Research
Report

An Elementary Test of the Normal 2PL Model Against the Normal 3PL Alternative

Shelby J. Haberman

Research &
Development

June 2006
RR-06-14

An Elementary Test of the Normal 2PL Model Against the Normal 3PL Alternative

Shelby J. Haberman
ETS, Princeton, NJ

June 2006

Abstract

A simple score test of the normal two-parameter logistic (2PL) model is presented that examines the potential attraction of the normal three-parameter logistic (3PL) model for use with a particular item. Application is made to data from a test from the Praxis$^{\text{TM}}$ series. Results from this example raise the question whether the normal 3PL model should be used routinely in preference to the normal 2PL model unless evidence exists that a substantial gain in description of data is achieved.

Key words: Score test, log penalty, maximum likelihood.

Acknowledgments

This paper has benefited from conversations with Paul Holland, Sandip Sinharay, and Matthias von Davier.

A simple variation on the traditional score test (Rao, 1973, p. 418) can be derived to check if the three-parameter logistic (3PL) model is an attractive alternative to the two-parameter logistic (2PL) model without actually fitting a 3PL model. This test is derived in Section 1.. In Section 2., its use is considered for data from the Praxis$^{\text{TM}}$ series of examinations. Implications of results for psychometric practice are considered in Section 3.. Although the specific application considered here does not appear to be readily found in the literature, similar attempts at model diagnosis have been employed in the past to detect other departures from the 2PL model (Glas, 1999).

Throughout this report, $n \geq 1$ examinees each take a test with $q \geq 3$ items, a random variable X_{ij} is 1 if item j is answered correctly by examinee i, and X_{ij} is 0 if item j is not answered correctly. Each vector $\mathbf{X}_i$ of responses X_{ij}, $1 \leq j \leq q$, is independent and identically distributed. The set Γ of possible values of $\mathbf{X}_i$ consists of all q-dimensional vectors such that each coordinate is 0 or 1. The distribution of $\mathbf{X}$ is characterized by the array $\mathbf{p}$ of probabilities

$$p(\mathbf{x}) = P(\mathbf{X}_i = \mathbf{x})$$

for $\mathbf{x}$ in Γ, so that $\mathbf{p}$ is in the simplex T of arrays $\mathbf{r}$ with nonnegative elements $r(\mathbf{x})$, $\mathbf{x}$ in Γ, with a sum of 1. The log likelihood function at $\mathbf{r}$ in T is then

$$\ell(\mathbf{r}) = \sum_{i=1}^{n} \log r(\mathbf{X}_i),$$

and

$$\hat{H}(\mathbf{r}) = -(nq)^{-1}\ell(\mathbf{r})$$

estimates the expected log penalty per item

$$H(\mathbf{r}) = -q^{-1}E(\log r(\mathbf{X}_1))$$

from probability prediction of $\mathbf{X}_1$ by use of $\mathbf{r}$. For a nonempty subset S of T, the maximum log likelihood $\ell(S)$ of $\ell(\mathbf{r})$ for $\mathbf{r}$ in S then leads to the minimum estimated expected log penalty per item $\hat{H}(S) = -(nq)^{-1}\ell(S)$ of $\hat{H}(\mathbf{r})$ for $\mathbf{r}$ in S (Gilula & Haberman, 1994, 1995). Here $\hat{H}(S)$ is an estimate of the minimum expected log penalty per item $H(S)$ of $H(\mathbf{r})$ for $\mathbf{r}$ in S. A member $\hat{\mathbf{p}}$ of S is a maximum-likelihood estimate of the probability array $\mathbf{p}$ (relative to S) if $\ell(\hat{\mathbf{p}}) = \ell(S)$.

In both the 2PL and 3PL models (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Hambleton, Swaminathan, & Rogers, 1991), a random ability variable θ_i is associated with each examinee i, and the X_{ij}, $1 \leq j \leq q$, are conditionally independent given θ_i. The pairs $(\theta_i, \mathbf{X}_i)$ are independent and identically distributed, and the distribution function of θ_i is D. In this report, the simple case will be considered in which D is assumed equal to the standard normal distribution function Φ. For each item j, the conditional probability $P_j(\theta)$ that $X_{ij} = 1$ given $\theta_i = \theta$ is positive and less than 1, so that $Q_j(\theta) = 1 - P_j(\theta)$ is also positive and less than 1. The function P_j is the item characteristic curve, and

$$\lambda_j = \log(P_j/Q_j)$$

is the item logit function (Holland, 1990), so that

$$P_j = \frac{\exp(\lambda_j)}{1+\exp(\lambda_j)} \tag{1}$$

and

$$Q_j = \frac{1}{1+\exp(\lambda_j)}. \tag{2}$$

Let $\boldsymbol{\lambda}$ have coordinates λ_j for $1 \le j \le q$, and let

$$\mathbf{u}'\mathbf{v} = \sum_{j=1}^{q} u_j v_j$$

for q-dimensional vectors u and v with respective coordinates u_j and v_j for $1 \le j \le q$. For

$$V = \prod_{j=1}^{q} Q_j = \prod_{j=1}^{q} \frac{1}{1+\exp(\lambda_j)}, \tag{3}$$

a variation on the Dutch identity yields

$$p(\mathbf{x}) = \int V \exp(\mathbf{X}_i'\boldsymbol{\lambda}) dD \tag{4}$$

(Holland, 1990).

The set S_{2n} that corresponds to the normal 2PL model consists of all arrays $\mathbf{p}$ in S such that (3) and (4) hold, $D = \Phi$, and

$$\boldsymbol{\lambda}(\theta) = \theta\mathbf{a} - \boldsymbol{\gamma} \tag{5}$$

for some q-dimensional vectors $\mathbf{a}$ and $\boldsymbol{\gamma}$ with respective coordinates $a_j > 0$ and γ_j for $1 \le j \le q$. For item j, the item discrimination is a_j, and the item difficulty is γ_j/a_j. The set S_{3n} for the normal 3PL model consists of all arrays $\mathbf{p}$ in S such that (3) and (4) hold, $D = \Phi$, and

$$\lambda_j(\theta) = \log\{[c_j + \exp(a_j\theta - \gamma_j)]/(1-c_j)\} \tag{6}$$

for some real $a_j > 0$, c_j in $[0, 1)$, and γ_j. The 3PL case reduces to the 2PL case if each $c_j = 0$. The a_j and γ_j can be interpreted as in the 2PL model, and c_j is a guessing probability. In the construction of the desired test statistic, the restriction set S_{3nk} is considered for $1 \le k \le q$ in which $\mathbf{p}$ in S_{3n} is in S_{3nk} if (3), (4), and (6) hold for $D = \Phi$ and for some $a_j > 0$, $c_j > 0$ in $[0, 1)$, and γ_j, $1 \le j \le q$, such that $c_j = 0$ if $j \ne k$. For use in comparison of estimated expected penalties, it is also helpful to note that the set S_{1n} for the normal Rasch model consists of $\mathbf{p}$ in S_{2n} such that (3) and(4) hold, $D = \Phi$, and (5) holds for some q-dimensional vectors $\mathbf{a}$ and $\boldsymbol{\gamma}$ with respective coordinates $a_j > 0$ and γ_j for $1 \le j \le q$ and $a_j = a_1$ for $j > 1$.

1. The Test Statistic

To construct the desired score test statistic, consider an item k from 1 to q. Consider the null hypothesis that the probability array $\mathbf{p}$ is in S_{2n}, so that the normal 2PL model holds, against the alternative that $\mathbf{p}$ is in S_{3nk}. Let $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\gamma}}$ be the respective maximum-likelihood estimates of the vectors $\mathbf{a}$ and $\boldsymbol{\gamma}$ under the 2PL model. For $1 \le j \le q$, let $\hat{a}_j$ be coordinate j of $\hat{\mathbf{a}}$, and let $\hat{\gamma}_j$ be coordinate j of $\hat{\boldsymbol{\gamma}}$. To construct the test, consider the $3q$-dimensional vector $\boldsymbol{\tau}$ with coordinates $\tau_j = a_j > 0$, $\tau_{q+j} = \gamma_j$, and $\tau_{2q+j} = c_j$ in $[0, 1)$ for $1 \le j \le q$. Let $\mathbf{p}_{\boldsymbol{\tau}}$ be the array in S_{3n} such that (3), (4), and (6) hold for $\mathbf{p} = \mathbf{p}_{\boldsymbol{\tau}}$, and let $H(\boldsymbol{\tau}) = \ell(\mathbf{p}_{\boldsymbol{\tau}})$. Let

$$h_i(\boldsymbol{\tau}) = \log p_{\boldsymbol{\tau}}(\mathbf{X}_i),$$

so that

$$H(\boldsymbol{\tau}) = \sum_{i=1}^{n} h_i(\boldsymbol{\tau}).$$

The test statistic requires partial derivatives of H. Let $h_{ij}(\boldsymbol{\tau})$ denote the partial derivative of h_i at $\boldsymbol{\tau}$ with respect to τ_j, and let $H_j(\boldsymbol{\tau})$ denote the partial derivative of H at $\boldsymbol{\tau}$ with respect to τ_j, so that

$$H_j = \sum_{i=1}^{n} h_{ij}.$$

Let $\hat{\boldsymbol{\tau}}^*$ be the $3q$-dimensional vector with coordinates $\hat{\tau}^*_j = \hat{a}_j$, $\hat{\tau}^*_{q+j} = \hat{\gamma}_j$, and $\hat{\tau}^*_{2q+j} = 0$ for $1 \le j \le q$. For item j, the score test statistic is $U_j = H_{2q+j}(\hat{\boldsymbol{\tau}}^*)$. To evaluate U_j, let

$$\hat{\boldsymbol{\lambda}}(\theta) = \theta\hat{\mathbf{a}} - \hat{\boldsymbol{\gamma}}$$

be the maximum-likelihood estimate of $\boldsymbol{\lambda}(\theta)$ under the 2PL model. Let $\hat{\lambda}_j$ be coordinate j of $\hat{\boldsymbol{\lambda}}$, and let

$$\hat{V} = \prod_{j=1}^{q}[1 + \exp(\hat{\lambda}_j)]^{-1}$$

be the maximum-likelihood estimate of V for the 2PL model. Let

$$\hat{P}_j = [1 + \exp(\hat{\lambda}_j)]^{-1}$$

be the maximum-likelihood estimate of P_j under the 2PL model. Use of the chain rule of differentiation and use of standard properties of exponential families (Berk, 1972) shows that

$$U_j = n^{-1}\sum_{i=1}^{n} U_{ij},$$

where

$$U_{ij} = \frac{\int \hat{P}_j^{-1}(X_{ij} - \hat{P}_j)\exp(\mathbf{X}_i'\hat{\boldsymbol{\lambda}})\hat{V}\phi}{\int \exp(\mathbf{X}_i'\hat{\boldsymbol{\lambda}})\hat{V}\phi}.$$

Comparison of the standard asymptotic variance formula for $n^{1/2}U_j$ (Aitchison & Silvey, 1958) with standard regression formulas (Rao, 1973, pp. 267–268) shows that the asymptotic variance σ_j^2 of $n^{1/2}U_j$ is the same as the mean-squared error from linear prediction of $h_{i(2q+j)}(\boldsymbol{\tau})$ by $h_{ik}(\boldsymbol{\tau})$, $1 \le k \le 2q$. Differentiation shows that, under the 2PL model with $c_j = 0$ for $1 \le j \le q$,

$$h_{ij}(\boldsymbol{\tau}) = -\frac{\int (X_{ij} - P_j)\exp(\mathbf{X}_i'\boldsymbol{\lambda})V\phi}{\int \exp(\mathbf{X}_i'\boldsymbol{\lambda})V\phi}$$

and

$$h_{q+j}(\boldsymbol{\tau}) = \frac{\int \theta(X_{ij} - P_j)\exp(\mathbf{X}_i'\boldsymbol{\lambda})V\phi}{\int \exp(\mathbf{X}_i'\boldsymbol{\lambda})V\phi}$$

for $1 \le j \le q$.

It is a straightforward matter to verify that σ_j^2 is consistently estimated by the residual mean-squared error s_j^2 from linear regression of U_{ij} onto $\hat{h}_{ik} = h_{ik}(\hat{\boldsymbol{\tau}}^*)$ for $1 \le k \le 2q$, where $1 \le i \le n$. The desired statistic for item j is then $t_j = n^{1/2}U_j/s_j$. The statistic t_j has an approximate standard normal distribution under the 2PL model, with the approximation increasingly accurate as the sample size becomes large. If $\mathbf{p}_{\boldsymbol{\tau}}$ is in S_{3nj} for some item j and if c_j is small, then $\ell(S_{3nj})$ is well-approximated by $\ell(S_{2n}) + t_j^2/2$ and $\hat{H}(S_{3nj})$ is well approximated by $\hat{H}(S_{2n}) + t_j^2/(2nq)$.

2. Application to Data

In the example under study, $n = 8,686$ and $q = 45$. The test statistics for each item are shown in Table 1. Despite a substantial sample size, many items have score statistics compatible with the normal 2PL model. For example, $|t_k| \leq 2$ in 15 cases. On the other hand, items not compatible with the 2PL model are readily found, for 26 items have t_k greater than 2, and 4 items have t_k less than -2. Even with allowances for multiple comparisons, 10 standardized values that exceed 4 are very unlikely to occur by chance if the model is true.

It should be emphasized that the test statistics do not imply that the 3PL model provides a description of the data that is much better than the description provided by the 2PL model. For some perspective on this point, consider some estimated log-penalty functions that can be derived for the data under study. The estimate $\hat{H}(S_{2n})$ for the normal 2PL model is 0.59157, while $\hat{H}(S_{3n})$ for the normal 3PL model is 0.59074. This gain of 0.00083 is quite modest. For comparison, the minimum estimated expected penalty per item for the normal 1PL model is 0.59639, so that the gain from use of the normal 2PL rather than the normal one-parameter logistic (1PL) model is 0.00482. The estimated expected log penalty under the trivial model that all X_{ij}, $1 \leq j \leq q$, are independent is 0.62467, so that the gain for the normal 1PL model over the independence model is 0.02828, a much larger gain than the gain from the normal 1PL model to the normal 2PL model.

Table 1
Results of Tests for Nonzero Guessing Probabilities

Item k	Score average U_k	Standard error s_k	Standardized value t_k
1	0.00043	0.00054	0.80478
2	0.00132	0.00020	6.50359
3	0.00073	0.00189	0.38535
4	-0.00281	0.00128	-2.18717
5	0.00780	0.00271	2.87903
6	0.00086	0.00052	1.66437
7	0.00071	0.00026	2.73552
8	0.01491	0.00258	5.77798
9	0.01644	0.00279	5.88547
10	-0.00106	0.00076	-1.39806
11	0.00102	0.00034	2.99920

(*Table continues*)

Table 1 (continued)

Item	Score average	Standard error	Standardized value
k	U_k	s_k	t_k
12	0.01944	0.00408	4.76351
13	0.00463	0.00084	5.50615
14	0.01994	0.00395	5.05235
15	0.00026	0.00015	1.73915
16	0.00116	0.00047	2.48215
17	0.00192	0.00031	6.17770
18	0.00766	0.00101	7.57526
19	0.00546	0.00263	2.07236
20	0.00238	0.00036	6.69724
21	0.00028	0.00053	0.52986
22	0.00112	0.00294	0.38055
23	-0.00017	0.00038	-0.46052
24	0.00112	0.00029	3.84572
25	0.00116	0.00066	1.76433
26	0.00752	0.00468	1.60910
27	0.00052	0.00044	1.18942
28	0.00906	0.00269	3.36499
29	-0.00001	0.00045	-0.02107
30	-0.00046	0.00056	-0.82410
31	0.00049	0.00072	0.68124
32	-0.00086	0.00031	-2.76761
33	-0.00161	0.00065	-2.47247
34	0.00712	0.00174	4.09183
35	0.01303	0.00378	3.44932
36	0.01091	0.00354	3.08384
37	-0.00298	0.00137	-2.17524
38	0.00398	0.00112	3.55390
39	0.01109	0.00440	2.51879
40	0.00481	0.00147	3.27537
41	0.00463	0.00175	2.64318
42	0.00287	0.00115	2.48429
43	0.00444	0.00132	3.37090
44	0.00400	0.00207	1.92916
45	0.00695	0.00193	3.60106

3. Conclusions

The analysis here suggests that routine use of the normal 3PL model may not necessarily be wise. For the data under study, the score test suggests that many guessing parameters are not clearly different from 0 even if the normal 3PL model holds. In addition, the gain in data description from use of a 3PL rather than a 2PL model appears small. Given the much greater computational difficulties associated with the 3PL model relative to those for the 2PL model, the question must be raised whether proponents of the 3PL model can demonstrate cases in which the gain from the 3PL model rather than the 2PL model is much larger than is observed here. The issue of guessing parameters not clearly positive is especially important from a computational perspective, for computations with the 3PL model are hardest when the guessing probabilities do not clearly differ from 0 (Hambleton et al., 1991, p. 44).

An alternative approach to testing a 2PL versus a 3PL model would involve a likelihood-ratio chi-square test statistic such as $2nq[\hat{H}(S_{3n}) - \hat{H}(S_{2n})]$; however, such a test involves two complications. The 3PL model must be fit, and, even if the null hypothesis holds and the sample size is large, the chi-square approximation is not satisfactory due to the requirement that the guessing parameters c_j be nonnegative.

References

Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, *29*, 813–828.

Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Annuals of Mathematics and Statistics*, *43*, 193–204.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273–294.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE.

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

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.