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

Shelby J. Haberman
An Elementary Test of the Normal 2PL Model Against the Normal 3PL Alternative

Shelby J. Haberman
ETS, Princeton, NJ

June 2006
As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2006 by Educational Testing Service. All rights reserved.

ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). PRAXIS is a trademark of ETS.
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\textsuperscript{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™ 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 \( X_i \) of responses \( X_{ij}, 1 \leq j \leq q \), is independent and identically distributed. The set \( \Gamma \) of possible values of \( X_i \) consists of all \( q \)-dimensional vectors such that each coordinate is 0 or 1. The distribution of \( X \) is characterized by the array \( p \) of probabilities

\[
p(x) = P(X_i = x)
\]

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

\[
\ell(r) = \sum_{i=1}^{n} \log r(X_i),
\]

and

\[
\hat{H}(r) = -(nq)^{-1}\ell(r)
\]

estimates the expected log penalty per item

\[
H(r) = -q^{-1}E(\log r(X_1))
\]

from probability prediction of \( X_1 \) by use of \( r \). For a nonempty subset \( S \) of \( T \), the maximum log likelihood \( \ell(S) \) of \( \ell(r) \) for \( r \) in \( S \) then leads to the minimum estimated expected log penalty per item \( \hat{H}(S) = -(nq)^{-1}\ell(S) \) of \( \hat{H}(r) \) for \( r \) in \( S \) (Gilula & Haberman, 1994, 1995). Here \( \hat{H}(S) \) is an estimate of the minimum expected log penalty per item \( H(S) \) of \( H(r) \) for \( r \) in \( S \). A member \( \hat{p} \) of \( S \) is a maximum-likelihood estimate of the probability array \( p \) (relative to \( S \)) if \( \ell(\hat{p}) = \ell(S) \).

In both the 2PL and 3PL models (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Hambleton, Swaminathan, & Rogers, 1991), a random ability variable \( \theta_i \) is associated with each examinee \( i \), and the \( X_{ij}, 1 \leq j \leq q \), are conditionally independent given \( \theta_i \). The pairs \( (\theta_i, X_i) \) are independent
and identically distributed, and the distribution function of $\theta_i$ is $D$. In this report, the simple case will be considered in which $D$ is assumed equal to the standard normal distribution function $\Phi$. 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

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

and

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

Let $\lambda$ have coordinates $\lambda_j$ for $1 \leq j \leq q$, and let

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

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

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

a variation on the Dutch identity yields

$$p(x) = \int V \exp(X_i'\lambda)dD \quad (4)$$

(Holland, 1990).

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

$$\lambda(\theta) = \theta a - \gamma \quad (5)$$

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

$$\lambda_j(\theta) = \log\left\{ [c_j + \exp(a_j \theta - \gamma_j)]/(1 - c_j) \right\} \quad (6)$$
for some real \( a_j > 0, c_j \) in \([0,1)\), and \( \gamma_j \). The 3PL case reduces to the 2PL case if each \( c_j = 0 \).

The \( a_j \) and \( \gamma_j \) 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 \( 1 \leq k \leq q \) in which \( p \) 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 \( \gamma_j, 1 \leq j \leq q \), such that \( c_j = 0 \) if \( j \neq k \). 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 \( p \) in \( S_{2n} \) such that (3) and (4) hold, \( D = \Phi \), and (5) holds for some \( q \)-dimensional vectors \( a \) and \( \gamma \) with respective coordinates \( a_j > 0 \) and \( \gamma_j \) for \( 1 \leq j \leq q \) 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 \( p \) is in \( S_{2n} \), so that the normal 2PL model holds, against the alternative that \( p \) is in \( S_{3nk} \). Let \( \hat{a} \) and \( \hat{\gamma} \) be the respective maximum-likelihood estimates of the vectors \( a \) and \( \gamma \) under the 2PL model. For \( 1 \leq j \leq q \), let \( \hat{a}_j \) be coordinate \( j \) of \( \hat{a} \), and let \( \hat{\gamma}_j \) be coordinate \( j \) of \( \hat{\gamma} \). To construct the test, consider the \( 3q \)-dimensional vector \( \tau \) with coordinates \( \tau_j = a_j > 0, \tau_{q+j} = \gamma_j \), and \( \tau_{2q+j} = c_j \) in \([0,1)\) for \( 1 \leq j \leq q \). Let \( p_\tau \) be the array in \( S_{3n} \) such that (3), (4), and (6) hold for \( p = p_\tau \), and let \( H(\tau) = \ell(p_\tau) \). Let

\[
    h_i(\tau) = \log p_\tau(X_i),
\]

so that

\[
    H(\tau) = \sum_{i=1}^{n} h_i(\tau).
\]

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

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

Let \( \hat{\tau}^* \) 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 \( 1 \leq j \leq q \). For item \( j \), the score test statistic is \( U_j = H_{2q+j}(\hat{\tau}^*) \). To evaluate \( U_j \), let

\[
    \hat{\lambda}(\theta) = \theta \hat{a} - \hat{\gamma}
\]

\[
    h_i(\tau) = \log p_\tau(X_i),
\]

so that

\[
    H(\tau) = \sum_{i=1}^{n} h_i(\tau).
\]

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

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

Let \( \hat{\tau}^* \) 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 \( 1 \leq j \leq q \). For item \( j \), the score test statistic is \( U_j = H_{2q+j}(\hat{\tau}^*) \). To evaluate \( U_j \), let

\[
    \hat{\lambda}(\theta) = \theta \hat{a} - \hat{\gamma}
\]
be the maximum-likelihood estimate of $\lambda(\theta)$ under the 2PL model. Let $\hat{\lambda}_j$ be coordinate $j$ of $\hat{\lambda}$, 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
\[
U_{ij} = \frac{\int \hat{P}_j^{-1}(X_{ij} - \hat{P}_j) \exp(X_i'\hat{\lambda})\hat{V}\phi}{\int \exp(X_i'\hat{\lambda})\hat{V}\phi}.
\]

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 $\sigma_j^2$ of $n^{1/2}U_j$ is the same as the mean-squared error from linear prediction of $h_{i(2q+j)}(\tau)$ by $h_{ik}(\tau)$, $1 \leq k \leq 2q$. Differentiation shows that, under the 2PL model with $c_j = 0$ for $1 \leq j \leq q$,
\[
h_{ij}(\tau) = -\frac{\int (X_{ij} - P_j) \exp(X_i'\lambda)V\phi}{\int \exp(X_i'\lambda)V\phi}
\]
and
\[
h_{q+j}(\tau) = \frac{\int \theta(X_{ij} - P_j) \exp(X_i'\lambda)V\phi}{\int \exp(X_i'\lambda)V\phi}
\]
for $1 \leq j \leq q$.

It is a straightforward matter to verify that $\sigma_j^2$ is consistently estimated by the residual mean-squared error $s_j^2$ from linear regression of $U_{ij}$ onto $\hat{h}_{ik} = h_{ik}(\hat{\tau}^*)$ for $1 \leq k \leq 2q$, where $1 \leq i \leq n$. 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 $p_\tau$ 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.

\[ \text{Table 1} \]
\[ \text{Results of Tests for Nonzero Guessing Probabilities} \]

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

(Table continues)
### Table 1 (continued)

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

### 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


