Appropriateness Measurement for Some Multidimensional Test Batteries

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

Model-based methods for the detection of individuals inadequately measured by a test have generally been limited to unidimensional tests. Extensions of unidimensional appropriateness indices are developed here for multi-unidimensional tests (i.e., multidimensional tests composed of unidimensional subtests). Simulated and real data were used to evaluate the effectiveness of the multitest appropriateness indices. Very high rates of detection of spuriously high and spuriously low response patterns were obtained with the simulated

data. These detection rates were comparable to rates obtained for long unidimensional tests (both simulated and real) with approximately the same number of items. For real data, similarly high detection rates were obtained in the spuriously high condition; slightly lower detection rates were observed for the spuriously low condition. Several directions for future research are described. *Index terms: appropriateness measurement, item response theory, multidimensional tests, optimal appropriateness measurement, polychotomous measurement.*

It is difficult to identify mismeasured examinees when short tests are analyzed individually. Drasgow, Levine, Williams, McLaughlin, and Candell (1987), for example, simulated low-ability examinees who were given answers to 30% of the items on a short (30-item) test. The most powerful appropriateness index for this form of aberrance (Levine & Drasgow, 1988) was able to detect only 45% of the simulated cheaters at a 1% false-positive rate. This detection rate is much lower than that obtained on simulated and actual long power tests such as the Scholastic Aptitude Test (Verbal). In a simulation study, for example, Drasgow, Levine, and McLaughlin (1987) identified 93% of low-ability simulees given the answers to 30% of the items on an 85-item test.

The solution to increasing detection rates on short tests does not lie in better appropriateness indices for individual tests, because no index can provide higher detection rates than an optimal index. This fact has prompted the development of item response theory (IRT) methods for efficient detection of inappropriate test scores in situations in which examinees complete several short tests. In this approach, information about aberrance is pooled across tests that measure distinct, but correlated, traits. This approach seems valuable for test batteries that contain a number of short power tests, such as the Armed Services Vocational Aptitude Battery (ASVAB).

Multitest extensions are described for six standardized, suboptimal appropriateness indices, and multitest optimal indices are outlined. The optimal and standardized indices were evaluated in two studies. The first used simulated ASVAB data to assure that all assumptions about the item responses (e.g., local independence, three-parameter logistic item response functions) were correct. In the second study, an actual ASVAB dataset was used to evaluate the performance of the appropriateness indices under realistic conditions.

A second approach to the use of additional information for the identification of aberrant response patterns is also considered. Specifically, a regression analysis can incorporate other information about

---

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 2, June 1991, pp. 171-191
© Copyright 1991 Applied Psychological Measurement Inc.
0146-6216/91/020171-21\$2.30

an examinee in order to predict expected performance on a test. Large departures from expected performance are then assumed to indicate aberrant response patterns.

The regression approach seems sensible in a situation in which one or more test scores are available that result from ordinary responses, as well as one or more possibly aberrant test scores. Suppose, for example, that one of the tests in a battery was compromised (i.e., available to some examinees prior to its administration), but other subtests were believed to be secure (i.e., not available to examinees). In this situation, it seems natural to predict scores on the compromised test from scores on the secure tests. Examinees who did not cheat on the compromised test should have predicted scores in reasonably close agreement with their obtained scores; cheaters should obtain scores on the compromised test that are much higher than predicted. Therefore, the standardized residual (i.e., the standardized error of prediction) should function as a useful indicator of cheating; honest examinees should ordinarily have standardized residuals less than 2.0, whereas many cheaters should have standardized residuals greater than 2.0.

The ASVAB provides a variant of the above scenario. Only 4 of its 10 subtests are used to determine eligibility for enlistment. A low-ability examinee might be tempted to cheat on the subtests that determine enlistment eligibility, but not on the other subtests. In this case, regression can use the test scores not included in the composite that determines eligibility to predict the eligibility composite. Persons who cheated on the tests included in the composite, but not on the other tests, would therefore be expected to have large positive standardized residuals and therefore be identifiable.

As another example of the regression approach, consider a test battery with both operational and experimental sections, and with examinees being told which sections are experimental. Here regression can use scores from the operational sections to predict performance on the experimental sections; a large negative standardized residual would be expected for an examinee who did not make a serious effort on the experimental section.

Little is known about the efficacy of the standardized residual approach to the identification of aberrant response patterns. In the second study described below, this approach is compared to IRT-based methods of appropriateness measurement.

Appropriateness Indices for Multi-Unidimensional Tests

A multidimensional test with locally independent items is called multi-unidimensional if it consists of unidimensional subtests. Therefore, a test with M subtests was considered for which there were M blocks of item scores $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m, \dots, \mathbf{u}_M$, with

$$\mathbf{u}_m = (u_{m,1}, u_{m,2}, \dots, u_{m,i}, \dots, u_{m,n_m}) \quad , \quad (1)$$

and with M corresponding abilities $\theta_1, \theta_2, \dots, \theta_M$. The probability that item score vector $\mathbf{u}$

$$\mathbf{u} = (u_{1,1}, \dots, u_{1,n_1}, u_{2,1}, \dots, u_{2,n_2}, \dots, u_{M,1}, \dots, u_{M,n_M}) \quad (2)$$

equals some specified vector of 0s and 1s, $\mathbf{u}^*$, given that $\theta_1 = t_1, \theta_2 = t_2, \dots$, and $\theta_M = t_M$ can be simplified in two steps. First, let

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m, \dots, \theta_M) \quad (3)$$

and

$$\mathbf{t} = (t_1, t_2, \dots, t_m, \dots, t_M) \quad . \quad (4)$$

Then

$$P(\mathbf{u} = \mathbf{u}^* | \boldsymbol{\theta} = \mathbf{t}) = \prod_{m=1}^M \prod_{i=1}^{n_m} P(u_{mi} = u_{mi}^* | \boldsymbol{\theta} = \mathbf{t}) \quad (5)$$

by local independence. Second,

$$P(\mathbf{u} = \mathbf{u}^* | \boldsymbol{\theta} = \mathbf{t}) = \prod_{m=1}^M \prod_{i=1}^{n_m} P(u_{mi} = u_{mi}^* | \theta_m = t_m) \quad (6)$$

because each subtest has been assumed to be unidimensional.

Functions of independent variables are also independent; therefore, multi-unidimensionality implies

$$P[f_1(\mathbf{u}_1) = f_1(\mathbf{u}_1^*), f_2(\mathbf{u}_2) = f_2(\mathbf{u}_2^*), \dots, f_M(\mathbf{u}_M) = f_M(\mathbf{u}_M^*) | \boldsymbol{\theta} = \mathbf{t}] = \prod_{m=1}^M P[f_m(\mathbf{u}_m) = f_m(\mathbf{u}_m^*) | \theta_m = t_m] \quad (7)$$

for arbitrary functions f_m defined on the M blocks of item scores. In this application, the functions are unidimensional appropriateness indices. Equation 7 is applied below to obtain indices for multi-unidimensional tests from various indices for unidimensional tests.

Suboptimal Indices

Standardized ℓ_0 indices. The index ℓ_0 (Levine & Rubin, 1979) is defined as the logarithm of the maximum of the likelihood function for the examinee's item response pattern. For a multi-unidimensional test, ℓ_0 can be written as a sum of unidimensional indices as follows:

$$\begin{aligned} \ell_0(\mathbf{u}^*) &= \log \max_{\mathbf{t}} P(\mathbf{u} = \mathbf{u}^* | \boldsymbol{\theta} = \mathbf{t}) = \log \max_{\mathbf{t}} \prod_{m=1}^M P(\mathbf{u}_m = \mathbf{u}_m^* | \theta_m = t_m) \\ &= \log \prod_{m=1}^M \max_{t_m} P(\mathbf{u}_m = \mathbf{u}_m^* | \theta_m = t_m) = \sum_{m=1}^M \ell_0(\mathbf{u}_m^*) \quad (8) \end{aligned}$$

Note that for unidimensional and multi-unidimensional tests, ℓ_0 can also be expressed as the logarithm of the likelihood function for pattern $\mathbf{u}^*$ evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\mathbf{u}^*)$ computed from the pattern

$$\ell_0(\mathbf{u}^*) = \log P[\mathbf{u} = \mathbf{u}^* | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{u}^*)] \quad (9)$$

In an attempt to obtain a statistic with roughly the same distribution in groups with different ability levels, ℓ_0 was "standardized" (Drasgow, Levine, & Williams, 1985; see also Molenaar & Hoijtink, 1990). Thus Equation 9 was replaced by

$$z(\mathbf{u}^*) = \frac{\ell_0(\mathbf{u}^*) - E[\ell_0(\mathbf{u}) | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{u}^*)]}{\text{Var}^{1/2}[\ell_0(\mathbf{u}) | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{u}^*)]} \quad (10)$$

That is, the index value for each pattern is replaced by a linear transformation of the value, with coefficients for the linear transformation depending on estimated ability. For the heuristic reasoning leading to this approach to standardization, see Drasgow, Levine, and Williams (1985). For a critical discussion of this approach, see Molenaar and Hoijtink (1990).

Using Equation 7, the conditional mean and variance of ℓ_0 for a multi-unidimensional test are easily expressed as functions of unidimensional conditional means and variances

$$E[\ell_0(\mathbf{u}) | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{u}^*)] = \sum_{m=1}^M E[\ell_0(\mathbf{u}_m) | \theta_m = \hat{\theta}_m(\mathbf{u}_m^*)] \quad (11)$$

and

$$\text{Var}[\ell_0(\mathbf{u}) | \theta = \hat{\theta}(\mathbf{u}^*)] = \sum_{m=1}^M \text{Var}[\ell_0(\mathbf{u}_m) | \theta_m = \hat{\theta}_m(\mathbf{u}_m^*)] \quad (12)$$

Formulas for unidimensional means and variances are given by Drasgow et al. (1985).

The IRT model used here for the analysis of dichotomous responses was the three-parameter logistic model (Birnbbaum, 1968). Therefore, the standardized ℓ_0 for this model was denoted z_3 . A generalization of the standardized ℓ_0 index for polychotomously scored responses was also evaluated in Studies 1 and 2. The multi-unidimensional standardized polychotomous index, denoted z_p , combines unidimensional log likelihoods, conditional means, and conditional variances in exactly the same manner as z_3 . Formulas for the unidimensional terms were given by Drasgow et al. (1985).

Standardized extended caution indices. Let $T2^{(m)}$ and $T4^{(m)}$ denote Tatsuoaka's (1984) second and fourth extended caution indices computed for the m th test. Tatsuoaka found that $E(T2^{(m)} | \theta_m) = 0$; she also provided expressions for $E(T4^{(m)} | \theta_m)$ and the conditional variances of $T2^{(m)}$ and $T4^{(m)}$. The standardized multitest extensions of these indices are

$$T2 = \frac{\sum T2^{(m)}}{[\sum \text{Var}(T2^{(m)} | \hat{\theta}_m)]^{1/2}} \quad (13)$$

and

$$T4 = \frac{\sum [T4^{(m)} - E(T4^{(m)} | \hat{\theta}_m)]}{[\sum \text{Var}(T4^{(m)} | \hat{\theta}_m)]^{1/2}} \quad (14)$$

Fit statistics. The squared standardized residual fit statistic described by Wright (1977) involves an item-by-item standardization of the dichotomously scored item responses. Let $P_{mi}(t_m)$ equal the probability of a correct response on item i of test m among examinees with ability t_m , and $Q_{mi}(t_m) = 1 - P_{mi}(t_m)$. A multitest extension of Wright's statistic is

$$F1 = \sum_{m=1}^M \sum_{i=1}^{n_m} \{ [u_{mi}^* - P_{mi}(\hat{\theta}_m)]^2 / P_{mi}(\hat{\theta}_m) Q_{mi}(\hat{\theta}_m) \} \quad (15)$$

The second fit statistic investigated was described by Rudner (1983). Here the statistic is

$$F2^{(m)} = R_m / V_m \quad (16)$$

where

$$R_m = \sum_{i=1}^{n_m} [u_{mi}^* - P_{mi}(\hat{\theta}_m)]^2 \quad (17)$$

and

$$V_m = \sum_{i=1}^{n_m} P_{mi}(\hat{\theta}_m) Q_{mi}(\hat{\theta}_m) \quad (18)$$

An extension to the multitest case is

$$F2 = \sum_{m=1}^M R_m / \left(\sum_{m=1}^M V_m \right) \quad (19)$$

Optimal Indices

The most powerful appropriateness index for a given form of aberrance is the likelihood ratio statistic (Levine & Drasgow, 1988). The likelihood ratio generally requires considerably more computation than do the suboptimal indices. However, the computations are not burdensome for two relatively short tests, and the index provides a useful benchmark. A multitest optimal index is described below, and a short-cut method used to compute it (Levine, 1991) is outlined.

The likelihood ratio statistic is the ratio of the probability of a response pattern computed from a model for aberrance divided by the probability computed with a model for ordinary behavior. Thus

$$LR(\mathbf{u}^*) = P_A(\mathbf{u}^*)/P_N(\mathbf{u}^*) \quad , \quad (20)$$

where $P_A(\mathbf{u}^*)$ is the probability of sampling an examinee with item response pattern $\mathbf{u}^*$ from a population of examinees conforming to a model for aberrant responding, and $P_N(\mathbf{u}^*)$ is the corresponding probability for ordinary examinees.

The probabilities P_A and P_N are computed by integrating likelihood functions. For a dichotomous, unidimensional model, for example,

$$P_N(\mathbf{u}^*) = \int \prod_i P_i(t)^{u_i} [1 - P_i(t)]^{1-u_i} f(t) dt \quad . \quad (21)$$

The numerator P_A is calculated as

$$P_A(\mathbf{u}^*) = \int P(\mathbf{u} = \mathbf{u}^* | \theta = t) f(t) dt \quad , \quad (22)$$

where the likelihood function $P(\mathbf{u} = \mathbf{u}^* | \theta = t)$ is derived from a model for aberrance. Levine and Drasgow (1988) provided a practical algorithm for computing likelihood functions for a large variety of types of aberrance.

For multi-unidimensional tests, the likelihood functions are functions of as many variables as there are subtests. Evaluating these integrals can be costly and time consuming for batteries with many subtests, but Levine (1991) described an algorithm for efficiently evaluating the integrals for unidimensional and multi-unidimensional tests. The key points for Levine's algorithm are:

1. Likelihood functions have an approximately normal shape—at least near their maxima—for nearly all examinees, and for the models for ordinary and aberrant behavior considered here. Thus, except for some rare response patterns, if $\ell(t)$ is the logarithm of the conditional probability of observing a specified response pattern given $\theta = t$, then for some constants $\alpha_1, \alpha_2, \alpha_3$,

$$|\alpha_1 t^2 + \alpha_2 t + \alpha_3 - \ell(t)| \quad (23)$$

will be small, provided $|t - \hat{\theta}|$ is small. (Of course, different values of α_1, α_2 , and α_3 must be used for each response pattern and model.)

2. The product of two functions with normal shape is again a function with normal shape. Equivalently, the sum of two quadratic polynomials is quadratic. This generalization holds for functions of more than one variable as well.
3. Ability distributions are generally normal, or are at least well approximated by normal distributions.
4. The integral of a normal density and, more generally, the integral of the exponential of a (negative definite) quadratic function can be calculated explicitly without numerical integration.

For one-dimensional tests, points 1 through 3 imply that the product of a normal ability density and an examinee's likelihood function generally will have a normal shape. Thus for one-dimensional

tests, polynomial regression is used to compute a quadratic approximation of the logarithm of the product of the density f and the examinee's likelihood function $P(\mathbf{u} = \mathbf{u}^* | \theta = t)$ to obtain values α_1 , α_2 , and α_3 such that

$$\log [P(\mathbf{u} = \mathbf{u}^* | \theta = t)f(t)] = \alpha_1 t^2 + \alpha_2 t + \alpha_3 \quad (24)$$

for t close to the maximum of the likelihood function. When there is only one subtest,

$$P(\mathbf{u} = \mathbf{u}^*) = \int P(\mathbf{u} = \mathbf{u}^* | \theta = t)f(t) dt \quad (25)$$

is approximated by

$$\int \exp(\alpha_1 t^2 + \alpha_2 t + \alpha_3) dt, \quad (26)$$

which can be evaluated explicitly without numerical quadrature. The numerator and the denominator of the likelihood ratio statistic are evaluated separately for each examinee in this way.

A pragmatic justification of the approximation lies in its high degree of agreement with numerical quadrature. Furthermore, high rates of detection of aberrant response patterns were obtained in the studies described below. A theoretical justification and additional details are given by Levine (1991; for a treatment of very long tests, see Chang, 1990).

The same reasoning was used in this study for multi-unidimensional tests, except that the likelihood functions for the unidimensional subtests were approximated separately, and then the multivariate density was introduced. Thus for a two-subtest battery, polynomial regression was used first to compute a function

$$\exp(\alpha_1 t^2 + \alpha_2 t + \alpha_3) \quad (27)$$

approximating the likelihood for the first subtest near its maximum, and a second function

$$\exp(As^2 + Bs + C) \quad (28)$$

approximating the second test likelihood function near its maximum. When the bivariate density $f(s, t)$ is normal with known means and covariance matrix, routine algebraic manipulations can be used to put the product

$$\exp(\alpha_1 t^2 + \alpha_2 t + \alpha_3) \exp(As^2 + Bs + C) f(t, s) \quad (29)$$

in the form

$$\exp(\text{quadratic function of } s \text{ and } t), \quad (30)$$

which can be integrated explicitly. For additional details, see Levine (1991).

Study 1: Simulated ASVAB Data

Purpose

Study 1 was conducted to answer several questions:

1. To what extent can aberrant responding on short tests be identified?
2. How effective are the suboptimal multitest appropriateness indices relative to optimal multitest appropriateness indices?
3. What are the upper limits on the detectability of certain benchmark forms of aberrance when information from several short tests is combined?

For the optimal indices to be truly optimal, all assumptions used to specify the index must be true. Therefore, data in Study 1 were simulated to satisfy perfectly all assumptions. In Study 2, an actual ASVAB dataset was used to evaluate the properties of the optimal and practical indices with data that failed to satisfy model assumptions in ways that were realistic.

Method

Data generation. A simulated ASVAB Arithmetic Reasoning (AR) test was the first of the two unidimensional tests. It is a 30-item, four-option, multiple-choice test. The simulated data were derived from a sample of $N = 2,978$ examinees that was taken from the *Profile of American Youth* (Office of the Assistant Secretary of Defense, 1982) dataset by selecting every fourth examinee. The LOGIST (version 2B) computer program (Wood, Wingersky, & Lord, 1976) was used to estimate three-parameter logistic item response functions (IRFs).

Option response functions (ORFs) for the incorrect options (with "omit" and "not reached" treated as a single incorrect option) were estimated by means of Levine's (1984; 1989a; 1989b) multilinear formula scoring (MFS) theory. Drasgow, Levine, Williams, McLaughlin, and Candell (1989) provided a detailed description of these analyses.

The 15-item Paragraph Comprehension (PC) test and the 35-item Word Knowledge (WK) test were pooled to form a second unidimensional test. The correlation between PC and WK is .82 (Ree, Mullins, Mathews, & Massey, 1982), and their correlation corrected for attenuation is .96. Consequently, fitting a unidimensional item response model to the 50-item combined test (WKPC) seemed justified.

As with the AR test, LOGIST was used to estimate IRFs, and MFS was used to estimate ORFs. Plots showing estimated curves and empirical proportions indicated good fits of both the IRFs and ORFs to the data. The IRFs and ORFs estimated from the AR and WKPC tests were used as the "true" IRFs and ORFs for Study 1.

For the first step in the simulation, a sample of 3,000 simulated response patterns was created and used as a test-norming sample. The IRFs previously estimated were used to determine probabilities of correct responses, and the MFS ORFs were used to determine the probabilities of incorrect options. θ s for the two tests were sampled from a bivariate standard normal distribution, with the correlation parameter set equal to .8 (after correcting for attenuation, the correlations of WK and PC with AR are about .8; see Ree et al., 1982). Thus, a vector (θ_1, θ_2) was sampled from a bivariate standard normal distribution with correlation .8 for each simulated response pattern; θ_1 and the AR IRFs and ORFs were used to simulate a polychotomously scored 30-item unidimensional test; θ_2 and the WKPC IRFs and ORFs were then used to simulate a polychotomously scored 50-item unidimensional test. The entire response vector of 80 items was taken as the data provided by one simulee.

The test norming sample was used to determine the item and test statistics required to compute the multitest practical appropriateness indices (z_3 , $T2$, $T4$, $F1$, $F2$), based on the three-parameter logistic model. This entailed two runs of LOGIST (one for the simulated AR and one for the simulated WKPC).

A sample of 4,000 ordinary response vectors and 16 samples of 2,000 aberrant response vectors each were created. The ordinary sample was generated in exactly the same manner as the test norming sample (except that different seeds were used for the random number generators). The 16 aberrant samples resulted from varying three factors: the type of aberrance (spuriously high or spuriously low), the severity of aberrance (mild or moderate), and the distribution from which simulated θ s were sampled.

Eight of the aberrant samples contained spuriously high response vectors, and the remaining eight samples contained spuriously low response vectors. Spuriously high response patterns were created

by generating ordinary response vectors as described for the ordinary sample, and then replacing separately a given percentage k of simulated responses (randomly sampled without replacement) with correct responses for each of the two simulated unidimensional tests separately. Spuriously low response patterns were also created by generating ordinary response vectors, randomly selecting without replacement a fixed percentage of items for each test, and replacing the responses to these items with random responses (i.e., a response was replaced by option A with probability .25, option B with probability .25, . . . , and option D with probability .25). Mildly aberrant response patterns were generated using $k = 15\%$ (i.e., 5 of 30 AR items and 8 of 50 WKPC items). Moderately aberrant response patterns were created using $k = 30\%$ (i.e., 9 of 30 AR items and 15 of 50 WKPC items).

Ability level of the aberrant sample was the third variable manipulated. A composite θ was computed for each examinee by the formula

$$\frac{\theta_1 + \theta_2}{[\text{Var}(\theta_1 + \theta_2)]^{1/2}} = (\theta_1 + \theta_2)/1.9 \quad (31)$$

This composite has a standard normal distribution. Composite θ s for the spuriously high samples were sampled from four parts of the standard normal distribution: very low (VL, 0th through 9th percentiles), low (LO, 10th through 30th percentiles), low average (LA, 31st through 48th percentiles), and high average (HA, 49th to 64th percentiles). These intervals were used because it was more important to detect spuriously high response patterns for low-ability examinees than for high-ability examinees. Similarly, it was more important to detect spuriously low responses by high-ability examinees. Consequently, composite abilities were sampled from four average- to high-ability strata for the spuriously low samples: very high (VH, 93rd percentile and above), high (HI, 65th through 92nd percentiles), and HA and LA. The ability percentiles used here corresponded approximately to the percentiles of the Armed Forces Qualification Test (AFQT) categories. (The AFQT is a composite of subtests from the ASVAB.)

Analysis. The practical appropriateness indices were computed for the 4,000 response vectors in the ordinary sample. The item and test statistics estimated from the test norming sample were used to compute all but one of these indices. This simulates the process by which practical appropriateness indices would be computed in many applications. The one exception was the standardized ℓ_0 index (z_p) computed from the polychotomously scored item responses. It was computed using the true ORFs and IRFs. This allowed for the estimation of ORFs from the test norming sample to be bypassed, which resulted in a significant reduction in computing time. (Despite the advantage gained by being computed from true rather than estimated ORFs, it is shown below that z_p falls short of some other indices; therefore, the advantage given to z_p is of little practical consequence.)

Optimal appropriateness indices were computed (using the true ORFs and IRFs) for the ordinary sample for four aberrant conditions: 15% spuriously high, 30% spuriously high, 15% spuriously low, and 30% spuriously low. Two optimal appropriateness indices were computed for each of these conditions. The first, LR_p , is the optimal index for *polychotomous scoring* of the item responses. The second index, LR_3 , results from using only the information in the *dichotomously scored* item responses. LR_3 is thus based on a submodel—the three-parameter logistic model—for the polychotomous data in which all incorrect responses were grouped together.

The practical appropriateness indices were computed for each of the 16 aberrant samples. In addition, the three-parameter logistic and polychotomous model 15% spuriously high optimal indices were computed for the four samples with this form of aberrance, the 30% spuriously high optimal indices were computed for the four samples with this form of aberrance, and so forth.

Optimal appropriateness indices are optimal for the specified form of aberrance (e.g., 15% spuriously high) in a population in which the ability density is normal for both the ordinary and aberrant populations, and each response vector is either ordinary or 15% spuriously high. In fact, the ordinary group had this ability distribution. By restricting the ability range for the aberrant group, the power of the index that was optimal for the population as a whole was determined in a particular subpopulation.

Evaluation criteria. The main criterion for evaluating the appropriateness indices was the proportion of aberrant response patterns correctly identified as aberrant when various proportions of ordinary response patterns were misclassified as aberrant. Specifically, the appropriateness index values for the 4,000 ordinary response patterns were sorted from most aberrant to least aberrant. Similarly, each set of 2,000 aberrant response patterns was sorted from most to least aberrant. Then each of the 6,000 index values was used as a cutting score to classify the 4,000 ordinary response patterns into false-positive and true-negative categories. The 2,000 aberrant response patterns were also classified into true-positive and false-negative categories. Dividing the number of true positives by 2,000 and the number of false positives by 4,000 provided a hit rate and a corresponding false-positive rate for each of the 6,000 cutting scores. See Hulin, Drasgow, and Parsons (1983, pp. 131-134) for an example illustrating the calculation of classification rates.

Hit rates at each false-positive rate were determined for all 16 aberrance conditions. This allowed for types of aberrant response patterns to be determined that had acceptably high detection rates, using both optimal and practical methods. The characteristics of response patterns that could not be detected were evident from examining the 16 aberrance conditions separately.

Results

Figure 1 shows rates of detection for the VL and LA groups subjected to the spuriously high manipulation. Results for the spuriously low manipulation applied to the HA and VH groups are shown in Figure 2. Both figures were constructed using a classification rule that produced a 5% false-positive rate. To conserve space, detection rates for the other conditions and other false-positive rates are omitted (these results, which are consistent with the trends seen in Figures 1 and 2, can be obtained from the authors).

Figures 1 and 2 show that multitest extensions of appropriateness indices provide better detection rates than their unidimensional counterparts. Figure 1a, which presents the results for the VL θ range, illustrates this: At a 5% false-positive rate for the 15% spuriously high condition, the polychotomous optimal index LR_p detected 46% of the aberrant response patterns when only the AR item responses were used, 57% when only the WKPC item responses were used, and 72% when the multitest extension (MT) was applied to the entire set of 80 items. Similar advantages of the multitest extensions can be seen in Figures 1 and 2.

For the 15% spuriously high manipulation applied to the VL group, Drasgow, Levine, and McLaughlin (1987) obtained a 50% detection rate for an optimal index for an 85-item *unidimensional* test when the false-positive rate was 1%. The *multitest* optimal index (MTO) in this study provided a 55% detection rate under these conditions. At false-positive rates of 3%, 5%, and 10%, MTO achieved detection rates of 67%, 72%, and 78% for the 15% spuriously high treatment; the detection rates in the earlier study were 64%, 70%, and 77%. For the 30% spuriously high treatment, MTO had detection rates of 88%, 92%, 94%, and 95% at false-positive rates of 1%, 3%, 5%, and 10%; the corresponding detection rates in the earlier study were 93%, 95%, 97%, and 98%. Thus, the polychotomous model MTO indices provided detection rates that were generally similar to the rates provided by the polychotomous model optimal indices for the long unidimensional test.

Figure 1
Rates of Detection of Spuriously High Response Patterns for the Simulation Dataset

Figure 2
 Rates of Detection of Spuriously Low Response Patterns for the Simulation Dataset

The hit rates for the multitest *practical* indices were less similar to those of practical indices on long unidimensional tests. The differences were particularly obvious for the spuriously high conditions, and can perhaps be best illustrated by comparing the detection rate of the best practical index to that of the optimal index. At a 1% false-positive rate for the 30% spuriously high treatment, the ratio in the present study was .85 (.75 for z_3 divided by .88 for LR_p). The corresponding ratio was .98 in the Drasgow, Levine, and McLaughlin (1987) study (.91 for $T2$ divided by .93 for LR_p). This ratio was .47 for the LA range in the present study; the ratio from the earlier study was .73.

These results reveal an important difference between unidimensional appropriateness measurement and multidimensional appropriateness measurement: z_3 , $T2$, and $T4$ have efficiently detected aberrant response patterns on long unidimensional tests in previous research; however, this finding was not replicated with the multitest extensions for some θ ranges. There were substantial differences in hit rates between practical and optimal multitest appropriateness indices (see Figures 1b, 1d, 2a, and 2c). Therefore, practical appropriateness indices were found to be relatively less effective in the multitest case than in the unidimensional case.

Study 2: Actual ASVAB Data

Purpose

Four separate issues were addressed in Study 2. The first was standardization: An appropriateness index is well standardized if its conditional distributions (given θ) vary very little across different ability levels. Standardization is important because it reduces the confounding between ability and measured appropriateness. Drasgow, Levine, and McLaughlin (1987) found z_3 and $F2$ to be well standardized when they were computed for a long *unidimensional* test. Study 2 examined the standardizations of the *multitest* extensions of z_3 , $F2$, and the other practical appropriateness indices.

In previous research (Drasgow et al., 1985; Levine & Drasgow, 1982), the unidimensional l_0 appropriateness index has provided similar rates of detection with real and simulated data. The second issue examined in Study 2 was whether the detection rates obtained for simulated multitest data generalize to actual multitest data.

For an optimal appropriateness index to be truly optimal, IRFs (and ORFs if the analysis is polychotomous) must be known and must fit the data, tests assumed to be unidimensional must truly be unidimensional, the correlation between ability on the two tests must be known, and the ability densities must be known. Because a real dataset was analyzed in Study 2, all of these conditions were violated. The third issue, therefore, was the *robustness* of optimal indices under conditions in which they are not truly optimal.

The final issue was the effectiveness of the standardized residual (SR) as an appropriateness index. It was created by regressing the total number-correct score from the AR and WKPC tests on the Math Knowledge (MK) and General Science (GS) tests,

$$\text{Predicted}(\text{AR} + \text{WKPC}) = \hat{\beta}_1 + \hat{\beta}_2\text{MK} + \hat{\beta}_3\text{GS} = 7.98 + 1.20\text{MK} + 1.88\text{GS} \quad , \quad (32)$$

and then standardizing the residual

$$\text{AR} + \text{WKPC} - \text{Predicted}(\text{AR} + \text{WKPC}) \quad , \quad (33)$$

as described by Cook and Weisberg (1982). The correlation between MK and AR (after correcting for attenuation) was .88, the corrected correlation between GS and WK was .94, and the corrected correlation between GS and PC was .90 (Ree et al., 1982). Large positive values of SR indicate spuriously high test scores, and large negative values of SR indicate spuriously low scores.

Method

Datasets. The *Profile of American Youth* sample provided the database for Study 2. The test norming sample consisted of responses of the $N = 2,978$ *American Youth* examinees analyzed in the first phase of Study 1. The AR and WKPC IRFs and ORFs estimated from this sample were used for all analyses in Study 2. Also, the statistics needed for the $T2$ and $T4$ appropriateness indices were obtained from this sample.

An ordinary sample of 2,716 response vectors was formed by selecting examinees 2, 6, 10, . . . , from the *American Youth* tape, and then deleting the data from the 262 examinees who failed to answer at least 77% of the items on both the AR and the WKPC tests. This criterion was based on the conclusion of Drasgow et al. (1985) that test scores of individuals who answer less than 77% of the test are very likely to be invalid measures of ability.

The remaining examinees on the *American Youth* tape (examinees 3, 4, 7, 8, 11, 12, . . .) were used to form six more samples. These were created by determining the frequency distribution of total score across both AR and WKPC tests (i.e., AR + WKPC), sorting into groups on the basis of the percentiles used for the AFQT categories, and removing examinees who answered fewer than 77% of the items on either the AR or WKPC tests. Score ranges and sample sizes are shown in Table 1.

Aberrant samples were formed exactly as in Study 1. Thus, 15% and 30% spuriously high treatments were applied to the four lowest ability groups, and 15% and 30% spuriously low treatments were applied to the four highest ability groups.

Analysis. Appropriateness indices were computed with IRFs and ORFs estimated in the first phase of Study 1. The correlation between θ_1 and θ_2 was assumed to be .8, and the ability density was assumed to be standard normal truncated to $(-5.0, 3.5)$ when computing optimal indices.

To evaluate index standardization, appropriateness indices were computed before the aberrance treatments on the six samples listed in Table 1. The resulting distributions of index scores for these samples of *ordinary* examinees who differed in ability were then compared to the distribution of the ordinary sample of 2,716 examinees. A well-standardized index will have similar frequency distributions in the ordinary sample and each of the six samples stratified on ability.

Finally, the 15% and 30% spuriously high treatments were applied to the four lowest ability groups shown in Table 1, and the 15% and 30% spuriously low treatments were applied to the four highest ability groups. All appropriateness indices were computed for the resulting 16 samples of aberrant response patterns.

Index standardization. Although each practical appropriateness index was standardized, the expressions for the conditional expectations and variances of the indices were obtained using the assumption that θ_1 and θ_2 were known. They were unknown, of course, so it was important to investigate the conditional distributions of the appropriateness indices for ordinary examinees. These conditional

Table 1
Score Ranges and Sample Sizes for θ Categories

θ Range	AR Plus WKPC Score Range	Sample Size
Very high	74 to 80	494
High	59 to 73	1,537
High average	50 to 58	941
Low average	39 to 49	959
Low	24 to 38	1,155
Very low	0 to 23	342

distributions were investigated using the responses of the examinees in the six samples shown in Table 1 prior to any aberrance manipulation.

Results

The accuracy of the standardizations of the practical indices is illustrated by Figure 3. This figure presents receiver operating characteristic (ROC) curves for seven practical appropriateness indices (see Hulin et al., 1983, for a detailed discussion of the use of ROC curves in appropriateness measurement). Abscissa values in all cases were determined from the ordinary sample of 2,716 examinees. For the top row of the figure, ordinate values were based on the responses of the 342 examinees in the VL range prior to any aberrance manipulation (i.e., this sample was simply an ordinary low-ability group). Ordinate values for the middle row of the figure were based on the LA sample, and the bottom row was determined from the VH sample. Only the lower-left quarter of each ROC curve is shown in order to conserve space and because the primary concern was with an index's standardization for low misclassification rates. Results for the other three ability ranges are not shown because they were consistent with the trends in Figure 3.

It is clear from Figure 3 that z_p , SR, and FI are not consistently well standardized. The z_3 index is reasonably well standardized across ability levels, although its performance for the highest ability level is somewhat disappointing. $F2$ is fairly well standardized across ability levels. The most surprising result, however, is the very accurate standardizations of the multitest extensions of $T2$ and $T4$. Their standardizations were not very good for a long unidimensional test (Drasgow, Levine, & McLaughlin, 1987); their standardizations were excellent in this study, except for the highest ability group.

Spuriously high conditions. Figure 4 presents detection rates (at a 5% false-positive rate) of the MT indices for the spuriously high manipulations applied to actual ASVAB data for two of the ability groups. A comparison of these results to the results for simulation data (Figure 1) reveals generally similar detection rates. The detection rates for the polychotomous model optimal index LR_p tended to be slightly decreased for the actual ASVAB data, but detection rates for the dichotomous model appropriateness indices were almost unchanged.

Of the practical appropriateness indices, z_3 was clearly the most effective for the lowest ability range (Figure 4a). $T2$ and $T4$ had rates very close to z_3 in the LA range. The other practical appropriateness indices (z_p , FI , $F2$, and SR) appeared to be less effective. Figure 4 shows that SR was very ineffective for detecting low-ability cheaters.

Although the detection rates for the spuriously high conditions were similar across the simulated and real datasets, there was an important difference that is not shown in Figure 4: Both the ordinary and the aberrant groups for the actual ASVAB datasets generally had larger index scores. For example, 1.6% of the 4,000 simulated ordinary examinees from Study 1 had z_3 scores less than -2.0, and 11.4% had z_3 scores less than -1.0. For the 2,716 *American Youth* examinees taken as the ordinary group, the corresponding rates were 3.4% and 16.2%. This trend was also apparent for $T2$, $T4$, and LR_3 . LR_3 , for example, had 4.2% and 12.9% of its values greater than 5 and 2, respectively, for the *American Youth* ordinary group, but only 1.8% and 7.7% for the Study 1 simulated ordinary group. Thus, the distributions of index scores for the *American Youth* ordinary group had more extreme values than did the distribution for the Study 1 simulated ordinary group. Detection rates of spuriously high examinees did not greatly decrease, however, because there were comparable shifts in the distributions of index scores for the aberrant samples.

Spuriously low conditions. Rates of detection (given 5% false positives) for two of the samples subjected to the spuriously low treatments are shown in Figure 5. The rates shown in this figure are

Figure 3
 Partial ROC Curves for Standardizations of the Practical Appropriateness Indices

Figure 4
Rates of Detection of Spuriously High Response Patterns for the American Youth Dataset

Figure 5
 Rates of Detection of Spuriously Low Response Patterns for the American Youth Dataset

somewhat lower than the comparable rates obtained with simulated data shown in Figure 2. The decreased detection rates occurred because the distributions of index scores for the *ordinary* sample from the *American Youth* dataset were shifted toward more extreme values. However, in contrast to the spuriously high conditions, the distributions of index scores for the spuriously low response patterns created from the *American Youth* dataset were basically unchanged from the simulated data conditions. Thus, the "signal" was unchanged, but the "noise" increased; therefore, the signal-to-noise ratio decreased.

Although the rates of detection of spuriously low response patterns were lower for the *American Youth* data than for the simulated data, some fairly high detection rates were nonetheless obtained. For example, LR_p detected 92% and 98% of the 15% and 30% spuriously low examinees, respectively, for the VH range at a 5% false positive rate.

Discussion

The transition from simulated data in Study 1 to real data in Study 2 was reasonably successful. Simulated low-ability cheaters in both studies were highly detectable by means of appropriateness measurement. For example, 88% of the *American Youth* examinees in the lowest ability range who were subjected to the 30% spuriously high treatment could be detected by the optimal LR_p index when the false positive rate was 5%, 93% could be detected by z_3 , and 83% could be detected by T_2 . The comparable detection rates for the simulated data were 94%, 90%, and 85%, which shows relatively little deterioration in detection rates for the real data.

There were moderate reductions in detection rates for the spuriously low conditions. For example, LR_p , LR_3 , z_3 , and T_2 identified 89%, 74%, 66%, and 55% of the simulation data in the 49% to 64% ability range, 15% spuriously low response patterns; for the *American Youth* data, the corresponding rates were 77%, 69%, 61%, and 50%.

Conclusions

Multitest extensions of practical and optimal appropriateness indices were found to provide very high rates of detection of aberrant response vectors in some important conditions. These high detection rates clearly justify the use of the well-standardized practical indices in many applied settings.

A simple approach to implementing a well-standardized index in an applied setting would be to use a simulation study to determine cutting scores that misclassify various proportions of ordinary response patterns. The user would then select the cutting score that misclassified ordinary responses at a rate considered to be tolerable. A more sophisticated analysis was illustrated by Drasgow and Guertler (1987), who used classification accuracy, the utility and disutility of correct and incorrect classifications, and the base rate of aberrance in a decision theoretic approach to determining a cutting score.

A limitation on the present results is imposed by the choice of the correlation (.80) between θ_1 and θ_2 . A reviewer suggested that item responses generated by two latent traits with such a high correlation could be tantamount to a single long unidimensional test, and that it might be interesting to evaluate the indices in the context of two uncorrelated tests. Two responses can be made to these comments. First, recent results show substantial drops in detection rates when simulated verbal and quantitative tests are incorrectly analyzed as a single unidimensional test (Drasgow, Levine, Williams, McCusker, Thomasson, & Lim, 1990, p. 24).

Second, and perhaps more important, a correlation of .80 was *deliberately* selected in order to study implications for major academic admissions and employee-selection tests. For example, some correlations (after correcting for attenuation) are: .73 between the SAT Verbal and Quantitative tests

(College Board, 1989-90, p. 19); .80 between the Enhanced ACT English and Mathematics tests (American College Testing Program, 1989, p. 44); .80 for the General Aptitude Test Battery Verbal Aptitude and Numerical Aptitude (Hartigan & Wigdor, 1989, p. 138; United States Employment Service, 1983, p. 18); and .77 for the ASVAB Arithmetic Reasoning and Word Knowledge tests (Ree et al., 1982, pp. 9, 21). Thus, the research presented here studied a situation that typifies tests currently in use.

In contrast to the high detection rates obtained by the best IRT appropriateness indices, some very low detection rates were obtained by the regression method. For example, only 2% of the *American Youth* very low ability 15% spuriously high response patterns were identified by SR at a 5% false positive rate; z_3 detected 69%. In fact, the results for SR are even worse than they appear: SR is based on a total of 130 items (30 AR items, 50 WKPC items, 25 MK items, and 25 GS items), whereas the IRT appropriateness indices used only 80 items. Considerably higher detection rates for the IRT indices would be expected if they used all 130 items. Consequently, it appears that IRT indices rather than the SR index should be used in operational settings.

In comparing the best practical appropriateness indices to optimal indices, it is clear that the practical indices were much better than the non-IRT SR measure, but that they sometimes fell considerably short of the detection rates of the optimal indices. Therefore, it seems reasonable to conclude that further research is needed to develop and validate better practical appropriateness indices.

Optimal indices were used as benchmarks for the evaluation of practical appropriateness indices. Such benchmarks are valuable for theoretical evaluations of a practical index because they allow for its efficacy to be determined in relation to the best possible index. At present, however, too little is known about the properties of optimal indices to advocate their use in applied settings. A study of the robustness of optimal indices to violations of their assumptions is currently under way.

In this and previous research (e.g., Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987; Levine & Drasgow, 1982), a relatively limited variety of test-taking anomalies have been considered. Specifically, all items have been viewed as equally likely to elicit spuriously high or spuriously low responses, all response patterns in an aberrant sample have been assumed to have exactly the same number of aberrant responses, and the aberrant group has been assumed to have the same ability distribution as the ordinary group. More general formulations of optimal indices can eliminate these restrictive conditions.

A variety of measurement problems might be studied with more general optimal indices, such as determining whether a response vector was generated by an ordinary examinee who barely exceeded some threshold ability, or by a very low-ability examinee who was cheating on a substantial number of items (e.g., 20 items). In this case, the ability density for the aberrant group would *not* be the standard normal density. Another problem that could be addressed by a more general optimal index arises when it is known that part or all of one subtest is no longer secure. The optimal index would test whether a response pattern was generated by an ordinary examinee or an examinee who had access to a specific subset of items. An optimal index could also be formulated that allows a range for the number of aberrant responses.

These refinements in optimal indices would allow for powerful detection of aberrant response patterns in practical settings. Now that a significant part of the theory necessary for more sophisticated optimal indices has been provided by Levine and Drasgow (1988), implementation of these theoretical notions is currently under way.

References

- American College Testing Program. (1989). *Preliminary technical manual for the Enhanced ACT Assessment*. Iowa City IA: Author.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- Chang, H. H. (1990, June). *Asymptotic posterior normality of IRT models*. Paper presented at the Office of Naval Research Contractor's Meeting on Model-Based Psychological Measurement, Portland OR, U.S.A.
- College Board. (1989-90). *ATP guide for high schools and colleges*. Princeton NJ: Educational Testing Service.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, 72, 10-18.
- 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., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59-79.
- 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.
- Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L. (1987). *Modeling incorrect responses to multiple-choice items with multilinear formula score theory* (Measurement Series 87-1). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.
- Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, 13, 285-299.
- Drasgow, F., Levine, M. V., Williams, B., McCusker, C., Thomasson, G. L., & Lim, R. G. (1990). *Evaluation of optimal appropriateness measurement for use in practical settings* (AFHRL-TP-89-41). Brooks Air Force Base TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington DC: National Academy Press.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones-Irwin.
- Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84-4). Champaign IL: University of Illinois, Department of Educational Psychology.
- Levine, M. V. (1989a). *Formula scoring, basic theory and applications* (Measurement Series 89-1). Champaign IL: University of Illinois, Department of Educational Psychology.
- Levine, M. V. (1989b). *Ability distributions, pattern probabilities, and quasi-densities* (Measurement Series 89-2). Champaign IL: University of Illinois, Department of Educational Psychology.
- Levine, M. V. (1991). *Approximating response pattern probabilities*. Unpublished manuscript.
- 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. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161-176.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-289.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika*, 55, 75-106.
- Office of the Assistant Secretary of Defense. (1982). *Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery*. Washington DC: Office of the Assistant Secretary of Defense, Manpower, Reserve Affairs, and Logistics.
- Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). *Armed Services Vocational Aptitude Battery: Item and factor analyses of Forms 8, 9, and 10* (AFHRL-TR-81-55, AD-A113465). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- 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.
- United States Employment Service. (1983). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of the general factors over specific factors in the prediction of job performance*. Washington DC: Division of Counseling and Test Development, Employment and Training Adminis-

tration, U.S. Department of Labor.
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: Ed-
ucational Testing Service.
Wright, B. D. (1977). Solving measurement problems
with the Rasch model. *Journal of Educational Mea-
surement*, 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 work reported here builds on theoretical research sup-
ported by Contract No. N00014-86K-0482, NR 442-1546,
from the Office of Naval Research, Michael V. Levine,
Principal Investigator. 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, Universi-
ty of Illinois, 603 East Daniel Street, Champaign IL
61820, U.S.A.