Estimating Item and Ability Parameters in Homogeneous Tests With the Person Characteristic Function

John B. Carroll
University of North Carolina, Chapel Hill

On the basis of monte carlo runs, in which item response data were generated for a variety of test characteristics, procedures for estimating item and ability parameters for homogeneous, unidimensional tests are developed on the assumption that values of the slope parameter *a* and the guessing parameter *c* are constant over items. The procedures focus on estimates of the *a* parameter, regarded as an important statistic for characterizing an ability. This parameter is estimated from person characteristic functions for different levels of the total raw score distribution. The procedures can be applied to datasets with relatively small or very large *N*s and with either relatively small or large numbers of items. They are illustrated with data from several cognitive ability tests. *Index terms: cognitive ability tests, homogeneous tests, item parameter estimation, item response theory, person characteristic function.*

Developments in item response theory (IRT) have taken two major directions. On the one hand, the three-parameter logistic (3PL) model developed by Birnbaum (1968) has been given much attention (Hambleton, 1983; Lord, 1980) and has given rise to computer programs, such as LOGIST and BILOG (Mislevy & Stocking, 1989), for estimating model parameters; these programs assume that items vary not only in difficulty *b*, but also in discrimination *a* and a *c* parameter. On the other hand, Wright and Stone (1979), and Andrich (1988), among others, have studied the simple logistic model (SLM) of Rasch (1960/1980), which is essentially a one-parameter special case of the 3PL model. The SLM model, as set forth by Wright and Panchapakesan (1969), has been the basis for computer programs such as BICAL (Wright & Mead, 1977), but estimation procedures for the model have easily been implemented on microcomputers. This model assumes that items vary only in difficulty; for all items the value of *a* is taken to be equal to 1, and the value of *c* is taken to be equal to 0. The seriousness of the debate over which of these models is to be accepted—either in general or for particular cases—is illustrated by Wright's (1984) highly critical review of several books (Hambleton, 1983; Hulin, Drasgow, & Parsons, 1983) that emphasized the usefulness of the 3PL model.

In view of this debate and the problems often encountered in using either of the models, Mosier's (1941) suggestion—that test theory be considered from the perspective of classical psychophysics—appears to be useful and informative. Thus, person responses would be examined as a function of item or task difficulty (i.e., the "stimulus"), and only secondarily as a function of person ability. This is somewhat in contrast to either the 3PL or the SLM models, which focus on probabilities of person responses as a function of person ability (even though both models do, of course, also consider item difficulties). This tendency on the part of conventional IRT is exemplified in the frequent use of the item characteristic curve (or item response function, IRF), which plots probability of correct response as a function of person ability for a particular item.

The function relating correct response proba-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 109–125

0146-6216/90/020109-17$2.10

bility to item difficulty inevitably requires the use of all (or nearly all) the items in a test, with their difficulties (or some function thereof) forming the baseline or abscissa of the curve. It would not be meaningful to plot such a curve for a single person because the probability values would consist only of 1s and 0s (for a dichotomously-scored test) and the "curve" could fluctuate violently from point to point. Nevertheless, it appears that Mosier (1941) used such curves in scoring tests by what he called a "constant process" derived from psychophysics; he found, however, that the reliability of scoring a test by determining threshold difficulty values with such curves was disappointing, as compared to that of traditional number-correct scoring.

It would seem more meaningful to consider a curve in which item probabilities are obtained by averaging them over a suitable group of persons—for example, persons making a particular test score, or any one of a small range of test scores, on the premise that the total test score is a sufficient statistic for estimating ability (Lord & Novick, 1968, p. 429). It is also possible to obtain a meaningful curve for a single individual by averaging probabilities over groups of items of similar difficulties. Weiss (1980, p. 444) mentions his "independent discovery" of Mosier's formulation, and called the resulting function the person characteristic curve. Carroll (1980, p. 449) also mentions early work with such a function, both with a pitch discrimination test (later developed and reported, Carroll, 1983), and an achievement test (Carroll & Schohan, 1953). Trabin and Weiss (1983) have studied the person response curve as a means of determining whether data from "real people" fit the IRT model, and others (e.g., Drasgow, Levine, & McLaughlin, 1987) have used such curves to measure what they call the "appropriateness" of test scores. For purposes of the present research, this model will be called the person characteristic function (PCF).

The PCF function studied here takes the same mathematical form as that of the standard 3PL function, on which the item characteristic curve is based, except that variation is examined over item difficulties b instead of over values of θ, the ability parameter, and it is assumed that the a and c parameters, respectively, are constant over all items:

$$p = c + \frac{1 - c}{1 + \exp[-1.7a(\theta - b)]} . \quad (1)$$

In most of the development in this paper the value of c is taken to be 0, but at a later stage a means of handling a nonzero value of c is suggested. In this way, the PCF preserves many of the features of the 3PL model; it differs from the 3PL model only in its assumption that a and c are constant over items.

Relation of the PCF Model to the Rasch Model

If c is taken to be zero and a is considered constant over all items, it is tempting to regard the PCF model as a simplified two-parameter model, or even as conforming to Rasch's SLM. It has similarities to the SLM because the latter dispenses with the a parameter by regarding it as constant over items, or more specifically, as equal to 1; actually, there is a "hidden," generally unnoticed, discrimination parameter in the SLM.

This can be demonstrated by generating item response data for different values of a according to the PCF model, and then estimating item difficulty and person parameters by the SLM—for example, by the maximum likelihood estimation procedures developed by Wright and Panchapakesan (1969). The variances of these parameters over items and over persons, respectively, increase markedly as a function of the value of a used in generating the item response data. Also, at least under these conditions, item difficulty values as estimated by the SLM have very high linear correlations with the logits and the normal deviate transforms of p_j, the probability of correct response to item j for all persons. (In 5 out of 6 cases examined, the correlation was slightly higher for the normal deviate transform.) Finally, the regression slopes of these correlations, which are the slopes of item difficulty value

on either the logit or the normal deviate of p, are clearly related to the value of a in the PCF model.

These results are exhibited in Table 1. Simulated item response data were generated for 100 persons whose θ values were distributed N(0,1). The test was assumed to contain 45 items, consisting of 5 items at each of 9 values of b that had a mean of 0 and standard deviation of 2. The items ranged from $b = -3.10$ to $b = +3.10$ in equal steps of approximately .77, or in theoretical p value from .999 to .001. The empirical, obtained p values had generally smaller ranges, depending on the value of a. For example, for $a = .4$, they ranged from .852 to .126, but for $a = 2$, they ranged from 1.0 to 0.0. Items with obtained $p = 0.0$ or 1.0 were, of course, excluded from the SLM computations. Table 1 shows statistics based on the SLM δ estimate of item difficulty on a logit scale, where $\text{logit}_p = \ln[p/(1-p)]$; and ξ, which is the normal deviate transform of p, the probability of correct response.

The results shown in Table 1 may be spurious to some extent, partly because the empirical item difficulties had different ranges, but the matter needs further study. The datasets certainly could have identical marginal values of p and frequencies of scores (X_j), but different patterns of cell values that would reflect different values of a. In any case, the results in Table 1 suggest that the SLM misses an important parameter. Lord and Novick (1968) point out that "Rasch's model in effect treats all items as having a discriminating power equal to the average discriminating power of the items" (p. 492).

The present paper shows that it is possible to estimate the a parameter by procedures that use the PCF. The a parameter, or an average over items, can of course be estimated through the full two- or three-parameter models, but estimation with the PCF procedures is probably simpler and more readily attainable for small samples or for datasets with large numbers of items.

The Meaning and Use of the *a* Parameter

In conventional use of two- or three-parameter models in which values of a are estimated separately for each item, this parameter is regarded as an index of item discrimination, and is useful as a guide for editing or eliminating poorly-discriminating items. In practice, values of a for items are rarely greater than about 2.0; in fact, LOGIST requires setting an upper bound for such values (Mislevy & Stocking, 1989). In the present context, however, where a is taken to be constant over the items of a test, it is regarded as indicat-

Table 1
Rasch Model Estimates and Related Data for Datasets Generated With Different Values of a Using the PCF Model

	a					
Statistic	.4	.7	1.0	1.3	1.6	2.0
Number of iterations in MLE computations	4	6	8	12	16	19
σ_δ	1.37	2.57	3.38	4.76	5.21	6.13
Logit_p						
Mean	−.05	−.13	−.09	−.37	−.05	.01
SD	1.22	2.06	2.39	2.71	2.74	2.58
ξ_p						
Mean	−.03	−.06	−.04	−.19	−.01	.00
SD	.73	1.16	1.31	1.45	1.47	1.39
$r(\delta,\text{logit})$	.99995	.99936	.99921	.99881	.99833	.99786
$r(\delta,\xi)$	.99989	.99984	.99984	.99923	.99996	.99810
Slopes: δ.logit	1.12	1.25	1.41	1.75	1.90	2.37
δ.ξ	1.87	2.23	2.57	3.27	3.55	4.40

ing something about the ability underlying the test, or at least something about the overall characteristics of the items comprising the test, especially after items with atypically low discriminations have been eliminated from the dataset.

It is hypothesized that abilities may vary significantly in the value of their characteristic *a* parameter—that is, some abilities may have a characteristically high value of *a*, such that PCFs are typically very steep for all individuals, and the probability of correct answering declines sharply as item difficulty increases. Other abilities would have a characteristically low value of *a*, such that PCFs have shallow slopes and the probability of correct answering declines very gradually as item difficulty increases. A high value of *a* might occur for abilities reflecting sensitivity to physical attributes of stimuli, as in the case of a pitch discrimination test. Lower values of *a* might occur for abilities concerned with a domain of knowledge, as in the case of a measure of knowledge of English vocabulary or of historical facts.

There is an analogy in the domain of psychophysics, where Stevens (1961) has shown that different aspects of sensory responses have different power functions. One of the chief motivations underlying the present research was to develop relatively simple methods for determining the characteristic values of *a* for different abilities. There was no intention of excluding use of standard two- and three-parameter models for such a purpose, but it was hoped that the PCF model could make the enterprise simpler. Even if the standard 2PL or 3PL model were used, focus of attention would be on the average values of *a* that are estimated over items or tasks, and on their variance, which would be expected to be relatively low for well-constructed tests.

Furthermore, use of the PCF (as opposed to the IRF) is seen as a more informative way of defining the nature of an ability, in that an ability is defined in terms of an individual's systematic variation in power of correct performance over a range of task difficulties. Characteristically, the PCF takes the form of a negative normal ogive or logistic curve; the slope of this function can be either gradual or steep, and is represented by the *a* parameter, which is hypothesized to be the same for all tasks and regions of the task difficulty continuum (and for all individuals). These considerations support the assumption, for estimation purposes, of a constant value of *a* over items and individuals. Tests of model fit would presumably indicate whether this assumption can be accepted.

It is not intended that the *a* parameter be used to differentiate abilities; indeed, two entirely different (uncorrelated) abilities could have similar *a* parameters. It would be of interest, however, if one ability were found to have a high value of *a*, and another ability were found to have a low value of *a*.

Method and Results

Overview

Simulated item response datasets were generated, with systematic variation of the parameters known to affect the characteristics of the datasets according to the PCF model and the assumptions underlying it. For each dataset, a number of statistics were determined that might be used to "back-estimate" the parameters used in generating the dataset, and the predictability of these statistics from the generating parameters was determined. These dependent variables were then used in multiple regressions as predictors of the generating parameters, and in particular the *a* parameter, because this parameter affects the estimation of other generating parameters. The result of this last step was a cross-validated procedure for predicting the generating parameters to a relatively high degree of accuracy. Finally, the procedure was applied to several sets of empirical data to investigate, at least provisionally, how well it could be relied on for such datasets.

Monte Carlo Runs

Method. Carroll, Meade, & Johnson (in press) developed and illustrated provisional procedures for deriving PCFs from test data.

They also reported on a series of monte carlo runs to investigate the properties of PCF parameters and their estimation. In that study research samples of test data were generated for a number of values of a, c, and N. Item parameters were restricted, however, to the case of 50 items with 10 items at each value of b_k = -2, -1, 0, 1, and 2. (Thus $\bar{b}$ was 0 and the variance of b was 2.0.) The present research extended this work by varying not only a but also the mean and standard deviation of item difficulties b_k (the normal deviate transforms of probabilities of correct response, p_k, for item set k), with the hope of obtaining more widely applicable procedures for estimating a.

By means of a program PCFGENA, written in BASIC (for Apple II+ or IIe microcomputers) and in TURBOBASIC (for IBM-PC or compatible microcomputers), 600 monte carlo runs were performed. 300 of these generated item response data (Sample A) for N = 100 persons, while the other 300 generated data (Sample B) for N = 400 in order to investigate the effect of N on the stability and accuracy of estimates. In order to study variance within a particular combination of generating parameters in each set of 300 runs, there were two runs (producing Samples A1 and A2, and Samples B1 and B2) for each of the 150 combinations of the following three parameters, each with several equally-spaced levels:

a: .3, .6, .9, 1.2, 1.5, 1.8

$\bar{b}$: -.8, -.4, 0, .4, .8

σ_b: .4, .8, 1.2, 1.6, 2.0.

For any given run, there were 9 sets of 5 items each, each set k with a constant value of b_k, equally spaced over a range, with the values of $\bar{b}$ and σ_b specified by the design. It was assumed that 9 sets would be sufficient to produce a suitable range of item difficulties, and that 5 items at each value of b_k would produce good estimates of empirical probabilities of correct responses. In every case, there were 45 items, which is in the range of typical lengths for many tests. It was assumed that the results could be generalized to cases with different numbers of items per set and with greater numbers of items overall.

The variation in parameters spanned the range that can be expected for typical test data, which included tests that were easy, of average difficulty, and difficult. It also included tests with varying ranges of item difficulty, and with score distributions of varying amounts of positive or negative skewness. An impression of the ranges of difficulty for the item sets can be gained from the following information: for $\bar{b} = 0$ and $\sigma_b = .4$, b_k for item sets ranged from -.62 to +.62, corresponding to p_k values of .268 to .732; for $\sigma_b = 2.0$, b_k for item sets ranged from -3.10 to 3.10, corresponding to p_k values of .001 to .999; for $\bar{b} = .8$ and $\sigma_b = .4$, b_k for item sets ranged from .18 to 1.42, centered at .80, corresponding to p_k values of .571 to .922, centered at .788; and for $\sigma_b = 2.0$, b_k for item sets ranged from -2.30 to 3.90, corresponding to p_k values of .011 to .99995. Item sets with extreme values of b_k were generally not functional in estimation of the PCF because they were likely to produce items that had values of $p_k = 0$ or 1 for all simulated examinees (simulees); data from such items were, in effect, edited out of the datasets.

The range of values of a would certainly include all except the most extremely deviant sets of actual test data, considering that values of a estimated by LOGIST for actual test data generally center around 1.0, and rarely exceed 1.8. They are also rarely smaller than .3, the lowest value used in the simulations, except possibly for items destined to be edited out of a dataset.

The value of the c parameter was set equal to 0 for each run, in order to avoid complications that would otherwise be introduced. (At a later stage discussed below, simulations were run with a value of c = .5, in order to investigate how to handle data with values of c greater than 0.)

Simulees were assumed to be normally distributed with respect to θ, with mean = 0 and variance = 1. For the given value of N, values of θ were assigned as the mean normal deviate

values of the successive N portions, each with area $1/N$, of the normal density function. Although it was possible to generate data for any desired values of mean θ and σ_θ, it was necessary to fix these values at 0 and 1, respectively, in order to obtain data from which univocal estimates of a could be made. This was because of the linear indeterminancy in the 3PL model (which also occurs in the PCF model) pointed out by Mislevy and Stocking (1989, p. 58).

Item reponses were generated by random number procedures—that is, a score of 1 for an item was assigned when a random number from a uniform 0,1 distribution equaled or exceeded the probability given by the PCF function; otherwise, a score of 0 was assigned. Care was taken to use random number generation procedures that were known to have excellent randomization properties. Use of the same PCF specified by Equation 1 for all items in a dataset guaranteed that all items could be assumed to measure the same ability.

Analysis and results. Depending on the value of N, 4,500 or 18,00 item responses of 0 or 1 were generated for each dataset. From these data, a number of statistics were derived. The mean, standard deviation, and skewness of the raw score distribution were computed, as well as the frequencies of particular scores. For purposes of generalizing the results to samples with different numbers of items, the means and standard deviations were also obtained in terms of proportion correct. The reliability was obtained as the square of the correlation of raw score with the true score assigned to each simulee. Item probabilities were averaged over the items in each of the nine item sets; variation among the items in each item set was ignored because it could only be due to random fluctuations. The frequency distribution of raw scores was partitioned into $n = 9$ (or sometimes fewer) groups, such that each group (except possibly the end groups that would not be used in PCF computations) contained as close as possible to N/n cases. This was done by an algorithm that insured that the cases would be partitioned in the same way regardless of whether counting might start from the bottom or the top of the distribution—the algorithm actually operated by counting outward in both directions from the score at the median.

The resulting score groups were called n-iles, or if n was 9 (as was usually the case), noniles. An $n \times 9$ matrix was then formed of the probabilities p_{jk} of correct response for each combination of n-ile and item set, with corresponding marginal values for rows and columns, and j was the subscript for n-ile. On the average, cell values were based on $5N/n$ item responses; row marginals were based on $45N/n$ responses, and column marginals for item sets were based on $5N$ responses, where n was usually equal to 9, but was occasionally 7 or 8 (for highly-skewed score distributions). Table 2 presents, for illustration, a typical matrix generated for $N = 100$, $a = 1.2$, $\bar{b} = .4$, and $\sigma_b = 1.6$. For the data in Table 2, $\bar{p}$ was .431, and σ_p was .180.

Next, the probabilities p_{jk} were transformed to normal deviates ξ_{jk}. The marginal probabilities p_k were also transformed in this way to produce values of b_k', defined as empirical item-set difficulty values. (For any datasets in which $c > 0$, these transformations were made only after correction for the c parameter by the formula $p' = (p - c)/(1 - c)$. The computer program for generating and analyzing data incorporated this formula, but of course did not change the resulting values of p_{jk}' from p_{jk} when $c = 0$.) Any probabilities that were equal to 1.0 or 0.0 (or $<$ 0.0 in the case of correction for chance) were assigned normal deviate transforms for $p = 1 - 1/(2N)$ or $1/(2N)$, respectively.

Values were transformed to normal deviates to examine the linear slopes of the values against empirical item difficulty values b_k', the normal deviate transform being regarded as an appropriate metric for this purpose. The resulting values after this operation on Table 2 are shown in Table 3. The table also shows the values of b_k (theoretical item-difficulty values) that were used in generating the data for the nine item sets and

Table 2
Proportion Correct for Score Groups and Item Sets, From Data Generated for $a = 1.2$, $\bar{b} = .4$, $\sigma_b = 1.6$, $N = 100$

Difficulty and Score Group	Item Set 1	2	3	4	5	6	7	8	9	Mean	N_j
b_k	-2.08	-1.46	-.84	-.22	.40	1.02	1.64	2.26	2.88		
Score											
0-2	.200	.000	.000	.000	.000	.000	.000	.000	.000	.022	1
3-10	.783	.450	.117	.067	.000	.000	.000	.000	.000	.159	12
11-14	.929	.843	.500	.171	.029	.000	.000	.000	.000	.275	14
15-17	1.00	.867	.680	.387	.160	.027	.013	.013	.000	.350	15
18-19	.980	.960	.820	.560	.300	.080	.000	.000	.000	.411	10
20-22	1.00	.985	.815	.631	.569	.169	.046	.000	.000	.468	13
23-25	.986	.986	.986	.886	.568	.329	.029	.000	.000	.532	14
28-30	1.00	.982	1.00	.891	.855	.491	.200	.127	.036	.620	11
31-42	1.00	1.00	1.00	1.00	.900	.920	.580	.320	.120	.760	10
Mean	.952	.872	.722	.550	.398	.226	.092	.050	.016	.431	100

Table 3
ξ_{jk} and v From Values in Table 2 for n-iles

b_k and n-ile	Item Set 1	2	3	4	5	6	7	8	9	Mean	ω
b_k	-2.08	-1.46	-.84	-.22	.40	1.02	1.64	2.26	2.88		
n-ile 1											
ξ	-.84	-2.58	-2.58	-2.58	-2.58	-2.58	-2.58	-2.58	-2.58	-2.01	-.788
v	*	0	0	0	0	0	0	0	0		
n-ile 2											
ξ	.78	-.13	-1.19	-1.50	-2.58	-2.58	-2.58	-2.13	-2.58	-1.00	-1.183
v	*	*	*	*	0	0	0	0	0		
n-ile 3											
ξ	1.47	1.01	.00	-.95	-1.90	-2.58	-2.58	-2.58	-2.58	-.60	-1.440
v	*	*	*	*	0	0	0	0	0		
n-ile 4											
ξ	2.58	1.11	.47	-.29	-.99	-1.93	-2.22	-2.22	-2.58	-.39	-1.490
v	1	*	*	*	*	0	0	0	0		
n-ile 5											
ξ	2.05	1.75	.92	.15	-.52	-1.41	-2.58	-2.58	-2.58	-.22	-1.485
v	1	1	*	*	*	*	0	0	0		
n-ile 6											
ξ	2.58	2.16	.90	.33	.17	-.96	-1.68	-2.58	-2.58	-.08	-1.446
v	1	1	*	*	*	*	0	0	0		
n-ile 7											
ξ	2.19	2.19	2.19	1.20	.22	-.44	-1.90	-2.58	-2.58	.08	-1.620
v	1	1	1	*	*	*	0	0	0		
n-ile 8											
ξ	2.58	2.09	2.58	1.23	1.06	-.02	-.84	-1.14	-1.79	.31	-1.351
v	1	1	1	*	*	*	*	*	0		
n-ile 9											
ξ	2.58	2.58	2.58	2.58	1.28	1.41	.20	-.47	-1.17	.71	-1.231
v	1	1	1	1	*	*	*	*	*		
b_k'	-1.66	-1.14	-.59	-.13	.26	.75	1.33	1.64	2.14	.17	-1.437

the values b_k' (empirical item-difficulty values) for the item sets. For the data in Table 3, $\bar{b}'$ was .290 with $\sigma_{b'} = 1.219$.

An ω statistic was also computed as the average value of ω_j, $j = 2, \ldots, n - 1$, where ω_j was a measure of the best-fitting linear slope of the values ξ_{jk}, as regressed on the values of b_k'. This statistic was an attempt to capture the effect of the a parameter on the PCFs represented in the vectors ξ_{jk} for given j. Because of the presence of extreme values and excessive variability in ξ_{jk} it was necessary to develop a special procedure for measuring this slope.

The values in a vector ξ_{jk} for a particular value of j were classified into three possible categories: (1) those exceeding 1.6449, corresponding to $p_{jk} > .95$; (2) those between 1.6449 and -1.6449, considered "directly usable points;" and (3) those less than -1.6449, corresponding to $p_{jk} < .05$. For all points in each of categories 1 and 3, the normal deviate transform of the mean value of p_{jk} was computed over those points and the mean of the corresponding values b_k'; all computations of means were weighted by the numbers of items involved. If all points were in category 1 or in category 3, a point was added to the right or left, respectively, consisting of a pair in which $b' =$ the normal deviate transform for $p = 1/(2N)$ or $1 - 1/(2N)$, respectively, and ξ is the normal deviate transform for $p = 1 - 1/(2N)$ or $1/(2N)$, respectively. These points were used, together with any and all points in category 2, to compute the best-fitting linear slope.

If, however, the slope was positive (which can happen by chance, although rarely), a further point was added at the right or left, respectively, as mentioned above, depending upon whether the mean ξ was greater or less than (or equal to) zero, and the slope was recomputed. Table 3 shows the classifications of values, v, as "1", "*", and "0" for categories 1 to 3, respectively, and the slopes so computed for each j. The mean of the slopes for $j = 2, 3, \ldots, n - 1$ (with each ω_j weighted by the value of N_j) is $\omega = -1.437$. This procedure excludes slopes computed for $j = 1$ or n because they are unreliable, usually being based on vectors in which all or most of the points are in category 1 or in category 3. The exclusion of such slopes is analogous to the exclusion of items with $p = 0$ or 1 in other IRT models.

Regression analysis showed that all these statistics were highly predictable from the generating parameters. They could not be perfectly predictable, of course, because of random fluctuations in the generated data. In some cases, transforming the parameter a to the function $\zeta = a/(1 + a^2)^{1/2}$ produced a slightly better prediction, apparently because of a closer fit to linearity. Table 4 shows multiple correlations for the monte carlo runs for Samples A and B, with asterisks indicating which variables had significant β weights in each regression equation. For the variables $\bar{p}$, skewness, and $\bar{b}'$, it made no difference whether a or ζ was used in the prediction, because the prediction relied almost exclusively on the generating parameter $\bar{b}$. For variable ω, significantly better prediction was obtained when the variable a was used, while better predictions for the remaining variables were obtained with the use of ζ. Generally, the predictions for Sample B, with $N = 400$ cases per observation, were slightly better than those for Sample A, with $N = 100$ cases per observation.

Prediction of Generating Parameters

Method and results. The fact that various statistics of datasets were highly predictable from the generating parameters gave reason to believe that these statistics could be successfully used in "back-estimation" of the generating parameters, not only for simulated datasets, but also for empirical datasets. Considerable effort was devoted to investigating ways of predicting the generating parameters by multiple linear regression techniques from the statistics enumerated in Table 4.

Of principal interest was the prediction of the a parameter, because with a reliable estimate of a it was possible to use this (or a function of it) in predicting the other two generating parameters. Although the reliability variable was initially investigated as a predictor, it was dropped from fur-

Table 4
Multiple Correlations and Statistical Significance of Generating Parameters a, ζ, b, and σ_b for Sample A (N = 100) and Sample B (N = 400) Based on 300 Generated Datasets for Each Sample

Dependent Statistic	Sample A					Sample B				
	a	ζ	b	σ_b	R	a	ζ	b	σ_b	R
$\bar{p}$										
Run 1	ns	-	*	ns	.9541	ns	-	*	ns	.9553
Run 2	-	ns	*	ns	.9541	-	ns	*	ns	.9553
σ_p										
Run 1	*	-	ns	*	.8939	*	-	ns	*	.8987
Run 2	-	*	ns	*	.9208	-	*	ns	*	.9251
Skewness										
Run 1	ns	-	*	ns	.7759	ns	-	*	ns	.8063
Run 2	-	ns	*	ns	.7759	-	ns	*	ns	.8064
Reliability										
Run 1	*	-	ns	*	.7319	*	-	ns	*	.7487
Run 2	-	*	ns	*	.8496	-	*	ns	*	.8714
$\bar{b}'$ Run 1	ns	-	*	ns	.9586	ns	-	*	ns	.9583
Run 2	-	ns	*	ns	.9586	-	ns	*	ns	.9583
$\sigma_{b'}$										
Run 1	*	-	ns	*	.9550	*	-	ns	*	.9575
Run 2	-	*	ns	*	.9669	-	*	ns	*	.9685
ω										
Run 1	*	-	ns	*	.9148	*	-	ns	*	.9438
Run 2	-	*	ns	*	.8896	-	*	ns	*	.9101

Note. * = t significant at $p < .0001$; ns = t not significant at $p = .1$; - = generating parameter not used in the multiple regression.

ther consideration for three reasons: first, as may be seen in Table 4, it was the least predictable of the statistics; second, it failed to add significant contributions beyond those of other variables; and third, there are well-known problems in choosing the appropriate measure of the reliability of scores in an empirical dataset. The statistics $\bar{p}$, skewness, and $\bar{b}'$ proved to be not useful in predicting a, even in interaction with other statistics (as shown by their insignificant regression weights). However, the absolute value of $\bar{b}'$ was useful, indicating that certain other statistics were affected as a function of the distance of $\bar{b}'$ from zero.

Apart from this, the following variables were left as possible predictors of a: σ_p, $\sigma_{b'}$, and ω. Because the results in Table 4 suggested that there was greater linearity in correlation when the ζ transform of a was used with certain statistics, emphasis was placed on predicting ζ, with prediction of a based on the transform of ζ to a, and a further prediction of a directly. Considering all results, the most generally useful predictions were obtained with five variables, some of them multiplicative: $X_1 = \sigma_p$, $X_2 = \sigma_{b'}$, $X_3 = \omega$, $X_4 = \sigma_{b'}{}^*\omega$, and $X_5 = |\bar{b}'\omega|/\sigma_{b'}$.

These investigations of the estimation of ζ and its transform a were conducted with both samples A and B, and their half-samples A1, A2, B1, and B2. Table 5 presents multiple regression results for ζ in these samples. Similar multiple regressions were computed for the prediction of a, but are not shown here; multiple correlations were approximately .97, lower than that for ζ.

By averaging raw regression coefficients for Samples A1, A2, B1, and B2, linear equations for predicting a and ζ from X_1 through X_5 were developed, as shown in Table 6, and they were

Table 5
Multiple Regression Prediction of ζ From the Equation $\hat{\zeta} = h + g_1X_1 + \ldots + g_5X_5$ From Dataset Statistics in Six Samples

		Sample					
Predictor		A1	A2	A	B1	B2	B
X_1							
	r	.593	.593	.593	.594	.599	.597
	β	.887	.916	.899	.910	.909	.910
	t	37.41	33.22	50.41	42.46	40.23	58.84
	g	3.21	3.25	3.22	3.28	3.29	3.29
X_2							
	r	.528	.523	.526	.526	.525	.526
	β	−.214	−.172	−.191	−.185	−.226	−.205
	t	−2.68	−2.25	−3.48	−3.00	−3.46	−4.59
	g	−.098	−.079	−.088	−.085	−.104	−.095
X_3							
	r	−.810	−.835	−.822	−.851	−.858	−.854
	β	.292	.310	.298	.299	.302	.301
	t	8.44	8.44	11.99	10.05	9.65	14.01
	g	.254	.265	.257	.262	.271	.266
X_4							
	r	−.669	−.665	−.667	−.664	−.662	−.663
	β	−1.182	−1.155	−1.165	−1.170	−1.212	−1.191
	t	−13.51	−13.62	−19.28	−16.74	−16.36	−23.55
	g	−.386	−.374	−.378	−.374	−.389	−.381
X_5							
	r	−.121	−.156	−.138	−.158	−.143	−.151
	β	−.183	−.171	−.177	−.170	−.172	−.171
	t	−16.82	−15.78	−23.21	−21.11	−20.46	−29.57
	g	−.045	−.044	−.044	−.044	−.044	−.044
h		.008	.014	.010	.007	.017	.012
R		.995	.994	.995	.997	.997	.997
SE $\hat{\zeta}$		.021	.021	.021	.016	.017	.016

Note. All values of b, t, and R significant at $p < .05$, or in most cases $< .001$.

Table 6
Equations Used in Estimating a, ζ, $\bar{b}$, and σ_b

Given:
$X_1 = \sigma_p$
$X_2 = \sigma_{b'}$
$X_3 = \omega$
$X_4 = X_2X_3$
$X_5 = |\bar{b}'|X_3/X_2$
Then
$X_6 = -.758 + 2.921X_1 - .921X_2 - .414X_3 - 1.218X_4 - .1X_5$,
(a preliminary estimate of a)
$X_7 = .012 + 3.257X_1 - .092X_2 + .263X_3 - .381X_4 - .044X_5$, (estimate of ζ)
If $X_7 > .95$ then $X_7 = .95$
If $X_7 < .05$ then $X_7 = .05$
$X_8 = X_7/(1 + X_7^2)^{1/2}$, (a further estimate of a derived from X_7)
$\hat{a} = .01 + .335X_6 + .655X_8$
$\hat{\zeta} = X_7$
$\text{Est}[\bar{b}] = (\bar{b}')/\hat{\zeta}$
$\text{Est}[\sigma_b] = (\sigma_{b'})/\hat{\zeta}$

applied to the data from all samples, with results shown in Table 7. The correlations of $\hat{a}$ with a were very high, and the mean errors from the use of this equation were negligible, ranging from −.001 to .004 over the 6 samples. From this evidence, the equation for estimating a could be considered to be adequately cross-validated (even though double cross-validation was not performed). On the other hand, standard errors of estimate were not insubstantial, being about .091 in Sample A and .066 in Sample B, which had more cases per observation.

The failure to predict the generating a parameter perfectly could probably be attributed mainly to inherent random fluctuations in generated data, and only little to imperfect estimation methods. The "back-predictions" of generating parameters were significantly higher than the predictions of individual statistics from the generating parameters shown in Table 4, a finding that suggests that the back-predictions were about as high as could be obtained with linear methodology, or for that matter, any methodology. The amount of inherent random fluctuation was indicated by the within-cell variances found in 3-way ANOVAs of the data, with 2 observations for each combination of generating parameters. Taking the square roots of these variances, the within-cell standard deviation for Sample A was .0554, and .0396 for Sample B. These values were approximately two-thirds of the corresponding standard errors of estimate obtained for the prediction of a, as shown in Table 7.

Table 7 also shows correlations of ζ, the nonlinear function of the generating a parameter, with the estimate of ζ ($\hat{\zeta}$); they are higher than the correlations for $\hat{a}$. However, the scaling of ζ relative to that of a made it undesirable to use as a parameter. Small errors in estimates of ζ can correspond to large differences in a at values of ζ that approach unity.

Finally, Table 7 shows that the remaining two generating parameters, $\bar{b}$ and σ_b can be well predicted by dividing b' and $\sigma_{b'}$, respectively, by the estimated value of ζ. This is justified by theory implicit in Lord and Novick's (1968) Equation 16.9.4 relating empirical item difficulties to

Table 7
Correlations of Estimates with Generating Parameters and Errors of Prediction for the Generating Parameters from Equations in Table 6, in Six Samples

	Sample					
Statistic	A1	A2	A	B1	B2	B
$\hat{a}$ with a						
r	.985	.984	.984	.992	.991	.992
Error of prediction: $\hat{a} - a$						
Mean	.000	.004	.002	.004	−.001	.002
SD	.089	.093	.091	.065	.067	.066
$\hat{\zeta}$ with ζ						
r	.995	.994	.995	.997	.997	.997
Est[$\bar{b}$] with $\bar{b}$						
r	.997	.996	.996	.999	.999	.999
Error of prediction: Est[$\bar{b}$] − $\bar{b}$						
Mean	−.003	−.013	−.008	−.001	−.001	−.001
SD	.046	.049	.048	.030	.027	.028
Est[σ_b] with σ_b						
r	.993	.990	.992	.996	.996	.996
Error of prediction: Est[σ_b] − σ_b						
Mean	−.018	−.010	−.014	−.018	−.016	−.017
SD	.071	.081	.076	.054	.057	.056

theoretical item difficulties when both are expressed in normal deviate form.

Table 8 presents evidence concerning errors of estimate at given values of the generating a parameter. Estimation error of a tends to increase somewhat with a, but estimation error for σ_b tends to decrease with a. Errors of estimate were consistently smaller for Sample B (with 400 cases per observation) than for Sample A (with only 100 cases per observation). These errors were relatively small, however; thus it was probably not worthwhile to attempt adjusting for them. There were no systematic trends in errors of estimates of $\bar{b}$ and σ_b as a function of these statistics.

The case of $c > 0$. All results presented above are based on the case of $c = 0$. When c is assumed to be greater than 0, the required computations can be based on values of p'_{jk}, corrected for chance success by the formula $p' = (p - c)/(1 - c)$, prior to transformations to values of ξ_{jk}. The correction was applied to all values p_{jk}, as well as to the mean and standard deviation of proportion-correct raw scores. (This correction was incorporated in relevant computer programs mentioned in this article; see Carroll, 1945, for formulas for correcting means and standard deviations for the c parameter.)

To investigate the effect of $c > 0$ on estimates of a, monte carlo runs were conducted for $c = .5$, for the same series of a parameters, $\bar{b}$, and σ_b as before, and for $N = 100$ and $N = 400$, but for only one case for each combination of parameters. Using the same equation to estimate a that was applied to the Sample A and Sample B data, the correlation between a and $\hat{a}$ was .850 for the run with $N = 100$ cases per observation, and it had a standard error of estimate of .272. The mean error was -.082, which was significantly different from zero ($t = -3.68$, $p < .001$).

The lower correlation was to be expected, in view of the greater instability of data corrected for guessing, but it was still a respectably high value. The larger errors occurred with small values of σ_b and higher values of $\bar{b}$. This oc-

Table 8
Errors of Parameter Estimates for Specified Values of a in Samples A and B (Statistics for a Based on 50 Observations; for $\bar{b}$ and σ_b Based on 60 Observations)

Parameter Estimated		a = .3	.6	.9	1.2	1.5	1.8	All
Sample A								
a								
	Mean	-.006	.035	.028	.009	-.008	-.046	.002
	SD	.038	.044	.049	.054	.114	.153	.091
$\bar{b}$								
	Mean	-.023	-.004	-.004	-.008	-.005	-.003	-.008
	SD	.081	.043	.035	.029	.037	.040	.048
σ_b								
	Mean	.029	-.046	-.019	-.011	-.014	-.023	-.014
	SD	.128	.080	.046	.041	.045	.056	.076
Sample B								
a								
	Mean	-.008	.022	.018	-.001	-.020	.000	.002
	SD	.030	.026	.034	.044	.075	.122	.066
$\bar{b}$								
	Mean	-.002	-.002	.000	.001	-.001	-.003	-.001
	SD	.045	.028	.021	.021	.021	.025	.028
σ_b								
	Mean	.021	-.037	-.020	-.016	-.016	-.034	-.017
	SD	.096	.050	.036	.031	.031	.037	.056

curred, for example, in tests for which there was less variation in item difficulty, permitting less accurate estimation of the PCF slopes, and for which the items were generally difficult, such that greater amounts of guessing would be expected. Comparable statistics for the run with $N = 400$ cases per observation were $r = .925$, mean error of estimate = -.067, and standard error of estimate = .196. As might be expected, the accuracy of estimate improved with the higher number of cases per observation.

The estimation of the c parameter is often problematic, particularly in the case where multiple-choice options vary in attractiveness. However, it was deemed desirable to use a c parameter in the estimations because some abilities can be measured with two-option items (e.g., with possible responses same and different, as in a pitch discrimination test) without serious response bias and with confident setting of $c = .5$. Also, even when multiple-choice distractors vary in attractiveness, inspection of the PCF table for low-scoring groups and difficult items suggests that values of p_{jk} descend to an asymptote; the value of this asymptote could be taken as an estimate of c as a parameter applying on the average to all items.

Applications

In practical use of the PCF model with empirical test data, a table of probabilities p_{jk} analogous to Table 2 can be constructed by grouping items and scores. The data should be edited first in order to exclude items with atypically low item-test biserial correlations. The items should be grouped into approximately five to nine sets of increasing difficulty, each set having several (2 or more) items with similar values of p; the sets need not be of equal size. The standard deviation of the proportion-correct total score distribution is computed. Then the total score distribution is divided into approximately nine score intervals, with approximately equal numbers of persons in each group, except for the lowest and highest score groups, which are not used in PCF computations and thus may contain fewer cases than the other score groups (or more cases, when distributions are highly skewed).

The PCF table (as it may be called), along with its marginals, is then transformed into normal deviate values like those in Table 3. If the value of c is other than zero, the values p'_{jk} are corrected for c before the transformation (with a lower bound of $p'_{jk} = 0$). The statistics $\bar{b}'$, $\sigma_{b'}$, and ω can then be computed on the basis of the PCF table. Estimates of a, ζ, $\bar{b}$, and σ_b can then be made by the formulas presented in Table 6. (A computer program for these operations is available.) If theoretical item difficulty values are desired for specific item sets k, or even for specific items i, they may be estimated by dividing the respective b'_k or b'_i values by $\hat{\zeta}$.

Because the scale specified by b applies both to item difficulty and to θ, values of b corresponding to proportion-correct scores (corrected for the c parameter if $c > 0$) can be estimated by using the values of b' at which the PCF is at a liminal value 0 for each score group j. Let ϕ_j be this value for score group j, computed in the course of determining the slope of the PCF for that score group. An estimate of the corresponding θ value can then be obtained by dividing ϕ_j by $\hat{\zeta}$. Values of ϕ_j do not always increase monotonically with j, however, because of random fluctuations in data. If a plot of ϕ_j against p' appears to be nonlinear, the estimation of θ from ϕ_j can be accomplished by fitting a curve through the points by inspection, and finding the corresponding θ values by dividing the Y coordinates of the fitted points by $\hat{\zeta}$. If the plot appears to be acceptably linear, θ values for given values of p' can be estimated by finding the least-squares linear function $\phi = f_{p'}$, and then constructing the corresponding function $b_{p'} = f_{p'}/\hat{\zeta}$.

A Small Set of Vocabulary Test Data

The data were from the administration of a wide-range vocabulary test to 143 students in sociology classes at a university. The test had 35 items, and all examinees tried every item. Item response data (in terms of scores of 0 or 1 on each item) were entered into a computer program that

(among other things) determined the item difficulties and biserial *r*s with total scores. Of the 35 items, 28 items were selected (by the program) as having biserial *r*s of .20 or greater, and *p* (proportion correct) values other than 0 or 1. The mean biserial *r* of these 28 items against the revised total score was .505; revised total scores ranged from 9 to 28, with a mean of 15.27 and SD of 3.84. The Kuder-Richardson reliability (formula 20) was .741.

The 28 items were divided into 8 sets, each comprised of from 2 to 4 items with closely similar *p* values, and a PCF table was constructed showing the probabilities correct for the 8 item sets at each of 9 total score levels. These values were entered into a computer program ZETA that produced estimated statistics by the equations listed in Table 6.

The value of *a* for these data was estimated as .662; this value implied that the PCF curves had slopes, with respect to *b*, that were relatively gradual. The value of $\bar{b}$ was estimated as –.533, indicating that the items were relatively easy on the average for the group. Items were assumed to have mean $\theta = 0$ and $\sigma_\theta = 1$. The value of σ_b, however, was estimated as 2.046; thus the items varied widely in difficulty.

The program ZETA contains an option by which one or more values of the *c* parameter can be specified. In the present case, the fact that the items were all 5-choice items suggested that *c* should be set equal to .2. The shapes of the PCF curves, however, did not imply that *c* should be set as high as .2, because low-scoring individuals (e.g., in the first 2 noniles) had probabilities of passing the more difficult items that approached zero. Therefore, *c* was set at zero to produce the estimates. When *c* was arbitrarily set equal to .05 (a value that might be justified as being the probability approached by the lowest scoring individuals for the most difficult items), the estimate of *a* was slightly higher at .751, and there were corresponding changes in the estimated values of b_k. Yet the value of $c = 0$ was judged to be most appropriate for these data because individuals generally did not guess their answers; they selected them on the basis of whatever knowledge they had of the choices, and in the absence of relevant knowledge they were unlikely to select the correct choice.

The estimate of *a* for these data was validated by using the estimated statistics $\hat{a}$, Est[$\bar{b}$], Est[σ_b], *c*, and values of $\hat{b}_k$ as generating parameters to produce one or more samples of data based on those parameters. Program PCFGENA was used to produce five such samples. The resulting total score distributions and the 8 item-set score distributions were compared, using χ^2 tests, with the corresponding distributions from the test data. The obtained frequencies were tested against the average frequencies over the five generated samples, with these average frequencies regarded as theoretical population values. In all cases except one (for item set 5, with $p < .025$), the values of χ^2 were not significant even at the $p = .10$ level, although the generated total score distributions tended to have more low scores (below 9) than the obtained distribution. The smaller number of low scores in the obtained distribution may have been due to possible selectivity and lower-tail truncation, as opposed to the normal distribution assumed in generating the responses. In any case, the results suggested that the estimates were acceptable.

Three Tests from the Woodcock-Johnson Psycho-Educational Battery—Revised

Item response data were available for 1,800 individuals tested with the norming versions of three subtests of the Woodcock-Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989). The subtests, Picture Vocabulary, Concept Formation, and Visual Closure, respectively, represent factors *Gc*, *Gf*, and *Gv* of the Cattell-Horn theory of intelligence (Horn, 1988). The data appeared to be particularly suitable for testing the PCF procedures developed here, for several reasons:

1. The tests are individually administered and require open-ended responses, rather than

choices among alternatives (as in typical paper-and-pencil tests); thus, the value of the c parameter could be assumed to be zero.
2. Complete data were available on a large number of cases selected over a wide range of ages and talents.
3. The test items were carefully devised, selected, revised, and arranged in order of difficulty by the tests' authors to measure abilities over a wide range.
4. It was highly likely that each test was unidimensional.

For purposes of testing the PCF model, the 1,800 cases for each subtest were divided into 8 samples of 225 cases each, taken successively from the data file. The order in which the cases appeared in the data file was essentially arbitrary; test score and age distributions were roughly similar over the samples. χ^2 and other tests indicated that the samples were actually significantly different in test score and age distributions, but these variations were not considered of importance for the present purposes. Indeed, it was of interest to see whether the parameter estimates would be relatively comparable over the subsamples, despite moderate variations in test score and age distributions.

Table 9 shows results of applying the PCF procedures to the eight subsamples for each test, and to the total samples. Despite some variation in the estimates of the a parameter, it is clear that the tests are different with respect to that parameter. There is also consistency in the estimates of the σ_b parameter. Variations in $\bar{b}$ reflect variations in the mean ability levels of the samples.

Table 9 also shows results from the application of the LOGIST program to these data (but only for the total samples). The value of mean $\hat{a}$ over the items for each of the three tests is roughly comparable to the value of a estimated by the PCF procedures. The considerable variance in LOGIST estimates of a over items is noteworthy; the PCF procedures assume a constant value of a over items. In any case, the two types of estimates rank the tests in the same way. These results tend to support the accuracy of the PCF estimates. (For further discussion of the results

Table 9
Estimates of a, $\bar{b}$, and σ_b for Eight Samples of Item Response Data for Three Subtests of the Woodcock-Johnson Psycho-Educational Battery—Revised

	Subtest								
	Picture Vocabulary			Concept Formation			Visual Closure		
Sample	a	$\bar{b}$	σ_b	a	$\bar{b}$	σ_b	a	$\bar{b}$	σ_b
1	1.83	.33	1.42	1.66	.11	.85	.89	−.40	2.45
2	1.82	−.04	1.40	1.72	−.25	.80	.87	−.62	2.46
3	1.64	−.11	1.44	1.51	−.36	.84	.93	−.83	2.38
4	1.89	−.24	1.26	1.86	−.33	.71	.91	−.88	2.30
5	1.90	−.21	1.27	1.71	−.40	.76	.82	−1.04	2.52
6	2.09	−.40	1.29	1.68	−.57	.86	.93	−1.10	2.44
7	1.77	−.26	1.37	1.60	−.54	.85	.87	−.95	2.47
8	1.92	−.34	1.28	1.58	−.54	.81	.88	−1.00	2.51
Mean	1.86	−.16	1.34	1.67	−.36	.81	.89	−.85	2.46
SD	.12	.21	.07	.10	.21	.05	.03	.22	.08
Pooled	1.82	−.16	1.32	1.71	−.35	.78	.87	−.88	2.47
LOGIST $\hat{a}$ (over items)									
Mean	1.93			1.61			1.06		
SD	.37			.57			.35		

for the Woodcock-Johnson data, see Carroll, in press.)

Discussion

The behavioral scaling (Carroll, in press) of psychological and educational tests requires methods through which test scores can be related to a theoretical item difficulty scale. The theoretical item difficulty scale is to be distinguished from the empirical difficulty scale because the latter describes the difficulties of items in a particular sample, whereas the former contains parameters assumed to generate item responses. Procedures using the person characteristic function, as studied here, appear to offer an appropriate solution to such behavioral scaling.

In application to typical empirical test data, the procedures assume that items are homogeneous with respect to the ability or abilities they measure. Factor-analytic techniques could be applied to datasets to confirm their unidimensionality. Alternatively, standard item analysis procedures (e.g., use of biserial *r*s against total scores) could be applied to select items that appear to measure best a single ability or a complex of abilities.

The procedures assume some grouping of items and of individuals. Individuals are grouped in terms of quantiles or n-iles (where n is normally no more than 9) in order to compute averaged PCF functions. Items are grouped in item sets to make for more accurate determination of PCF functions. Although the procedures can produce estimates if the item sets consist of only one item each, the sampling fluctuations in proportions correct might be too extreme to yield satisfactory estimations, especially if N is small.

The theoretical difficulty scale b has a linear indeterminacy or arbitrary metric in the sense that its origin, 0, is relative to a given sample. On the other hand, its units are tied to transforms of probability values, and in this sense they may be considered on an equal-interval scale. Furthermore, given that the items administered to different samples (with different levels, or distributions of ability) are all or partly identical and measure the same ability, it would be possible to refer the items to a common theoretical difficulty scale for all samples, and thus to equate different raw score scales on the basis of the theoretical item difficulty scale.

The PCF estimation procedures depend on linear multiple regression techniques. The possibility that maximum likelihood procedures would produce superior estimates has not been investigated. One advantage of the present procedures is that they are easily and immediately computable without iteration, both for small and large datasets.

The PCF procedures also depend on the assumption of a normal distribution of ability underlying any sample studied. The robustness of the procedures with respect to violation of this assumption should be investigated.

The research reported here emphasizes the possible importance of the a parameter as a way of characterizing an ability, and it is a parameter that is not well represented in the Rasch model. The PCF procedures need to be applied to a wide variety of ability tests to confirm or disconfirm the hypothesis that abilities differ in an important way with respect to this parameter.

References

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park CA: Sage.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading MA: Addison-Wesley.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10,* 1–19.

Carroll, J. B. (1980). Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 449–452). Minneapolis MN: Psychometric Methods Program, Department of Psychology, University of Minnesota.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift in honor of Frederic*

M. Lord (pp. 257–283). Hillsdale NJ: Erlbaum.

Carroll, J. B. (in press). Test theory and the behavioral scaling of test performance. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale NJ: Erlbaum.

Carroll, J. B., Meade, A., & Johnson, E. S. (in press). Test analysis with the person characteristic function: Implications for defining abilities. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* Hillsdale NJ: Erlbaum .

Carroll, J. B., & Schohan, B. (1953). *Construction of comprehensive achievement examinations for Navy officer candidate programs.* Pittsburgh PA: American Institute for Research.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59–79.

Hambleton, R. K. (Ed.) (1983). *Applications of item response theory.* Vancouver: Educational Research Institute of British Columbia.

Horn, J. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.) (pp. 645–685). New York: Plenum.

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

Mosier, C. I. (1941). Psychophysics and mental test theory, II. The constant process. *Psychological Review, 48,* 235–249.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut. [Expanded edition, University of Chicago Press, 1980].

Stevens, S. S. (1961). Toward a resolution of the Fechner-Thurstone legacy. *Psychometrika, 26,* 35–47.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.

Weiss, D. J. (1980). Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 444–448). Minneapolis MN: Psychometric Methods Program, Department of Psychology, University of Minnesota.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery–Revised: Tests of Cognitive Ability.* Allen TX: DLM Teaching Resources.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Educational Review, 3,* 281–288.

Wright, B. D., & Mead, R. J. (1977). *BICAL: Calibrating items and scales with the Rasch model.* Chicago: University of Chicago, Department of Education.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Acknowledgments

Thanks are due to Richard C. Woodcock for supplying item response data from the Woodcock-Johnson Psycho-Educational Battery—Revised, and to Albert E. Beaton for conducting analyses of these data by the program LOGIST. The author is also appreciative of comments by the Editor and anonymous reviewers on earlier versions of this paper.

Author's Address

Send requests for reprints and further information to John B. Carroll, 409 North Elliott Road, Chapel Hill, NC 27514, U.S.A.