Graduate Management Admission Council®

Nonparametric Online Item Calibration

Fumiko Samejima

University of Tennesee

Keynote Address Presented June 7, 2007

Abstract

In estimating the operating characteristic (OC) of an item, in contrast to parametric estimation, nonparametric estimation directly approaches the entire conditional probabilty curve, without assuming any mathematical form. Samejima proposed several nonparametric methods in the 1970s and 1980s, under several multi-year research contracts with the Office of Naval Research. Later, one of them, the conditional p.d.f. approach, was adapted to the environment of computerized adaptive testing, utilizing its strengths. In the present research, the truncated logistic model, which leads to higher accuracy in estimating the item characteristic functions (ICFs) of dichotomous items, was used. The results of simulations showed that this method faithfully depicted even complicated changes in nonmonotonic ICFs using as few as 1,202 hypothetical examinees.

Acknowledgment

This research was partly funded by the Law School Admission Council, 1999–2001. Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Samejima, F. (2007). Nonparametric online item calibration. In D. J. Weiss (Ed.).
Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.
Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Fumiko Samejima, 7910 High Heath Drive, Knoxville, TN 37919, U.S.A.
Email: fsamejim@utk.edu

Nonparametric Online Item Calibration

Parametric Versus Nonparametric Estimation of Operating Characteristics

Operating Characteristic of a Discrete Item Response

Let θ be the latent trait, K_g be any *discrete* response to item g and k_g denote its realization. The *operating characteristic* (OC) $P_{k_g}(\theta)$, of a specific discrete item response k_g is defined by

$$P_{k_g}(\theta) \equiv \text{prob.}[\, K_g = k_g \mid \theta \,] . \tag{1}$$

When item g is a dichotomous item, the binary item score is denoted by U_g, with u_g (= 0 or 1) as its realization. The *item characteristic function* (ICF), $P_g(\theta)$, is defined by

$$P_g(\theta) \equiv \text{prob.}[\, U_g = 1 \mid \theta \,] . \tag{2}$$

The three-parameter logistic model (3PL) (Birnbaum, 1968), whose ICF is defined by

$$P_g(\theta) = c_g + (1 - c_g)\ [1 + \exp\{-D\, a_g\, (\theta - b_g)\}]^{-1} , \tag{3}$$

with D usually set equal to 1.702, is the most widely used model for the multiple-choice test item in paper-and-pencil testing as well as in computerized adaptive testing CAT).

Differences Between Parametric and Nonparametric Estimations

In parametric estimation of the OC or ICF, a mathematical form is presumed, and thus the estimation is reduced to that of item parameters (e.g., a_g, b_g and c_g in the three-parameter logistic ,model, or 3PL), whereas in nonparametric estimation no mathematical forms are assumed for the OC or ICF, and researchers let their research data discover itself. Thus parametric estimation has an advantage of simplicity, while developing a method of nonparametric estimation is naturally more difficult and challenging. From the truly scientific standpoint, however, the latter is more appropriate because it will make us avoid molding our data into a specific mathematical formula that might not be relevant. Thus it is advisable to use nonparametric estimation more often, especially on the initial stage of research.

Lord (1970) has developed a nonparametric estimation method and applied it for SAT Verbal Test data, and concluded that Birnbaum's 3PL may be supported for those test items. One restriction of Lord's nonparametric method is that it works only for a large set of data, like those collected at the Educational Testing Service. Later, based on separate rationales, Samejima (1981, 1984b, 1998), Levine (1984), Ramsay (1991), etc., developed approaches and methods that do not require such a huge set of data.

The Conditional P.D.F. Approach

Samejima (1998) proposed three different nonparametric approaches, and concluded that the *conditional p.d.f. approach* (CPDFA) might be the most realistic approach in the sense that it works well without taking too much CPU time, and it is possible to handle many *target* items together until the last stage where the procedure is branched to estimating OCs of separate target items. For its rationale and mathematical logic, the reader is directed to Samejima (1998) and/or her many ONR research reports (most of which are still available on request).

The final outcome of the CPDFA is:

$$\hat{P}_{k_g}(\theta) = \frac{\Sigma_{s\in k_g} \hat{W}_{k_g}(\tau;\hat{\tau}_s)\ \hat{\phi}(\tau \mid \hat{\tau}_s)}{\Sigma_{s=1}^{N} \hat{W}_{k_g}(\tau;\hat{\tau}_s)\ \hat{\phi}(\tau \mid \hat{\tau}_s)} \quad . \tag{4}$$

where s is an individual examinee, τ denotes a strictly increasing transformation of ability θ, $\hat{\tau}_s$ is the maximum likelihood estimate (MLE) of ability τ for

individual s, $\hat{\phi}(\tau \mid \hat{\tau}_s)$ denotes the estimated conditional density function of τ, given $\hat{\tau}_s$, and

$\hat{W}_{k_g}(\tau;\hat{\tau}_s)$ is the estimated *differential weight function* (DWF) such that

$$W_{k_g}(\tau;v) \equiv \frac{prob.[k_g \mid \tau]}{prob.[k_g \mid v]} \quad . \tag{5}$$

The procedure of using Equation 4 as the estimate of the OC or ICF is called the *Differential Weight Procedure* (DWP). When DWF is set equal to unity for all τ and $\hat{\tau}_s$, Equation 4 becomes

$$\hat{P}_{k_g}(\theta) = \frac{\Sigma_{s\in k_g}\ \hat{\phi}(\tau \mid \hat{\tau}_s)}{\Sigma_{s=1}^{N}\ \hat{\phi}(\tau \mid \hat{\tau}_s)} \quad . \tag{6}$$

and the procedure of using Equation 6 as the estimate of OC or ICF is called the *Simple Sum Procedure* (SSP).

It is noted that Equation 5 is the OC or ICF itself, so there is no way to estimate it except for using the outcome of SSP as its estimate. Thus if the outcome of SSP (Equation 6) turns out to be close enough to the true OC or ICF, it can serve as a good estimate of DWF, and it is expected that the outcome of DWP will become a better estimate of the OC or ICF than that of the SSP.

Simulated Data and Research Procedure

Selection of a Set of 300 Core Items (Old Test)

From the 2,131 dichotomous items that LSAC previously administered, that are represented by three estimated parameters, a_g, b_g and c_g in Equation 3 of the 3PL, 300 core items were selected to serve as the item pool for online item calibration. Thus, an effort must be made to select core items that provide large enough amounts of test information for a wide range of ability to make estimation of the ICFs of new, target items accurate.

Because the c_g (guessing parameter) in the 3PL provides nothing but noise, first, all items whose estimated c_g were 0.2 or greater were discarded from the original 2,131 items, and only the remaining 1,452 items were considered.

Second, because the first a_g b_g parameters in the 3PL no longer serve as the discrimination and difficulty indices, respectively, some was needed device to make the appropriate selection of the core items. Figure 1a illustrates how the c_g *parameter* in the 3PL makes the item easier than the parameter b_g indicates, and also less discriminating than the parameter a_g indicates,

compared with the ICF in the logistic model (2PL) (solid line) and that in the normal ogive model (dotted line), in both of which b_g is the value of θ at which the ICF is equal to 0.5, and a_g is proportional to the slope of the ICF at $\theta = b_g$.

Thus, the ICF of each item is approximated by that of the 2PL, as illustrated in Figure 1b, for the interval of θ higher than the critical value θ_g (Samejima, 1973) below which the basic function does not decrease monotonically, (or, almost equivalently, the item response information function (Samejima, 1969, 1972, 1973) for the correct answer assumes negative values,) to avoid multi-modal likelihood functions (Samejima, 1973, Yen et. al., 1991).

Figures 2a and 2b, both from Samejima (1973), illustrate the non-monotonic basic function of $u_g = 1$ (Figure 2a) in the 3PL, and examples of multi-modal likelihood functions for response patterns that include such an item response. (For details, see Samejima, 1973.)

Thus, in the truncated 2PL the ICF can be written as

$$P_g(\theta) \begin{cases} = 0 & \text{for } -\infty < \theta < \theta_g \\ = [\,1 + \exp[-D\, a_g^*(\theta - b_g^*)\,]^{-1} & \text{for } \theta_g \leqslant \theta < \infty \end{cases} \tag{7}$$

where

$$\theta_g = \frac{1}{2D\, a_g} \log c_g + b_g \ . \tag{8}$$

Based on the truncated 2PL model specified above, 300 core items were selected in such a way that their difficulty parameters in the truncated 2PL distributed as evenly as possible for a wide range of ability, and also their discrimination parameters in the truncated 2PL were as high as possible. Table 1 illustrates the first page of the table that was actually used for the core item selection. In this table, items are arranged in the order of the difficulty parameter b_g^* in the truncated 2PL.

Hypothetical Examinees and CAT Using the Core Item Set As the Item Pool

Birnbaum (1968) has shown that the test information function, $I(\theta)$, can be obtained as the sum of all item information functions for dichotomous items, and Samejima (1969, 1972, 1973) showed the same conclusion for general discrete items that includes graded response items. (Note that this is an outcome, not the definition of the test information function.) Samejima (1973) also defined the item information function $I_g(\theta)$ as the regression of the item response information function, $I_{kg}(\theta)$, and she has shown that her definition of $I_g(\theta)$ includes Birnbaum's item information function as a special case. Because in this paper both core items and target items were dichotomous items, it is sufficient to use Birnbaum's item information function such that

$$I_g(\theta) = \frac{[P_g'(\theta)]^2}{P_g(\theta)\,(1 - P_g(\theta))} \ . \tag{9}$$

where $P'_g(\theta)$ is the first derivative of $P_g(\theta)$ with respect to θ.

Figure 1

FIGURE 1a: Illustration of the Meanings of Item Parameters, a_g and b_g in 3PL, in Comparison with Those in 2PL.

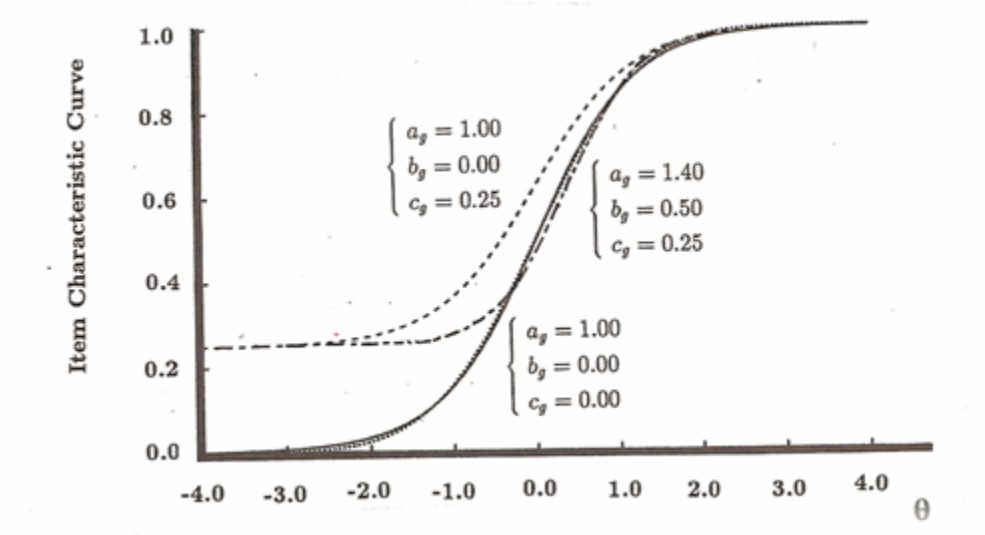


FIGURE 1b: Approximation of an ICF in 3PL by 2PL for the Interval, $(\max[\theta_g, P_g^{-1}(0.05)], P_g^{-1}(0.95))$ to Provide *Truncated 2PL*.

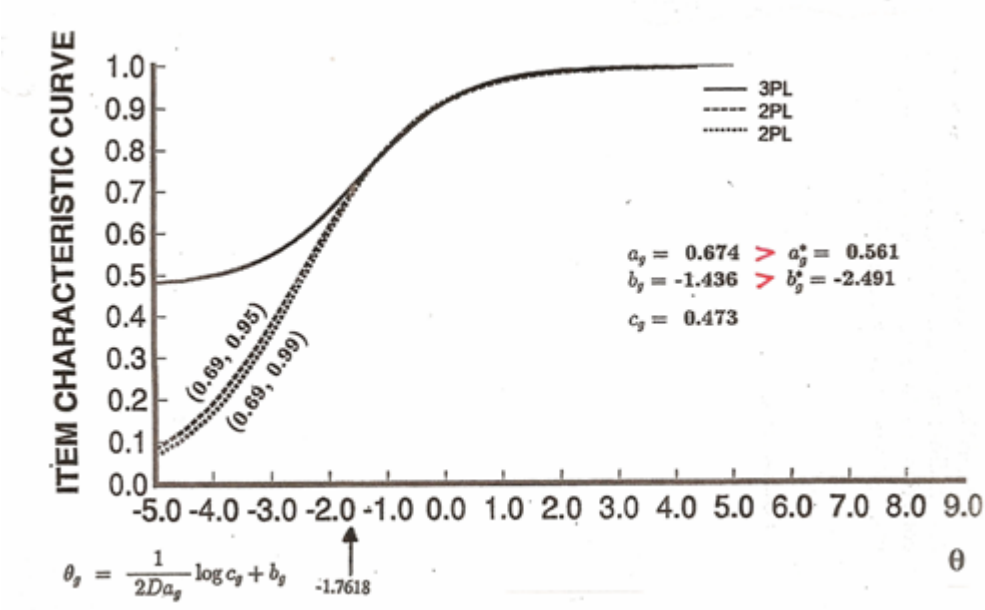


Figure 2

Figure 2a: Non-Monotone Basic Function for $U_g = 1$ in 3PL

(Cited from: Psychometrika, 38, page 229.)

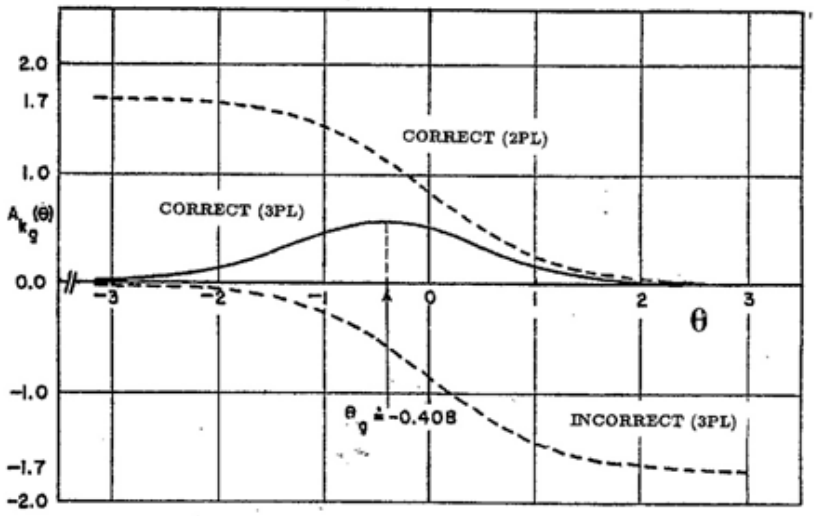


Figure 2b: Examples of Bimodal Likelihood Functions in 3PL

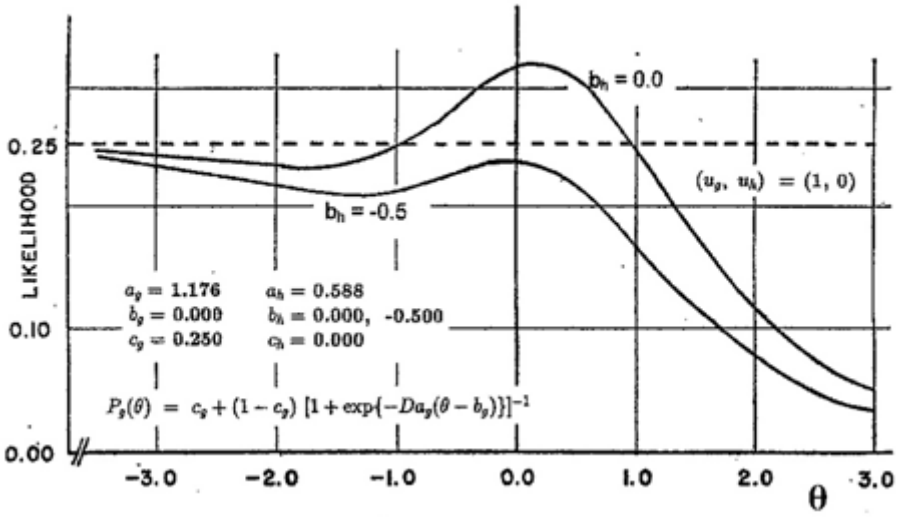


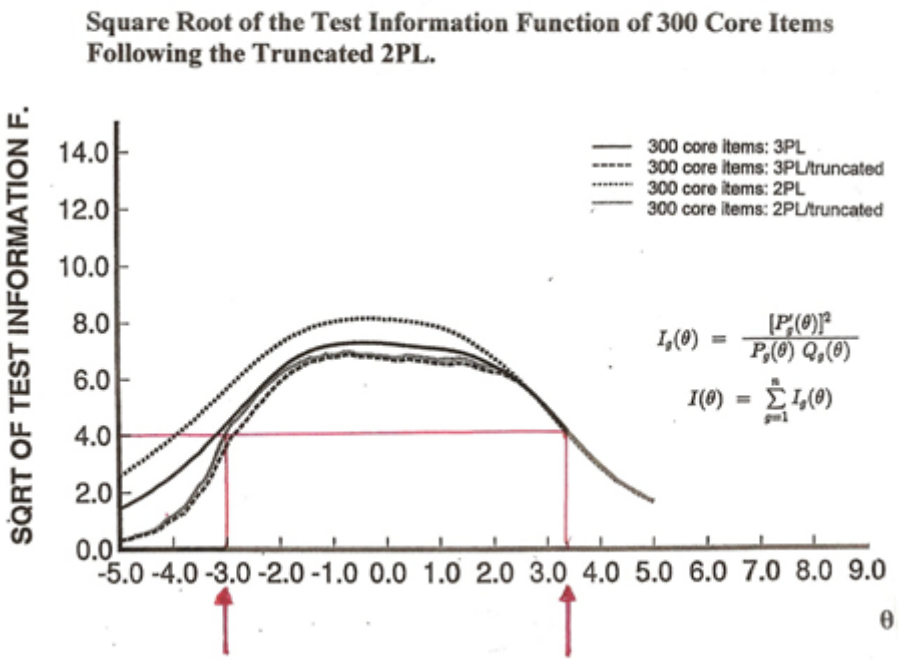
Table 1. 1,452 Items Selected From LSAC's 2,131 Dichotomous Items Discarding Those With $c_g \geq 0.20$ (First Page Only)

			a_g^*	b_g^*	c_g	θ_g	$\int_{-\infty}^{\infty} I_g(\theta)\, d\theta$ 3PL	$\int_{-\infty}^{\infty} I_g(\theta)\, d\theta$ 2PL
1	1157	ITEM 1898	0.37922	-3.37052	0.14800	-4.24932	0.42311	0.47951
2	1158	ITEM 1559	0.41675	-3.34507	0.14800	-4.14356	0.47367	0.54447
3	852	ITEM 2002	0.41888	-3.26209	0.13150	-4.18898	0.48993	0.55530
4	942	ITEM 494	0.31169	-3.25877	0.13340	-4.48193	0.34308	0.37620
5	304	ITEM 252	0.41655	-3.13899	0.06410	-4.77929	0.53560	0.56146
6	1056	ITEM 1434	0.32341	-3.07683	0.13980	-4.19128	0.36094	0.40534
7	788	ITEM 1576	0.34300	-3.07028	0.12590	-4.25503	0.39535	0.43896
8	1123	ITEM 1620	0.53026	-3.01856	0.14360	-3.67440	0.63764	0.77751
9	972	ITEM 1616	0.63202	-2.96038	0.13700	-3.54339	0.78017	0.97277
10	535	ITEM 1570	0.58076	-2.92310	0.10350	-3.77208	0.75381	0.88050
11	1022	ITEM 1375	0.75491	-2.91284	0.13860	-3.39963	0.93425	1.20736
12	541	ITEM 1560	0.38495	-2.89372	0.10350	-4.17246	0.47282	0.52333
13	1149	ITEM 1457	0.46732	-2.86878	0.14630	-3.59723	0.55524	0.67466
14	859	ITEM 1972	0.34535	-2.85185	0.13150	-3.97468	0.40102	0.45553
15	754	ITEM 1625	0.29168	-2.85021	0.12200	-4.29547	0.32897	0.36208
16	658	ITEM 659	0.63989	-2.81357	0.11790	-3.49912	0.81826	1.00126
17	534	ITEM 1573	0.53616	-2.80178	0.10350	-3.71399	0.69669	0.80793
18	804	ITEM 1717	0.36215	-2.77801	0.12830	-3.88151	0.42766	0.48959
19	532	ITEM 1575	0.31860	-2.73944	0.10350	-4.28575	0.37878	0.41395
20	632	ITEM 1833	0.83397	-2.72219	0.11550	-3.25660	1.08295	1.36914
21	536	ITEM 1569	0.57216	-2.70413	0.10350	-3.56350	0.74879	0.88333
22	927	ITEM 1668	0.69019	-2.70278	0.13270	-3.25834	0.86600	1.10450
23	1124	ITEM 1619	0.69422	-2.66388	0.14360	-3.16421	0.85503	1.11496
24	930	ITEM 1473	0.74811	-2.65927	0.13270	-3.17491	0.93874	1.21535
25	834	ITEM 2024	0.37422	-2.63598	0.13080	-3.68118	0.44599	0.51954
26	742	ITEM 1649	0.39941	-2.62283	0.12200	-3.67885	0.48807	0.56661
27	810	ITEM 1703	0.77231	-2.60784	0.12830	-3.11898	0.98341	1.26347
28	825	ITEM 656	0.60168	-2.60266	0.12940	-3.25980	0.75586	0.94644
29	904	ITEM 1749	0.37012	-2.59253	0.13250	-3.62815	0.44078	0.51442
30	755	ITEM 1569	0.80818	-2.58879	0.12200	-3.10541	1.04024	1.33046
31	1103	ITEM 1773	0.51824	-2.58863	0.14080	-3.27735	0.63356	0.79072
32	672	ITEM 1690	0.57367	-2.56040	0.11890	-3.30726	0.73606	0.89689
33	746	ITEM 1642	0.39911	-2.55848	0.12200	-3.60606	0.49086	0.56985
34	438	ITEM 216	0.55615	-2.55600	0.08730	-3.56802	0.74945	0.86427
35	857	ITEM 1981	0.71065	-2.54985	0.13150	-3.09999	0.89319	1.15337
36	542	ITEM 1558	0.32934	-2.54612	0.10350	-4.03708	0.39992	0.44215
37	533	ITEM 1574	0.44730	-2.54074	0.10350	-3.63740	0.57481	0.66081
38	298	ITEM 1182	0.56296	-2.52462	0.06310	-3.74802	0.79431	0.87924
39	970	ITEM 1618	0.93067	-2.51388	0.13700	-2.90625	1.16667	1.55601
40	885	ITEM 1556	0.48096	-2.50653	0.13190	-3.30425	0.59564	0.72619
41	1283	ITEM 491	0.37195	-2.50376	0.16760	-3.23687	0.42187	0.52252
42	882	ITEM 1932	0.74369	-2.50372	0.13190	-3.01817	0.94043	1.21721
43	1134	ITEM 1563	0.31721	-2.50352	0.14410	-3.59699	0.36087	0.42193
44	890	ITEM 1464	0.52370	-2.50254	0.13190	-3.23799	0.65220	0.80688

Figure 3 presents, by smallest dots, the square root of the test information function for the total set of 300 core items thus selected following the truncated 2PL. It can be seen in Figure 3 that the core item set provided standard errors of estimation as small as 0.25 or less for the interval of θ, (–3.0, 3.4), taking the reciprocal of the square root of test information as an approximated standard error of estimation.

1,202 examinees were hypothesized, with θ distributed uniformly for the interval of θ, (–3.0, 3.0). Using the set of core items as the item pool, by the monte carlo method a sequence of binary item scores (response pattern) of each hypothetical examinee was produced in the CAT environment, that is, after each presentation of a core item, based on the current response pattern, the examinee's MLE of θ was evaluated, and an item with the largest amount of item information at the current MLE of θ out of the remaining core items in the item pool was selected and presented.

Figure 3



In so doing, following the truncated 2PL, the amount of item information below the critical value θ_g was set equal to 0 for each core item, so that the item would never be presented if the examinee's current MLE of θ was below the critical value θ_g. Thus the use of the truncated 2PL instead of the 3PL for the core items avoided the possibility for some response patterns to have multi-modal likelihood functions (Samejima, 1973), which happen more often than many researchers are aware (Yen at al, 1991).

Five stopping rules were used, that is, presentation of new items was stopped and the testing for the examinee in question was ended when the estimated standard error of estimation (SE), i.e., the reciprocal of the square root of the test information function at the current MLE of θ, reached or exceeded 0.25, 0.32, 0.40, 0.50 , respectively, and also regardless of the estimated SE when the number of presented core items reached 40 for the sake of comparison. Note that in CAT each examinee takes a customized subset of the core items, and thus the amount of test information was evaluated as the sum total of $I_g(\theta)$'s in Equation 9 over the subset of individually customized items.

Figures 4a and 4b illustrate the number of the core items selected and presented to each of the 601 examinees in half of the total group (Figure 4a) and the square root of the amount of test information evaluated at the true value of θ (not at its MLE) (Figure 4b) when the SE = 0.32 stopping rule was used. It was observed that for the intervals of θ, (–3.0, –1.7) and (2.3. 3.0), substantially larger numbers of items had to be presented compared with the interval between the two, because of the scarceness of informative items in these intervals.

Figure 4

FIGURE 4a: Number of Items Presented to Each Examinee When the SE = 0.32 Stopping Rule Was Used, in Comparison with 40 Items. (Outcomes of the First Half Group of 601 Examinees Only.)

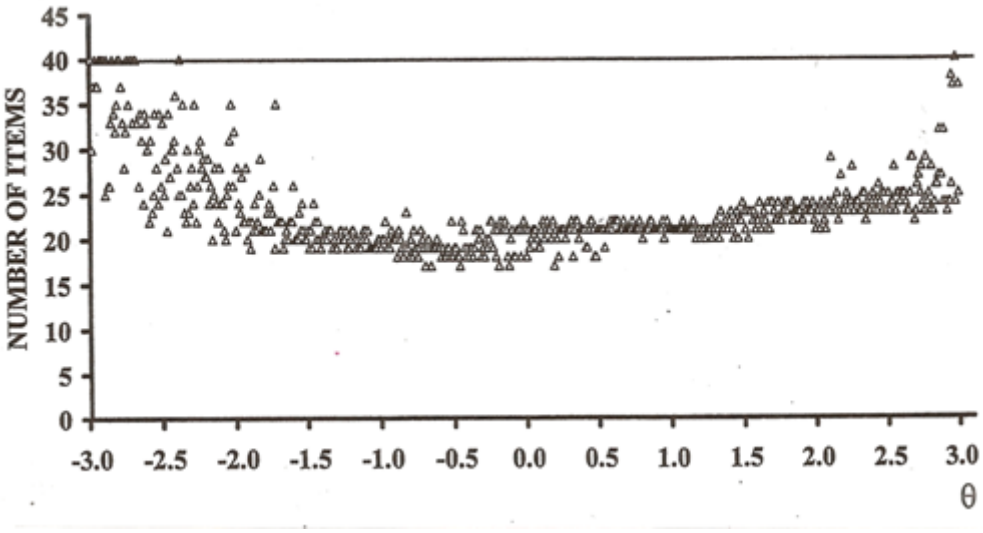
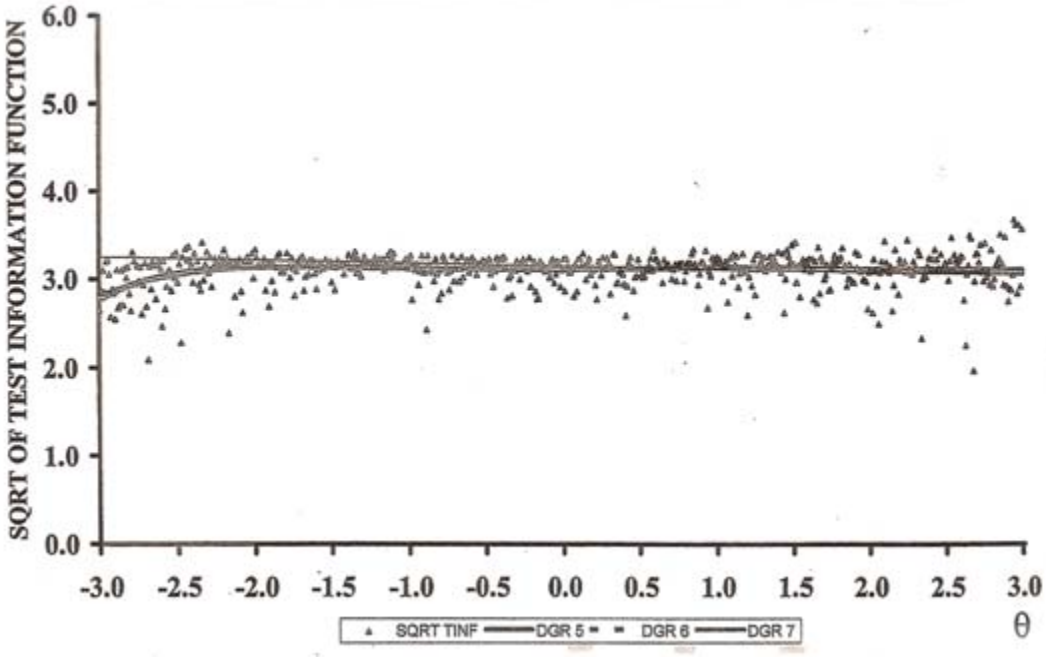


FIGURE 4b: Square Root of the Amount of Test Information for the SE=0.32 Stopping Rule Plotted against the True Values of θ . Least Squared Polinomials of Degree 5, 6, 7 are Drawn. (Outcomes of the First Half Group of 601 Examinees Only.)



Concurrently with the above CAT, 25 *target items* that did not belong to the core item set and whose ICFs were to be nonparametrically estimated were administered to each hypothetical examinee non-adaptively, scored, and kept separately until a later stage.

Transformation of θ to τ That Has a Constant Information Function

On the right hand side of Equation 4 or Equation 6)there is no θ, but τ is used that has a one-to-one mapping with θ. This transformation is given by

$$\tau = \frac{1}{C_1}\int_{-\infty}^{\theta}[I(t)]^{1/2}dt + C_0 \quad . \tag{10}$$

where C_0 is an arbitrary constant to adjust the origin of τ, and C_1 is another arbitrary constant. Let $I^*(\tau)$ denote the test information function based on the transformed ability, τ. It can be seen from the transformation formula Equation 10 that

$$\sqrt{I^*(\tau)} = C_1 \quad . \tag{11}$$

Equation 11 indicates that the test information function of τ has a constant value, and its square root equals C_1 in Equation 10. By a general characteristic of MLE, the MLE of τ is given by

$$\hat{\tau}_v = \tau(\hat{\theta}_v) \quad . \tag{12}$$

where V is a response pattern, or sequence of item scores of the customized items scores, and v denotes its realization, $\hat{\theta}_v$ is the MLE of the examinee's ability based on response pattern is v, and $\hat{\tau}_v$ is the MLE of τ for the response pattern $V = v$. Because an examinee s has one and only one v, hereafter, θ_v and θ_s are used interchangeably, and so are τ_v and τ_s, $\hat{\theta}_v$ and $\hat{\theta}_s$, and $\hat{\tau}_v$ and $\hat{\tau}_s$, respectively.

An advantage of transforming θ to τ by Equation 10 is that: (1) $F(\hat{\tau} \mid \tau)$, the conditional distribution of $\hat{\tau}$ given τ, can be approximated by $N(\tau, C_1^{-1/2})$, and (2) because of this the first through fourth conditional moments of τ, given $\hat{\tau}$, can be approximated by

$$E(\tau \mid \hat{\tau}_v) = \hat{\tau}_v + \frac{1}{C_1^2}\frac{d}{d\hat{\tau}_v}\log g(\hat{\tau}_v) \quad . \tag{13}$$

$$Var.(\tau \mid \hat{\tau}_v) = \frac{1}{C_1^2}\left\{1 + \frac{1}{C_1^2}\frac{d^2}{d\hat{\tau}_v^2}\log g(\hat{\tau}_v)\right\} \quad . \tag{14}$$

$$E[\{\tau - E(\tau \mid \hat{\tau}_v)\}^3 \mid \hat{\tau}_v] = \frac{1}{C_1^6}\left\{\frac{d^3}{d\hat{\tau}_v^3}\log g(\hat{\tau}_v)\right\} \quad . \tag{15}$$

$$\begin{aligned} E[\{\tau - E(\tau \mid \hat{\tau}_v)\}^4 \mid \hat{\tau}_v] = {} & \frac{1}{C_1^4}\{3 + \frac{6}{C_1^2}[\frac{d^2}{d\hat{\tau}_v^2}\log g(\hat{\tau}_v)] \\ & + \frac{3}{C_1^4}[\frac{d^2}{d\hat{\tau}_v^2}\log g(\hat{\tau}_v)]^2 + \frac{1}{C_1^4}[\frac{d^4}{d\hat{\tau}_v^4}\log g(\hat{\tau}_v)]\} \end{aligned} \tag{16}$$

where $g(\hat{\tau})$ is the probability density function of $\hat{\tau}$.

A close look of Equations 13 through 16 discloses that if the probability density function, $g(\hat{\tau})$, is well approximated by a four-times differentiable function, all the four moments on the left-hand side of those equations can be evaluated. Because $\hat{\theta}_v$ for each of the 1,202 hypothetical examinees are observable, using Equations 10 and 12, 1,202 values of $\hat{\tau}$ were obtained, and using the method of moments the least squared polynomial can be obtained as an estimated $g(\hat{\tau})$ (Samejima & Livingston, 1979).

Figure 5

FIGURE 5a: Frequency Distribution of MLE of π for the 1,202 Hypothetical Examinees. Stopping Rule: SE=0.32.

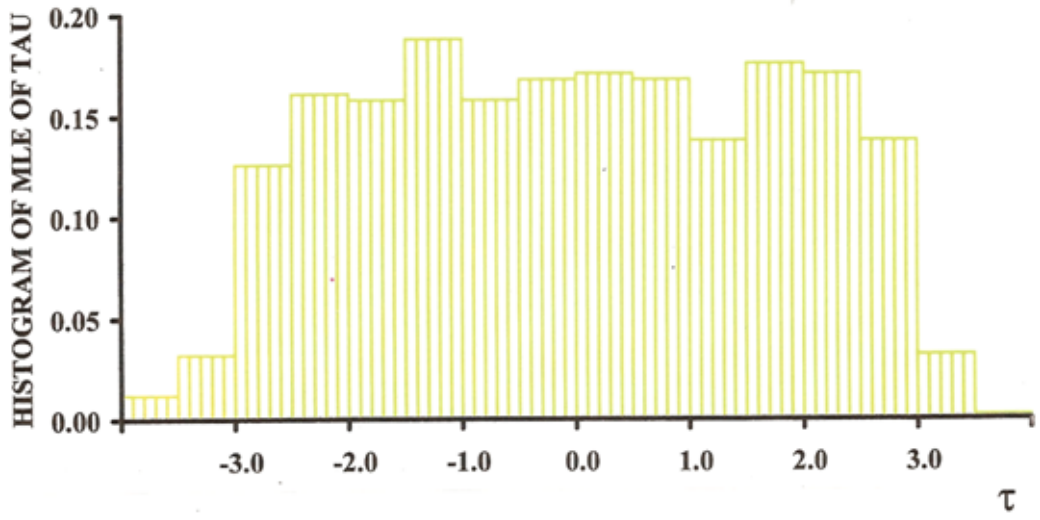


FIGURE 5b: Least Squares Polynomials of Degrees 3 and 4 for the Set of 1,202 MLEs of τ Obtained with the SE=0.32 Stopping Rule, Computed by the Method of Moments.

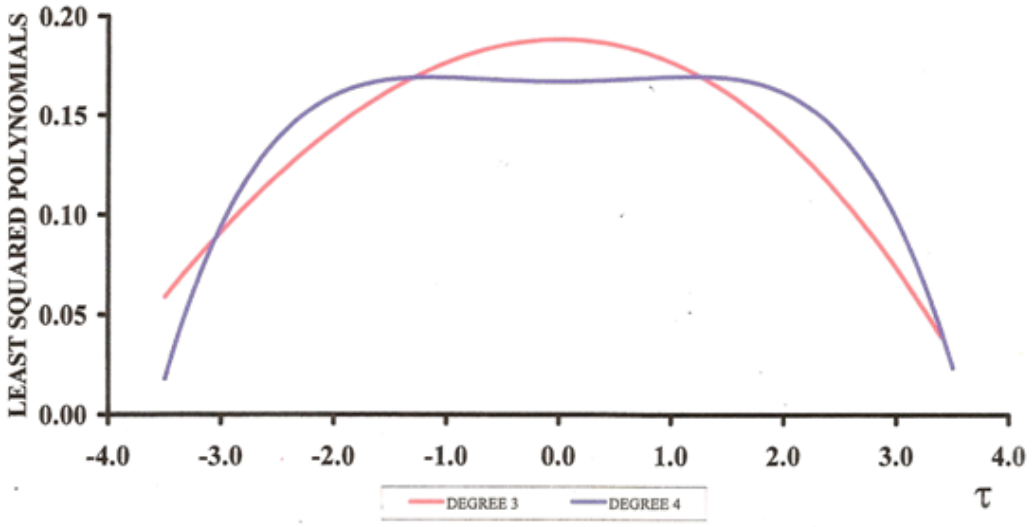


Figure 5a presents the frequency distribution of the 1,202 values of $\hat{\tau}_v$ with a small interval width of 0.10 . The polynomials of degree 3 and 4 fitted by the method of moment are shown in Figure 5b. It is noted that in the latter figure these two curves are substantially different from each other. Hereafter, we shall call it the Degree 3 Case when the polynomial of degree 3 is used as the estimate of $g(\hat{\tau})$ in Equations 13 through 16 for each of the 1,202 $\hat{\tau}_s$s , and the Degree 4 Case when the polynomial of degree 4 is used.

Using Pearson's criterion κ and other indices (Elderton & Johnson, 1969) such that

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} \quad . \tag{17}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad , \tag{18}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad . \tag{19}$$

where μ_2 , μ_3 and μ_4 are second, third, fourth moments about the mean μ_1^* . In this research, these moments were replaced by the conditional moments obtained for each $\hat{\tau}_s$ by Equations 13 through 16, respectively, in each of the Degree 3 and 4 Cases. The values of these indices branch each of the 1,202 conditional distributions to one of the Pearson system distributions, and its approximated probability density function, $\phi(\tau \mid \hat{\tau}_s)$, can be calculated as a function of τ .

Note that the entire procedure was done separately for each of the Degree 3 and 4 Cases and the outcomes were compared with each other.

Table 2 presents the first page of the output showing the process of selecting one of the Pearson's distributions in the Degree 3 Case, in the ascending order of τ_s. In this table, Type 8 indicates a normal distribution and Types 9 and 10 are junk distributions for convenience, and Type 1 and 2 are asymmetric and symmetric Beta distributions, respectively, following Pearson's naming. Frequency distributions of those types are shown as Table 3, for each of the Degree 3 and 4 Cases.

When $\beta_1 = 0$ and $\beta_2 = 3$, they indicate a normal distribution for $\phi(\tau \mid \hat{\tau}_s)$. Table 3 makes it clear that the majority of conditional distributions are approximated by normal distributions, in both Degree 3 and 4 Cases.

Pearson's Type 2 distribution, that is, a symmetric Beta distribution, also has $\beta_1 = 0$. In Table 2, there are seven Type 2 distributions observed. A close examination of their values of β_2, however, clarifies that these values are very close to 3, ranging from 2.940 to 2.974.

Even Pearson's Type 1 distribution, an asymmetric Beta distribution with $\beta_1 > 0$, that appears ten times in Table 2, with the exception of $s = 32$ that has $\beta_1 = 0.896$ and $\beta_2 = 0.786$, for all the other nine β_1 ranges from 0.006 to 0.051, and also β_2 ranges from 2.673 to 2.924, indicating that they are close to normal distributions. Both Type 1 and Type 2 (and Type 9 and Type 10, if any) distributions show up at very low and very high values of $\hat{\tau}_s$, and they are seldom seen at intermediate values of $\hat{\tau}_s$. This result is consistent with the author's many other results of the CPDFA in non-CAT environments.

Table 2: First Page of the Output Showing the Pearson's Indices and the Type of Pearson Distribution Assigned to Each of the 1,202 Examinees for the Degree 3 Case and Stopping Rule SE = 0.32 (1 = Asymmetric Beta Distribution, 2 = Symmetric Beta Distribution, 8 = Normal Distribution)

SUBJECT ID2 (ISBN)	ID1 (ISBJ)	MLE (TAU)	MEAN	MOMENTS ABOUT MEAN 2	3	4	BETA1	BETA2	CRITERION	TYPE	ID3
1	1001	-3.37808	-3.26115	0.08687	0.00413	0.02106	0.026	2.791	-0.040	1	1
2	1002	-2.68474	-2.62809	0.09833	0.00064	0.02887	0.000	2.986	-0.011	8	2
3	1003	-2.89571	-2.82734	0.09668	0.00101	0.02780	0.001	2.974	-0.015	2	3
4	1004	-3.24389	-3.14532	0.09114	0.00260	0.02606	0.009	2.897	-0.029	1	4
5	1005	-2.98576	-2.91120	0.09569	0.00126	0.02715	0.002	2.964	-0.018	2	5
6	1006	-2.98584	-2.91128	0.09569	0.00126	0.02715	0.002	2.964	-0.018	2	6
7	1007	-3.11459	-3.02945	0.09383	0.00177	0.02590	0.004	2.942	-0.022	2	7
8	1008	-3.44971	-3.32029	0.08358	0.00545	0.01867	0.051	2.673	-0.048	1	8
9	1009	-3.22719	-3.13055	0.09155	0.00247	0.02434	0.008	2.905	-0.028	1	9
10	1010	-3.34304	-3.23146	0.08818	0.00363	0.02199	0.019	2.829	-0.036	1	10
11	1011	-2.30745	-2.26556	0.10000	0.00032	0.02994	0.000	2.995	-0.007	8	11
12	1012	-3.22215	-3.12608	0.09167	0.00243	0.02443	0.008	2.907	-0.028	1	12
13	1013	-2.32067	-2.27836	0.09996	0.00033	0.02992	0.000	2.994	-0.007	8	13
14	1014	-2.35281	-2.30944	0.09985	0.00035	0.02985	0.000	2.994	-0.007	8	14
15	1015	-2.71483	-2.65670	0.09813	0.00068	0.02875	0.000	2.985	-0.012	8	15
16	1016	-3.17655	-3.08539	0.09267	0.00212	0.02511	0.006	2.924	-0.025	1	16
17	1017	-2.82183	-2.75795	0.09734	0.00086	0.02823	0.001	2.979	-0.014	8	17
18	1018	-2.54936	-2.49873	0.09906	0.00049	0.02935	0.000	2.990	-0.009	8	18
19	1019	-2.83458	-2.76996	0.09724	0.00088	0.02816	0.001	2.979	-0.014	8	19
20	1020	-2.87778	-2.81055	0.09685	0.00097	0.02791	0.001	2.975	-0.015	8	20
21	1021	-2.78565	-2.72381	0.09763	0.00079	0.02842	0.001	2.981	-0.013	8	21
22	1022	-2.96285	-2.88995	0.09597	0.00119	0.02733	0.002	2.967	-0.017	2	22
23	1023	-2.79682	-2.73436	0.09755	0.00081	0.02836	0.001	2.981	-0.013	8	23
24	1024	-2.40480	-2.35966	0.09967	0.00038	0.02973	0.000	2.993	-0.008	8	24
25	1025	-2.75650	-2.69623	0.09785	0.00074	0.02856	0.001	2.983	-0.012	8	25
26	1026	-3.00375	-2.92786	0.09547	0.00132	0.02700	0.002	2.962	-0.018	2	26
27	1027	-2.70736	-2.64960	0.09818	0.00067	0.02878	0.000	2.985	-0.012	8	27
28	1028	-3.43189	-3.30580	0.08448	0.00507	0.01934	0.043	2.709	-0.046	1	28
29	1029	-3.12171	-3.03592	0.09371	0.00181	0.02582	0.004	2.940	-0.023	2	29
30	1030	-2.75131	-2.69131	0.09789	0.00073	0.02858	0.001	2.983	-0.012	8	30
31	1031	-3.21464	-3.11941	0.09184	0.00238	0.02455	0.007	2.910	-0.027	1	31
32	1032	-3.66588	-3.47833	0.06414	0.01538	0.00323	0.896	0.786	-0.987	1	32
33	1033	-2.64561	-2.59080	0.09856	0.00059	0.02902	0.000	2.988	-0.011	8	33
34	1034	-2.18768	-2.14942	0.10034	0.00027	0.03016	0.000	2.996	-0.006	8	34
35	1035	-2.75348	-2.69337	0.09787	0.00074	0.02857	0.001	2.983	-0.012	8	35

TABLE 3: Frequency Distributions of Pearson Type Distributions for Each of Degrees 3 and 4 Cases. Stopping Rule: SE=0.32.

Degree 3 Case	
TYPE	FRQC.
1	53
2	78
3	0
4	0
5	0
6	0
7	0
8	1,070
9	1
10	0
TOTAL	1,202

Degree 4 Case	
TYPE	FRQC.
1	143
2	108
3	0
4	0
5	0
6	0
7	0
8	911
9	34
10	6
TOTAL	1,202

Each of type numbers 1-7 indicate Pearson type number except:
8: Normal distribution
9: Others
10: Undefined

For the above reasons, and because of the confirmation that mixture of a small number of non-normal distributions practically does not affect the outcome of SSP nor DWP in the non-CAT environment, normal distributions were used for the estimated $\phi(\tau \mid \hat{\tau}_s)$ for all $\hat{\tau}_s$ s, based on only the first two conditional moments. Dominance of the normal distribution is no surprise; it can be interpreted by relating it to the Dutch Identity (Holland, 1990).

Substituting those approximated conditional densities, $\phi(\tau \mid \hat{\tau}_s)$, into Equation 6. the estimated ICF is obtained as the outcome of SSP for each of the twenty-five target items; using this outcome as the DWF in Equation 4, the estimated ICF was obtained as the outcome of DWP for each target item.

In addition, the *criterion operating characteristic* is defined as the outcome obtained by using the true ICF as the DWF in Equation 4. This criterion indicates the limitation of the whole procedure adopted in the research; that is, the outcome of SSP or DWP cannot exceed the criterion operating characteristic in accuracy of estimating the true ICF, unless part or all of the procedures used in the research is revised. Note that only the use of simulated data makes it possible to obtain this criterion.

Since in the present research dichotomous items are exclusively used, it will be called c*riterion ICF* and abbreviated by Cr.ICF.

Results

As it turned out, the estimated ICFs were practically the same for each target item in the Degree 3 and 4 Cases, although these two polynomials of degrees 3 and 4 approximating $g(\tau_v)$ are substantially different (Figure 5b). Figures 6a through 6d illustrate comparisons of the outcomes of SSP, between Degrees 3 and 4 Cases, and between two stopping rules, SE = 0.32 and 40 item. It can be seen that for the same stopping rule the estimated ICFs are practically identical with each other. As was expected, however, outcome of the 40-item stopping rule and that of the SE = 0.32 stopping rule were substantially different, and the latter was closer to the true ICF than the former, in each of the Degree 3 and 4 Cases.

Figures 7a through 7h present the true ICF, the outcome of SSP, that of DWP, and the Cr.ICF for four nonmonotonic ICFs (7a through 7d) and for four monotomic ICFs, including very accurately estimated ICFs and less accurately estimated ICFs. In Figure 7a, the true ICF of item T183 had a relatively simple nonmonotonicity, and the Cr.ICF was very close to the true ICF. The outcome of SSP was already close to the true ICF and Cr.ICF, but the outcome of DWP showed a slightly closer fit. The true ICF of item T111 had a little more complicated nonmonotonicity, but tendencies similar to those observed for item T183 were also recognized, and the same was true with item T148, although its Cr.ICF haf larger discrepancies from the true ICF for the interval of θ, (–0.1, 1.2). Regardless of these differences, in all three sets of outcomes nonmonotonicities in the true ICF were well detected.

FIGURE 6: Comparison of the SSP Outcomes of the Target Item T103 between Degree 3 Case and Degree 4 Case, and between the Stopping Rule: SE=0.32 and 40 Item Stopping Rule.

FIGURE 6a: Degree 3 Case, Stopping Rule: SE=0.32.

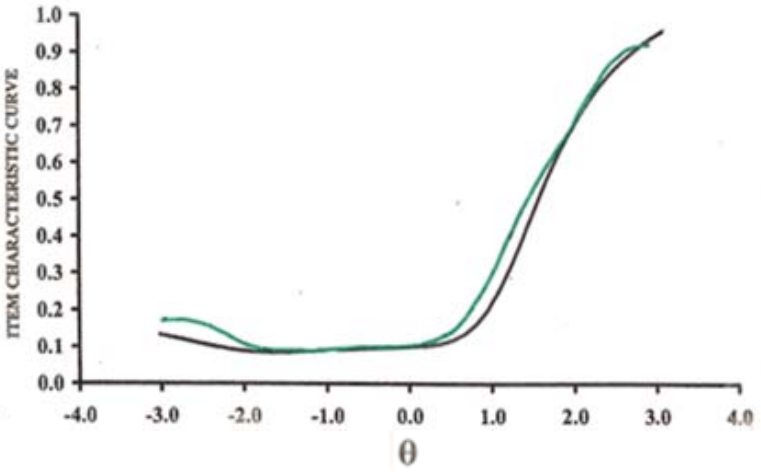


FIGURE 6b: Degree 3 Case, Stopping Rule: 40 Items.

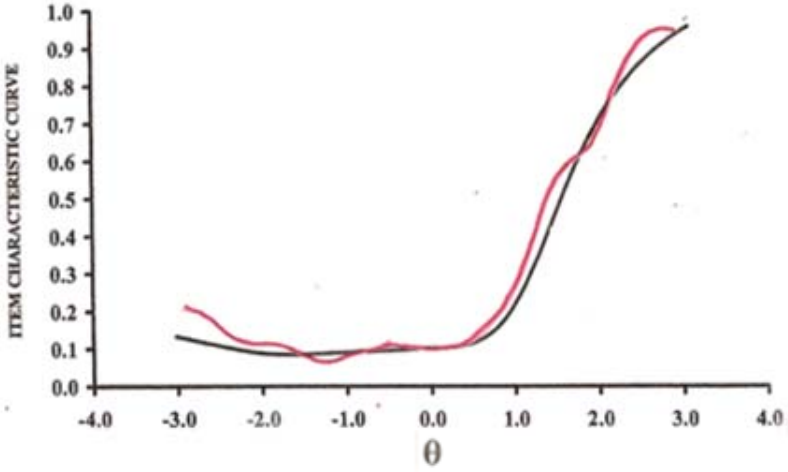


FIGURE 6c: Degree 4 Case, Stopping Rule: SE=0.32

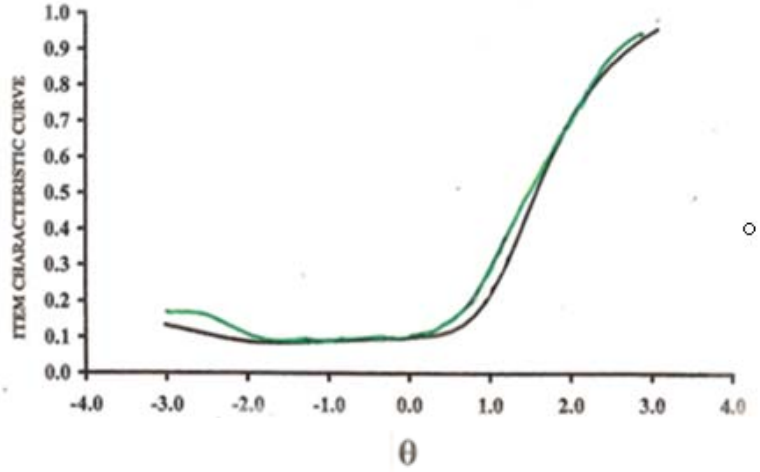
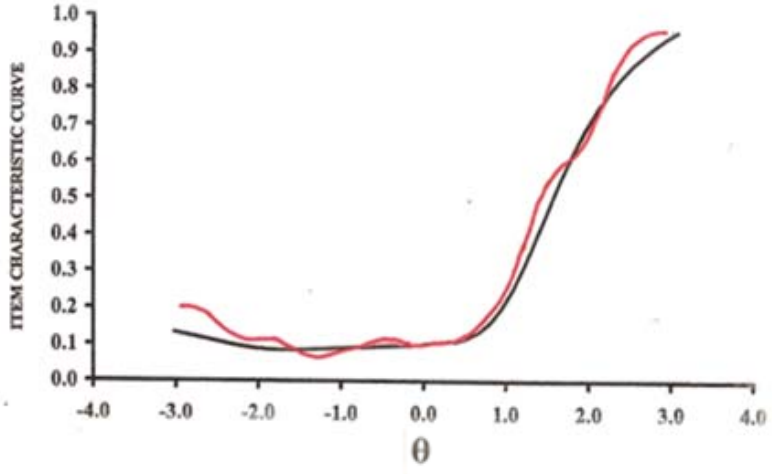


FIGURE 6d: Degree 4 Case, Stopping Rule: 40 Items.



Outcomes for item T012 in Figure 6d are of interest. The true ICF of this item had a relatively flat part for the interval of θ (–3.0, 0.6), and a very steep part for the interval of θ, (0.6, 3.0). Although the Cr.ICF for the latter interval of θ was close enough to the true ICF, and the outcome of DWP showed some improvement over that of SSP, in the former interval of θ the Cr.ICF itself had substantially large windings, and these windings are exaggerated in the outcome of DWP compared with that of SSP. One conceivable reason for these windings is the way τ-transformation was made (Figure 4b), another might be the relatively small number, 300, of items in the item pool, and consequently the standard error of estimation in the stopping rule had to be as large as 0.32. It is necessary to pursue all conceivable reasons, and further refinement of the present method is desired.

If the windings are considered as errors when they occur in the relatively flat part of the true ICF, smoothing of that part of estimated ICF might be legitimate. For the outcome of DWP for item T012, smoothing was tried by fitting the least squared polynomials of degree 3 and 4, as shown in Figure 7d, though these two curves were practically identical. If the estimated ICF by DWP is replaced by one of these curves for the interval of θ, (–3.0, 0.0), a substantial improvement will occur in the outcome of DWP.

True ICFs of the items B465, B101, B439 and B424 in Figures 7e through 7h are monotonic, that were taken from 3PL ICFs. The first two are examples of accurate estimation in the sense that the Cr.ICF was very close to the true ICF, and the outcomes of both SSP and DWP were close to these two. There are slight windings in Cr.ICF and also in the outcomes of SSP and DWP, especially for item B101, however, and they were exaggerated in the outcome of DWP caused by the use of the outcome of SSP as DWF.

Figures 7g and 7h are example of not so slight windings in the Cr.ICF, and they are more exaggerated in the outcome of DWP. In Figure 7h, assuming those windings are errors, smoothing by the least squared polynomials of degree 3 (green) and degree 4 (light blue) of the DWP outcome are drawn.

Estimated ICFs in these eight figures are the outcomes in the Degree 3 Case. As was mentioned earlier, the outcomes in the Degree 4 Case turned out to be very similar to those in the Degree 3 Case.

Discussion and Conclusions

In general, Cr.ICFs were close to their respective true ICFs, and both the SSP and DWP outcomes were close to both curves, indicating the success of the present nonparametric online item calibration approach.

Both SSP and DWP detected nonmonotonicity of ICFs accurately when it existed. In general, when nonmonotonicity existed, the outcomes of DWP tended to be closer to both the true ICF and Cr.ICF than those estimated by SSP.

Sometimes windings were observed in the estimated ICFs when the true ICF, or part of it, was relatively flat, or slowly and monotonically increasing in θ. They were observed even in the criterion ICFs, indicating that further improvement of the method is necessary, and/or increase in the sample size might be needed. However, assuming that windings for relatively flat part(s) of the ICF were errors, smoothing can be effectively made, using the least squared polynomial of a relatively small degree (e.g., 3 or 4) by the method of moments. Irregularity in some estimated

FIGURE 7: Examples of Outcomes of SSP (light blue), of DWP (dark red) Using the Outcome of SSP as DWF, Compared with the True ICF (black) and the Criterion ICF (red). (Non-Monotone ICFs)

FIGURE 7a

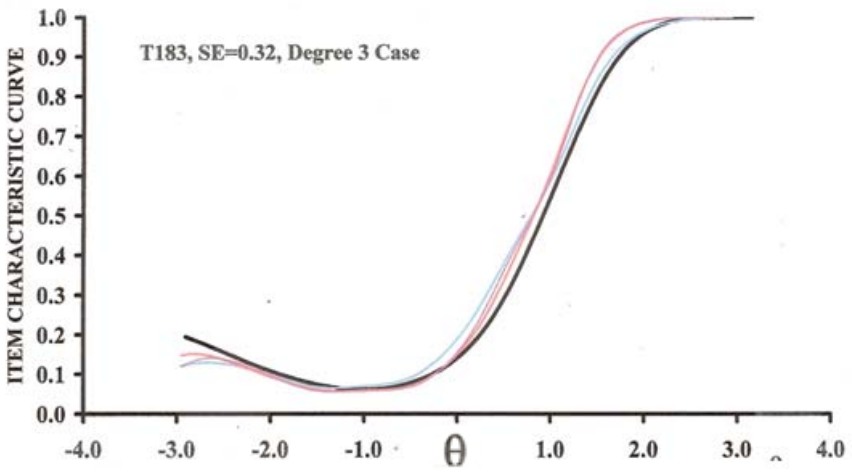


FIGURE 7c

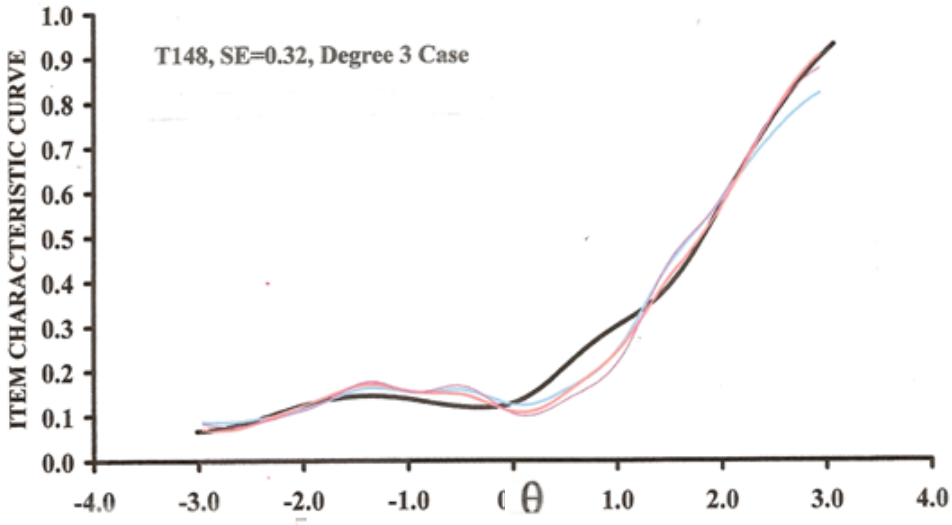


FIGURE 7b

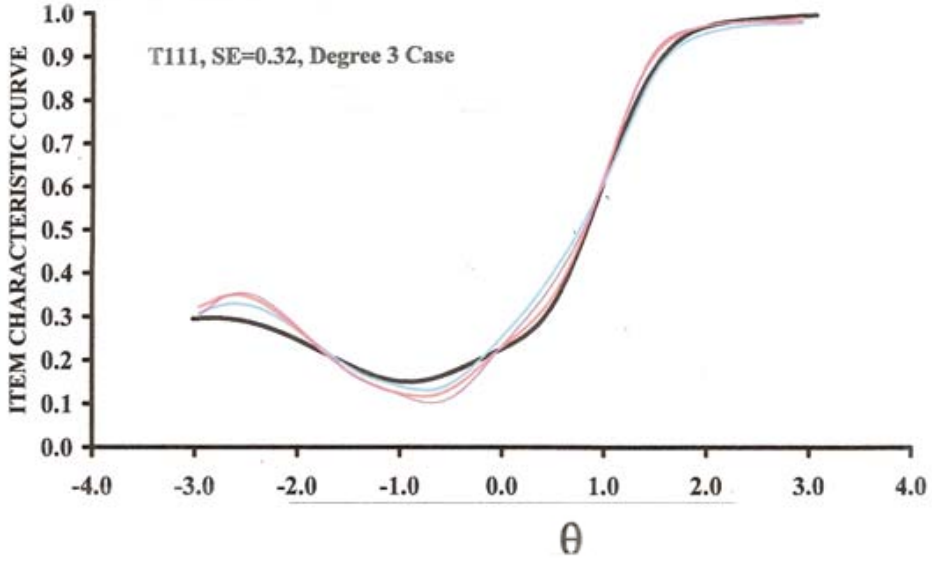


FIGURE 7d

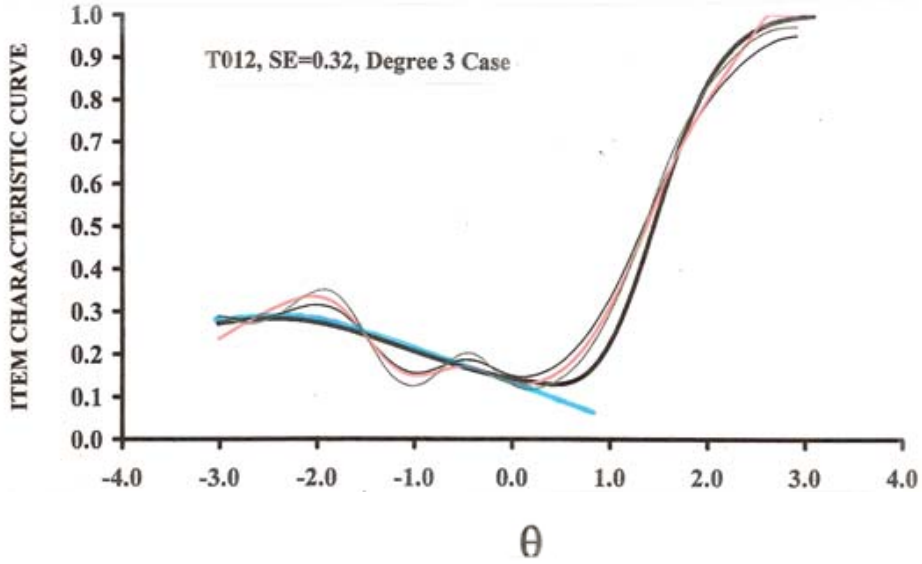


FIGURE 7e

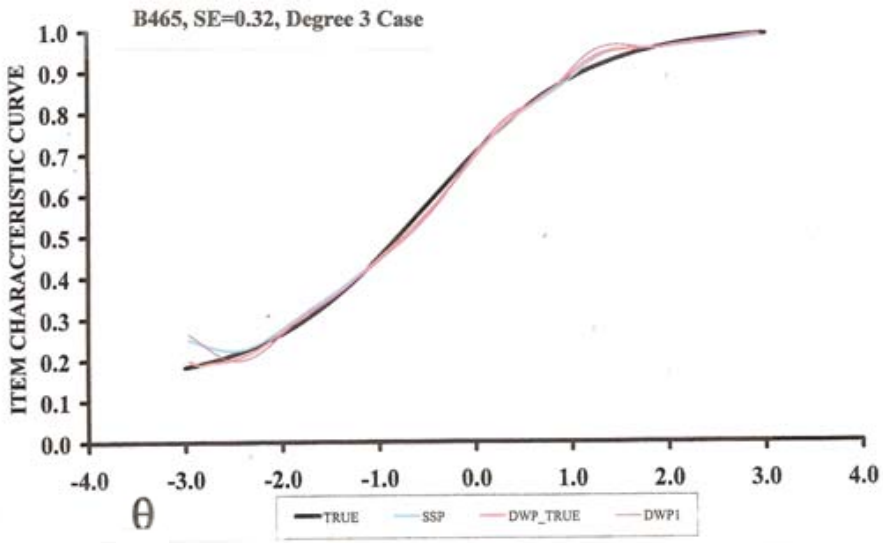


FIGURE 7f

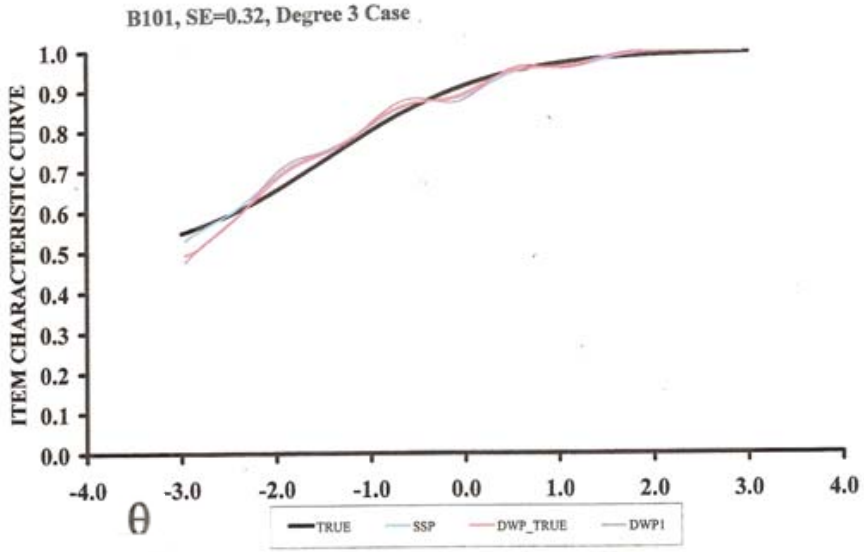


FIGURE 7g

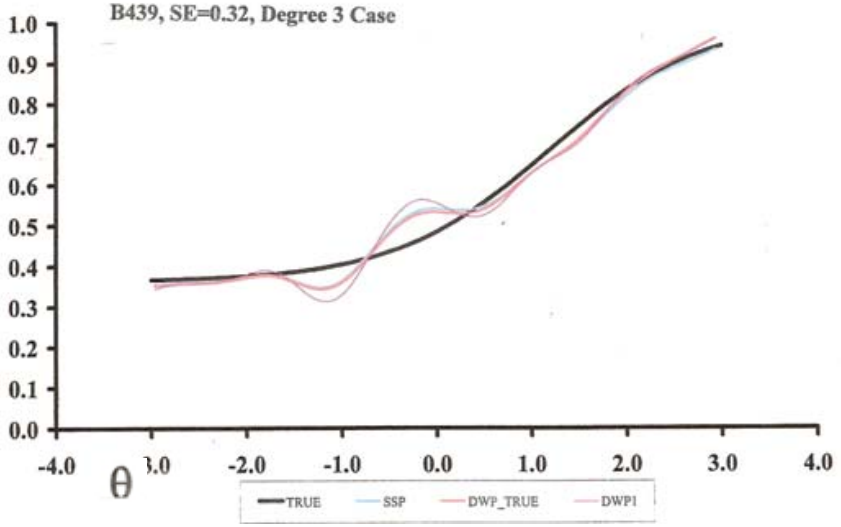
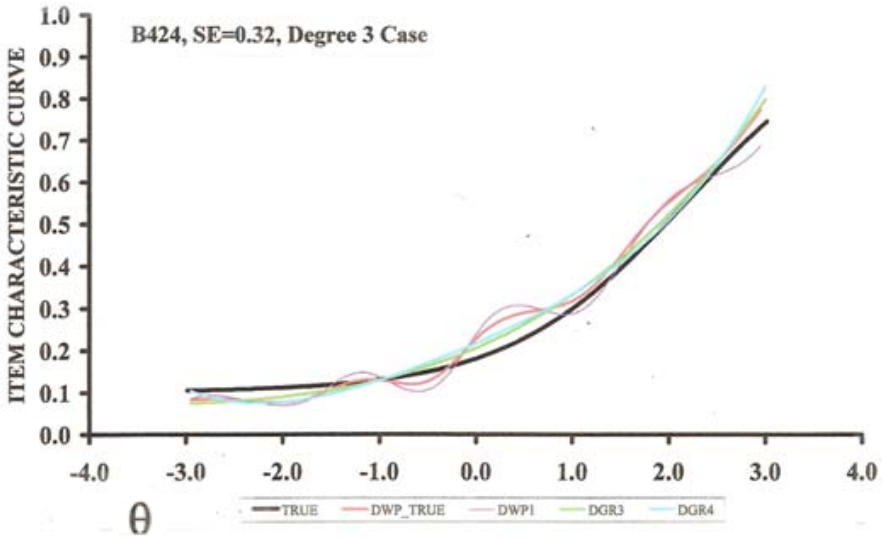


FIGURE 7h



ICFs might be ameliorated if the number of core items is increased, so that the stopping rule of, say, SE = 0.2 or less becomes possible.

The present online item calibration is perfectly applicable for graded responses both for core items and target items and, moreover, their inclusion in the core item set and use of them for items that are presented at the early stage of CAT to branch the examinees roughly over the ability scale should make CAT more efficient. (Note that, in general, a graded response item provides larger amounts of information than a dichotomous item; Samejima, 1969.)

It is interesting to note that, even if a researcher uses parametric estimation of OCs or ICFs the results might have characteristics of the outcomes of nonparametric estimation. A good example is the failure in recovering the values of the three parameters in the 3PL that has repeatedly happened in the past. A strong reason for this failure is that the third parameter, c_g , in Equation 3 is nothing but noise and it is practically impossible to estimate its value because inclusion of more examinees on lower levels of θ will create greater errors in the estimates of their ability levels .

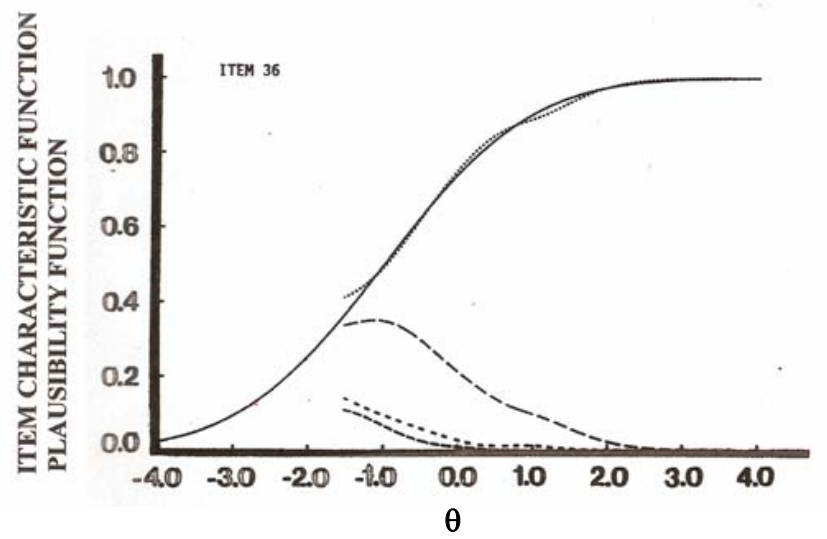
Both SSP and DWP are useful in clarifying and using information provided by incorrect alternative answers of multiple-choice items, in addition to that provided by the correct answer. Figure 8 presents two examples of the estimated operating characteristics of the incorrect alternative answers by SSP, called *plausibility functions*, discovered from the data collected for the 11 level Vocabulary Subtest of the Iowa Tests of Basic Skills (Samejima, 1984a, 1994). These examples indicate that incorrect alternative answers might have differential information that could be used in ability estimation, online or off-line. It is advisable to include such incorrect answers, the plausibility of which appeal to the examinees of different levels of ability.

Figure 9 presents the ICF and the plausibility function of the most plausible incorrect alternative answer of a hypothetical multiple-choice test item following Samejima's (1979) model, illustrating that both choices can be used in ability estimation, using a truncated normal ogive or logistic model for the correct answer (upper graph) and similarly for the most plausible incorrect answer.

The truncated 2PL or normal ogive model has many other effective applications, especially in the CAT environment. Although nonparametric item calibration is important, there has not been strong enough interest among researchers and practitioners. The present research is not really completed, but there are several things that need further investigation. The author invites the reader to participate in such investigations, as well as to use the present method for his/her own research and inform the author of its outcomes.

Figure 8

Examples of Plausibility Functions of Four Distractors Estimated by CPDFA, Based on the Empirical Data on the Level 11 Vocabulary Test of the Iowa Tests of Basic Skills.



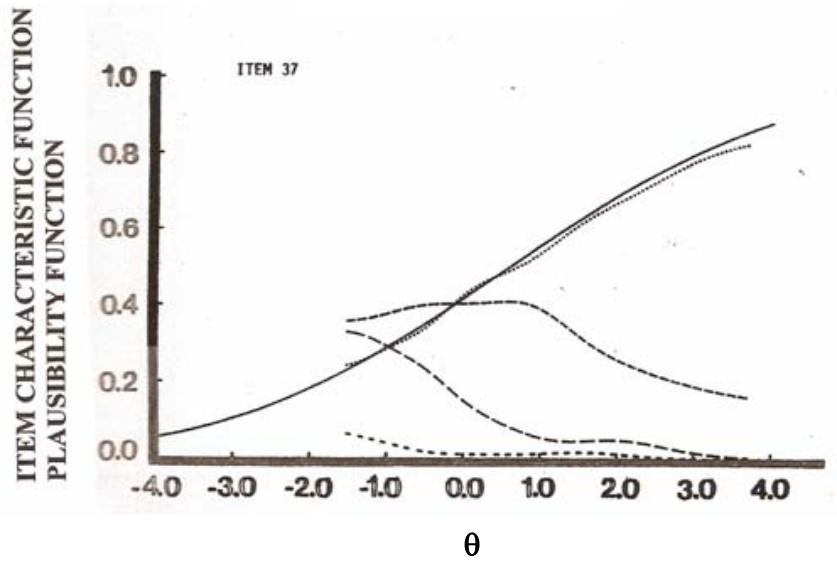
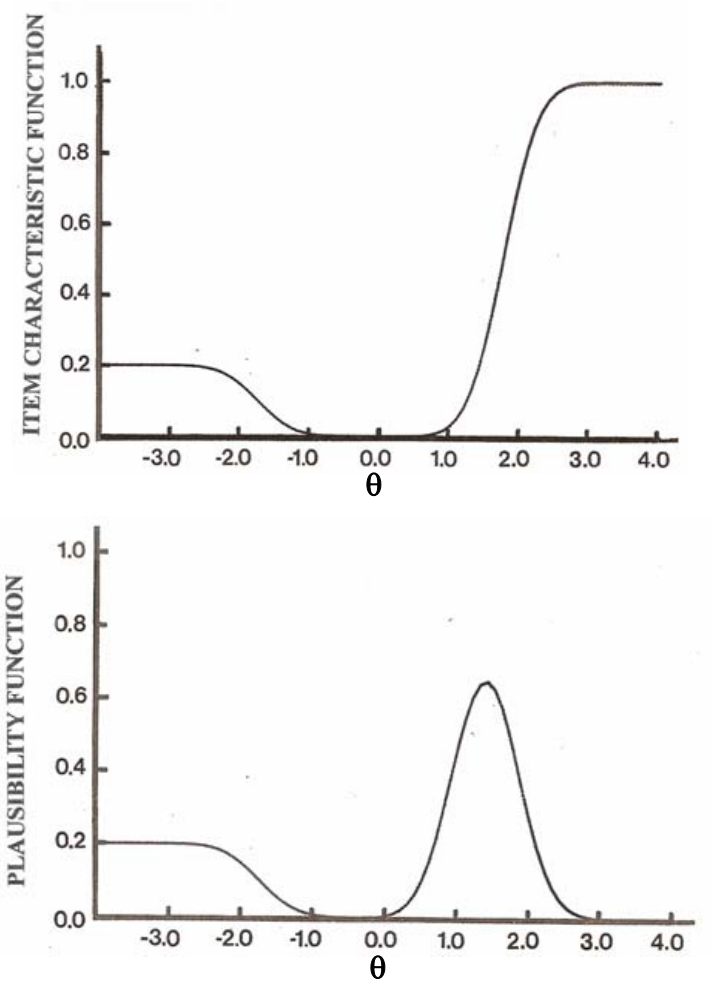


Figure 9

Operating Characteristics of the Correct Answer (above) and the Most Plausible Incorrect Answer (below) of an Hypothesized Five- Choice Item, Following Samejima's Model for the Multiple-Choice Item.



References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Contributed chapters in Lord, F. M. & Novick, M. R., *Statistical theories of mental test scores*, Chapters 17–20, Reading, MA: Addison-Wesley.

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

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55,* 5–18.

Levine, M. V. (1984). *An introduction to multilinear formula score theory*. Champain, IL: University of Illinois, Office of Naval Research Report 84–4.

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their Mathematical formulas: A confrontation of Birnbaum's logistical model. *Psychometrika*, **35**, 43–50.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric characteristic curve estimation. *Psychometrika*, *56,* 611–630.

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

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

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, *38,* 221–233.

Samejima, F. (1979). *A new family of models for the multiple-choice items*. Knoxville, TN: University of Tennessee, Office of Naval Research Report 79-47.

Samejima, F. (1981). *Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item*. Knoxville, TN: University of Tennessee, Office of Naval Research Final Report.

Samejima, F. (1982). *Information loss caused by noise in models for dichotomous items*. Knoxville, TN: University of Tennessee, Office of Naval Research Report 82-1.

Samejima, F. (1984a). *Plausibility functions of Iowa Vocabulary Test items estimated by the simple sum procedure of the conditional p.d.f. approabh*. Knoxville, TN: University of Tennessee, Office of Naval Research Report 84-1.

Samejima, F. (1984b). *Advancement of latent trait theory.* Knoxville, TN: University of Tennessee, Office of Naval Research Final Report.

Samejima, F. (1994). Nonparametric estimation of the plausibility functions of the distractors of vocabulary test items . *Applied Psychological Measurement, 18,* 35–51.

Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika*, *63,* 111–131.

Samejima, F. & Livingston, P. S. (1979). *Method of moments as the least square solution for fitting a polynomial*. Knoxville, TN: University of Tennessee, Office of Naval Research Report 79-2.

Yen, W. M., Burket, G. R. & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model, *Psychometrika*, *56,* 39–54.