Nonparametric Estimation of the Plausibility Functions of the Distractors of Vocabulary Test Items

Fumiko Samejima

The University of Tennessee

The Level 11 Vocabulary Subtest of the Iowa Tests of Basic Skills was analyzed using a two-stage latent trait approach and an empirical dataset of 2,356 examinees. First, each of the 43 multiple-choice test items was scored dichotomously; then, assuming the (two-parameter) normal ogive model the item parameters were estimated. The operating characteristics of the correct answer and of the three distractors were estimated using a nonparametric approach called the simple sum procedure of the conditional probability density function approach combined with the normal approach method. Differential information was provided by the distractors, and these operating characteristics were named the plausibility functions of the distractors. The operating characteristic of the correct answer of each item estimated by assuming the normal ogive model was compared with the nonparametrically estimated operating characteristic for model validation. It was concluded that the nonparametric approach leads to efficient estimation of the latent trait. *Index terms: distractors, item response theory, latent trait models, multiple-choice test items, nonparametric estimation, plausibility functions of distractors.*

Multiple-choice test items have been used in many ability and achievement tests. Test results have been analyzed using classical test theory, and in recent years using latent trait theory. The most widely used latent trait models include the Rasch (1960) model and the three-parameter logistic model (Birnbaum, 1968). In each case, an item response is scored correct or incorrect, depending on whether the examinee selected the correct alternative or one of the incorrect alternatives, called *distractors*. Thus these models, which belong to the category of the *equivalent distractor model* (Samejima, 1984), ignore information that might be provided by each distractor.

Bock (1972) proposed the multinomial model, which deals with a set of discrete item responses. The model discovers the implicit order of these nominal responses, and Samejima (1972) pointed out that it can be considered to be a member of the *heterogeneous case* of the graded response model. [Both Masters' (1982) partial credit model and Muraki's (1992) generalized partial credit model, which deal with ordered response categories and are special cases of Bock's multinomial model, belong to the heterogeneous case of the graded response model.] Bock applied his multinomial model for multiple-choice test items and clarified the implicit order of the distractors.

Samejima (1979) proposed a family of models for multiple-choice test items, which assumes that the examinee selects a distractor intentionally and, therefore, each incorrect alternative, as well as the correct alternative, provides unique, differential information. The examinee's guessing behavior is still taken into consideration, but only as the last resort when the examinee has no idea which alternative is the most plausible answer to the item. The Bock-Samejima model (Samejima, 1979) is based on Bock's multinomial model, but includes a consideration of random guessing. Many other models have

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 1, March 1994, pp. 35-51

0146-6216/94/010035-17$2.10

been proposed for polychotomous item responses (e.g., Andrich, 1978; Samejima, 1969, 1972), but they are basically for explicitly ordered polychotomous, or graded, items.

Plausibility Functions

The *plausibility function* of each distractor—defined as the conditional probability assigned to the selection of that particular distractor at a given trait level—may provide differential information, and the examinee's trait level may be estimated more efficiently and accurately using information obtained from the distractors as well as from the correct answers. Thus, the multiple-choice test item may no longer be a "blurred image" of the free-response test item—that is, a free-response model made fuzzier by noise. Models such as the above belong to the category of the *informative distractor model* (Samejima, 1984).

The plausibility functions of the distractors must be estimated to determine if each distractor provides unique information. The operating characteristic (OC) of a discrete item response is the conditional probability, given the latent trait θ, that the examinee's response will fall into that specific response category. A nonparametric method for estimating the OCs should be used to discover the true shapes of the plausibility functions of the distractors without any prior assumptions, rather than requiring them to fit some predetermined mathematical form.

Lord (1959) developed a nonparametric method for estimating the OCs of discrete item responses for large datasets. Levine (1984) developed a nonparametric method that does not require a large dataset, which is applicable to the kind of data analyzed here. Samejima's methods for estimating the OCs of discrete item responses were introduced previously (Samejima, 1977, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f, 1980, 1981, 1988, 1990a, 1990b, 1991). The methods of estimating the OCs of discrete item responses are characterized by two features: (1) no mathematical form for the OCs of discrete item responses is assumed, and (2) estimation is efficient enough to use a relatively small dataset (several hundred to a few thousand examinees provide reasonably good estimations).

Approaches to Estimating OCs

Two main approaches can be used to estimate the OCs of discrete item responses: the bivariate probability density function (BPDF) approach and the conditional PDF (CPDF) approach. In the BPDF approach, the bivariate distribution of the latent trait τ (which is transformed from the original θ) and its maximum likelihood estimate, $\hat{\tau}$, is approximated for each of the subgroups of examinees who share the same discrete item response to a specified, *target* item, for which the OCs are to be estimated. Thus, the procedure must be repeated as many times as the number of discrete item response categories for each target item.

In contrast, the CPDF approach approximates the conditional distribution of τ, given $\hat{\tau}$, for the total group of examinees, and the result is branched into separate discrete item response subgroups for each target item. The CPDF approach is further categorized into four procedures: the simple sum (SS) procedure, the weighted sum procedure, the proportioned sum procedure, and the differential weight procedure. Each of the four procedures has four methods of approximating the conditional distribution of τ, given $\hat{\tau}$: the Pearson system method, the two-parameter beta method, the normal approach (NA) method, and the lognormal approach method.

Let θ assume any real number, and k_g be any discrete response category to item g. Assume that there is a set of test items with known characteristics measuring θ. This set of test items is referred to as the *Old Test*.

It has been shown (Samejima, 1991) that, although the estimator of the OC in the BPDF approach is based on consistency (i.e., as the number of examinees, N, tends to positive infinity, the estimate of the OC converges in probability to the true OC), the estimator in the SS procedure of the CPDF approach is not based on consistency; therefore, to modify the estimator an additional process is needed, which is used in the differential weight procedure of the CPDF approach (Samejima, 1990a, 1990b). If the Old Test has a sufficient amount of test information for the θ interval of interest, however, then the estimated OC in the SS procedure of the CPDF approach will become very close to the true OC (Samejima, 1991). Because the SS procedure requires substantially less computing time than both the BPDF approach and the differential weight procedure of the CPDF approach, and because it does not have to deal with subgroups of small numbers of examinees in approximating the joint bivariate distributions of τ and $\hat{\tau}$ as in the BPDF approach, if the above condition is satisfied (i.e., there is a sufficient amount of test information of the Old Test for the θ interval of interest), then this procedure will be preferred over the BPDF approach and the other three procedures (the weighted sum, proportioned sum, and differential weight procedures) of the CPDF approach. An additional process of model validation, which is discussed below, is desirable, however.

The two-parameter beta method and the NA method are simpler versions of the Pearson system method. These two methods use only the first two estimated conditional moments of τ, given $\hat{\tau}$. This is an advantage over the Pearson system method, which also requires the third and fourth conditional moments. Estimation for the third and fourth conditional moments is less accurate than that of the first two conditional moments. Here the SS procedure of the CPDF approach combined with the NA method was used.

The SS Procedure of the CPDF Approach Combined With the NA Method

Let $I(\theta)$ denote the test information function of the Old Test of n items. The transformation of θ to τ is

$$\tau = C_1^{-1}\int_{-\infty}^{\theta}[I(t)]^{1/2}dt + C_0 \ , \tag{1}$$

where C_0 is an arbitrary constant that adjusts the origin of τ, and C_1 is an arbitrary constant that equals the square root of the test information function of τ, $I^*(\tau)$, so that

$$C_1 = [I^*(\tau)]^{1/2} \tag{2}$$

for all τ. This transformation is tentative and is used to simplify the mathematics; eventually τ will be transformed back to θ. The transformation will be simplified if a polynomial approximation to the square root of $I(\theta)$ is used, which is accomplished using the method of moments (Samejima & Livingston, 1979). Thus, Equation 1 can be changed to the form

$$\tau \doteq C_1^{-1}\sum_{k=0}^{m}\alpha_k(k+1)^{-1}\,\theta^{k+1} + C_0 = \sum_{k=0}^{m+1}\alpha_k^*\theta^k \ , \tag{3}$$

where α_k ($k = 0, 1, \ldots, m$) is the kth coefficient of the polynomial of degree m approximating the square root of $I(\theta)$, and α_k^* is the new kth coefficient that is given by

$$\alpha_k^* \begin{cases} = C_0 & k = 0 \\ = (C_1 k)^{-1}\alpha_{k-1} & k = 1, 2, \ldots, m+1 \ . \end{cases} \tag{4}$$

The first through fourth conditional moments of τ, given $\hat{\tau}$, can be obtained from the density function, $g^*(\hat{\tau})$, of $\hat{\tau}$ and C_1 only, a benefit resulting from the transformation of θ to τ (Samejima, 1977, 1981).

The two coefficients, β_1 and β_2, and Pearson's criterion κ (e.g., Elderton & Johnson, 1969; Johnson & Kotz, 1970) are obtained by

$$\beta_1 = \mu_3^2\,\mu_2^{-3} \ , \tag{5}$$

$$\beta_2 = \mu_4 \mu_2^{-2} \ , \tag{6}$$

and

$$\kappa = \beta_1(\beta_2 + 3)^2[4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)]^{-1} \ , \tag{7}$$

where

$$\mu_2 = \mathrm{Var}(\tau \,|\, \hat{\tau}_s) \ , \tag{8}$$

$$\mu_3 = \mathrm{E}\{[\tau - \mathrm{E}(\tau \,|\, \hat{\tau}_s)]^3 \,|\, \hat{\tau}_s\} \ , \tag{9}$$

and

$$\mu_4 = \mathrm{E}\{[\tau - \mathrm{E}(\tau \,|\, \hat{\tau}_s)]^4 \,|\, \hat{\tau}_s\} \ . \tag{10}$$

In the SS procedure of the CPDF approach, the OC, $P_{k_g}(\theta)$, of the discrete item response k_g of a target item g is estimated by

$$\hat{P}_{k_g}(\theta) = \hat{P}^*_{k_g}[\tau(\theta)] = \sum_{s \in k_g} \phi(\tau \,|\, \hat{\tau}_s) \left[\sum_{s=1}^{N} \phi(\tau \,|\, \hat{\tau}_s) \right]^{-1} , \tag{11}$$

where $s = 1, 2, \ldots, N$ indicates an individual examinee, and $\phi(\tau \,|\, \hat{\tau}_s)$ denotes the conditional density of τ, given $\hat{\tau}_s$. This conditional density is estimated using the estimated conditional moments of τ, given $\hat{\tau}_s$. In the NA method, $\phi(\tau \,|\, \hat{\tau}_s)$ is approximated by the normal density function, using the first two estimated conditional moments of τ, given $\hat{\tau}_s$, as its two parameters.

Method

Data

A subset of the data collected for the Iowa Tests of Basic Skills, Form 6, Levels 9-14 (Hieronymus & Lindquist, 1971) was used. The level numbers, 9 through 14, correspond to the age of examinees for whom the subset of test items is suitable. Thus, the tests are designed for fourth through ninth graders. There are 11 tests, and each test focuses on a different basic skill. The number of items in the separate tests varies from 74 to 178. The six level subtests of each test are not disjoint; some test items are used in two, or even three, adjacent levels.

The dataset used here was sampled from a dataset collected in three different school systems in Iowa from 1971 to 1977, using the Level 11 Vocabulary Subtest. There were 2,460 examinees in the original dataset. Close examination of these original data revealed, however, that certain examinees omitted too many items. They were excluded, and the resulting revised data are based on 2,364 examinees. (For details see Samejima & Trestman, 1980.) The vocabulary subtest was selected due to the results of the factor analysis, which are discussed below. The Level 11 Vocabulary Subtest had 43 test items, and each item had four alternatives (i.e., one correct answer plus three distractors.)

The Old Test

A suitable substitute for the Old Test was needed, because there was no set of test items measuring the same vocabulary ability with known characteristics. Therefore, the Level 11 Vocabulary subtest was used twice—first as the Old Test and later as the set of target test items.

In the first stage, each item was scored as correct or incorrect (i.e., dichotomously scored), and the dichotomous normal ogive model was assumed. The OC of the correct answer, $P_g(\theta)$, of item g in the normal ogive model is given by

$$)^{-1/2}\int_{-\infty}^{a_g(\theta-b_g)} \exp(-u^2/2)\,du \;, \tag{12}$$

the item discrimination parameter, and b_g is the item difficulty parameter. This function e conditional probability that an examinee will answer item g correctly, given θ. The model :d tentatively, and the item parameters were estimated for the 43 test items. In the second 43 items were treated as multiple-choice test items with polychotomous item responses. :re estimated for the four alternatives of each item, and the estimated plausibility func- : three distractors and the nonparametrically estimated OC of the correct answer were 'his latter result was compared with the hypothesized normal ogive function as part of validation process, which is discussed below.

ıodel was validated, then the estimated OCs of the distractors were accepted as their ›lausibility functions. If the model was not validated, the invalidated test items were :ither more suitable models were assumed or they were discarded. A new Old Test was nd the estimation process was repeated.

ıeter Estimation for the Old Test Items Using the Normal Ogive Model

[952) method was used in the first stage of this study. It was assumed that the response ɔf the 2,364 examinees on the Level 11 Vocabulary subtest had a multinormal distribution ıt distribution. If a single common factor accounted for the 43 response tendencies, then would be operationally defined as the vocabulary ability for the subtest. As a result of ariate normal assumption, the θ distribution for these 2,364 examinees would become d the origin and the unit of the θ scale would be set equal to the mean and the standard f this normal distribution, respectively.

ıchoric correlation coefficient was obtained for each pair of test items, using a program the author. These correlations were adjusted for unbiasedness, which means that each ess in absolute value than the sample correlation. Factor analysis was conducted using ʌ computer program (Frane & Hill, 1974) for a principal factor solution. The diagonal f the interitem correlation matrix, $\mathbb{R}$, were replaced by estimated communalities, using estimation of each of the n communalities, with the squared multiple correlation of each :h all other variables as its initial estimate. [The same procedure was applied for each of :vel 11 subtests, and the resulting sets of eigenvalues are shown elsewhere (Samejima, 1984), he Level 11 Reading Comprehension subtest, for which the interitem correlation matrix tisfy the requirement of being positive semidefinite.]

: latent trait dimension had been operationally defined, the item parameters were esti- he normal ogive model (see Equation 12). The estimated item discrimination parameter, n difficulty parameter, $\hat{b}_g$, were obtained by

$$\rho_g^2)^{-1/2} \tag{13}$$

$$, \tag{14}$$

', where ρ_g is the factor loading of item g on the first common factor, and $\hat{\gamma}$ is the normal responding to the proportion correct, p_g, of item g on the Old Test.

Results

ıcture and Estimated Item Parameters

ritem correlation matrix is given elsewhere (Samejima, 1984) and consisted of all nonnegative correlation coefficients, which indicated the existence of a dominating common factor analogous to Spearman's general factor. The results of the principal factor solution (Table 1) showed that the first eigenvalue was substantially larger than any other eigenvalue, and the second eigenvalue was

Table 1
Principal Factor Loadings and Eigenvalues of the First Six Common Factors of the 43 Item Response Tendencies

Item	Factor 1	2	3	4	5	6
1101	.192	-.023	.350	.042	.091	-.230
1102	.398	-.250	-.024	.393	-.262	.063
1103	.513	-.210	-.003	.171	.029	.006
1104	.655	-.166	.024	-.074	.145	-.079
1105	.638	-.051	-.109	.048	.065	-.175
1106	.549	-.283	-.008	.053	.071	.121
1107	.469	-.245	.108	-.035	.133	.225
1108	.556	-.105	.009	.142	.077	-.147
1109	.558	-.190	.013	-.070	.006	-.051
1110	.523	-.042	-.298	-.077	-.177	.015
1111	.685	-.273	.004	-.022	.058	-.061
1112	.443	.021	-.112	-.217	.009	.048
1113	.441	.060	.217	-.077	-.074	-.127
1114	.316	.037	-.012	.033	.053	.063
1115	.388	-.133	.071	-.092	.037	.083
1116	.439	-.118	-.045	-.122	.330	.029
1117	.691	.022	-.081	-.139	-.097	-.052
1118	.684	-.087	.258	-.047	-.102	.090
1119	.554	.080	.004	-.065	-.112	-.082
1120	.553	-.084	-.014	.011	.249	-.104
1121	.537	.062	.225	.004	-.014	.120
1122	.517	.206	.237	.075	.032	.190
1123	.668	-.106	-.087	-.068	-.043	-.075
1124	.555	.005	-.052	.050	-.012	.025
1125	.511	-.199	-.127	.130	-.055	.243
1126	.522	.050	-.048	-.034	.057	.033
1127	.327	.064	-.111	-.225	-.119	.011
1128	.684	.094	-.227	.052	-.037	-.112
1129	.561	.052	.139	-.107	-.178	-.070
1130	.633	.106	.276	-.106	-.135	.041
1131	.647	.105	-.195	.011	.042	.021
1132	.497	.272	-.040	.129	.014	.055
1133	.539	.053	-.065	.093	-.282	-.003
1134	.400	.028	.078	.172	-.044	-.234
1135	.347	.265	-.042	.095	.162	-.035
1136	.564	.245	-.109	-.009	.092	-.123
1137	.373	.089	-.126	-.145	-.072	.103
1138	.331	.035	.017	.094	-.002	-.157
1139	.518	.189	.110	.060	-.070	.204
1140	.461	.237	.022	-.086	-.002	-.019
1141	.384	.204	-.043	-.017	.187	.290
1142	.510	.161	.048	-.007	.059	-.095
1143	.142	.257	-.054	.368	.136	.046
Eigenvalue	11.4174	1.0398	.7704	.6788	.6395	.6288

not much larger than the third or other eigenvalues, as was expected from the interitem correlation matrix.

The factor loading matrix of the 43 response tendencies for the first six common factors also is shown in Table 1. Because of the confidentiality of the test items, the items in all tables are listed in order of the estimated difficulty parameter $\hat{b}_g$, and an 11 (which stands for Level 11) has been added as a prefix to each item number. Thus, Item 1124 here does not mean Item 24 of the Level 11 Vocabulary Test.

All factor loadings on the first common factor were positive and, except for Items 1101 and 1143, were greater than .300 and ranged from .316 to .691. The largest cluster of factor loadings on a common factor, excluding those on the first common factor, was on the fourth factor (i.e., .393 for Item 1102 and .368 for Item 1143). Most of the factor loadings on the other common factors were less than .300 in absolute value. Based on these results, the first common factor was operationally defined as vocabulary ability (the θ), and the entire set of items was used as the Old Test. Table 2 shows the proportion correct (p_g), the normal deviate ($\hat{\gamma}_g$) corresponding to the proportion correct p_g, $\hat{a}_g$, and $\hat{b}_g$, for each item g of the Old Test.

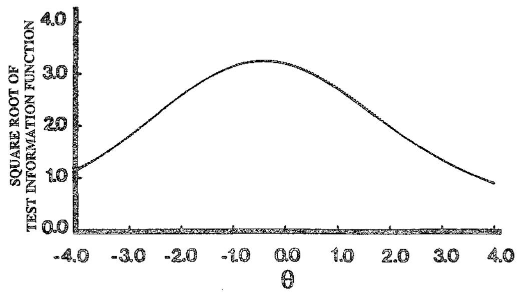
Test Information Function of the Old Test and the Transformation of θ to τ

Figure 1 shows the square root of $I(\theta)$ of the Old Test (solid line). The Old Test is most informative around $\theta = -.4$, which was slightly below the average θ level of the examinees. This was expected because 30 out of the 43 items had negative b_gs, and 28 items had difficulty parameters between −1.0 and .2 (see Table 2). Figure 1 also shows the polynomial of degree 7 obtained by the method of moments (dotted line), using the θ interval (−5.0, 5.0). The formula for this polynomial is given by

Table 2
Estimated Item Discrimination Parameter $\hat{a}_g$, Estimated Item Difficulty Parameter $\hat{b}_g$, Proportion Correct p_g, and Normal Deviate $\hat{\gamma}_g$ for Each of the 43 Old Test Items

Item	$\hat{a}_g$	$\hat{b}_g$	p_g	$\hat{\gamma}_g$	Item	$\hat{a}_g$	$\hat{b}_g$	p_g	$\hat{\gamma}_g$
1101	.196	−4.257	.79315	−.81740	1123	.898	−.357	.59433	−.23870
1102	.434	−2.331	.82318	−.92755	1124	.667	−.342	.57530	−.18988
1103	.598	−1.210	.73266	−.62088	1125	.594	−.340	.56895	−.17370
1104	.867	−1.077	.75973	−.70543	1126	.612	−.318	.56599	−.16617
1105	.829	−1.000	.73816	−.63768	1127	.346	−.250	.53257	−.08173
1106	.657	−.987	.70601	−.54177	1128	.938	−.179	.54865	−.12225
1107	.531	−.948	.67174	−.44472	1129	.678	−.170	.53807	−.09557
1108	.669	−.900	.69162	−.50045	1130	.818	−.042	.51058	−.02652
1109	.672	−.867	.68570	−.48370	1131	.849	.054	.48604	.03500
1110	.614	−.821	.66624	−.42955	1132	.573	.126	.47504	.06261
1111	.940	−.803	.70897	−.55038	1133	.640	.217	.45347	.11690
1112	.494	−.781	.63536	−.34608	1134	.436	.258	.45897	.10303
1113	.491	−.731	.62648	−.32254	1135	.370	.398	.44501	.13828
1114	.333	−.676	.58460	−.21368	1136	.683	.402	.41032	.22672
1115	.421	−.626	.59602	−.24306	1137	.402	.526	.42217	.19635
1116	.489	−.569	.59856	−.24962	1138	.351	.577	.42428	.19096
1117	.956	−.557	.64975	−.38465	1139	.606	.595	.37902	.30806
1118	.938	−.485	.62986	−.33148	1140	.519	.748	.36506	.34497
1119	.665	−.468	.60237	−.25949	1141	.416	.782	.38198	.30028
1120	.664	−.420	.59179	−.23215	1142	.593	1.007	.30372	.51373
1121	.637	−.398	.58460	−.21368	1143	.143	4.175	.27665	.59282
1122	.604	−.376	.57699	−.19420					

Figure 1
Square Root of the Test Information Function $[I(\theta)]^{1/2}$ (Solid Line) and Its Approximation (Dotted Line)



$$[I(\theta)]^{1/2} \doteq 3.1915950 - 0.23604972\theta - 0.273225550\theta^2 + 0.026248259\theta^3 + 0.012315578\theta^4 - 0.0011485951\theta^5 - 0.00022787645\theta^6 + 0.0000183226697\theta^7 \, . \tag{15}$$

This polynomial and $[I(\theta)]^{1/2}$ of the Old Test are almost identical for $-4.0 < \theta < 4.0$. This was expected because the polynomial obtained by the method of moments is also the least squares solution for the specified degree of the polynomial and the interval of θ (Samejima & Livingston, 1979). [The method of moments was used for four different intervals of θ—(–4.0, 4.0), (–4.5, 4.5), (–5.0, 5.0), and (–5.5, 5.5): The interval (–5.0, 5.0) provided the best fit.]

The polynomial for transforming θ to τ was obtained by Equations 3 and 4, using the coefficients shown in Equation 15 for α_k ($k = 0, 1, 2, \ldots, 7$) in Equations 3 and 4 and setting the two constants, C_0 and C_1, to 0.0 and 4.0, respectively. The resulting polynomial of degree 8 is given by

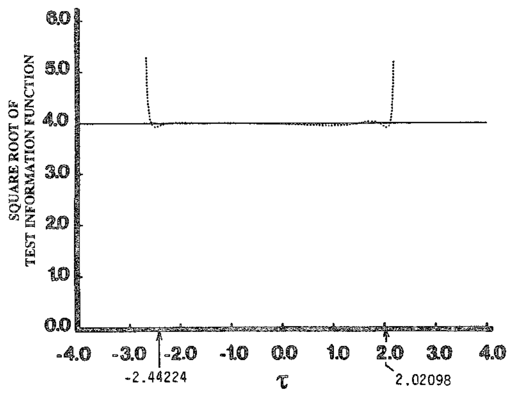
$$\tau(\theta) \doteq 0.00000000 + 0.79789874\theta - 0.029506215\theta^2 - 0.022768792\theta^3 + 0.00164051 62\theta^4 + 0.00061577891\theta^5 - 0.000047858127\theta^6 - 0.0000081384446\theta^7 + 0.00000057258428\theta^8 \, . \tag{16}$$

The square root of the test information function, $I^*(\tau)$, for the transformed latent trait τ is given by

$$[I^*(\tau)]^{1/2} = [I(\theta)]^{1/2} \frac{d\theta}{d\tau} \, . \tag{17}$$

Figure 2 presents $[I^*(\tau)]^{1/2}$ (solid line) and its approximation, which was obtained using the approxi-

Figure 2
Square Root of the Test Information Function $[I^*(\tau)]^{1/2}$ Approximated by the Polynomial Transformation of θ to τ (Dotted Line)



mated polynomial for $[I(\theta)]^{1/2}$ from Equation 17 and by using the derivative of θ with respect to τ from Equation 16 (dotted line). Because the θ interval (–4.0, 4.0) corresponds to the τ interval (–2.44244, 2.02098), the latter interval is shown by arrows in Figure 2. For this interval of τ, the approximated square root of the test information function was almost constant and was very close to 4.0.

The maximum likelihood estimate of θ, denoted by $\hat{\theta}_s$ for examinee s, was obtained for each examinee from his/her response pattern to the Old Test items. $\hat{\theta}_s$ can be directly transformed to the maximum likelihood estimate of the transformed τ, denoted by $\hat{\tau}_s$, using Equation 16. This was done for the 2,356 examinees whose $\hat{\tau}_s$s were within the interval (–3.75, 3.75). The eight examinees who had $\hat{\tau}_s$s that did not fall within that interval were excluded from the original data of 2,364 examinees.

Figure 3 presents the frequency distribution of the 2,356 $\hat{\tau}_s$s using an interval width of .25, and the polynomials of degrees 3 (dotted line) and 4 (dashed line), obtained by the method of moments. In these two cases, the interval of $\hat{\tau}$ used in the method of moments was (–1.91742, 1.95366). The lowest and the highest values of $\hat{\tau}_s$ among the 2,356 $\hat{\tau}_s$s were used as the two endpoints of the interval. These two polynomials of degrees 3 and 4 are given by

$$\hat{g}^*(\hat{\tau}) \doteq 0.42358084 - 0.046813019\hat{\tau} - 0.13270786\hat{\tau}^2 + 0.020014202\hat{\tau}^3 , \qquad (18)$$

and

$$\hat{g}^*(\hat{\tau}) \doteq 0.45023559 - 0.044232853\hat{\tau} - 0.20387563\hat{\tau}^2 + 0.018406862\hat{\tau}^3 + 0.022176405\hat{\tau}^4 , \qquad (19)$$

respectively. Figure 3 shows that these two polynomials fit the frequency distribution very well.

Estimated Plausibility Functions of the Distractors

The two cases from Equations 18 and 19 were used as estimated density functions, $\hat{g}^*(\hat{\tau})$, and will be referred to as the degree 3 case and degree 4 case, respectively. The first through fourth conditional moments of τ, given $\hat{\tau}_s$, were obtained for each individual $\hat{\tau}_s$ for the degree 3 and the degree 4 cases. From those results, β_1, β_2, and Pearson's criterion κ also were computed using Equations 5, 6, and 7, respectively.

Figure 3
Frequency Distribution of the 2,356 $\hat{\tau}$s Based on the Old Test (Solid Line) and the Two Polynomials of Degrees 3 (Dashed Line) and 4 (Dotted Line) Obtained by the Method of Moments

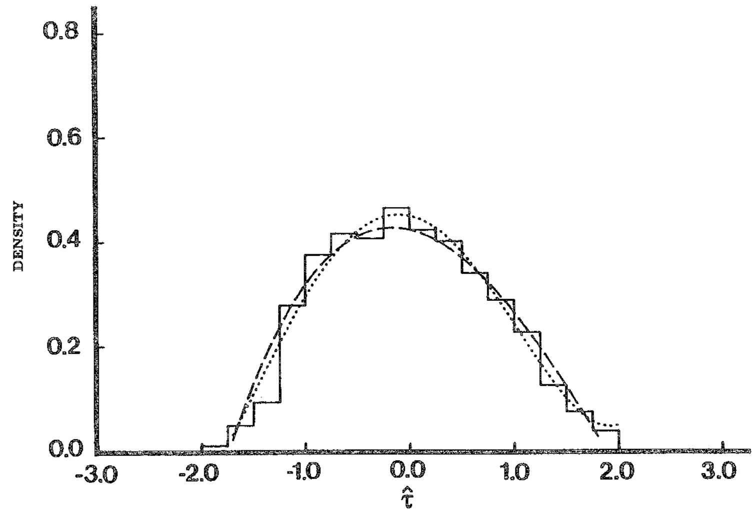


Table 3 presents the frequency distribution of the 2,356 $\hat{\tau}_s$s with respect to the types of the conditional distribution of τ, given $\hat{\tau}_s$, in both the degree 3 and 4 cases. These types—Types 1 through 7—are Pearson's types (e.g., Elderton & Johnson, 1969; Johnson & Kotz, 1970) assigned by evaluating the values of the criterion κ. In both the degree 3 and 4 cases more than 60% of the conditional distributions were assigned to the normal distribution, and most of the others belonged to the beta distribution (i.e., either Pearson's Type 1 or 2). There were some examinees whose conditional distributions of τ were undefined, either due to negative estimated even conditional moments (UND1) or to negative estimated conditional probability densities (UND2). Those examinees were excluded for the rest of the research.

Table 3
Frequency Distribution of the 2,356 $\hat{\tau}_s$s With Respect to the Pearson Types of Their Conditional Distributions of τ

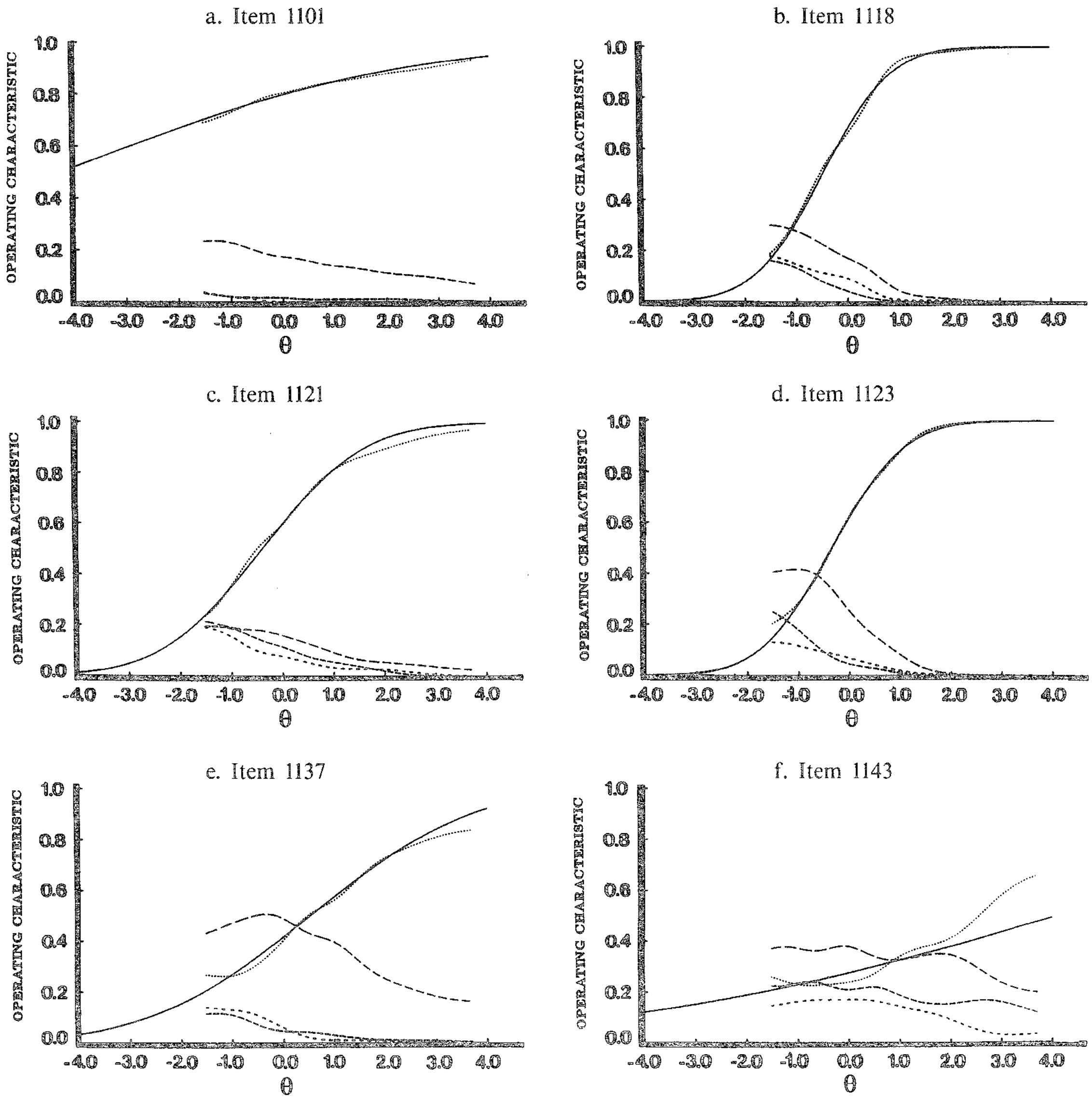
Type	Degree 3 Case	Degree 4 Case
1	362	380
2	402	220
3	0	0
4	6	69
5	0	1
6	1	8
7	0	89
Normal	1,458	1,536
UND1	112	47
UND2	15	6
Total	2,356	2,356

The above result justified the selection of the NA method in both the degree 3 and 4 cases. A close examination of the values of these indexes showed that for many examinees whose conditional distributions of τ belonged to types of Pearson's distributions other than normality, the values of β_1 and β_2 were very close to 0.0 and 3.0, respectively; these are values that characterize the normal distribution. (Note that $\beta_1 = 0$ for any distribution of Pearson's Type 2.) This result was not unexpected, because it can be shown (Samejima, 1981) that, if τ distributes normally or uniformly, the conditional distribution of τ, given $\hat{\tau}$, follows a complete or truncated normal distribution; thus, if the (unconditional) distribution of τ is close to either a normal or uniform distribution, which is likely to happen in practice, then the conditional distribution will become close to a normal distribution.

Because the results of the degree 3 and 4 cases were very similar, only the degree 4 case is considered further. It is worth noting, however, that the results of the degree 3 case were as good as the results of the degree 4 case, in spite of the fact that the degree of the polynomial approximating $g^*(\hat{\tau})$ was one less and as small as 3. A slight disadvantage of the degree 3 case was that 127 examinees were excluded from the rest of the research, but only 53 were excluded in the degree 4 case.

The OCs of the four alternatives were obtained for each of the 43 vocabulary test items using the SS procedure of the CPDF approach combined with the NA method. The 43 resultant estimated OCs are shown elsewhere (Samejima, 1984); six examples are presented in Figure 4 (dotted and dashed lines). Figure 4 also shows the normal ogive function (solid line) specified by the two estimated item parameters given in Table 2. Because of the confidentiality of the test, the order of alternatives has been changed for each item, and the correct answer is always called alternative A and its estimated

Figure 4
Estimated Operating Characteristic of the Correct Answer From the Normal Ogive Model (Solid Line) and by the Nonparametric Approach (Dotted Line), and the Estimated Plausibility Functions of Distractors B (Shortest Dashes), C (Medium Dashes), and D (Longest Dashes)



OC is represented in Figure 4 by dots, and the incorrect alternatives are called alternatives B, C, and D and their estimated plausibility functions are presented by dashes of three increasing lengths.

The results of these items in Figure 4 were selected as typical results, except for those of Item 1143. Item 1143 provided an exceptionally poor fit of the normal ogive function to the nonparametrically estimated OC of the correct answer. Most of the other 42 items provided at least reasonably good fits, as the results of Items 1101, 1118, 1121, 1123, and 1137 demonstrate.

The results provide four general observations:

1. Most of the nonparametrically estimated OCs of the correct answer were very close to the corresponding parametrically estimated normal ogive functions. (The only exceptions were those for Item 1143 and, to a lesser extent, for Items 1114 and 1115, which are not shown in Figure 4.)
2. Most of the sets of estimated plausibility functions of distractors indicated that they belonged to the informative distractor model rather than the equivalent distractor model, which means that each separate distractor provides its own differential information. The closest configuration of the estimated plausibility functions to the equivalent distractor model was for Item 1121. For some items, pairs of distractors showed almost identical estimated plausibility functions, (e.g., alternatives B and C of Item 1101).
3. Certain distractors attracted the examinees who fell in specific ranges of θ. Table 4 identifies those distractors in the columns labeled *informative*. Note that the distractors listed in Table 4 are only those that proved to be informative in the present analyses, in which the estimated OCs were truncated at the lower levels of θ.
4. Some distractors had unusually flat estimated plausibility functions. They are listed in Table 4 in the columns labeled *flat*. These distractors attracted examinees of low and high levels of θ almost equally.

Item analysis would benefit from using this type of information on the distractors. Items can be improved by identifying more informative sets of distractors.

There is no way to know about the information provided by the distractors on θ levels lower than -1.6; it is suspected that there are many more informative distractors than those identified here. They could be discovered by analyzing data collected for younger examinees in future research.

Table 4
Distractors Whose Estimated Plausibility Functions Were Informative or Unusually Flat

Item	Informative	Flat	Item	Informative	Flat	Item	Informative	Flat
1101		D	1116	C		1130	B	
1102			1117	B		1131	C	
1103	D		1118	D		1132	C	B
1104			1119	B		1133	B	
1105			1120	B		1134	C	C
1106	D		1121			1135		B,D
1107	C	B	1122			1136	C	
1108	B		1123	D		1137	D	
1109			1124	B		1138	C,D	
1110	C		1125	C		1139	C	
1111			1126	C		1140	B	
1112	D		1127	B,C		1141		B
1113		D	1128	C		1142	B,C	
1114	D	B,C	1129	C		1143	D	B,C,D
1115	C	B,D						

Figure 4 illustrates that this method can identify the implicit order of the distractors of some items. For example, in Item 1137, alternative D appeared to be the second best answer, and the order between alternatives B and C was not obvious. It may be appropriate to use graded scores for this item: 0 for either B or C, 1 for D, and 2 for A. The cumulative OC, $P^*_{x_g}(\theta)$, which indicates the conditional probability, given θ, with which the examinee obtains the item score x_g or greater for each $x_g = 0$, 1, 2, was obtained and plotted against θ, based on 17 items (Samejima, 1984). The results suggest some promise in using Samejima's family of models for the multiple-choice test item (Samejima, 1979) for these types of items.

Model Validation

There is sufficient evidence to support the set of assumptions used in the present research. For most Old Test items the normal ogive model was validated by the goodness of fit of the normal ogive functions to the corresponding nonparametrically estimated OCs of the correct answer. The normal ogive model was invalidated for Item 1143 (and to a lesser extent for Items 1114 and 1115), however.

To proceed further with model validation, the following procedure was followed. For each pair of items, using their estimated normal deviates, $\hat{\gamma}_g$ (see Table 2) and the estimated tetrachoric correlation between their response tendencies, the two-by-two contingency table was produced from the bivariate normal distribution. The χ^2 statistic of the four frequencies in the actual contingency table was computed using the frequencies in the contingency table thus produced as the theoretical values. Because the empirically obtained normal deviates and tetrachoric correlations were used to obtain the theoretical frequencies for each pair of items, the resultant χ^2 statistic should have a negligibly small value, in order for the bivariate normal assumption to be validated.

Most of the χ^2 values were very small. For example, Table 5 shows the χ^2 statistics for Items 1118 and 1117 against each of the other 42 items. Table 5 also shows these values for Items 1101 and 1143. Because the degrees of freedom were 0, Item 1143 had unusually large χ^2 statistics, as is obvious in Table 5. Note that this item provided the poorest fit of the normal ogive function to the nonparametrically estimated OC of the correct answer (see Figure 4f). In contrast, Item 1101 provided a surprisingly good fit (see Figure 4a). Items 1101 and 1143 were the easiest and most difficult items—they had the largest difficulty parameters in absolute value (see Table 2). Many other items had small χ^2 values against other items, except for those against Items 1143 and 1101. Those results are given elsewhere (Samejima, 1984).

To investigate the relationship between the χ^2 values and the goodness of fit of the normal ogive function to the nonparametrically estimated OC of the correct answer, all 43 Old Test items were categorized into four classes of fit: good, fair, poor, and very poor. Table 6 presents the item numbers thus categorized and the frequencies of χ^2 values greater than or equal to .01. A negative correlation between the goodness of fit and the frequency of these χ^2 values was observed, as expected. On the other hand, there was a substantial number of items that had high frequencies of large χ^2 statistics, and yet showed good fit of the normal ogive functions. This indicates the robustness of the entire procedure used here.

Further examination indicated that most χ^2 values greater than or equal to .01 would disappear if those items that had 13 or more such χ^2 values were excluded. In fact, some items would have no large χ^2 values (LCVs) if only two items, 1143 and 1101, were excluded. These items were classified as Category 1. Some other items would have no LCVs if 5 more items that had 19 to 21 LCVs (Items 1114, 1127, 1135, 1102, and 1138) were excluded (Category 2). Similarly, for Category 3 the 5 additional items that had 13 to 16 LCVs (Items 1137, 1116, 1115, 1134, and 1141) were excluded to eliminate LCVs. The remaining items that were not excluded nor already categorized were classified as Category 4.

Table 5
Four Examples of the Sets of χ^2 Statistics for Good (Items 1118 and 1117) and Poor (Items 1101 and 1143) Fit to Bivariate Normality

Item and χ^2 Values									
Item 1118									
.00918	.00206	.00260	.00256	.00174	.00152	.00124	.00168	.00110	.00336
.00280	.00298	.00198	.00764	.00189	.00609	.00321	.00181	.00414	.00152
.14038	.00129	.00196	.00369	.00181	.00331	.00593	.00269	.00219	.00555
.00314	.00159	.00289	.00665	.00524	.00218	.00182	.00358	—	.00147
.00127	.00224	.00204							
Item 1117									
.01794	.00124	.00174	.00200	.00176	.00142	—	.00107	.00124	.00600
.00222	.00348	.00228	.00588	.00205	.00635	.00257	.00452	.00411	.00217
.07956	.00126	.00221	.00188	.00162	.00336	.00399	.00215	.00227	.00363
.00305	.00177	.00363	.00485	.00401	.00254	.00174	.00350	.00124	.00193
.00144	.00298	.00281							
Item 1101									
—	.02309	.18162	.70985	.01323	.01246	.01794	.15688	.04920	.16785
.01971	.13859	.01066	.02627	.03445	.09895	.02763	.02447	.01853	.03045
.88370	.02561	.05016	.03250	.04378	.02991	.73564	.02281	.03418	.20828
.03535	.03341	.01863	.04813	.13741	.01430	.01466	.11926	.00918	.02732
.02255	.03192	.02357							
Item 1143									
.88370	.04498	.22610	.13053	.13433	.29440	.07956	.02398	.71174	.03705
.05161	.03860	.20463	.01998	.02249	.30588	.01853	.06475	.02390	.06416
—	.50696	.01871	.10999	.02130	.23743	.04560	.02542	.03617	.09646
.02215	.02188	.03187	.01408	.02610	.07153	.15061	.05051	.14038	.04156
.09809	.01736	.02086							

Table 7 presents the item numbers of these 31 categorized items in relation with the goodness of fit of the normal ogive function to the corresponding nonparametrically estimated OC of the correct answer. There was a substantial difference between the goodness of fit of Category 3 and those of Categories 1 and 2, as expected. These results further support the procedure used here.

Discussion and Conclusions

Most of the test items did not follow the equivalent distractor model. In fact, many distractors were informative, and the results suggested that most of these items belong to the informative distractor model.

The results support the set of assumptions adopted here. First, the normal ogive model assumed for the Old Test items in the first stage was well supported by the goodness of fit of the OCs of the correct answer estimated by assuming the normal ogive model to those estimated by the nonparametric approach. The results support not only the assumed normal ogive model but also the multivariate normality assumption assumed for the 43 response tendencies. Secondly, the use of the NA method was well supported by the estimated conditional moments of τ, given $\hat{\tau}_s$.

It may be more appropriate to repeat the entire procedure with a revised Old Test, which excluded Item 1143 and possibly Items 1114 and 1115 that do not show very good fits of the normal ogive functions. Considering that there were only three items with poor or very poor fit, however, the repetition with the new Old Test could not be expected to provide substantially improved results.

Because of the confidentiality of the test, the content of the alternatives that were informative or that provided unusual configurations of the estimated plausibility functions cannot be discussed.

Table 6
Item Numbers Categorized with Respect to Their Frequency of the χ^2 Statistics $\geq$ .01 and to the Goodness-of-Fit Between the Normal Ogive Curve and the Estimated OC of the Correct Answer

	Fit			
Frequency	Good	Fair	Poor	Very Poor
1	1118			
2	1104, 1117, 1123, 1128, 1130	1131		
3	1105, 1111, 1126	1129		
4	1106, 1108, 1109	1136		
5	1122, 1124	1120, 1132, 1133, 1140		
6	1110, 1119	1107, 1121, 1142		
7		1125, 1139		
8		1113		
9	1103, 1112			
13	1141	1134	1115	
15	1116			
16		1137		
19	1138			
20	1102			
21	1127, 1135		1114	
41	1101			
42				1143
Total	25	15	2	1

In many cases, however, possible reasons for these results could be identified, although in some cases the results were puzzling. In any case, these results provide valuable information for item analysis, and for improvement of the set of distractors to eventually make all of them informative.

The methodology used here appears to be promising, and should be useful in future research. A next step should be to determine how to use the differential information obtained from the distractors and the correct answers, in order to increase the efficiency of the estimation of the latent trait. At this stage, to simplify the mathematics, parameterization of the nonparametrically estimated plausibility functions of the distractors is advisable. If no existing parametric models (e.g., Samejima's family of models for the multiple-choice test item) fit, then Ramsay & Wang's (1993) hybrid item response theory models will be useful. They are very general models that have substantial flexibility, and their OCs provide very good fit to functions of various shapes.

Table 7
Item Numbers of the 31 Items for Which χ^2 Values Became Less Than .01 When Certain Items Were Excluded

	Fit	
Category	Good	Fair
1	1104, 1117, 1118, 1123, 1128, 1130	1131
2	1105, 1106, 1122, 1124, 1126	1129
3	1108, 1109, 1110, 1111, 1119	1107, 1113, 1120, 1121 1132, 1133, 1136, 1139, 1140
4	1103, 1112	1125, 1142
Total	18	13

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

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.

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

Elderton, W. P., & Johnson, N. L. (1969). *Systems of frequency curves.* New York: Cambridge University Press.

Frane, J. W., & Hill, M. (1974). *Annotated computer output for factor analysis.* Los Angeles CA: BMDP Statistical Software.

Hieronymus, A. N., & Lindquist, E. F. (1971). *Iowa Tests of Basic Skills* (Levels Edition), Form 6. Boston: Houghton Mifflin.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions* (Vol. 2). Boston: Houghton Mifflin.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Office of Naval Research Rep. 84-4, N00014-83-K- 0397). Champaign: University of Illinois, Model-Based Measurement Laboratory.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M. (1959). An approach to mental test theory. *Psychometrika, 24,* 283–302.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Ramsay, J. O., & Wang, X. (1993, June). *Hybrid IRT models.* Paper presented at the 1993 Psychometric Society meeting, Berkeley CA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson and Lydiche.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometric Monograph,* No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph,* No. 18.

Samejima, F. (1977). *Estimation of the operating characteristics of item response categories I: Introduction to the two-parameter beta method* (Office of Naval Research Rep. 77-1, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978a). *Estimation of the operating characteristics of item response categories II: Further development of the two-parameter beta method* (Office of Naval Research Rep. 78-1, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978b). *Estimation of the operating characteristics of item response categories III: The normal approach method and the Pearson system method* (Office of Naval Research Rep. 78-2, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978c). *Estimation of the operating characteristics of item response categories IV: Comparison of the different methods* (Office of Naval Research Rep. 78-3, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978d). *Estimation of the operating characteristics of item response categories V: Weighted sum procedure in the conditional P.D.F. approach* (Office of Naval Research Rep. 78-4, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978e). *Estimation of the operating characteristics of item response categories VI: Proportioned sum procedure in the conditional P.D.F. approach* (Office of Naval Research Rep. 78-5, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1978f). *Estimation of the operating characteristics of item response categories VII: Bivariate P.D.F. approach with normal approach method* (Office of Naval Research Rep. 78-6, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Office of Naval Research Rep. 79-4, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1980). *Estimation of the operating characteristics when the test information of the Old Test is not constant I: Rationale* (Office of Naval Research Rep. 80-2, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1981). *Final report: Efficient methods of estimating the operating characteristics of item response categories and a challenge to a new model for the multiple-choice item* (Office of Naval Research Final Research Rep. of N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1984). *Plausibility functions of Iowa Vocabulary test items estimated by the simple sum procedure of the conditional P.D.F. approach* (Office of Naval Research Rep. 84-1, N00014-81-C-0569).

Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1988). *Final report: Advancement of latent trait theory* (Office of Naval Research Final Research Rep. of N00014-81-C-0569). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1990a). *Differential weight procedure of the conditional P.D.F. approach for estimating the operating characteristics of discrete item responses* (Office of Naval Research Report 90-4, N00014-87-K-0320). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1990b). *Final report: Validity study in multidimensional latent space and efficient computerized adaptive testing* (Office of Naval Research Final Research Rep. of N00014-87-K-0320). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1991, June). *An efficient nonparametric method for estimating the operating characteristics of discrete responses.* Paper presented at the 1991 Psychometric Society meeting, New Brunswick NJ.

Samejima, F., & Livingston, P. S. (1979). *Method of moments as the least squares solution for fitting a polynomial* (Office of Naval Research Rep. 79-2, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F., & Trestman, R. L. (1980). *Analysis of Iowa data I: Initial study and findings* (Office of Naval Research Rep. 80-1, N00014-77-C-0360). Knoxville: University of Tennessee, Department of Psychology.

Acknowledgments

The research was mostly supported by the Office of Naval Research through contracts N00014-77-C-0360, N00014-81-C-0569, N00014-87-K-0320, and N00014-90-J-1456. The dataset was obtained courtesy of William Coffman, then director of the Iowa Testing Programs and professor at the University of Iowa. The author is obliged to him and his successor, Leonard S. Feldt.

Author's Address

Send requests for reprints or further information to Fumiko Samejima, 310B Austin Peay Building, The University of Tennessee, Department of Psychology, Knoxville TN 37996-0900, U.S.A. Internet: samejima@psych1.psych.utk.edu.