	08	36	15	32	38	22	20	27
03	89	51	18	42	22	37	45	30	26	33
04	91	55	26	47	27	42	50	37	31	40
05	93	59	34	52	35	47	55	42	36	43
07	95	64	44	60	45	53	60	54	46	50
10	97	70	56	66	56	60	67	66	57	59

consists of 4,000 response vectors that were generated from θ values sampled from the standard normal distribution truncated to $[-2.05,2.05]$. Note that the detection rates for z_3, F2, and T2 are fairly close to the rates for LR. It is clear from Table 3 that the 30% spuriously high treatment is very detectable: LR, z_3, and T2 all have detection rates of 90% or more when the error rate is 1%. Even the relatively moderate 15% spuriously high treatment (which affects at most 13 items on the 85-item test) is fairly detectable: LR and z_3 have detection rates of 50% and 46% at a 1% error rate. O/E and JK, which were shown to be well-standardized, have little power. At a 1% error rate, O/E and JK respectively detect only 22% and 33% of the 30% spuriously high response vectors.

Table 4 presents the results for the 15% and 30% spuriously high treatment applied to the moderately low ability range (10th through 30th percentiles). It should be more difficult to detect aberrant response vectors in this ability range than in the low ability range because the expected number of responses changed due to the aberrance manipulation is smaller. Surprisingly, the detection rates for LR do not decrease sharply: At a 1% error rate, the detection rates are 50% versus 45% for 15% spuriously high, and 93% versus 89% for 30% spuriously high. The detection rates decline more rapidly for z_3 (46% vs. 30% for 15% spuriously high; 90% vs. 75% for 30% spuriously high) and F2 (34% vs. 21%; 85% vs. 73%). The rates of decline of T2 and T4 are intermediate. T2 declines from 37% to 33% for 15% spuriously high and from 91% to 81% for the 30% treatment. T4 declines from 30% to 25% and from 87% to 79%.

Table 7
Proportion of Aberrant Response Patterns
Generated from 65-92% Ability Range Detected by
Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously Low										
001	55	26	05	17	00	17	12	00	03	09
005	66	38	19	32	01	26	20	00	12	16
01	68	44	30	37	03	36	26	01	21	26
02	73	52	47	46	06	45	36	02	32	37
03	75	58	59	53	09	50	42	03	38	43
04	77	62	65	56	13	55	47	05	42	47
05	78	65	70	60	18	58	51	06	46	50
07	81	70	76	67	26	63	56	10	52	55
10	83	76	81	72	36	69	63	16	58	62
30% Spuriously Low										
001	80	54	00	40	01	44	45	04	04	12
005	89	66	08	58	09	54	55	13	15	27
01	91	71	18	62	19	64	63	24	31	44
02	94	78	42	72	32	74	72	32	48	59
03	95	83	59	77	40	77	77	41	55	64
04	96	85	69	80	47	80	80	47	61	71
05	97	87	75	83	55	83	82	53	67	74
07	98	89	82	87	63	86	86	63	75	80
10	98	92	88	91	74	90	89	72	81	85

The trends seen in Tables 3 and 4 continue in Table 5, which presents the results for the 15% and 30% spuriously high treatments applied to the low average ability range (31st to 48th percentiles). In Table 5 the LR index provides detection rates that are roughly 50% higher than the best practical indices. For example, at a 1% error rate LR has a detection rate of 34% for the 15% treatment; z_3, F2, T2, and T4 have detection rates of 18%, 15%, 23%, and 20%, respectively. The detection rates are 78% versus 51%, 53%, 51%, and 57% for the 30% spuriously high condition at a 1% error rate.

Table 6 presents the results for the 15% and 30% spuriously low treatment applied to the high average ability sample (between the 49th and 64th percentiles). It is evident that the practical appropriateness indices are quite ineffective relative to the optimal index. At a 1% error rate LR has a 47% detection rate for the 15% treatment; the highest rate of any of the practical indices is a very low 16%. The pattern of results for the 30% condition is similar. Here the LR detection rate is an impressive 79% when the error rate is 1%; the next best index (z_3) detects only 35% of the aberrant sample.

The practical appropriateness indices have detection rates that are closer to the rates of the optimal statistic in Table 7, which presents the results for the 15% and 30% spuriously low samples with θs in the 65th through 92nd percentiles. This trend is continued in Table 8, which presents the results for the

Table 8
Proportion of Aberrant Response Patterns Generated from 93-100% Ability Range Detected by Appropriateness Indices at Selected ROC Curve Points

False Alarm Rate	Appropriateness Indices									
	LR	z_3	F1	F2	S	T2	T4	IOV	O/E	JK
15% Spuriously Low										
001	73	55	26	39	01	31	23	00	06	12
005	80	68	59	57	10	42	33	00	15	20
01	81	72	71	62	17	54	41	01	21	30
02	84	78	82	72	27	63	52	02	33	43
03	86	82	88	77	36	67	57	03	38	49
04	86	84	90	80	43	71	63	05	43	54
05	88	87	91	82	50	74	66	06	47	56
07	89	90	93	86	60	79	71	11	56	64
10	91	92	94	89	69	84	77	16	64	72
30% Spuriously Low										
001	93	88	06	78	10	83	79	09	27	47
005	96	93	38	88	32	90	86	21	53	65
01	97	95	59	91	47	94	90	31	68	78
02	98	97	81	94	63	96	94	41	82	88
03	98	98	92	96	72	97	95	51	87	90
04	98	98	95	97	76	98	96	59	89	93
05	99	98	96	98	82	98	97	63	91	94
07	99	98	98	98	88	98	98	72	94	96
10	99	99	98	99	93	99	98	80	96	97

spuriously low treatments applied to the highest ability category (percentiles 93 and above). At a 1% error rate in Table 8, for example, LR detects 81% of the 15% spuriously low response patterns; z_3, F2, and T2 have detection rates of 72%, 62%, and 54%. For the 30% treatment, the rate for LR is 97%; z_3, F2, and T2 have rates of 95%, 91%, and 94%.

Distributions of three indices. Drasgow and Guertler (in press) have presented a utility theory approach to the use of appropriateness measurement in practical settings. Their approach requires the densities of an index in normal and aberrant samples. Consequently, normal distributions were fitted to the distributions of z_3, F2, and T4 by equating the first two moments of the normal distribution to the empirical moments. These analyses were based on the first 1,000 response vectors from the normal sample and each of the 12 aberrant samples. The fitted means and standard deviations are presented in Table 9. As a crude measure of fit, Kolmogorov-Smirnov test statistics were computed to compare the empirical distributions to normal distributions with the observed moments. No significant ($\alpha = .05$) departures of empirical distributions from the corresponding fitted normal distributions were found. As the Kolmogorov-Smirnov test is sometimes conservative when fitted moments are substituted into the theoretical distribution (Massey, 1951), these results should be viewed with some caution.

Table 9
Means and Standard Deviations of Empirical Distributions of z_3, F2, and T4 (N=1,000)

Condition and Ability Range	Severity of Aberrance: 15% z_3	15% F2	15% T4	30% z_3	30% F2	30% T4
Spuriously High						
00-09%						
Mean	-2.32	1.28	1.56	-4.00	1.49	3.22
S.D.	1.13	0.14	0.94	1.22	0.15	1.07
10-30%						
Mean	-1.85	1.23	1.39	-3.32	1.43	3.04
S.D.	1.11	0.14	0.98	1.19	0.15	1.10
31-48%						
Mean	-1.38	1.19	1.22	-2.47	1.36	2.38
S.D.	1.03	0.14	1.02	1.21	0.17	1.19
Spuriously Low						
49-64%						
Mean	-1.02	1.13	0.65	-1.58	1.19	1.20
S.D.	1.03	0.13	0.99	1.14	0.14	0.98
65-92%						
Mean	-1.85	1.23	1.17	-2.74	1.34	2.12
S.D.	1.16	0.16	1.11	1.19	0.15	1.08
93-100%						
Mean	-3.01	1.37	1.78	-4.28	1.54	3.50
S.D.	1.30	0.17	1.14	1.32	0.17	1.24
Normals*						
0-100%						
Mean	0.09	0.99	-0.14			
S.D.	0.97	0.12	0.86			

*To conserve space results for the normal sample are listed under the columns headed 15% aberrant responses.

Discussion

There has been a growing interest in appropriateness measurement among both researchers and testing practitioners. To date, however, there has been little critical study of the various indices available. The results of the research summarized here clearly indicate that there are important differences in the properties of appropriateness indices. Figures 1 through 3 show that some indices are poorly standardized (e.g., IOV), and that even a "standardized" index may not be well-standardized (F1). Table 2 illustrates the problems that are caused by poorly standardized indices. IOV, for example, should not be used as an appropriateness index because it is confounded with ability.

A well-standardized index is not, however, necessarily a good appropriateness index. The O/E and JK indices were shown to be reasonably well-standardized in Figures 1 through 3, but Tables 3 through 8 clearly show them to be ineffective in detecting aberrant response patterns.

Perhaps the most important finding of this study is that z_3, F2, and T2 provide rates of detection of some forms of aberrance that are nearly optimal but have inadequate rates of detection for other forms

of aberrance. In particular, these three indices have near-optimal rates of detection when the spuriously high treatment is applied to very low θ response vectors and when the spuriously low treatment is applied to very high θ response vectors. Unfortunately, these indices have rates of detection far below optimal when the spuriously high and low treatments are applied to response vectors with nearly average θ values.

These results indicate that new indices that are more powerful than z_3, F2, and T2 must be devised to identify inappropriate scores from examinees whose abilities are near average. It may be necessary to construct two indices: one for spuriously low response patterns and one for spuriously high response patterns. This psychometric necessity would be quite useful for practitioners because it would allow them to *diagnose* the cause of aberrance as well as to detect aberrant response patterns.

References

Cliff, N. (1979). Test theory without true scores? *Psychometrika, 44,* 373–393.

Drasgow, F., & Guertler, E. (in press). Detecting inappropriate test and scale scores in practical settings. *Journal of Applied Psychology.*

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. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika, 65,* 457–487.

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* (pp. 104–122). Vancouver: Educational Research Institute of British Columbia.

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

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42–56.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675–685.

Levine, M. V., & Drasgow, F. (1984). *Performance envelopes and optimal appropriateness measurement* (Report No. 84–5). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory. (ERIC Document Reproduction Service No. ED 263 126)

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

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46,* 68–78.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., pp. 80–203). Reading MA: Addison-Wesley.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20,* 207–219.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho. (In Japanese).

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

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81–96.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76–6). Princeton NJ: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Acknowledgments

This research was supported by Contract No. F41689-84-D-0002 from the U.S. Air Force Human Resources Laboratory to the Human Factors and Logistics Division of Universal Energy Systems Inc., Dayton OH, U.S.A. The authors thank Malcolm Ree, Randolph Park, and James Earles for their helpful suggestions.

Author's Address

Send requests for reprints or further information to Fritz Drasgow, Department of Psychology, University of Illinois, 603 E. Daniel Street, Champaign IL 61820, U.S.A.