Choice of Test Model for Appropriateness Measurement

Fritz Drasgow

University of Illinois at Urbana-Champaign

Several theoretical and empirical issues that must be addressed before appropriateness measurement can be used by practitioners are investigated in this paper. These issues include selection of a latent trait model for multiple-choice tests, selection of a particular appropriateness index, and the sample size required for parameter estimation. The three-parameter logistic model is found to provide better

detection of simulated spuriously low examinees than the Rasch model for the Graduate Record Examination, Verbal Section. All three appropriateness indices proposed by Levine and Rubin (1979) provide good detection of simulated spuriously low examinees but poor detection of simulated spuriously high examinees. A reason for this discrepancy is provided.

Item response theory (IRT) has been used by researchers to solve a variety of measurement problems that are very difficult to address using the methods of classical test theory (CTT). A number of examples have been provided by Lord (1980), Marco (1977), and Wright (1977). In this paper application of IRT to another problem apparently beyond the scope of CTT is considered. Levine and Rubin (1979) and Levine and Drasgow (in press-a, in press-b) refer to this application as "appropriateness measurement."

The goal of appropriateness measurement is to detect individuals for whom multiple-choice test scores are not representative measures of ability. A test score can be a misleading index of ability or achievement for many reasons. These reasons include atypical schooling, copying some answers from a more able neighbor, and exceptional creativity (see Levine & Rubin, 1979). The detection by appropriateness measurement of test scores that are unrepresentative measures of an examinee's actual ability is limited to test-taking anomalies producing recognizably unusual response patterns. For example, atypical education of a very bright individual may lead to an answer sheet containing several correct responses to very difficult items despite several incorrect responses to easy items.

An example of a quantitative measure of the atypicality of a response vector is provided by the ℓ_o appropriateness index described by Levine and Rubin (1979). Let u_i denote the i^{th} dichotomously scored (0 if incorrect, 1 if correct) item response, $P_i(\theta)$ denote the probability of a correct response to item i among examinees with ability θ , and $Q_i(\theta) = 1 - P_i(\theta)$. Then, using the local independence assumption of IRT, the ℓ_o index is the logarithm of the likelihood function evaluated at the maximizing value of θ :

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 3, Summer 1982, pp. 297-308
© Copyright 1982 Applied Psychological Measurement Inc.
0146-6216/82/030297-12\$1.60

$$\ell_o = \log \max_{\theta} \prod_i P_i(\theta)^{u_i} Q_i(\theta)^{(1-u_i)}. \quad [1]$$

The ℓ_o index can be motivated by the following observation (see Levine & Drasgow, in press-a, for a more extensive discussion). Suppose that some test-taking anomaly yields an answer sheet with incorrect responses to a number of easy items and correct responses to several difficult items. Then, the likelihood of this response pattern will be very small for all values of θ , and consequently the maximum of the likelihood function in Equation 1 will be very small. For this reason small values of ℓ_o suggest aberrance.

Levine and Rubin's (1979) Gaussian model provides two other quantifications of response vector atypicality. In this model an examinee has ability θ_i on item i , where θ_i is assumed to be randomly sampled from a normal (θ_g, σ^2) distribution. Levine and Rubin show that the likelihood of the responses $u_1, \dots, u_n$ given person parameters θ_g and σ^2 is

$$\begin{aligned} \text{Prob}(u_1, u_2, \dots, u_n \mid \theta_g, \sigma^2) \\ = \prod_i \int P_i(t)^{u_i} Q_i(t)^{(1-u_i)} \phi(t; \theta_g, \sigma^2) dt \end{aligned} \quad [2]$$

where

$$\phi(t; \theta_g, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} (t - \theta_g)^2 / \sigma^2\right]. \quad [3]$$

The analogue of ℓ_o in this model is denoted ℓ_n :

$$\ell_n = \log \max_{\theta_g, \sigma^2} \text{Prob}(u_1, u_2, \dots, u_n \mid \theta_g, \sigma^2). \quad [4]$$

Note that Equation 1 is the special case of Equation 4 for which all $\theta_i = \theta_g$.

The maximum likelihood ratio test statistic

$$LR = \ell_n - \ell_o \quad [5]$$

provides one measure of response vector atypicality, and the maximum likelihood estimate $\hat{\sigma}$ of the standard deviation of the θ_i distribution provides a second measure. Both LR and $\hat{\sigma}$ should tend to be relatively large when an examinee's responses appear to have been generated by more than a single ability θ_g . It is precisely these examinees for whom total test score is likely to provide an inappropriate measure of "ability." Levine and Rubin (1979) provide theoretical justifications for Gaussian model appropriateness indices, and Levine and Drasgow (in press-b) discuss computational methods.

Using real and simulated data, Levine and Rubin (1979) and Levine and Drasgow (in press-a) have found that ℓ_o , LR , and $\hat{\sigma}$ all can provide good detection of examinees with simulated *spuriously high* and *spuriously low* multiple-choice test scores. The present research continues and extends these earlier studies by addressing several theoretical and practical problems that must be examined before

multiple-choice test makers can implement appropriateness measurement. These issues include choice of an IRT model for multiple-choice tests, examination of the relative effectiveness of various appropriateness indices, and examination of the sample size that is required for parameter estimation.

Choice of IRT Model for Multiple-Choice Items

One of the most important considerations in appropriateness measurement is the choice of a functional form for P_i in Equations 1 and 4. The three-parameter logistic model (Lord, 1968, 1980) has been studied primarily in the context of multiple-choice items. In this model, the probability of a correct response to item i among examinees with ability θ is given as

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1 + \exp[-Da_i(\theta-b_i)]}, \quad [6]$$

where

a_i is the item discriminating power,

b_i is the item difficulty,

c_i is the lower asymptote of the item characteristic curve (ICC), and

D is a scaling constant usually set equal to 1.7.

This model has been criticized (see Wright, 1977) on the grounds that it has unavoidable difficulties in parameter estimation that preclude proper estimation of item discriminating power. In addition, practical experience has shown that the lower asymptote parameter can be very difficult to estimate by the method of maximum likelihood (Lord, 1968).

Estimators for the Rasch or one-parameter logistic model (Rasch, 1960), in contrast, have been shown to have optimal statistical properties. The Rasch model can be written as

$$P_i(\theta) = 1/\{1 + \exp[-D(\theta-b_i)]\}. \quad [7]$$

from which it is evident that this model is the special case of the three-parameter logistic model where $a_i = 1$ and $c_i = 0$ for all items. Andersen (1970, 1972) proved that conditional maximum likelihood (CML) estimation of item parameters for the Rasch model yields consistent estimates, where "conditional" refers to conditioning on the minimal sufficient statistic (number correct). Wright and Douglas (1977) observed that a test of more than 10 or 15 items cannot be calibrated by their CML method due to rapid accumulation of round-off error. Consequently, Wright and Douglas developed an unconditional maximum likelihood (UML) estimation procedure that simultaneously estimates item and person parameters. When a correction factor was applied to UML estimates, Wright and Douglas found that the resulting parameter estimates were indistinguishable from the CML estimates.

More efficient methods of controlling round-off error in CML estimation have been developed recently. Gustafsson (1980b) and Fischer (1981) have summarized this research. For the purpose of testing the goodness of fit of the Rasch model in a particular data set, it is important to use CML estimation (Gustafsson, 1980a, 1980b). In the present research, however, there is no assertion that the Rasch model provides a good fit to multiple-choice tests. Instead, the parameter estimates are used only to compute appropriateness indices. Consequently, due to the results of Wright and Douglas (1977), it is immaterial whether CML or UML estimates corrected for bias are used.

The choice between the Rasch model and the three-parameter logistic model for use in appropriateness measurement should be influenced by at least two considerations. First, when computing the appropriateness indices in Equations 1 and 4, how robust is the Rasch model to nonzero item characteristic curve lower asymptotes and varying item discriminations? Second, although three-parameter logistic parameter estimates have not been shown to be optimal, are these estimates sufficiently accurate for the purposes of appropriateness measurement? The trade-off between Rasch model robustness and three-parameter logistic model estimation accuracy is examined empirically in this paper.

A Modification of ℓ_o

The index ℓ_o is a sum of negative numbers, one for each item answered. Thus, ℓ_o is influenced by test length and number of items omitted. Dividing ℓ_o by the number m of items answered provides an index that is expected to be less sensitive to test length and omitting rate than ℓ_o . Moreover, the index

$$\ell_g = \exp(\ell_o/m) \quad [8]$$

can be interpreted as the geometric mean response probability. For these reasons ℓ_g was used in place of ℓ_o .

The Personal Biserial

In addition to the ℓ_g , $\hat{\sigma}$, and LR appropriateness indices, Donlon and Fischer's (1968) "personal biserial" correlation coefficient, r_{perbis} , is examined. These authors proposed r_{perbis} as a measure of the extent to which an individual's pattern of correct and incorrect responses matches the item difficulties of a normative group. A personal biserial correlation is computed between the dichotomously scored item responses of a particular individual and the group-determined item difficulties. In contrast, an item-total biserial correlation is computed across examinees for a given item. A low or even negative r_{perbis} can result when an individual answers many easy items incorrectly while responding correctly to several difficult items.

The measure of item difficulty used by Donlon and Fischer is the Educational Testing Service delta. The delta for an item is "... that value on the baseline of a normal curve with mean 13.0 and standard deviation 4 above which the area under the curve is equal to the proportion passing the item among those who reach it" (Donlon & Angoff, 1971, p. 23). The personal biserial can be computed from

$$r_{perbis} = \frac{\bar{\Delta}_r - \bar{\Delta}_c}{s} \frac{p}{h}, \quad [9]$$

where

$\bar{\Delta}_r$ is the mean delta for items reached,

$\bar{\Delta}_c$ is the mean delta for items answered correctly,

s is the standard deviation of the deltas for items reached,

p is the number of items answered correctly divided by the number of items reached, and

h is the ordinate of the normal (0,1) density function evaluated at the point where the cumulative distribution function equals p .

Method

Data

The Graduate Record Examination Board provided a magnetic tape with the item responses to the Graduate Record Examination (GRE) given by 10,000 examinees. Examinees' responses to items on the Verbal section of the GRE (GRE-V) constituted the data base for all analyses in the present study.

An attempt was made to control the distribution of ability in the normal and aberrant groups. This was done to establish that the detection rate achieved by an index is not an artifact of the relation between index score and ability. The reason for controlling ability distributions across normal and aberrant groups is described in more detail in the discussion section.

Two homogeneous groups of examinees were obtained by inspecting the responses of Examinees 5,001 through 10,000. A moderately low ability group was formed by selecting examinees with no more than 10 omitted items and a formula score (the number of correct responses minus one-fourth the number of incorrect responses) in the interval 30.5 to 37.0, inclusive. This interval corresponds to scaled scores from 420 to 460. A total of 115 examinees met these two specifications.

A high ability group was created by utilizing the same criterion for omitting rate; examinees in this group were required to have formula scores in the interval 58.5 to 63.75, inclusive. The scaled scores corresponding to these limits are 600 and 630. A total of 200 examinees formed the high ability group.

Aberrant examinees were created by a procedure similar to the method employed by Levine and Rubin (1979). A spuriously high response vector was generated from each examinee in the moderately low ability group by randomly selecting 20 item responses from the 95-item test without replacement. These items were then rescored to be correct, regardless of the examinee's original responses. If an omitted item was encountered, it was skipped and not counted toward the 20-item total. Hence, an omitted item was never rescored to be correct or incorrect. There were 115 simulated spuriously high response vectors generated from the moderately low ability group.

A spuriously low response vector was created from each examinee in the high ability group by randomly selecting 20 item responses without replacement. Each of these items was rescored as correct with probability 1/5 and incorrect with probability 4/5. Again, omitted items were not altered nor counted toward the total of 20 tampered items. There were 200 simulated spuriously low examinees generated from the high ability group.

The two types of aberrance simulated here may each be given interpretations that are quite realistic. For example, consider an individual who has memorized the answers to some exam questions. This examinee will obviously obtain a spuriously high score. Alternatively, consider a high ability individual for whom English is a recently acquired second language. There may be items containing idiomatic expressions that confuse this examinee. If he or she answers such items, the probability of a correct response is quite small and his or her score will be spuriously low.

The aberrance-inducing modifications systematically affected the distribution of formula scores. Table 1 contains the cumulative proportions of formula scores after the spuriously high and spuriously low modifications. These distributions are roughly comparable; consequently, a single group of normal examinees was selected from Examinees 5,001 through 10,000. The distribution of formula scores for the matched normal group also appears in Table 1. Examination shows that the match between normal and spuriously low groups is fairly good: The largest difference in cumulative proportion between the two frequency distributions is .05. The match between normal and spuriously high groups is less accurate, with the largest difference being .14.

Table 1
Formula Score Cumulative Proportions for
Aberrant and Normal Groups

Formula Score	Group		
	Spuriously High	Spuriously Low	Normal
39	.01	.01	.00
40	.03	.02	.02
41	.07	.04	.04
42	.17	.08	.08
43	.25	.20	.16
44	.30	.27	.24
45	.41	.37	.35
46	.57	.50	.47
47	.73	.64	.59
48	.83	.75	.71
49	.87	.81	.78
50	.94	.87	.86
51	.95	.91	.90
52	.98	.93	.94
53	1.00	.96	.98
54	1.00	.98	1.00
55	1.00	.99	1.00
56	1.00	1.00	1.00
57	1.00	1.00	1.00
Sample Size	115	200	255

Estimation

BICAL (Wright & Mead, 1975) was used to estimate the parameters of the one-parameter logistic model. UML estimates corrected for bias were obtained so that parameter estimates would be equivalent to CML estimates. Estimation for the three-parameter logistic model was done by LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976).

A computer program written by Gruvaeus and Jöreskog (1970) was used to estimate θ_g and σ for the Gaussian models. Initially, the program uses steepest descent to minimize the user's criterion function. Then, usually after two or three steepest descent iterations, the program utilizes the Fletcher-Powell (1963) method. Convergence is quite rapid once Fletcher-Powell minimization begins, usually requiring no more than three iterations. Finally, FORTRAN IV computer programs were written to calculate the delta measure of item difficulty and to compute r_{perbis} .

Results

Large Samples

The computations in this section are based on item parameter estimates obtained from analyzing the responses of 3,000 normal examinees (Examinees 1 through 3,000).

To evaluate the effectiveness of an appropriateness index in detecting aberrant response vectors, index scores for the normal and aberrant samples were calculated. Then, a cutting score t was specified. For convenience, assume that smaller index scores indicate aberrance. The proportion of aberrant response vectors with index scores less than t was determined. This proportion is referred to as the "hit rate." The proportion of index scores less than t in the normal sample was also determined. This is the "false alarm rate."

Hit rates for several false alarm rates are presented in Table 2. Detectability of spuriously low examinees was quite good for both the Rasch and three-parameter logistic models and for all appropriateness indices. Note that the three-parameter model performs slightly better than the Rasch model at most false alarm rates. The detection rates for all indices based on IRT models clearly dominate

r_{perbis}

Because Table 2 points to a modest superiority for the three-parameter model, comparison of the different indices will focus on the results for the three-parameter model. Table 2 shows that $\hat{\sigma}$ and LR provide almost equivalent detection of aberrance. At a 3% false alarm rate LR is more effective, at a 5% false alarm rate $\hat{\sigma}$ is more effective, and at other false alarm rates hit rates are quite close. The ℓ_g index yields slightly poorer detection for almost all false alarm rates. The most obvious conclusion from Table 2 is that all three appropriateness indices are effective in discriminating between normal and aberrant examinees.

Table 3 shows that detection is much more difficult for spuriously high examinees. Although over 60% of the aberrant examinees are detected when the false alarm rate is 25%, at low false alarm rates there is only mediocre detection. Table 3 indicates that for false alarm rates of less than 10% (1) the differences between the Rasch model and three-parameter logistic models are small, (2) there is little difference between appropriateness indices, and (3) comparatively poor detection results.

Table 2
Proportion of Spuriously Low Examinees
Detected at Several False Alarm Rates

Model and Index	False Alarm Rate					
	.01	.03	.05	.07	.10	.25
Rasch						
ℓ_g	.075	.250	.465	.535	.610	.870
$\hat{\sigma}$	.135	.265	.500	.580	.640	.885
LR	.095	.280	.465	.585	.630	.885
Three-Parameter Logistic						
ℓ_g	.125	.315	.495	.625	.690	.900
$\hat{\sigma}$	.190	.265	.580	.660	.740	.895
LR	.195	.330	.520	.630	.720	.890
r_{perbis}	.070	.270	.330	.415	.610	.865

Table 3
Proportion of Spuriously High Examinees
Detected at Several False Alarm Rates

Model and Index	False Alarm Rate					
	.01	.03	.05	.07	.10	.25
Rasch						
ℓ_g	.000	.070	.139	.217	.287	.635
$\hat{\sigma}$	.009	.061	.148	.191	.296	.643
LR	.000	.070	.139	.209	.304	.635
Three-Parameter Logistic						
ℓ_g	.000	.061	.139	.235	.304	.643
$\hat{\sigma}$	.026	.026	.156	.217	.296	.626
LR	.009	.070	.165	.226	.278	.643
r_{perbis}	.017	.087	.122	.156	.322	.609

Small Samples

The purpose of this analysis is to exacerbate the difficulties of estimating item parameters for the three-parameter logistic model by decreasing the test norming sample to only 500 examinees. This is a substantially smaller sample than the 1,000 examinees that Lord (1968) has suggested may be required for proper convergence of LOGIST. In contrast, a sample of 500 seems adequate for the Rasch model.

The sample of 500 was created by taking every 10th examinee from the GRE data tape, starting with Examinee 1 and ending with Examinee 4,991. LOGIST and BICAL were then utilized to compute item parameter estimates. Appropriateness indices were calculated using these item parameter estimates for the normal and spuriously low response vectors analyzed in the preceding section. These response vectors were employed so that any decrease in detectability of aberrance due to inaccurate estimates of item parameters would be apparent. Results for spuriously high examinees were not obtained because of their low detection rates, seen in Table 3.

Table 4 shows that detection of spuriously low examinees is still quite good. In fact, this table shows few differences from Table 2. For low false alarm rates, there is little evidence that appropriateness measurement has been degraded by estimating item parameters in a small sample. Comparing Table 4 to Table 2 indicates that although ℓ_g , $\hat{\sigma}$, and LR for the three-parameter model provide slightly different patterns of detection, the overall detection rates are very similar. For example, $\hat{\sigma}$ has a lower hit rate in Table 4 than in Table 2 for a false alarm rate of 1%; but at a false alarm rate of 3%, detection is slightly better. Overall, virtually no power has been lost by using item parameter estimates derived from a small sample. LR actually does somewhat better in Table 4 than in Table 2 at false alarm rates of 3% and higher.

From Table 4 it is apparent that the three-parameter model yields generally better aberrance detection than does the Rasch model for false alarm rates of 10% and less. This is particularly evident at false alarm rates of 1%, 5%, and 7%. For the three-parameter model, ℓ_g , $\hat{\sigma}$, and LR are all roughly comparable. However, $\hat{\sigma}$ and LR may be slightly more effective.

The small sample results are quite surprising. Originally, it was anticipated that estimating item parameters for the three-parameter model in a small sample would seriously degrade appropriateness measurement. In contrast, estimation for the Rasch model should be much less affected by the small sample size. Hence, it was expected that the Rasch model would provide greater aberrance detection than the three-parameter model. Table 4 indicates that this prediction is not correct, at least for an item parameter estimating sample of 500 examinees. The anticipated effect might still emerge in smaller samples.

Summary and Discussion

Appropriateness measurement does not seem greatly affected by the differences between the Rasch and three-parameter logistic models. Careful inspection of the results presented in the last two sections lends support to the conclusion that appropriateness indices based on the three-parameter logistic model yield generally better detection of spuriously low examinees at low false alarm rates. These results suggest that the three-parameter model, despite its estimation problems, is more useful for the purposes of appropriateness measurement for GRE-V data. Note that these results provide support for the various estimation methods programmed in LOGIST (Lord, 1968, 1974). In contrast, the Rasch model seems to be sensitive to the violations of its assumptions that occur in this data set.

It is also apparent that the three-parameter logistic model does not yield drastically superior aberrance detection. If cost is a major consideration to a potential user of appropriateness measurement, then indices based on the Rasch model may well be preferred. Initial item parameter estimation for the three-parameter model requires roughly seven to eight times as much execution time (on an IBM 360/75 computer) as item parameter estimation for the Rasch model. Further, the execution

Table 4
Proportion of Spurious Low Examinees Detected
at Several False Alarm Rates

Model and Index	False Alarm Rate					
	.01	.03	.05	.07	.10	.25
Rasch						
ℓ_g	.065	.250	.430	.520	.655	.875
$\hat{\sigma}$	.115	.250	.460	.580	.635	.895
LR	.085	.235	.430	.585	.635	.880
Three-Parameter Logistic						
ℓ_g	.120	.305	.545	.610	.700	.900
$\hat{\sigma}$	.170	.290	.580	.660	.715	.910
LR	.175	.340	.560	.645	.730	.895

time needed to compute appropriateness indices for the three-parameter model is roughly two to three times as great as the execution time required for the Rasch model.

Estimating item parameters in a moderately small sample did not appear to degrade appropriateness measurement. One implication of this finding is that appropriateness measurement appears feasible in much smaller samples than previously believed and thus may be useful to practitioners with data sets of modest size.

Although the personal biserial was able to detect spuriously low examinees reasonably well, this index was clearly dominated by all IRT-based appropriateness indices. The Rasch model ℓ_g index is not substantially more difficult to compute than r_{perbis} and provides higher hit rates at false alarm rates of 5% and 7%. Thus, a researcher desiring a convenient, inexpensive appropriateness index should apparently utilize the Rasch model ℓ_g index rather than r_{perbis} .

In a study designed to assess the effectiveness of appropriateness indices, it is important to control the distribution of ability. This is because the distributions of some indices are systematically related to ability. Drasgow and Levine (1982) have derived equations that show the mean of ℓ_o given θ changes substantially as θ varies. In particular, Drasgow and Levine found that the expected value of ℓ_o among normal Scholastic Aptitude Test—Verbal section (SAT-V) examinees with $\theta = 2.0$ was approximately -30 , whereas the expected value among examinees with $\theta = 0.0$ was about -45 . Thus, an examinee with $\theta = 0.0$ and $\ell_o = -45$ would appear quite aberrant if compared to an examinee with $\theta = 2.0$ and $\ell_o = -30$, even though both examinees have average index scores at their respective ability levels.

The implications of the relation between ℓ_o (and, hence, ℓ_g) and θ for the research presented here are twofold. First, since the match between formula score distributions for the normal group and the spuriously low group is fairly good, the high detection rates obtained for this aberrant group are *not* a statistical artifact. Second, note that the differences between formula score distributions for the spuriously high group and the normal group will tend to *favor* detection of these aberrant examinees. Consequently, the low detection rates in Table 3 are not an artifact.

In recent research (Drasgow & Levine, 1982), the distribution of ℓ_o has been investigated. When a test provides a substantial amount of information about ability (so that $\hat{\theta} \cong \theta$), it is possible to obtain approximate expressions for the mean and variance of ℓ_o . Moreover, the product in Equation 1 can be rewritten as a sum of n terms (one term for each item), which would be statistically independent if $\hat{\theta} = \theta$. It has been found that ℓ_o follows a distribution that is close to the normal distribution for SAT-V examinees. Thus, it is possible to compute a standardized ℓ_o index, which approximately follows a normal (0, 1) distribution. Research investigating this index is currently in progress.

It is unfortunate that the appropriateness indices examined in this research did not effectively detect spuriously high response vectors. At first glance, this result appears to contradict Levine and Rubin's (1979) finding that spuriously high response vectors were *more* detectable than spuriously low response vectors for simulated SAT-V data. However, a careful inspection of estimated item parameters indicates an important difference between the two examinations. In the present study using GRE-V data, only three items had lower asymptotes estimated to be less than or equal to .10 and a total of seven items had values of $\hat{c}$, less than or equal to .15. Hence, there were very few items for which it was extremely probable that low to moderate ability examinees would respond incorrectly. Thus, according to the three-parameter logistic model, it is likely that even low ability examinees will answer several difficult items correctly (as many low ability examinees actually did). Consequently, the spuriously high tampering procedure did not yield especially unusual vectors of item responses, because normal examinees of low to moderate ability commonly answered several difficult items correctly. In contrast, many of the items simulated in the Levine and Rubin study (obtained from Lord's, 1968, fit-

ting of the SAT-V exam) had small values of c_i . In particular, there were 20 items with c_i less than or equal to .10 and 42 items with c_i less than or equal to .15. Further, of the 20 items with c_i less than or equal to .10, 17 items had both a_i and b_i greater than unity. Thus, there were a substantial number of difficult ($b_i > 1$), discriminating ($a_i > 1$) items with near-zero lower asymptotes. Low to moderate ability examinees have very little chance of responding correctly to these items. If several are answered correctly (implying a spuriously high score), the examinee is conspicuous and, as found by Levine and Rubin (1979), detectable by appropriateness measurement.

Finally, it is important to realize that generalizations from the empirical results obtained here concerning choice of an IRT model for appropriateness measurement should be made with some caution. It appears reasonable to predict that the three-parameter logistic model will provide superior detection of spuriously low examinees on tests where almost all items have lower asymptotes of .2 or greater. The relative merits of the three-parameter and Rasch models with respect to appropriateness measurement are less clear for a test like the SAT-V. Because the SAT-V has many items with near-zero lower asymptotes, the Rasch model might be quite effective for the purposes of appropriateness measurement.

References

- Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 1970, 32, 283-301.
- Andersen, E. B. The numerical solution to a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, 1972, 34, 42-54.
- Donlon, T. F., & Angoff, W. H. The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program*. New York: College Entrance Examination Board, 1971.
- Donlon, T. F., & Fischer, F. E. An index of an individuals' agreement with group-determined item difficulties. *Educational and Psychological Measurement*, 1968, 28, 105-113.
- Drasgow, F., & Levine, M. V. *Advances in appropriateness measurement*. Manuscript in preparation, 1982.
- Fischer, G. H. On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, 1981, 46, 59-77.
- Fletcher, R., & Powell, M. J. D. A rapidly convergent descent method for minimization. *The Computer Journal*, 1963, 2, 163-168.
- Gruvaeus, G. T., & Jöreskog, K. G. *A computer program for minimizing a function of several variables* (Research Bulletin 70-14). Princeton NJ: Educational Testing Service, 1970.
- Gustafsson, J.-E. Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 1980, 33, 205-233. (a)
- Gustafsson, J.-E. A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, 1980, 40, 377-385. (b)
- Levine, M. V., & Drasgow, F. Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, in press. (a)
- Levine, M. V., & Drasgow, F. Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press, in press. (b)
- Levine, M. V., & Rubin, D. F. Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 1979, 4, 269-290.
- Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, 28, 989-1020.
- Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 1974, 39, 247-264.
- Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum, 1980.
- Marco, G. L. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 1977, 14, 139-160.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research, 1960.
- Wood, R. L., & Lord, F. M. *A user's guide to LOGIST* (Research Bulletin 76-4). Princeton NJ: Educational Testing Service, 1976.

- Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Bulletin 76-6). Princeton NJ: Educational Testing Service, 1976.
- Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, 14, 97-116.
- Wright, B. D., & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 1977, 37, 47-60.
- Wright, B. D., & Mead, R. J. *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1975.

Acknowledgments

This paper describes research performed in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Illinois at Urbana-Champaign. I express my gratitude to my dissertation director, Michael V. Levine, for his most valuable advice and guidance and for providing access to results he obtained in research partially supported by grants from Educational Testing Service, the Graduate Records Examination Board (No. 75-3), and the College Entrance Examination Board (PJ 82201). I also thank the Graduate Records Examination Board for allowing use of GRE data.

Author's Address

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