Optimal Detection of Certain Forms of Inappropriate Test Scores

Fritz Drasgow and Michael V. Levine
University of Illinois

Optimal appropriateness indices, recently introduced by Levine and Drasgow (1984), provide the highest rates of detection of aberrant response patterns that can be obtained from item responses. In this article they are used to study three important problems in appropriateness measurement. First, the maximum detection rates of two particular forms of aberrance are determined for a long unidimensional test. These detection rates are shown to be moderately high. Second, two versions of the standardized ℓ_0 appropriateness index are compared to optimal indices. At low false alarm rates, one standardized ℓ_0 index has detection rates that are about 65% as large as optimal for spuriously high (cheating) test scores. However, for the spuriously low scores expected from persons with ill-advised testing strategies or reading problems, both standardized ℓ_0 indices are far from optimal. Finally, detection rates for polychotomous and dichotomous scorings of the item responses are compared. It is shown that dichotomous scoring causes serious decreases in the detectability of some aberrant response patterns. Consequently, appropriateness measurement constitutes one practical testing problem in which significant gains result from the use of a polychotomous item response model.

An examinee's score on a multiple-choice test may occasionally fail to provide a representative measure of ability or achievement. Examinees' scores are spuriously high when examinees copy answers from high ability neighbors or when they have been given the answers to some questions by informants. Spuriously low scores result from alignment errors (answering the $(i + 1)$st item in the answer sheet space provided for the ith answer) over a block of items, atypical education, unusually creative interpretations of normally easy items, language difficulties, suboptimal omitting strategies, and a variety of other sources.

It is important to detect aberrant response patterns. Test scores that provide unrepresentative measures of abilities can lead to misclassification errors that are expensive to test users. Furthermore, these errors can have profound implications for test takers. Individuals with spuriously high tests scores may be selected for jobs or academic programs for which they are incompetent, and spuriously low test scores may cause deserving individuals to be denied professional and academic opportunities.

Levine and Rubin (1979) used the term *appropriateness measurement* to denote model-based methods for detecting inappropriate test scores. Appropriateness measurement methods usually attempt to identify

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 1, March 1986, pp. 59–67

0146-6216/86/010059-09$1.70

inconsistent response patterns that have correct answers to difficult items co-occurring with incorrect answers to easy items (Drasgow, 1982; Drasgow, Levine, & Williams, 1985; Harnisch & Tatsuoka, 1983; Levine & Drasgow, 1982; Levine & Rubin, 1979; Rudner, 1983; Tatsuoka, 1984). Such response patterns are expected to result from the types of test-taking anomalies described above.

Previous Research

Simulated spuriously high and spuriously low response patterns have been shown to be detectable by appropriateness measurement (Levine & Rubin, 1979). High detection rates have been obtained despite model misspecification and item parameters estimated with error from samples that include substantial proportions of inappropriate response patterns (Levine & Drasgow, 1982). Very high detection rates have been found when response patterns of low ability examinees have been modified to simulate cheating on 20% to 30% of the items on an 85-item test and when response patterns of high ability examinees have been modified to simulate spuriously low responding on similar proportions of items (Drasgow et al., 1985).

The standardized ℓ_0 index (Drasgow et al., 1985), denoted z_3, is easy to compute and has been found to provide relatively powerful detection of inappropriate response patterns. This index is computed from the dichotomously scored item responses, coded $u_i = 1$ if correct and $u_i = 0$ if incorrect, in the following way. First, the logarithm of the likelihood function at the maximum likelihood estimate $\hat{\theta}$ of θ, is computed as

$$\ell_0 = \sum[u_i \log P_i(\hat{\theta}) + (1-u_i)\log Q_i(\hat{\theta})] \quad , \tag{1}$$

where $P_i(\theta)$ is the probability of a correct response among examinees with ability θ and $Q_i(\theta) = 1 - P_i(\theta)$. Then,

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

where $M(\hat{\theta})$ and $S(\hat{\theta})$ are the conditional expected value and variance of ℓ_0, respectively, given $\theta = \hat{\theta}$, that is,

$$M(\hat{\theta}) = \sum[P_i(\hat{\theta})\log P_i(\hat{\theta}) + Q_i(\hat{\theta})\log Q_i(\hat{\theta})] \quad , \tag{3}$$

and

$$S(\hat{\theta}) = \sum P_i(\hat{\theta})Q_i(\hat{\theta})\{\log [P_i(\hat{\theta})/Q_i(\hat{\theta})]\}^2 \quad . \tag{4}$$

Although the three-parameter logistic model has been used in previous research (hence "z_3"), any of the logistic, normal ogive, or other parametric models can be used in the above expressions.

The observed rates of detection of aberrant response patterns found in earlier studies seem to be high enough for practical applications of appropriateness measurement. Nonetheless, it has not yet been determined whether an appropriateness index such as the z_3 index is powerful in an absolute sense. More specifically, none of the appropriateness indices compared to z_3 has been found to provide substantially higher rates of detection of spuriously high response patterns. Is this because no such index exists? Does a dramatically superior appropriateness index exist, but remain undiscovered? Comparative studies of available appropriateness indices (Drasgow et al., 1985; Harnisch & Tatsuoka, 1983; Rudner, 1983) cannot answer these fundamental questions.

Optimal Appropriateness Measurement

Means for determining the statistical power of the most powerful appropriateness indices for a given form of aberrance were recently presented by Levine and Drasgow (1984). They denoted the likelihood of a response vector **u** under a model for normal response patterns, such as the three-parameter logistic, by $P_{Normal}(\mathbf{u})$, and the likelihood under a model for aberrant response patterns by $P_{Aberrant}(\mathbf{u})$. The Neyman-Pearson Lemma can then be used to show that the highest rate of detection of aberrant response patterns is obtained by classifying **u** as aberrant when

$$P_{Aberrant}(\mathbf{u}) \geqslant \text{constant } P_{Normal}(\mathbf{u}) \quad , \tag{5}$$

where the constant is selected to achieve a specified Type I error rate (i.e., a specified rate of misclassifying normal response patterns as aberrant).

In many cases it is easy to compute $P_{Normal}(\mathbf{u})$. For example, assume that (1) the item responses are scored dichotomously, (2) $P_i(t)$ is the probability of a positive response on item i among normal examinees with ability t, (3) local independence holds, and (4) the ability density is denoted $f(t)$. Then the conditional likelihood of **u** given ability t is

$$P_{Normal}(\mathbf{u}|t) = \prod_{i=1}^{n} P_i(t)^{u_i} [1 - P_i(t)]^{1-u_i} \quad , \tag{6}$$

and the unconditional likelihood is

$$P_{Normal}(\mathbf{u}) = \int P_{Normal}(\mathbf{u}|t) f(t) dt \quad , \tag{7}$$

where the range of integration can be taken as the support of f. When f and the P_i are smooth functions, the integral in Equation 7 can be easily and accurately approximated using numerical methods.

Levine and Drasgow (1984) showed that $P_{Aberrant}(\mathbf{u})$ can be obtained for some forms of aberrance by taking the conditioning-integrating argument one step further. To illustrate their approach, suppose that the form of aberrance under investigation is the 10% spuriously high condition. Here it is assumed that some examinees are given the answer key to 10% of the test items which, for the statistical analysis, are taken as randomly sampled. The likelihood of **u** can be written conditioning on both t and a particular 10% of the test items. Then, the likelihood $P_{Aberrant}(\mathbf{u}|t)$ is determined by averaging all the conditional probabilities obtained by selecting different aberrant items. Finally, $P_{Aberrant}(\mathbf{u})$ is obtained by integration as in Equation 7 or by a quadrature formula.

Direct computation of $P_{Aberrant}(\mathbf{u}|t)$ for long tests would be extremely time-consuming. For example, there are about 400 billion ways of choosing 10% of the test items on an 85-item test. Note also that to compute $P_{Aberrant}(\mathbf{u})$, $P_{Aberrant}(\mathbf{u}|t)$ must be evaluated for several values of t, so trillions of additions would be required. Of crucial import to the present research is a recursive formula presented by Levine and Drasgow (1984). The formula reduces the number of operations required to evaluate $P_{Aberrant}(\mathbf{u}|t)$ to less than one thousand for an 85-item test with 10% of the items spuriously high. Under these conditions, it has been found that $P_{Aberrant}(\mathbf{u})$ can usually be computed in less than one-tenth of a second of CDC Cyber 175 CPU time, and costs approximately one cent at the University of Illinois.

Although optimal appropriateness indices are reasonably inexpensive to compute on a university mainframe computer, they are currently intended only as research tools. Fundamentally different expressions for $P_{Aberrant}(\mathbf{u})$ are obtained for spuriously high and spuriously low response patterns. Furthermore, the expression for $P_{Aberrant}(\mathbf{u})$ for, say, the 10% spuriously high condition with an 85-item test is different from the expression for the 12% spuriously high condition, because the former conditions on the 400 billion ways of choosing 10% of the test items and the latter conditions on the 3 trillion ways of choosing 12% of the test items. Although the recursive formula can be modified to accommodate a distribution of aberrant response types, the goal at this time is to use optimal indices to answer such research questions as:

1. What is the maximum detection rate achievable by appropriateness measurement for the various benchmark types of inappropriateness?
2. How closely do selected practical, multipurpose indices approach the optimal detection rates?

3. Can substantially better appropriateness measurement be achieved by polychotomous model indices, that is, indices that use the information in wrong answers?

Three Problems in Appropriateness Measurement

In the present research optimal indices were used to examine three questions. First, what are the upper limits on detection rates in appropriateness measurement? An answer to this question is important because it would indicate whether acceptably high detection rates are possible for important forms of aberrance on a given test. If only low rates of detection are possible for some serious forms of aberrance, then it may be prudent to lengthen the test, reduce decision makers' reliance on the test, change the testing process so that the aberrance inducing processes are inhibited (e.g., increase proctoring if copying is a problem), and so forth.

The second issue studied here concerns the virtues of the z_3 index: How powerful is the z_3 index in relation to the optimal index? If optimal indices provide only slightly higher detection rates, then z_3 would be a good candidate for use in practical settings because it is very easy to compute. In previous studies with dichotomously scored item responses z_3 was found to provide detection rates that are comparable to the rates obtained by much more sophisticated indices (e.g., Levine & Rubin's, 1979, *LR* index). Nonetheless, these studies do not indicate whether z_3 provides rates of detection that are nearly optimal or whether all the indices under consideration were far from optimal.

The third problem studied concerns the loss of information about aberrance that is incurred when multiple-choice items are scored dichotomously rather than polychotomously. The polychotomous item response models currently available seem less tractable than their dichotomous counterparts. Furthermore, there are only a few examples of substantial gains from polychotomous analyses. Can substantial increases in the rates of detection of aberrant response vectors be made using a polychotomous analysis? Toward this end Drasgow et al. (1985) generalized the z_3 index to the polychotomous case. They expected this new index, labeled z_h, to be superior to z_3 because a polychotomous model provides a more fine-grained description of item responses than a dichotomous model. They found, however, that z_3 provided substantially higher rates of detection for spuriously high response patterns than did z_h. Does any index for the polychotomous data yield markedly higher rates of detection than z_3? Finding a polychotomous model optimal index to be clearly superior to dichotomous model indices (including the dichotomous model optimal index) would underscore the value of polychotomous item response models and the need for better polychotomous model appropriateness indices.

Study 1: Dichotomous Item Responses

Method

The problem in Study 1 was defined as: What is the upper limit on the detectability of mildly aberrant response vectors? How close to optimal is the z_3 index? Three data sets were created. The first was a normal sample of 4,000 response vectors. It was generated using three-parameter logistic item characteristic curves (ICCs) with parameters set equal to the item parameter estimates obtained from Levine and Drasgow's (1982) analysis of the 85-item April 1975 Scholastic Aptitude Test-Verbal section (SAT-V). The ability parameters were randomly sampled from a normal (0,1) distribution truncated to the $[-2.05, 2.05]$ interval. (The ability distribution was truncated so that the results of Study 1 would be more comparable to the results of Study 2.)

The second sample contained 2,000 response vectors subjected to the 10% spuriously high treatment. It was generated by first creating 2,000 normal response vectors by the method described above. Then nine items were randomly selected without replacement from each response vector and rescored as correct. The third sample also contained 2,000 response vectors. They were initially created as were the normal response vectors and then they were subjected to the 10% spuriously low treatment. Here nine items were randomly selected without replacement from each response vector and then each of these items was independently rescored as correct with probability .2 and as incorrect with probability .8.

The three-parameter logistic maximum likelihood estimate of ability was obtained for each response vector. The simulation item parameters were taken to be known quantities. (Levine & Drasgow, 1982, showed that values of ℓ_0 computed from estimated item parameters were almost identical to values computed from the actual item parameters for a long unidimensional test. For the optimal index to be truly optimal, it *must* be computed using the actual item parameters.) Then, the z_3 index was computed for the response vectors in all three samples. The optimal index for the 10% spuriously high condition was computed for each response vector in the normal sample and the spuriously high sample by the algorithm described by Levine and Drasgow (1984). The 10% spuriously low optimal index was similarly computed for the spuriously low sample and the normal sample.

Results

Detection rates for various false alarm rates (Type I error rates) for the 2,000 response vectors subjected to the 10% spuriously high treatment are shown in Table 1. It is clear that the z_3 index is close to optimal: Detection rates for z_3 are about 90% as large as the detection rates for the optimal index at corresponding false alarm rates.

Table 1 also shows that reasonably high detection rates were obtained with the optimal index. For example, 24% of the spuriously high response patterns were detected when the false alarm rate was 5%. These results must be evaluated in light of the fact that, on the average, the 10% spuriously high treatment changed about four incorrect responses to correct. Consequently, the signal in this context (4 altered item responses) is small relative to the noise (81 item responses not altered).

The results for the 2,000 response vectors subjected to the 10% spuriously low treatment are also shown in Table 1. The detection rates for the 10% spuriously low optimal index are somewhat lower than the rates for the 10% spuriously high optimal index. Detection rates for z_3 show larger decreases from the spuriously high condition to the spuriously low condition. Consequently, detection rates for z_3 are not very close to optimal; at low false alarm rates z_3 detects only about 65% as many spuriously low response vectors as the optimal index.

Study 2: Polychotomous Item Responses

Method

The problem in Study 2 was defined as: Is the power to detect aberrant response patterns substantially diminished when multiple-choice items are scored dichotomously? The polychotomous responses (five-option multiple-choice items with omitting allowed) generated for this study were created from the histograms constructed by Levine and Drasgow (1983). They used the three-parameter logistic model to estimate abilities for 49,470 examinees from the April 1975 administration of the SAT-V, formed 25 ability groups by using the 4th, 8th, . . ., 96th percentile points from the normal distribution, and then determined the frequency of option selection (treating skipped and not reached items as a single response category) for each ability group. Probabilities of option responses were then computed by linear interpolations between ability category medians (i.e., the 2nd, 6th, . . ., 98th percentile points from the normal distribution). Notice that response probabilities computed from these histograms for the correct options are strongly related to three-parameter logistic probabilities of correct responses.

Table 1
Detection Rates for Response Vectors Generated by the Three-Parameter Logistic Model for the 10% Spuriously High and 10% Spuriously Low Treatments

Prop. of Normals	Proportion Detected by		Ratio of
Misclassified	Optimal Index	z_3	Detection Rates
10% Spuriously High			
.01	.090	.087	.97
.02	.144	.120	.83
.03	.175	.148	.85
.04	.199	.179	.90
.05	.242	.202	.83
.06	.258	.220	.85
.08	.302	.278	.92
.10	.342	.308	.90
.15	.446	.388	.87
.20	.530	.453	.86
.25	.599	.510	.85
10% Spuriously Low			
.01	.076	.046	.61
.02	.100	.065	.65
.03	.130	.079	.61
.04	.150	.097	.64
.05	.170	.114	.67
.06	.181	.126	.70
.08	.210	.160	.76
.10	.234	.187	.80
.15	.312	.250	.80
.20	.376	.314	.83
.25	.444	.372	.84

Note: Item responses were generated for 2,000 simulated spuriously high examinees, 2,000 simulated spuriously low examinees, and 4,000 simulated normal examinees.

Three data sets were generated in Study 2: 4,000 normal response patterns, 2,000 normal response patterns subjected to the 10% spuriously high treatment, and 2,000 normal response patterns subjected to the 10% spuriously low treatment. Abilities were randomly sampled from the normal (0,1) distribution truncated to the [−2.05, 2.05] interval (because it was not possible to interpolate below the 2nd percentile or above the 98th percentile). Finally, the spuriously high treatment was performed exactly as in Study 1. In the spuriously low treatment, each of the five options was selected with probability .2. Omitted responses were never modified in the spuriously low treatment nor counted toward the total of nine items.

The polychotomous generalization z_h of z_3 was then computed for each response vector. The requisite quantities for z_h are

$$\ell_h = \sum_{i=1}^{n} \sum_{j=1}^{6} \delta_j(v_i) \log P_{ij}(\hat{\theta}) \quad , \tag{8}$$

$$M_h(\hat{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{6} P_{ij}(\hat{\theta}) \log P_{ij}(\hat{\theta}) \quad , \tag{9}$$

and

$$S_h(\hat{\theta}) = \sum_{i=1}^{n} \{ \sum_{j=1}^{6} \sum_{k=1}^{6} P_{ij}(\hat{\theta}) P_{ik}(\hat{\theta}) \log P_{ij}(\hat{\theta}) \log [P_{ij}(\hat{\theta}) / P_{ik}(\hat{\theta})] \} \quad , \tag{10}$$

where v_i is the response to the ith item, coded $v_i = 1$ for option A, . . ., $v_i = 5$ for option E, and $v_i = 6$ for omit,

$P_{ij}(\hat{\theta})$ is the histogram probability of $v_i = j$ at ability $\hat{\theta}$, and

$\delta_j(v_i) = 1$ if $v_i = j$ and 0 otherwise.

Since the histogram model likelihood function does not have a continuous first derivative and is therefore difficult to maximize, three-parameter logistic ability estimates (with the restriction $-2.05 \leq \hat{\theta} \leq 2.05$) were used in the above equations. The polychotomous model optimal index for the 10% spuriously high treatment was computed for the normal and spuriously high samples and the optimal index for the 10% spuriously low treatment was computed for the normal and spuriously low samples.

In many applications multiple-choice item responses are scored dichotomously. To determine the effects on appropriateness measurement of this simplification of the data, the item responses were scored dichotomously, and then z_3 and the three-parameter logistic optimal index were computed. Note that the three-parameter logistic optimal indices are not truly optimal in these analyses because the model used to analyze the data (the three-parameter logistic model) is not identical to the model used to create the data (the histogram model).

Results

Table 2 presents the results for the 10% spuriously high condition. This table shows why the detection rates of spuriously high response vectors by z_3 were found to be much higher than the rates for z_h in the Drasgow et al. (1985) study: z_h is a very ineffectual index for detecting spuriously high response patterns. At low false alarm rates, the z_h index yields detection rates only about one-fourth to one-third as high as the rates for the polychotomous model optimal index. In contrast, the detection rates for z_3 are about 65% of optimal.

Table 2 also shows that the three-parameter logistic model optimal index yielded detection rates that were roughly 75% as large as the rates for the truly optimal polychotomous model index. Thus, polychotomous scoring of simulated multiple-choice item responses allows moderately higher rates of detection of spuriously high response vectors.

Table 2 also presents the results for the 10% spuriously low condition. For this treatment the polychotomous model optimal index detects almost 41% of the aberrant response vectors at a 5% false alarm rate. This high detection rate is obtained because the spuriously low modification changed more responses here than in the dichotomous case. In particular, the spuriously low treatment can change an incorrect response that is appropriate (i.e., the favorite distractor at the examinee's ability t) to an incorrect response that is inappropriate (i.e., a distractor rarely selected by examinees with abilities near t).

Table 2 shows that dichotomous scoring seriously reduces the detectability of spuriously low response vectors: The three-parameter logistic optimal index has detection rates substantially lower than the polychotomous model optimal index. At a 1% false alarm rate, the three-parameter logistic optimal index detects less than half as many aberrant response vectors as does the polychotomous model optimal index.

It is also evident in Table 2 that neither the z_3 nor the z_h index is particularly effective. Both of these

Table 2
Detection Rates for Response Vectors Generated by the Histogram Model for the 10% Spuriously High and 10% Spuriously Low Treatments

Prop. of Normals	Appropriateness Index			
Misclassified	EC_h	z_h	EC_3	z_3
10% Spuriously High				
.01	.130	.032 (.25)	.087 (.67)	.082 (.63)
.02	.194	.056 (.28)	.144 (.74)	.121 (.62)
.03	.244	.073 (.30)	.184 (.75)	.154 (.63)
.04	.282	.090 (.32)	.223 (.79)	.183 (.65)
.05	.330	.108 (.33)	.256 (.78)	.204 (.62)
.06	.356	.125 (.35)	.282 (.79)	.230 (.65)
.08	.402	.155 (.39)	.343 (.85)	.270 (.67)
.10	.442	.197 (.45)	.380 (.86)	.317 (.72)
.15	.520	.272 (.52)	.488 (.94)	.400 (.77)
.20	.591	.352 (.60)	.556 (.94)	.482 (.82)
.25	.667	.410 (.61)	.620 (.93)	.547 (.82)
10% Spuriously Low				
.01	.231	.069 (.30)	.114 (.49)	.083 (.36)
.02	.298	.108 (.36)	.168 (.56)	.114 (.38)
.03	.344	.138 (.40)	.196 (.57)	.148 (.43)
.04	.381	.160 (.42)	.238 (.62)	.174 (.46)
.05	.406	.184 (.45)	.260 (.64)	.192 (.47)
.06	.438	.218 (.50)	.272 (.62)	.218 (.50)
.08	.504	.266 (.53)	.312 (.62)	.256 (.51)
.10	.546	.314 (.58)	.353 (.65)	.292 (.54)
.15	.640	.400 (.63)	.454 (.61)	.386 (.60)
.20	.706	.471 (.67)	.540 (.76)	.454 (.64)
.25	.758	.540 (.71)	.598 (.79)	.514 (.68)

Note: Item responses were generated for 2,000 simulated spuriously high examinees, 2,000 simulated spuriously low examinees, and 4,000 simulated normal examinees. EC_h = optimal histogram appropriateness index; EC_3 = optimal three-parameter logistic appropriateness index. Each number in parentheses is the ratio of a nonoptimal index's hit rate and the optimal index's hit rate.

indices have detection rates that are only about 40% as large as the optimal index at low false alarm rates. Note, however, that z_3 is relatively effective in relation to the dichotomous model optimal index.

Discussion

In the studies described above forms of aberrance that have been difficult to detect with non-optimal appropriateness indices were deliberately chosen (see Drasgow et al., 1985; Levine & Rubin, 1979). If forms of aberrance that were easy to detect had been used, such as the 30% spuriously high and low treatments, there would have been little improvement possible. The forms of aberrance studied here certainly allowed for improvement on the detection rates of the z_3 and z_h indices.

The results indicate that optimal indices perform from slightly to dramatically better than previously studied indices. The least improvement was found for truly dichotomous response vectors subjected to the spuriously high manipulation. In this condition the z_3 index yielded detection rates roughly 85% to 90% as large as possible. It seems reasonable to conclude that the z_3 index is satisfactory for this situation. The z_3 index is less satisfactory for detecting response vectors that were generated by the three-parameter logistic model and then subjected to the spuriously low treatment. At low false alarm rates, z_3 detects about 65% as many aberrant response vectors as detected by the optimal index.

When item responses were created to simulate a multiple-choice test (in particular, the SAT-V), optimal appropriateness measurement using the polychotomous form of the items provides substantially improved rates of detection. The most striking results were obtained in the spuriously low condition in which both the z_3 and z_h indices were clearly unsatisfactory. There are two implications of this result. First, it clearly shows that appropriateness measurement is one practical testing problem in which substantial gains can be made through the use of polychotomous item response models. The second implication of this research is that further work is needed on practical appropriateness indices based on polychotomous models. It is clear that z_h is less than satisfactory for detecting spuriously high response patterns on multiple-choice tests.

In summary, the results show that moderately high detection rates can be obtained for mildly aberrant response patterns. The rates of detection are higher for polychotomous item responses than for dichotomous item responses. The z_3 index is close to optimal for dichotomous item responses generated from three-parameter logistic ICCs and subjected to the spuriously high treatment, but it is less satisfactory when dichotomous item responses are subjected to the spuriously low treatment. Finally, detection rates were substantially lower when simulated multiple-choice items were analyzed dichotomously. These results emphasize the importance of developing powerful, practical polychotomous model indices.

References

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6,* 297–308.

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.

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.

Levine, M. V. (1985). *Multilinear formula scoring: Estimation of option characteristic curves.* Manuscript in preparation.

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* (Measurement Series 84-5). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–289.

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

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

Acknowledgments

This research was supported by contracts N00014-79C-0752, NR 154-445, and N00014-83K-0397, NR 150 518, from the Office of Naval Research, Michael V. Levine, Principal Investigator.

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.