APPROPRIATENESS MEASUREMENT WITH POLYCHOTOMOUS ITEM RESPONSE MODELS AND SS. (U) ILLINOIS UNIV AT URBANA MODEL BASED MEASUREMENT LAB F. DRASGOW ET AL. APR 84 UNCLASSIFIED MEASUREMENT SER-84-1 N00014-79-C-0752 F/G 5/10 NL
Appropriateness Measurement
with Polychotomous Item Response Models
and Standardized Indices

Fritz Drasgow, Michael V. Levine
and Esther A. Williams

Department of Educational Psychology
University of Illinois
Champaign, IL 61820

April 1989

The report is a public release; distribution unlimited.

Copyright © 1989 by the United States Government.
**Title:** Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices

**Authors:** Fritz Drasgow, Michael V. Levine and Esther A. Williams

**Performing Organization Name and Address:** Model Based Measurement Laboratory, University of Illinois, Urbana, IL 61820

**Controlled Office Name and Address:** Personnel and Training Research Programs, Office of Naval Research (code 442PT), Arlington, VA 22217

**Abstract:** The test scores of some examinees on a multiple-choice test may not provide satisfactory measures of their abilities. The goal of appropriateness measurement is to identify such individuals. Earlier theoretical and experimental work considered examinees answering all, or almost all, test items. This article reports research that extends appropriateness measurement methods to examinees with moderately high nonresponse rates. These methods treat nonresponse as if it were a deliberate option choice and then attempt to measure the "appropriateness" of the pattern of option choices. Earlier studies used latent trait theory, item response theory, multiple choice test, appropriateness measurement, person fit, polychotomous models, appropriateness index, wrong answers.
only the dichotomous pattern of right and "not right" answers. A general polychotomous model is introduced along with a technique called "standardization" designed to reduce the observed confounding between measured appropriateness and ability. A standardized appropriateness index based on a polychotomous model yielded higher rates of detection of simulated spuriously low examinees than the analogous index based on a dichotomous model. However, the converse was true for simulated spuriously high examinees. Standardization was found to reduce greatly the interaction between ability and measured appropriateness.
Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices
Fritz Drasgow, Michael V. Levine and Esther A. Williams
University of Illinois

Running head: Appropriateness Measurement

Address correspondence to: Fritz Drasgow
Department of Psychology
University of Illinois
603 E. Daniel Street
Champaign, IL 61820
USA
Abstract

The test scores of some examinees on a multiple-choice test may not provide satisfactory measures of their abilities. The goal of appropriateness measurement is to identify such individuals. Earlier theoretical and experimental work considered examinees answering all, or almost all, test items. This article reports research that extends appropriateness measurement methods to examinees with moderately high nonresponse rates. These methods treat nonresponse as if it were a deliberate option choice and then attempt to measure the "appropriateness" of the pattern of option choices. Earlier studies used only the dichotomous pattern of right and "not right" answers. A general polychotomous model is introduced along with a technique called "standardization" designed to reduce the observed confounding between measured appropriateness and ability. A standardized appropriateness index based on a polychotomous model yielded higher rates of detection of simulated spuriously low examinees than the analogous index based on a dichotomous model. However, the converse was true for simulated spuriously high examinees. Standardization was found to reduce greatly the interaction between ability and measured appropriateness.
1. Introduction

An examinee's score on a standardized multiple choice test may fail to provide a useful measure of ability for various reasons. The score may be too high because the examinee began the test with memorized answers to several questions or because the examinee copied answers to several questions from a much brighter examinee. The score may be too low because the examinee (a) made an alignment error over a block of items, answering, say, the eleventh item in the tenth space, the twelfth item in the eleventh space, ...; (b) interpreted several very easy items in creative ways and came to well-reasoned, albeit scored-as-incorrect, answers; (c) was tested in an unfamiliar language; (d) failed to answer items on which he/she was able to eliminate several incorrect options; or, (e) worked with extreme care and consequently never reached easy items on a power test.

In all of these examples, the examinee often produces an unusual pattern of answers with relatively many easy items answered incorrectly and hard items answered correctly. Appropriateness measurement (Levine and Rubin, 1979; Levine and Drasgow, 1982, 1983a; Drasgow, 1982; Hulin, Drasgow and Parsons, 1983, Chapter 4) is a model-based attempt to control test pathologies by recognizing unusual patterns. A model is fit to the item response patterns of a large sample of presumably normal examinees. Subsequently, individual examinees and their response patterns can be ordered according to how well they are fit by the group model.
Earlier appropriateness measurement work was based on models for dichotomous data and therefore was limited in two important ways. The pattern of nonresponse, which may have high diagnostic value, was ignored. In fact, earlier studies were forced to exclude examinees with high rates of nonresponse or introduce ad hoc corrections for omitting. Secondly, the earlier studies failed to take cognizance of which wrong option was chosen and therefore probably were not as sensitive to some irregularities as they might have been.

The work reported in this paper is intended to advance appropriateness measurement in two ways. A method is introduced for extending appropriateness measurement to samples of examinees with moderately high rates of nonresponse. Simultaneously methods sensitive to option choice are introduced. It will be seen that in pursuing these goals, progress has been made in comparing the appropriateness of scores at different ability levels.
2. Review of Appropriateness Measurement Terminology, Findings and Problems

The goal of appropriateness measurement is to identify examinees with inappropriate scores solely from their response patterns. This is done in two steps. First, a general psychometric model is fit to a large sample of nominally normal examinees. Then an index of goodness of fit or appropriateness index is used to measure the degree to which each individual examinee's response pattern fits the model used to characterize normal behavior.

In the first large scale, systematic appropriateness measurement study, Levine and Rubin (1979) showed that under ideal conditions certain test anomalies were detectable. They modified simulated item response data to create answer sheets with spuriously high and spuriously low scores. Three types of appropriateness indices were found to classify normal and moderately aberrant examinees rather well. However, their study was limited to simulated data conforming the "three parameter logistic model" (Birnbaum, 1968). Furthermore, their use of simulation item parameters (rather than estimated item parameters) left open the question of how well appropriateness measurement would perform in applications requiring parameter estimation.

Levine and Drasgow (1982, 1983a) extended the basic results to more realistic conditions. They used actual and simulated data to study systematically the effects of the unavoidable inclusion of aberrant examinees in samples of nominally normal examinees and of errors in estimating item parameters. They found good aberrance detection with actual and simulated data despite model misspecification and parameter estimation error. However, like Levine and Rubin they only considered examinees who answered all or nearly all items and they ignored which wrong answer was chosen when a wrong answer was chosen.
The research reported in this paper extends earlier appropriateness measurement studies to (1) examinees with substantial nonresponse rates by using (2) polychotomous models and (3) standardized indices. Nonresponse to an A-option multiple choice item is coded as the "choice" of an A+1st option. In this way, every item is, in a technical sense, answered. A polychotomous psychometric model is developed to quantify the probability of each option choice (including the A+1st) at each ability level. Aberrance is measured by the goodness of fit of the polychotomous model.

Concern for omitting has focused our attention on conditional distributions of indices and the problems of comparing appropriateness index values at different ability levels. Examinees omitting different items are, in effect, taking different tests. A low appropriateness index value in the presence of substantial omitting may be less indicative of aberrance than a higher index value for another examinee with a different nonresponse pattern. Thus we have attempted to introduce a "common metric" for indices. Our strategy has been to divide examinees into relatively homogeneous groups by using a gross feature of the response pattern, to approximate the distribution of the index values within each group, and to use the approximated conditional distributions to define a transformation of indices to a common distribution. The "gross feature" is the maximum likelihood ability estimate, which we expected to reflect omitting rates. The common distribution was the standard normal. This process, which we call standardization, has been useful in controlling the confounding of ability and appropriateness.
3. Option Response Functions and a Constant Ability, Polychotomous Model

As a descriptive model for normal test taking behavior we have used the most general unidimensional, locally independent constant ability model that generalizes the three-parameter logistic model. It can be shown that any unidimensional model with three parameter logistic item characteristic curves and conditionally independent item responses is a special case of this model, which we call the histogram model.

To express the assumptions of the histogram model let

\[ V = < V_1, V_2, \ldots, V_n > \]

denote the random vector of option choices and

\[ \chi = < \chi_1, \chi_2, \ldots, \chi_n > \]

be a vector of constants indicating option choices. It is assumed that for a unidimensional ability random variable \( \theta \)

\[
\text{Prob}\{V_i = v_i, V_2 = v_2, \ldots, V_n = v_n | \theta = t\} = \prod_{i=1}^{n} \text{Prob}\{V_i = v_i | \theta = t\}
\]

(1)

for all \( \chi \) and real \( t \). Furthermore, it is assumed that if \( v_i^* \) is the correct option choice for the \( i^{th} \) item, then for some \( a_i, b_i, c_i \) and for all \( t \)

\[
\text{Prob}\{V_i = v_i^* | \theta = t\} = c_i + (1-c_i) \left[ 1 + \exp[-a_i (t-b_i)] \right]^{-1}
\]

(2)
This model is tentatively introduced, not as a plausible model for test taking behavior, but as an admittedly crude descriptive model for test data that may or may not adequately support the extension of appropriateness measurement techniques to polychotomous data with high omitting rates. The functions

\[ P_{ij}(t) = \text{Prob}\{\text{option } j \text{ is chosen on item } i | \theta = t\}, \quad j = 1, \ldots, A+1 \]  

(3)

generalize the item response function of item response theory and are called option response functions. Their estimation is discussed in the following section.

The likelihood of a response pattern can be easily expressed in terms of the option response functions. According to this model, the probability of sampling an examinee with response pattern \( v = v \) from the subpopulation of all examinees with ability \( \theta = t \) is

\[ P(V = v | \theta = t) = \prod_{i=1}^{n} \sum_{j=1}^{A+1} \delta_j(v_i) P_{ij}(t), \]  

(4)

where the first \( A+1 \) positive integers are used as scores for option choices and \( \delta_j(k) = 1 \) if \( k = j \) and zero otherwise.

This equation has been used to compute polychotomous maximum likelihood ability estimates. The dichotomous model ability estimates \( \hat{\theta}_d \) are obtained by maximizing the dichotomous model likelihood function

\[ \prod_{i=1}^{n} \left[ u_i P_i(t) + (1-u_i)Q_i(t) \right] \]  

(5)

where \( u_i \) is one or zero according to whether \( v_i \) is the correct option, \( P_i(\cdot) \) is the option response function of the correct option given in equation (2), and \( Q_i(t) = 1 - P_i(t) \).
4. Option Curve Estimation

Various techniques have been proposed for estimating option response functions. Bock (1972) selects a parametric form for the functions and computes maximum likelihood estimates. The results of Lord (1969, 1970), Samejima (1981), and Levine (1982) on ability density estimation are relevant since option characteristic curves can be represented as ratios of ability density estimates. In particular, the probability of sampling an examinee choosing the $j$th option from the subpopulation of examinees with ability $t$ can be written as

$$P_{ij}(t) = \frac{f_{ij}(t)}{f(t)}$$

where $P_{ij}$ is the proportion of examinees choosing option $j$ for item $i$, $f(t)$ is the probability density of $\theta$ and $f_{ij}(t)$ is the $\theta$ density in the subpopulation selecting option $j$ for item $i$. Thus, if ability distributions can be estimated (from the dichotomously scored data), then option response curves can be estimated with no further specification of their form.

An option response function is simply the regression item option score on ability, i.e.

$$P_{ij}(t) = \mathbb{E}(\delta_j(V_i)|\theta=t)$$

(6)

where, as before, $\delta_j(k) = 1$ if $j = k$ and zero otherwise.

We have taken the simple expedient of using large sample estimates of the regression of option score on estimated ability $\hat{\theta}$

$$\mathbb{E}(\delta_j(V_i)|\theta=t)$$

as $P_{ij}(t)$ estimates in this initial study. To obtain these estimates,
maximum likelihood ability estimates ($\hat{\theta}_d$'s) were computed for a large sample of examinees from dichotomously scored data. The examinees were grouped according to their $\hat{\theta}_d$'s. The proportion choosing each option for each ability group was used as an estimate of a point on an option response function. Linear interpolation was used between estimated points. Numerical details are given in Levine and Drasgow (1983b).

These crude estimates of option response functions are not consistent and will lead to systematic errors in ability estimates. Nonetheless, they permit us to begin an evaluation of appropriateness measurement strategies without first undertaking a major parameter estimation task.
5. The Indices and their Standardizations

In this report we are exclusively concerned with generalizations of the linear function of item scores

\[ L_o = \sum_{i=1}^{n} u_i \log P_{i} (\hat{d}) + (1 - u_i) \log Q_{i} (\hat{d}) . \]  

Here \( u_i \) is the dichotomous item score which is one if item \( i \) is answered correctly and zero if it is answered incorrectly and \( \hat{d} \) maximizes the dichotomous model likelihood function. \( L_o \) has the advantage of being fairly easy to compute. In comparative studies it was found to perform roughly as well as more elaborate indices (Drasgow, 1982; Levine & Rubin, 1979).

\( L_o \) is the maximum of the logarithm of the dichotomous model likelihood function. The obvious generalization of \( L_o \) to the histogram model is the maximum for the polychotomous model log likelihood function

\[ \max_{\hat{d}} \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_j (v_i) \log P_{ij} (\hat{d}) . \]

In as much as our histogram model likelihood function does not have a continuous first derivative and was complicated to maximize, the generalization

\[ L_{o,h} = \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_j (v_i) \log P_{ij} (\hat{d}) \]

was used. \( L_{o,h} \) is the logarithm of the histogram model likelihood function evaluated at the dichotomous model maximum likelihood ability estimate \( \hat{d} \).
As discussed in detail in Section 7 below, the distribution of $\ell_0$ was found to depend on ability. Therefore, two new indices were defined:

$$Z_3 = [\ell_0 - E_3(\hat{\theta}_d)] / \sigma_3(\hat{\theta}_d)$$ (10)

and

$$Z_{h} = [\ell_0,h - E_h(\hat{\theta}_d)] / \sigma_h(\hat{\theta}_d).$$ (11)

In these formulas $E_3$, $E_h$, $\sigma_3$ and $\sigma_h$ are conditional means and standard deviations for the three-parameter logistic and histogram model. $E_3(t)$ is the conditional expected value of the random variable $X_3(t)$

$$X_3(t) = \sum_{i=1}^{n} U_i \log P_i(t) + (1-U_i) \log Q_i(t)$$ (12)

computed using the three-parameter logistic model. Thus,

$$E_3(t) = E\{X_3(t) | \theta = t\} = \sum_{i=1}^{n} P_i(t) \log P_i(t) + Q_i(t) \log Q_i(t).$$ (13)

$\sigma_3(t)$ is the square root of the conditional variance

$$\sigma_3^2(t) = \text{Var}\{X_3(t) | \theta = t\} = \sum_{i=1}^{n} P_i(t)Q_i(t) \log \left( \frac{P_i(t)}{Q_i(t)} \right)^2.$$ (14)

Similarly

$$E_h(t) = E\{ \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_i(\ell_v) \log P_{ij}(t) | \theta = t\} = \sum_{i=1}^{n} \sum_{j=1}^{A+1} P_{ij}(t) \log P_{ij}(t)$$ (15)

and

$$\sigma_h^2(t) = \text{Var}\{ \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_i(\ell_v) \log P_{ij}(t) | \ell_v = t\} = \sum_{i=1}^{n} \sum_{j=1}^{A+1} \sum_{k=1}^{A+1} P_{ijk}(t) \log P_{ijk}(t) \log \left( \frac{P_{ijk}(t)}{P_{ik}(t)} \right).$$ (16)
These transformations were found to reduce greatly the dependence of \( z_o \) and \( z_{o,h} \) on ability. Their rationale is discussed in Section 7 below.
6. Data and Parameter Estimation

Responses of approximately 75,000 examinees to the April, 1975 Scholastic Aptitude Test-Verbal section (SAT-V) were obtained from the College Entrance Examination Board. A spaced sample of 3,000 response vectors was formed by selecting the responses of every twentieth examinee, beginning with the first examinee. Item responses were then scored as correct, incorrect, omitted, or not-reached and the LOGIST computer program (Wood & Lord, 1976; Wood, Wingersky & Lord, 1976) was used to estimate item and person parameters of the three-parameter logistic model. Version 2.8 of LOGIST and its default options were used. Convergence was obtained before the maximum number of iterations was reached.

These item parameter estimates were then used to construct histograms summarizing the pattern of option selection at various ability levels for the 49,470 examinees following the first 25,000 examinees in the data set. The ability estimate (θd) was computed for each of these examinees by the method of maximum likelihood for the dichotomously scored data. The item parameters estimated by LOGIST were used in these calculations. Omitted and not-reached items were ignored in the calculations; Lord's (1974) modified likelihood function was not used. Examinees were sorted into 25 ability strata based on estimated ability. The fourth, eighth, twelfth, ..., 96th percentile points from the standard normal distribution were used as cutting scores to form the ability categories. The frequencies of option selection were used as cutting scores to form each of the 85 SAT-V items in each ability category. Finally, these frequencies were converted to proportions. The proportion choosing option j of item i for, say, the lowest ability group was taken as the estimate of Pij(θ) for θ = -2.054, the second percentile of the standard normal distribution.
Proportions choosing option \( j \) of item \( i \) for the other 24 ability groups were taken as estimates of the values of \( P_{ij} \) at the 6th, 10th, \ldots, 98th percentile points of the standard normal distribution. Linear interpolation between estimated values of \( P_{ij} \) was used when an ability estimate was between percentile points. No estimate of \( P_{ij}(\theta) \) was defined outside the interval \([-2.054, 2.054]\).

We chose to use 25 ability groups because this number appeared to be the best compromise between: (1) the desire to reduce sampling fluctuations by including a large number of examinees in each ability category; and (2) the desire to reduce bias by averaging over a short range of abilities. Graphs of estimated curves and further details on the estimation procedure are in Levine and Drasgow (1983b).
7. Investigation of Appropriateness Indices in Samples of Normal Examinees

7.1. Samples with Unrestricted Omitting

To examine the distributions of the various appropriateness indices, three-parameter logistic maximum likelihood estimates of ability were computed for the first 500 response vectors in the data set. A total of 464 examinees had ability estimates $\hat{\theta}_d$ in the interval $-2.054 \leq \hat{\theta} \leq 2.054$. Then $\xi_0$ for the three parameter logistic and histogram models, $z_3$, and $z_h$ were computed for this sample of 464 nominally normal examinees.

A scatterplot of the three-parameter logistic $\xi_0$ index and $\hat{\theta}$ for the first 150 examinees is presented in Figure 1. The darkened circles plotted in Figure 1 are conditional means of $\xi_0$ for the subset of the 464 examinees with $\hat{\theta}_d \in [t-.3,t+.3]$, $t = -2.0, -1.8, ..., +2.0$. The dependence of $\xi_0$ on estimated ability is apparent in this figure. The mean $\xi_0$ for examinees with $\hat{\theta}_d$ less than -1.62 is approximately -36. For examinees with $\hat{\theta}_d$ near -1.0, the mean $\xi_0$ is about -42. Mean $\xi_0$ rises as ability increases, until it reaches roughly -30 for examinees with $\hat{\theta} > +1.64$. Thus an $\xi_0$ score of -42 is quite low in one group of normal examinees and average in a less able group of normal examinees.

---

Insert Figure 1 about here
Figure 1 shows that some adjustment of the three-parameter logistic $I_0$ index is necessary: the regression of $I_0$ on $\hat{\theta}_d$ is not a horizontal line. Because the conditional distribution of $I_0$ given $\hat{\theta}_d$ varies as a function of $\hat{\theta}$, it is difficult to interpret the magnitude of $I_0$ directly.

The histogram $I_{0,h}$ index is plotted against $\hat{\theta}_d$ in Figure 2 for the first 150 examinees. The darkened circles are conditional means computed in the same way as in Figure 1. The dependence of histogram $I_0$ is even more apparent in this figure.

If $\hat{\theta}_d$ were equal to $\theta$, the local independence assumption could be used to reduce the dependence of $I_0$ on ability. According to the local independence assumption, in the subpopulation of examinees with ability $\theta = t$ the item scores $u_i$ are independent. Therefore the sum $X_3(t)$ given in equation (12) is approximately normal with mean and variance $E_3(t)$ and $\sigma^2_3(t)$ given in formulas (13) and (14). Therefore, $z_3$ in equation (10) would be approximately normal ($0,1$) for both low and high ability examinees. This was the rationale supporting the indices $z_3$ and $z_{h}$. The final expressions in equations (13) through (16) provide approximations to the actual moments when parameter estimates are substituted for parameters. Of course $\hat{\theta}_d$ is
not equal to $\theta$, so the standardization is, at best, approximate and the above argument is merely heuristic. Nonetheless, it will be shown shortly the transformed $z$ indices are much less sensitive to distribution of ability than the distributions of the untransformed $z$ indices.

Figure 3 presents the scatterplot of $z_3$ and $\hat{\theta}_d$ for the first 250 examinees and conditional means obtained from the entire sample. For most values of $\hat{\theta}_d$, the conditional means are, as desired, close to the line $y = 0$. For examinees with extreme values of $\hat{\theta}_d$, however, the conditional means are slightly less than zero.
The scatterplot of $z_h$ and $\hat{\theta}_d$ is presented in Figure 4 for the first 250 examinees. Again, there is little relation between the standardized index and $\hat{\theta}_d$. The most striking feature of this plot is the abnormally large number of examinees with very small values of $z_h$. In the entire sample of 464 examinees, there were 20 examinees with index scores less than -2.40; the expected number of scores in this range for a standard normal variable is only 3.8.

To determine the cause of the unusually frequent small index scores, response vectors of examinees with $z_h$ less than -2.4 were inspected. Interestingly, many of these response vectors had very large numbers of omitted items. Of the 20 examinees with the smallest values of $z_h$, 11 examinees (55 percent) omitted 30 or more items. In contrast, only 16 of the 444 examinees with $z_h$ greater than -2.4 omitted 30 or more items, or 3.6 percent. Further inspection of the 27 examinees with 30 or more omits showed that their mean $\hat{\theta}_d$ was -.14 and their mean $z_h$ was -1.87. Thus, this group appears quite ordinary with respect to ability, but has very atypical response patterns in that they omit more than 35 percent of the test.

A second group of nominally normal examinees also had very low index scores. There were seven examinees (in the sample of 464) who reached less than 77 percent of the items on the test. Their average $z_h$ index score was -2.41.
To further investigate the relation between high omitting and $z_h$, the response vectors of examinees 501 through 1,000 were analyzed. Note that this sample serves to replicate findings from the first sample. A total of 456 of these examinees had estimated abilities in the range from -2.05 to 2.05, 23 omitted 35 percent or more of the test items, 6 reached less than 77 percent of the test items and 16 had $z_h$ values of less than -2.4.

The relations between omitting and $z_h$ that were noted for the first sample of examinees were confirmed in this second sample. In particular, the mean $z_h$ for examinees who omitted more than 35 percent of the test was -1.85. Seven of these examinees had $z_h$ values of less than -2.4. Even stronger results occurred for the six examinees who reached less than 77 percent of the test items: all six had $z_h$ scores of less than -2.4. The mean $z_h$ for this group was -3.08.

A total of 430 examinees in the second SAT-V sample omitted less than 35 percent of the test and reached 77 percent or more of the test items. In this group, 6 examinees had $z_h$ scores of less than -2.4. The expected number for a standard normal population is 3.5. In contrast, of the 26 examinees who omitted more than 35 percent of the test, or reached less than 77 percent of the test items, 10 had $z_h$ values of less than -2.4; the expected number is .2.

It is not surprising that high omitting rates and not finishing the exam cause $z_h$ to indicate aberrance. Perhaps the most important point to note about the relation between high omitting and $z_h$ is that it is not high omitting per se that causes very extreme $z_h$ values. Instead, it is the too frequent omitting of easy items, or items with very effective
distractors. For example, examinees who reached less than 77 percent of
the test items did not attempt several easy items; across the last 10
items on each of the two SAT-V subsections, there were 5 items with \( \hat{b}_i \)
values less than -1.0, and 9 items with \( \hat{b}_i \) values less than 0.0.

Because examinees who omit more than 35 percent of the test or reach
less than 77 percent of test items appear to receive spuriously low test
scores, they are excluded from all subsequent analyses.

7.2. Restricted Omitting Sample

To investigate further the distributions of \( z_h \) and \( z_3 \), a large
sample of nominally normal examinees was formed. First, three parameter
logistic maximum likelihood estimates of ability were computed for
examinees 10,001 to 14,000 on the SAT-V tape. Then, examinees who met
the following three criteria were included in the sample:

1. Less than 35 percent of the test items were omitted;
2. 77 percent or more of the test items were reached;
3. Estimated ability was in the range \(-2.05 \leq \hat{\theta} \leq 2.05\).

The \( z_h \) and \( z_3 \) appropriateness indices were computed for the 3478 examinees
who satisfied these criteria.

Figure 5 presents the cumulative frequency distributions for \( z_h \) and
\( z_3 \) in the sample of 3478 nominally normal examinees. The cumulative
distribution function of the standard normal distribution is also presented.

From Figure 5, it is apparent that the distributions of \( z_h \) and \( z_3 \) are
slightly asymmetric: there are relatively few examinees with index scores
between -2.0 and 0.0, and relatively many with index scores between 0.0 and
1.2. Both empirical distributions are significantly different from the
standard normal distribution \((\alpha = .01)\) by the Kolmogorov-Smirnov test.
For the purposes of appropriateness measurement, it is not essential that $z_h$ and $z_3$ follow standard normal distributions. It is important that each index be distributed similarly across values of $\theta_d$. Table 1 presents information concerning the left tails of the distributions of $z_3$ and $z_h$ within five mutually exclusive ability intervals. The left tails of the conditional distributions of $z_3$ are relatively similar across the five ability intervals. The largest difference between cumulative proportions at any cutting score is only .03. The left tails of the conditional cumulative proportions of $z_h$ exhibit less invariance; here the largest difference is .054.

The relatively large differences in conditional distributions of $z_h$ for different ability levels may result from the presence of truly aberrant response patterns in the sample of nominally normal response vectors rather than inaccuracies in the standardization approximation. (It will be shown in Section 8 that $z_3$ and $z_h$ are more sensitive to some types of aberrance for examinees of very high or very low ability.) To investigate this possibility, the research described in the present subsection was replicated using simulation data.
7.3. Samples Simulated According to the Three Parameter Logistic Model and Histogram Model

Samples of 4,000 simulated examinees were generated using the three parameter logistic model and histogram model. Hypothetical probabilities of correct responses on dichotomously scored items and hypothetical probabilities of option selection on polychotomously scored items were computed using the three parameter logistic model ICC estimates and histogram option response function estimates described in Section 6. For each simulated examinee, an ability was sampled from the standard normal distribution truncated to the interval \([-2.05, 2.05]\). Responses to 85 items were then simulated as 85 independent multinomials with response probabilities obtained by substituting sampled ability in the three parameter logistic ICCs and histogram option response functions.

Ability was estimated for each response vector by the methods described in Sections 7.1 and 7.2. Response vectors for which \(|\hat{\theta}_d| > 2.05\) were discarded so that the results described in this section would be comparable to the results presented in Section 7.2. The \(Z_3\) and \(Z_h\) indices were computed using estimated ability. Simulated examinees were then sorted into five ability categories on the basis of \(\theta\) (not \(\hat{\theta}_d\)).

The cumulative proportions of appropriateness index scores for the five ability intervals are shown in Table 2. Note that (1) there was no model misspecification or parameter estimation problem here because the item parameters and option response probabilities used to compute index values were identical to those used to generate response vectors; and (2) there were no truly aberrant response vectors present. Although the cumulative proportions in Table 2 tend to be somewhat smaller than the corresponding proportions in
Table 1, the overall pattern is similar. Again, the largest difference in conditional proportions for the three parameter logistic is .03. The largest difference for the $z_h$ proportions is .046, which again suggests that there is less invariance of the conditional $z_h$ distribution than for the $z_3$ distribution.

Shown in Table 3 are the conditional proportions of index values obtained when simulated examinees are sorted into ability categories on the basis of $\hat{\theta}_d$ rather than $\theta$. The cumulative proportions for $Z_3$ are similar to the proportions shown in Table 2. Curiously, $Z_h$ shows more invariance across ability categories in Table 3 than in Table 2.

7.4. Summary

The standardized $l_0$ indices, $Z_h$ and $Z_3$, have empirical distributions that are reasonably close to the standard normal distribution. The Kolmogorov-Smirnov tests indicate that $Z_h$ and $Z_3$ do not exactly follow the standard normal distribution. Furthermore, Tables 1 and 2 indicate that the distributions of $Z_h$ and $Z_3$ are not completely independent of estimated ability. However, Figure 5 and Tables 1, 2, and 3 suggest that these effects are fairly small. In addition, these tables show that high rates of detection of aberrant response vectors will not result solely from differences in ability distributions. For this reason, $Z_h$ and $Z_3$ are used as appropriateness indices in the next section.
8. Appropriateness Measurement with Standardized $\hat{\theta}$ Indices

8.1. Overview

In this section, we compare the distributions of the two appropriateness indices in samples of normal examinees to the distributions in samples of examinees whose response vectors have been modified to simulate spuriously high and spuriously low examinees. The power of an appropriateness index is indicated by the extent to which the index separates the normal and aberrant groups.

8.2. Normal and Aberrant Groups

The normal group consists of the 3,478 nominally normal, low omitting examinees with $-2.05 < \hat{\theta} < 2.05$ previously described.

The aberrant groups were formed by the following process. First, only examinees with less than 35 percent of the test omitted and 77 percent or more test items reached were considered. Then, starting with examinee 1,001 on the SAT-V tape, 300 examinees with estimated ability in each of the five ability categories were selected from the next 2,000 records. The $\hat{\theta}$ categories, termed quintiles, are:

<table>
<thead>
<tr>
<th>Quintile</th>
<th>$\hat{\theta}$ range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>[-2.05, -.80]</td>
</tr>
<tr>
<td>Q2</td>
<td>(-.80, -.24]</td>
</tr>
<tr>
<td>Q3</td>
<td>(-.24, .24]</td>
</tr>
<tr>
<td>Q4</td>
<td>(.24, .80]</td>
</tr>
<tr>
<td>Q5</td>
<td>(.80, 2.05]</td>
</tr>
</tbody>
</table>

These quintiles of response vectors were then subjected to various types of tampering to simulate aberrance.
The k% spuriously high modification consisted of randomly selecting k% of the examinee's original responses without replacement. Then each response was rescored as correct, regardless of the original response. Note that omits were treated as any other response category and rescored as correct if selected. Ten, 20, and 30% modifications were applied to each of the five quintiles.

The k% spuriously low modification was slightly more complex. First, each examinee's response vector was inspected to determine the proportion, q, of omitted items. Then k% of the examinee's original responses were selected randomly and without replacement. Each item was rescored as an omit with probability q. Options A through E were selected with probability (1-q)/5. Note that this procedure reflects the examinee's original propensity to omit items. Again, 10, 20, and 30% modifications were applied to the quintiles.

After tampering with the response vectors in a quintile, ability was estimated for each modified response vector using the three parameter logistic model. Then $z_h$ and $z_3$ were computed for the modified response vectors in the quintile.

8.3. ROC Curves

We used ROC curves to display the effectiveness of an index for detecting simulated aberrance. Here, a value of the appropriateness index, say $t$, is specified. Then the proportion of normal and aberrant response vectors with index values less than $t$ are determined. Let

$$x(t) = \text{proportion of normal examinees with index values } \leq t$$

$$y(t) = \text{proportion of aberrant examinees with index values } \leq t$$

Plotting the $< x(t), y(t) >$ pairs for several values of $t$ produces an
ROC curve. An ROC curve that indicates good detection of aberrance is one that rises sharply from the origin to the upper left hand corner of the plot. A random classification system would produce an ROC curve that lies along the 45 degree diagonal line. To conserve space, we only plot ROC curves for low false alarm rates: \( x(t) \leq .20 \). An elementary description of the use of ROC curves in appropriateness measurement is given by Hulin, Drasgow, and Parsons (1983).

8.4. Results for the Spuriously Low Modification

Figure 6 presents the ROC curves for the spuriously low modification, with panel (A) corresponding to the \( z_h \) index and panel (B) corresponding to \( z_3 \) index. The 30% modification is indicated by circles, the 20% modification by squares, and the 10% modification by a solid line. The 45 degree diagonal line is also plotted.

The panels in Figure 6 portray an orderly, coherent pattern of detectability. In each case, tampering with more items leads to greater detectability. This is indicated by the ROC curves for the 30% modifications always rising more sharply than the other two curves, and the 20% modification ROC curves rising more sharply than the 10% curves.

It is clear that detectability increases with increasing ability. For example, the lowest detection rates occur for the first quintile where examinees had estimated ability in the range -2.05 to -.80 prior to tampering. It is obvious that the spuriously low treatment would have relatively little
effect on the responses of these low ability examinees. In contrast, examinees in quintile 5 all had estimated ability in the range .80 to 2.05 prior to tampering. Here the effects of the spuriously low modification on each examinee's response vector are much larger, and this is reflected in very high detection rates. Note that detectability increases evenly as pre-tampering ability increases; there is not a particular ability level below which appropriateness measurement is completely ineffective and above which appropriateness measurement is quite effective.

Despite the crude estimates of the histogram model's option response functions, it is clear that the $Z_h$ index is substantially superior to the $Z_3$ index. ROC curves for $Z_h$ generally rise more sharply than the corresponding $Z_3$ ROC curves, and hence provide better aberrance detection.

A reason that the histogram model affords better aberrance detection is straightforward: aberrant responses, as conceptualized and simulated here, are essentially random. Thus easy items can be missed and some extremely improbable ($P_{ij}$ less than .01) incorrect options are selected. The dichotomous test model is sensitive to incorrect responses to easy items, but is insensitive to the pattern of incorrect option selection. In contrast, the $Z_h$ index is affected by the selection of an incorrect option.

8.5. Results for the Spuriously High Modification

Figure 7 presents the results for the spuriously high modification. Again, the 30, 20, and 10% modifications are indicated by circles, squares, and solid lines, respectively. Clearly, tampering with more items increases detectability. As expected, detectability decreases with increasing ability in Figure 7. Providing the answer key for, say, 20% of the exam to bright examinees has a relatively small impact on their answers, and is not likely to be detectable by appropriateness measurement techniques. In contrast, providing low
ability examinees with answers to 20% of the exam will have substantial
effects on their responses. As seen in Figure 7, this type of spuriously
high score is detectable, especially by the $z_3$ index.

Perhaps the most interesting result obtained from the spuriously high
modification is the finding that the $z_3$ index is clearly superior to the
$z_h$ index. This result appears counter-intuitive because the histogram model,
which provides a fuller description of the test-taking behavior of normal
examinees, should provide more power in detecting departures from normal
test-taking behavior. We believe that the superiority of the dichotomous
test model for detecting spuriously high examinees is chiefly a result of
the particular class of appropriateness indices under consideration (i.e.,
the $x_0$ class). This class of indices is subject to a "swamping" effect
when utilized to detect spuriously high response vectors on polychotomously
scored multiple choice tests. Other, more sophisticated indices may not be
affected similarly.

The swamping effect is perhaps best described by example. Table 4
presents the frequency distribution of the terms that compose $x_0$ and $x_{0,h}$
for the first examinee in Quintile 2:

$$x_{0,3}^{(i)} = u_i \log P_{i}^{(\hat{\theta}_d)} + (1-u_i) \log Q_{i}^{(\hat{\theta}_d)}$$

and

$$x_{0,h}^{(j)} = \sum_{j=1}^{A+1} \delta_j(y_j) \log P_{ij}^{(\hat{\theta}_d)}$$

respectively.
Note that prior to tampering there was only one term less than -2.0 for the three parameter logistic model but there were 17 such terms for the histogram model. After tampering, there were 5 terms less than -2.0 for the dichotomous model (an increase of 400%) and 22 terms for the histogram model (an increase of 29%). It is interesting to note that three of the smallest four \( e_0(i) \) terms for the logistic model after tampering were items rescored as correct during tampering. In contrast, none of the three smallest histogram terms had been subjected to tampering, and only two of the smallest 11 terms had been affected by tampering.

Insert Table 4 about here

This example illustrates the swamping effect. A normal number of relatively rare incorrect option selections and mistakes on easy items—as noted above, 9 for the examinee in Table 4—camouflages correct answers to hard items produced by the spuriously high modification. This occurs because the probabilities for some incorrect options are very nearly zero in the histogram model, and most examinees choose a few of these improbable incorrect options during the 85 item SAT-V exam. Swamping occurs much less in the three parameter logistic model because the model does not differentiate between the various incorrect options.
9. Discussion

From the research presented in this article it is apparent that standardization substantially reduces the confounding between measured appropriateness and ability: The conditional distributions of \( z_3 \) and \( z_h \) are more nearly invariant across ability levels than are the conditional distributions of \( l_o \) and \( l_{o,h} \). Thus, standardized index scores for examinees of different abilities can be compared more easily when making classification decisions.

It must be emphasized that our implementation of the standardization concept (i.e. transforming index values to make the conditional distribution of an appropriateness index independent of estimated ability) can be improved in many ways. An improved estimate of the conditional distribution of an index can be obtained, if not analytically then by simulation. A more "robust" estimate of ability can be obtained by reducing the relative contribution of improbable responses to the estimate (Wainer & Wright, 1980; Jones, 1982). Our studies show that standardization is needed and is easily implemented.

In pilot studies for the research described by Levine and Rubin (1979), it was found that \( l_o \) provided very low rates of detection of aberrant response patterns in samples of examinees with unrestricted omitting. Levine and Rubin found much higher detection rates when samples were restricted to low rates of omitting. To handle higher rates of omitting we have introduced polychotomous models. It is interesting to note, however, that standardization of the dichotomous model appropriateness index \( l_o \) allows high detection rates in samples with only weak restrictions on omitting rates. These detection rates seem to be nearly as high as detection rates for \( l_o \) in samples with low omitting rates.
Improved detection of response patterns modified by the spuriously low treatment was also obtained from use of the $z_h$ index. This index is sensitive to the pattern of incorrect option selection and consequently facilitates identification of examinees who choose unusual options when they respond incorrectly.

The $z_h$ index was not as effective as the $z_3$ index in identifying spuriously high response patterns. Thus, we are left with the question of whether our particular choice of a polychotomous model appropriateness index was unfortunate: Would a different polychotomous model appropriateness index provide much higher detection rates? In our current research, we have identified an optimal appropriateness index for spuriously high response patterns. Our preliminary results indicate that $z_3$ detects spuriously high patterns at a rate much closer to the optimal index than $z_h$, but not so well as to discourage refinements of $z_h$ and the formulation of an alternative polychotomous index.

Omitted items are ignored when computing the standardized dichotomous appropriateness index. The standardized polychotomous index, in contrast, treats nonresponse as the selection of the $(A+1)$th option on an $A$ option multiple choice item. This "option" is then treated in a fashion similar to the other options when computing $z_h$. Examinees who omitted a large number of items or who failed to reach many items frequently received very low $z_h$ scores. Because it seems likely that these examinees would receive higher test scores (which are a linear function of number correct minus one-fourth of number incorrect) if they answered more items, it appears that $z_h$ has identified one form of naturally occurring spuriously low test score.

The very low appropriateness scores observed among nominally normal examinees with high nonresponse rates can be seen as casting doubt on the unidimensional, local independence assumptions of the histogram model. It seems likely that there are substantial individual differences between exam-
inexes in rates of responding and willingness to guess or use partial information. These departures from unidimensionality, though obvious in retrospect, constitute an serendipitous finding of considerable practical importance. Excessively conservative examinees who are reluctant to use partial information, examinees who perseverate on difficult items and other able, low scoring examinees with high nonresponse rates indeed do have inappropriately low number right scores. It seems desirable to identify and counsel them. From the testing organization's point of view, it seems wise to exclude them from item parameter estimation samples since their presence may introduce additional sampling error (and possibly bias) in item parameter estimates.
References


Bock, R.D. (1972) Estimating item parameter and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.


Acknowledgements

This work was supported by ONR contracts N00014-79C-0752, NR 154 445 and N00014-83K-0397, NR 150 518. We are grateful to the College Entrance Examination Board for providing access to the Scholastic Aptitude Test data.
Table 1
Cumulative Proportions of Appropriateness Index
Scores at Various Cutting Scores for 3478 SAT-V Examinees

<table>
<thead>
<tr>
<th>Ability Interval*</th>
<th>Low</th>
<th>Mod. Low</th>
<th>Ave.</th>
<th>Mod. High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutting Index Z3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>Curve</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2.58</td>
<td>.005</td>
<td>.014</td>
<td>.006</td>
<td>.012</td>
<td>.007</td>
</tr>
<tr>
<td>-1.96</td>
<td>.025</td>
<td>.045</td>
<td>.025</td>
<td>.030</td>
<td>.028</td>
</tr>
<tr>
<td>-1.64</td>
<td>.050</td>
<td>.069</td>
<td>.047</td>
<td>.050</td>
<td>.055</td>
</tr>
<tr>
<td>-1.30</td>
<td>.097</td>
<td>.106</td>
<td>.079</td>
<td>.089</td>
<td>.085</td>
</tr>
<tr>
<td>Cutting Index Zn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>Curve</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2.58</td>
<td>.005</td>
<td>.006</td>
<td>.002</td>
<td>.006</td>
<td>.018</td>
</tr>
<tr>
<td>-1.96</td>
<td>.025</td>
<td>.035</td>
<td>.016</td>
<td>.008</td>
<td>.013</td>
</tr>
<tr>
<td>-1.64</td>
<td>.050</td>
<td>.055</td>
<td>.017</td>
<td>.020</td>
<td>.024</td>
</tr>
<tr>
<td>-1.30</td>
<td>.097</td>
<td>.094</td>
<td>.050</td>
<td>.040</td>
<td>.046</td>
</tr>
<tr>
<td>Total N in Ability Interval</td>
<td>650</td>
<td>643</td>
<td>643</td>
<td>672</td>
<td>870</td>
</tr>
</tbody>
</table>

* The ability intervals are: low = [-2.05, -0.80], moderately low = (-0.80, -0.24), average = (-0.24, 0.24], moderately high = (0.24, 0.80], and high = (0.80, 2.05].
### Table 2

Cumulative Proportions of Appropriateness Index Scores at Various Cutting Scores for Simulated Examinees

<table>
<thead>
<tr>
<th>Ability Interval*</th>
<th>Cutting Normal Index Score Curve</th>
<th>Low</th>
<th>Mod. Low</th>
<th>Ave.</th>
<th>Mod. High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z₃</td>
<td>-2.58</td>
<td>.005</td>
<td>.007</td>
<td>.005</td>
<td>.003</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>-1.96</td>
<td>.025</td>
<td>.032</td>
<td>.013</td>
<td>.015</td>
<td>.018</td>
</tr>
<tr>
<td></td>
<td>-1.64</td>
<td>.050</td>
<td>.047</td>
<td>.030</td>
<td>.035</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>-1.30</td>
<td>.097</td>
<td>.083</td>
<td>.058</td>
<td>.081</td>
<td>.063</td>
</tr>
<tr>
<td>Total N in Ability Interval</td>
<td>748 815 762 780 795</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z₄h</td>
<td>-2.58</td>
<td>.005</td>
<td>.004</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>-1.96</td>
<td>.025</td>
<td>.013</td>
<td>.005</td>
<td>.008</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>-1.64</td>
<td>.050</td>
<td>.042</td>
<td>.016</td>
<td>.022</td>
<td>.024</td>
</tr>
<tr>
<td></td>
<td>-1.30</td>
<td>.097</td>
<td>.079</td>
<td>.033</td>
<td>.054</td>
<td>.064</td>
</tr>
<tr>
<td>Total N in Ability Interval</td>
<td>745 816 762 780 791</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Simulated examinees were sorted into ability intervals on the basis of θ.

*The ability intervals are: low = [-2.05, -0.80], moderately low = (-0.80, -0.24],
  average = (-0.24, 0.24], moderately high = (0.24, 0.80], and high = (0.80, 2.05].
Table 3
Cumulative Proportions of Appropriateness Index
Scores at Various Cutting Scores for Simulated Examinees

<table>
<thead>
<tr>
<th>Ability Interval*</th>
<th>Cutting Normal Index</th>
<th>Low</th>
<th>Mod. Low</th>
<th>Ave.</th>
<th>Mod. High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Z₃</td>
<td>-2.58</td>
<td>.005</td>
<td>.005</td>
<td>.005</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.96</td>
<td>.025</td>
<td>.024</td>
<td>.022</td>
<td>.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.64</td>
<td>.050</td>
<td>.038</td>
<td>.046</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.30</td>
<td>.097</td>
<td>.062</td>
<td>.097</td>
<td>.080</td>
</tr>
<tr>
<td>Total N in Ability Interval</td>
<td>742</td>
<td>794</td>
<td>765</td>
<td>752</td>
<td>847</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z₄</td>
<td>-2.58</td>
<td>.005</td>
<td>.004</td>
<td>.000</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.96</td>
<td>.025</td>
<td>.011</td>
<td>.008</td>
<td>.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.64</td>
<td>.050</td>
<td>.034</td>
<td>.020</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.30</td>
<td>.097</td>
<td>.065</td>
<td>.049</td>
<td>.064</td>
</tr>
<tr>
<td>Total N in Ability Interval</td>
<td>759</td>
<td>754</td>
<td>793</td>
<td>747</td>
<td>841</td>
<td></td>
</tr>
</tbody>
</table>

Note: Simulated examinees were sorted into ability intervals on the basis of $\hat{\theta}$.

* The ability intervals are: low = [-2.05, -0.80], moderately low = (-0.80, -0.24], average = (-0.24, 0.24], moderately high = (0.24, 0.80], and high = (0.80, 2.05).
Table 4

Frequency Distribution of $l_o(i)$ Terms for the First Examinee in Quintile 2

<table>
<thead>
<tr>
<th>Interval</th>
<th>Three Parameter Logistic Model</th>
<th>Histogram Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution for Original Responses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1.0, 0.0]</td>
<td>74</td>
<td>37</td>
</tr>
<tr>
<td>(-2.0, -1.0]</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>(-3.0, -2.0]</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>(-4.0, -3.0]</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Omit</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>$Z_3 = 1.22$</td>
<td>$Z_h = 0.57$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution After 20% Spuriously High Modification</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1.0, 0.0]</td>
<td>70</td>
<td>39</td>
</tr>
<tr>
<td>(-2.0, -1.0]</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>(-3.0, -2.0]</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>(-4.0, -3.0]</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Omit</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>$Z_3 = -1.96$</td>
<td>$Z_h = -1.09$</td>
<td></td>
</tr>
</tbody>
</table>
Figure Captions

1. Three parameter logistic $\ell_0$ plotted against $\hat{\theta}_d$ for 150 nominally normal examinees.

2. Histogram $\ell_{0,h}$ plotted against $\hat{\theta}_d$ for 150 nominally normal examinees.

3. Standardized three parameter logistic appropriateness index $z_3$ plotted against $\hat{\theta}_d$ for 250 nominally normal examinees.

4. Standardized histogram appropriateness index $z_h$ plotted against $\hat{\theta}_d$ for 250 nominally normal examinees.

5. Cumulative proportions for the standard normal distribution and the standardized $z_3$ and $z_h$ appropriateness indices.

6. ROC curves for the spuriously low manipulations. Panel (A) presents results for the $z_3$ appropriateness index and panel (B) presents results for the $z_h$ appropriateness index.

7. ROC curves for the spuriously high manipulations. Panel (A) presents results for the $z_3$ appropriateness index and panel (B) presents results for the $z_h$ appropriateness index.
Figure 2
Cumulative Proportion vs. z Score

- ▲ Standard Normal
- □ z₃
- ○ zₗ

Fig. 5
Proportion of Spuriously Low Response Vectors Detected

Proportion of Normal Response Vectors Misclassified

(a)

Proportion of Spuriously Low Response Vectors Detected

Proportion of Normal Response Vectors Misclassified

(b)
Proportion of Spuriously High Response Vectors Misclassified

(a) Proportion of Normal Response Vectors Misclassified

(b) Proportion of Normal Response Vectors Misclassified
Dr. Ed Allen  
Navy Personnel R&D Center  
San Diego, CA 92152

Dr. Nick Bond  
Office of Naval Research  
Liaison Office, Far East  
APO San Francisco, CA 96503

Lt. Alexander Bory  
Applied Psychology  
Measurement Division  
NAHRL  
NAS Pensacola, FL 32508

Dr. Robert Carroll  
NAVDP 115  
Washington, DC 20370

Dr. Stanley Collyer  
Office of Naval Technology  
800 N. Quincy Street  
Arlington, VA 22217

CDR Mike Curran  
Office of Naval Research  
800 N. Quincy St.  
Code 270  
Arlington, VA 22217

Dr. John Ellis  
Navy Personnel R&D Center  
San Diego, CA 92252

DR. PAT FEDERICO  
Code P13  
NPRDC  
San Diego, CA 92152

Dr. Cathy Fernandes  
Navy Personnel R&D Center  
San Diego, CA 92152

Dr. Moreau J. Kerr  
Chief of Naval Technical Training  
Naval Air Station Memphis (75)  
Millington, TN 38054

Dr. Leonard Kroeker  
Navy Personnel R&D Center  
San Diego, CA 92152

Dr. William L. Maloy (02)  
Chief of Naval Education and Training  
Naval Air Station  
Pensacola, FL 32506

Dr. Kneale Marshall  
Chairman, Operations Research Dept.  
Naval Post Graduate School  
Monterey, CA 93940

Dr. James McBride  
Navy Personnel R&D Center  
San Diego, CA 92152

CDR Ralph McCumber  
Director, Research & Analysis Division  
Navy Recruiting Command  
4015 Wilson Boulevard  
Arlington, VA 22205

Dr. George Moeller  
Director, Behavioral Sciences Dept.  
Naval Submarine Medical Research Lab  
Naval Submarine Base  
Groton, CT 06340

Dr. William Montague  
NPRDC Code 13  
San Diego, CA 92152

Library, Code P201L  
Navy Personnel R&D Center  
San Diego, CA 92152

Technical Director  
Navy Personnel R&D Center  
San Diego, CA 92152

Commanding Officer  
Naval Research Laboratory  
Code 2627  
Washington, DC 20370

Personnel & Training Research Group  
Code 442PT  
Office of Naval Research  
Arlington, VA 22217

1 LT Frank C. Petho, MSC, USN (Ph.D)  
CMET (H-432)  
NAS  
Pensacola, FL 32508
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Bernard Rimland (OIC)
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Carl Ross
CNET-PDC
Building 90
Great Lakes NTC, IL 60088

1 Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152

1 Dr. Robert B. Smith
Office of Chief of Naval Operations
OF-98TH
Washington, DC 20350

1 Dr. Richard Snow
Liaison Scientist
Office of Naval Research
Branch Office, London
Box 39
FPD New York, NY 09510

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Frederick Steinheiser
CNO - OP15
Navy Annex
Arlington, VA 20370

1 Mr. Brad Syapson
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. James Tweeddale
Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Edward Weil
Office of Naval Research (Code 411S&F)
800 North Quincy Street
Arlington, VA 22217

1 Dr. Douglas Wetzle
Code 12
Navy Personnel R&D Center
San Diego, CA 92152
Civilian Agencies

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Thomas A. Wara
U. S. Coast Guard Institute
F. O. Substation 1B
Oklahoma City, OK 73169

1 Dr. Frank Withrow
U. S. Office of Education
400 Maryland Ave. SW
Washington, DC 20202

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Private Sector

1 Dr. James Algina
University of Florida
Gainesville, FL 326

1 Dr. Erling D. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 Psychological Research Unit
NBH-3-44 Attn
Northbourne House
Turner ACT 2601
AUSTRALIA

1 Dr. Alan Baddeley
Medical Research Council
Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF
ENGLAND

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69976
Israel

1 Dr. R. Darrell Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Robert Brenner
American College Testing Programs
P. O. Box 168
Iowa City, IA 52243

1 Dr. Glenn Bryan
6208 Pce Road
Bethesda, MD 20817

1 Bundesministerium der Verteidigung
Referat P 11 4
Psychological Service
Postfach 1328
D-5300 Bonn 1
F. R. of Germany
Private Sector

1 Dr. Ernest R. Cadotte
307 Stokely
University of Tennessee
Knoxville, TN 37916

1 Dr. John B. Carroll
409 Elliott Rd.
Chapel Hill, NC 27514

1 Dr. Horan Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. Allan M. Collins
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Lynn A. Cooper
LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavekant 2
2334 EN Leyden
The NETHERLANDS

1 CTB/McGraw-Hill Library
2500 Garden Road
Monterey, CA 93940

1 Dr. Dattpradad Divgi
Syracuse University
Department of Psychology
Syracuse, NY 3210

1 Dr. He-Ki Dong
Ball Foundation
Roe 314, Building B
BOC Roosevelt Road
Glen Ellyn, IL 60137

1 Dr. Fritz Drasgow
Department of Psychology
University of Illinois
603 E. Daniel St.
Champaign, IL 61820

Private Sector

1 Dr. Susan Embertson
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
Lawrence, KS 66045

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5625 Callaghan
Suite 225
San Antonio, TX 78236

1 Dr. Leonard Feldt
Lindquist Center for Measurement
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Dexter Fletcher
University of Oregon
Department of Computer Science
Eugene, OR 97405

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01002
Private Sector

1 Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3939 O'Hara Street
PITTSBURGH, PA 15260

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 DR. JAMES E. GREENO
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. Delwyn Harnisch
University of Illinois
242b Education
Urbana, IL 61801

1 Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 90010

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
603 East Daniel Street
Champaign, IL 61820

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Jack Hunter
2122 Coolidge St.
Lansing, MI 48906

Private Sector

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Douglas H. Jones
Advanced Statistical Technologies Corporation
10 Trafalgar Court
Lawrenceville, NJ 08648

1 Professor John A. Keats
Department of Psychology
The University of Newcastle
N.S.W. 2308
AUSTRALIA

1 Dr. Scott Kelso
Mansfield Laboratories, Inc
270 Crown Street
New Haven, CT 06510

1 CDR Robert S. Kennedy
Canyon Research Group
1040 Woodcock Road
Suite 227
Orlando, FL 32803

1 Dr. William Koch
University of Texas-Austin
Measurement and Evaluation Center
Austin, TX 78703

1 Dr. Stephen Kosslyn
1236 William James Hall
33 Kirkland St.
Cambridge, MA 02138

1 Dr. Alan Lesgold
Learning R&D Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801
Private Sector

1. Dr. Charles Lewis
   Faculteit Sociale Wetenschappen
   Rijksuniversiteit Groningen
   Oude Boteringestraat 23
   9712GC Groningen
   Netherlands

2. Dr. Robert Linn
   College of Education
   University of Illinois
   Urbana, IL 61801

3. Mr. Phillip Livingston
   Systems and Applied Sciences Corporation
   6201 Kendilornt Avenue
   Riverdale, MD 20840

4. Dr. Robert Lockman
   Center for Naval Analysis
   200 North Beauregard St.
   Alexandria, VA 22311

5. Dr. Frederick M. Lord
   Educational Testing Service
   Princeton, NJ 08541

6. Dr. James Lumsden
   Department of Psychology
   University of Western Australia
   Nedlands W.A. 6009
   AUSTRALIA

7. Dr. Don Lyon
   F. D. Box 44
   Higley, AZ 85236

8. Dr. Gary Marco
   Stop 31-E
   Educational Testing Service
   Princeton, NJ 08451

9. Dr. Scott Maxwell
   Department of Psychology
   University of Notre Dame
   Notre Dame, IN 46556

10. Dr. Samuel T. Mayo
    Loyola University of Chicago
    820 North Michigan Avenue
    Chicago, IL 60611

11. Dr. Robert McKinley
    American College Testing Program
    P.O. Box 168
    Iowa City, IA 52243

12. Dr. Barbara Means
    Human Resources Research Organization
    300 North Washington
    Alexandria, VA 22314

13. Professor Jason Millman
    Department of Education
    Stone Hall
    Cornell University
    Ithaca, NY 14853

14. Dr. Allen Munroe
    Behavioral Technology Laboratories
    1845 Elisa Ave., Fourth Floor
    Redondo Beach, CA 90277

15. Dr. M. Alan Nacewander
    University of Oklahoma
    Department of Psychology
    Oklahoma City, OK 73109

16. Dr. Donald A. Norman
    Cognitive Science, E-015
    Univ. of California, San Diego
    La Jolla, CA 92037

17. Dr. Melvin R. Novick
    SS6 Lindquist Center for Measurement
    University of Iowa
    Iowa City, IA 52242

18. Dr. James Olson
    NICAT, Inc.
    1875 South State Street
    Orem, UT 84057

19. Dr. Jesse Orlansky
    Institute for Defense Analyses
    1801 N. Beauregard St.
    Alexandria, VA 22311

20. Wayne M. Patience
    American Council on Education
    GED Testing Service, Suite 20
    One Dupont Circle, NW
    Washington, DC 20036
Private Sector

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 Dr. James M. Pellegrino
University of California, Santa Barbara
Dept. of Psychology
Santa Barbara, CA 93106

1 Dr. Paul D. Reckase
ACT
P. O. Box 150
Iowa City, IA 52243

1 Dr. Thomas Reynolds
University of Texas-Dallas
Marketing Department
P. O. Box 688
Richardson, TX 75080

1 Dr. Andrew M. Rose
American Institutes for Research
1555 Thomas Jefferson St., NW
Washington, DC 20005

1 Dr. Ernst R. Rothkopf
Bell Laboratories
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
400 Ela Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 Frank L. Schmidt
Department of Psychology
Bldg. 66
George Washington University
Washington, DC 20052

1 Dr. Walter Schneider
Psychology Department
605 E. Daniel
Champaign, IL 61820

Private Sector

1 Lowell Schoer
Psychological & Quantitative Foundations
College of Education
University of Iowa
Iowa City, IA 52242

1 Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigar
Fujusawa 251
JAPAN

1 Dr. Edwin Shirley
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. William Sims
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 21A, Yale Station
New Haven, CT 06520

1 Martha Stocking
Educational Testing Service
Princeton, NJ 08540

1 Dr. Peter Stollf
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 David E. Stone, Ph.D.
Reston, VA 20261

1 Dr. William Stollf
University of Illinois
Department of Mathematics
Urbana, IL 61801

1 Dr. Patrick Suppes
Institute for Mathematical Studies in
The Social Sciences
Stanford University
Stanford, CA 94305
Private Sector

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Kikusui Tatsuoka
Computer Based Education Research Lab
252 Engineering Research Laboratory
Urbana, IL 61801

1 Dr. Maurice Tatsuoka
220 Education Bldg
1510 S. Sixth St.
Champaign, IL 61820

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Douglas Towne
Univ. of So. California
Behavioral Technology Labs
1845 S. Elena Ave.
Redondo Beach, CA 90277

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. V. R. R. Uppuluri
Union Carbide Corporation
Nuclear Division
P. O. Box Y
Oak Ridge, TN 37830

1 Dr. David Vale
Assessment Systems Corporation
2233 University Avenue
Suite 310
St. Paul, MN 55114

1 Dr. Kurt Van Lehn
Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304

1 Dr. Howard Maimer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

Private Sector

1 Dr. Michael T. Waller
Department of Educational Psychology
University of Wisconsin--Milwaukee
Milwaukee, WI 53201

1 Dr. Brian Waters
InRAEC
300 North Washington
Alexandria, VA 22314

1 Dr. David J. Weiss
660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 Dr. Donald O. Weitzman
Mitre Corporation
1820 Welley Madison Blvd
McLean, VA 22102

1 Dr. Christopher Wickens
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90007

1 German Military Representative
ATTN: Wolfgang Wildegrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

1 Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

1 Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940
END
FILMED