Adjusted Rasch Person-Fit Statistics

Dimiter M. Dimitrov
George Mason University

Richard M. Smith
Data Recognition Corporation

Two frequently used parametric statistics of person-fit with the dichotomous Rasch model (RM) are adjusted and compared to each other and to their original counterparts in terms of power to detect aberrant response patterns in short tests (10, 20, and 30 items). Specifically, the cube root transformation of the mean square for the unweighted person-fit statistic, \( t \), and the standardized likelihood-based person-fit statistic \( Z_3 \) were adjusted by estimating the probability for correct item response through the use of symmetric functions in the dichotomous Rasch model. The results for simulated unidimensional Rasch data indicate that \( t \) and \( Z_3 \) are consistently, yet not greatly, outperformed by their adjusted counterparts, denoted \( t^* \) and \( Z_3^* \), respectively. The four parametric statistics, \( t, Z_3, t^*, \) and \( Z_3^* \), were also compared to a non-parametric statistic, \( H^T \), identified in recent research as outperforming numerous parametric and non-parametric person-fit statistics. The results show that \( H^T \) substantially outperforms \( t, Z_3, t^*, \) and \( Z_3^* \) in detecting aberrant response patterns for 20-item and 30-item tests, but not for very short tests of 10 items. The detection power of \( t, Z_3, t^*, \) and \( Z_3^* \), and \( H^T \) at two specific levels of Type I error, .10 and .05 (i.e., up to 10% and 5% false alarm rate, respectively), is also reported.
Issues of data fit are of major concern with the application of latent trait measurement models. Previous research has dealt with issues of item fit (e.g., Andersen, 1973; Smith, 1991a; Smith, Schumacker, and Bush, 1998; Wollenberg, 1982), person fit (e.g., Drasgow, Levin, and McLaughlin, 1987, 1991; Karabatsos, 2003; Klauser, 1995; Molenaar and Hoijtink, 1990; Tatsuoka, 1984; Sato, 1975; Wright, 1977), or both (e.g., Reise, 1990; Smith, 1982, 1986, 1988, 1991b, 1994, 2000; Wright and Linacre, 1994; Wright and Panchepakesan, 1969; Wright and Stone, 1979). A historical review of fit indices in measurement with the Rasch model is provided by Smith (2000). It should be noted that person measurement disturbances can be classified as either random or systematic (Smith, 1991b). Random disturbances are associated with factors such as guessing or carelessness in responding to items, whereas systematic disturbances may occur, for example, with deficiency in a content area, test bias, or instructional differences.

To detect aberrant response patterns (measurement disturbances), previous research has developed numerous person-fit statistics: (1) norm-based indices such as the caution index (Sato, 1975) and the consistency index (Cliff, 1979), (2) indices based on item response theory such as the standardized index $Z_i$ (Hulin, Drasgow, and Parsons, 1983), (3) Rasch model indices such as the cube root transformation of the mean square statistic (e.g., Smith, 1982; Wright and Stone, 1979), and (4) covariance-based person-fit indices (e.g., $ECI_{2z}$ and $ECI_{4z}$; Tatsuoka, 1984; $H^c$, Sijtsma, 1986; Sijtsma and Mejer, 1992). Issues of the relative efficiency of these person-fit statistics within the context of the Rasch model have also been addressed in recent research. For example, Li and Olejnik (1997) compared five person-fit statistics on their power to detect aberrant response patterns within the framework of the Rasch model: Wright’s unweighted and weighted total person-fit statistics with the cube root transformation, $t$ (Smith, 1982, 1991b; Wright and Stone, 1979), the standardized likelihood-based person-fit statistic, $Z_i$ (Hulin, Drasgow, and Parsons, 1983), and two standardized covariance-based person-fit indices, $ECI_{2z}$ and $ECI_{4z}$ (Tatsuoka, 1984). Using simulated data for 30- and 60-item (unidimensional and two-dimensional) tests, Li and Olejnik (1997), among other things, report that (a) the chance of identifying misfitting response patterns increases with test length regardless of test dimensionality (one- or two-dimensional test), type of misfit (spuriously high or spuriously low), or the person-fit index and (b) the $Z_i$ index is as good or better that the other four person-fit statistics. Other studies of fit in the Rasch model indicate that the unweighted version of the (item or person) fit indices are generally more powerful in detecting measurement disturbances (e.g., Smith, 2000).

Karabatsos (2003) compared 36 (parametric and non-parametric) person-fit statistics within the context of the Rasch model using simulated data sets with variations in the type and percentage of aberrant responding examinees and three test lengths (17 items, 33 items, and 65 items). He found that, overall, the non-parametric person-fit statistic $H^c$ (Sijtsma, 1986; Sijtsma and Mejer, 1992) is best in detecting aberrant response patterns, followed by the parametric statistic $D(0)$ (Trabin and Weiss, 1983) and the non-parametric statistic $C$ (Sato, 1975). Regarding the two parametric person-fit statistics adjusted and compared in this article, $t$ and $Z_i$, Karabatsos (2003) reported that $t$ performed slightly better than $Z_i$ (with notations $ZUB$ and $l_z$, respectively, in his paper) on the two types of aberrant responses targeted here with the comparison of $t$ and $Z_i$: random guessing and cheating. He also found, among other things, that (a) lucky-guessers are slightly easier to detect than cheating examinees and (b) detection rates increase with test length. It should be noted that Karabatsos (2003) determined the ability of person-fit statistics to detect aberrant response patterns by estimating the area under the Receiver Operating Characteristic (ROC) curves produced by these statistics; (more details on the ROC method are provided the next section).

As this study addresses issues of parametric person-fit indices within the Rasch model (Rasch, 1960, 1980) it should be noted that the calculation of such indices in previous research (e.g.,
Karabatsos, 2003; Li and Olejnik, 1997) is based on the estimated Rasch probability for correct item response:

\[ P(\hat{\theta}_n) = \frac{\exp(\hat{\theta}_n - \hat{\delta}_i)}{1 + \exp(\hat{\theta}_n - \hat{\delta}_i)} , \]

where \( \hat{\theta}_n \) is an estimate of the person parameter (trait score of person \( n \)) and \( \hat{\delta}_i \) is an estimate of the item parameter (difficulty of item \( i \)) \( (i = 1, ..., L; n = 1, ..., N) \). However, Van den Wollenberg (1980) argues that \( P(\hat{\theta}_n) \) is somewhat at fault in estimating the probability for correct item response and more accurate estimation of the probability for correct response on item \( i \) for a person with a raw score \( r \) is provided, instead, through the use of symmetric functions

\[ \hat{\pi}_r i = \hat{\varepsilon}_{i} (\hat{\gamma}_{r-1}^{(i)}) / \hat{\gamma}_r , \]

where \( \hat{\varepsilon}_i \) is the Rasch “easiness” of item \( i \) \( [\hat{\varepsilon}_i = \log(-\hat{\delta}_i)] \), \( \hat{\gamma}_r \) is the basic symmetric function, and \( \hat{\gamma}_{r-1}^{(i)} \) is its first derivative (Van den Wollenberg, 1982, Equation 14). As a reminder, \( \hat{\gamma}_r \) is the sum of the products of the elements in all combinations of \( r \) (out of \( n \)) \( \hat{\varepsilon}_i \) values. For a 3-item test, for example, the symmetric functions are

\[
\begin{align*}
\hat{\gamma}_0 &= 1, \\
\hat{\gamma}_1 &= \hat{\varepsilon}_1 + \hat{\varepsilon}_2 + \hat{\varepsilon}_3, \\
\hat{\gamma}_2 &= \hat{\varepsilon}_1\hat{\varepsilon}_2 + \hat{\varepsilon}_1\hat{\varepsilon}_3 + \hat{\varepsilon}_2\hat{\varepsilon}_3, \text{ and} \\
\hat{\gamma}_3 &= \hat{\varepsilon}_1\hat{\varepsilon}_2\hat{\varepsilon}_3 = 1.
\end{align*}
\]

The first derivative, \( \hat{\gamma}_{r-1}^{(i)} \) is the symmetric function of order \( r-1 \) in all parameters except \( \hat{\varepsilon}_i \). For more information on symmetric functions in the Rasch model, the reader may refer, for example, to Baker (1992, p. 125). In this article, symmetric functions are calculated through the use of a recursive algorithm described by Gustafson (1980).

Purpose of This Study

The purpose of this study is to investigate the effect of replacing the estimated probability \( P(\hat{\theta}_n) \) from Equation 1 with its counterpart \( \hat{\pi}_{ri} \) from Equation 2 on the detection power of two frequently used parametric person-fit statistics: the cube root transformation index, \( t \) (Smith, 1982, 1991b; Wright and Stone, 1979), and the standardized likelihood-based person fit index, \( Z_3^* \) (Hulin, Drasgow, and Parsons, 1983). The comparison of \( t \) and \( Z_3^* \) is conducted with simulations of two specific aberrant behavior, random guessing and cheating, in the responses of examinees. The person-fit statistics \( t, Z_3, Z_3^* \), and \( Z_3^* \), are also compared with the non-parametric person-fit statistic \( H^* \) (Sijtsma, 1986; Sijtsma and Mejer, 1992) which has been found to outperform 35 other person-fit statistics including \( t \) and \( Z_3 \) (Karabatsos, 2003).

Method

The person-fit index \( Z_3 \) is estimated through the equation

\[ Z_3 = \frac{l_0 - E(l_0|\hat{\theta}_n)}{\text{VAR}(l_0|\hat{\theta}_n)^{1/2}} , \]

where \( \hat{\theta}_n \) is the maximum likelihood estimate of \( \theta_n \), \( l_0 \) is the natural logarithm of the likelihood for the response pattern of a person at \( \hat{\theta}_n \), \( E(l_0|\hat{\theta}_n) \) is the expected value of \( l_0 \) at \( \hat{\theta}_n \) and \( \text{VAR}(l_0|\hat{\theta}_n) \) is its conditional variance. Specifically, \( l_0 \) is estimated through the equation (Levin and Rubin, 1979)

\[ l_0 = \sum_{i=1}^{L} [x_i \log P_i(\hat{\theta}_n) + (1 - x_i) \log(1 - P_i(\hat{\theta}_n))]. \]

where \( x_i \) is the score (1 = correct, 0 = incorrect) for person \( n \) on item \( i \), the expected value of \( l_0 \) is
The other index of measurement disturbances used in this study, \( t \), represents a cube root transformation that converts the mean square, \( MS \), of the unweighted total person fit statistic to approximate unit normals (e.g., Smith, Schumacker, and Bush, 1998):

\[
\begin{align*}
t &= \left[(MS^{1/3} - 1)(3/S)\right] + (S/3),
\end{align*}
\]

where \( MS \) for a response pattern (at \( \hat{\theta}_n \)) in the Rasch model is estimated by the equation

\[
MS = \frac{1}{L} \sum_{i=1}^{L} \left\{ x_m - P_l(\hat{\theta}_n) \right\}^2,
\]

and its standard deviation, \( S \), is estimated by

\[
S = \frac{1}{L} \left\{ \sum_{i=1}^{L} \frac{1}{P_l(\hat{\theta}_n)} - 4L \right\}^{1/2}.
\]

The values of the adjusted person-fit indices \( t^* \) and \( Z^*_j \) are obtained by replacing \( P_l(\hat{\theta}_n) \) with \( \hat{\pi}_{ji} \) in Equations 5 through 10. Note that while the value of \( P_l(\hat{\theta}_n) \) is affected by random errors associated with the estimation of both \( \theta \) and \( \delta \), the value of \( \hat{\pi}_{ji} \) is affected only by random errors associated with the estimation of \( \delta \) because the factors in the right-hand side of Equation 2 depend only on \( \hat{\delta}_i \).

The non-parametric person-fit statistic \( H^T \) (Sijtsma, 1986; Sijtsma and Mejer, 1992), compared to \( t, Z_j, t^* \), and \( Z^*_j \) in this study, is calculated for the response pattern of examinee \( n \) as

\[
H^T_n = \frac{\sum (\beta_{nm} - \beta_n \beta_m)}{\min\{\beta_n(1-\beta_m), (1-\beta_n)\beta_m\}},
\]

where:

\[
\beta_n = \text{proportion correct item responses for examinee } n,
\]

\[
\beta_m = \text{proportion correct item responses for examinee } m,
\]

\[
\beta_{nm} = J^{-1} \sum_{j=1}^{J} X_{nj} X_{mj} \text{ is the covariance between the scores of examinee } n (X_{nj}) \text{ and examinee } m (X_{mj}) \text{ on a test of } J \text{ binary items; } (j = 1, ..., J).
\]

Data Simulations

The data sets were simulated in four replications for large data sets \( (J = 9,000) \) with a 3 x 2 x 2 crossed design which included (a) three test lengths: 10 items, 20 items, and 30 items, (b) two types of aberrant response behavior: guessing and cheating for low ability examinees; and (c) two levels of aberrance severity: 20% and 40% of the most difficult items were subject to guessing or cheating. With each simulated data set, the person-fit statistics were calculated using the estimates of item parameters, \( \hat{\delta}_i \), obtained with Rasch analysis of the simulated data sets. The item parameters used in the simulations were selected to be equally distant (with a mean of \( 0 \)) within the interval \([-2, 2]\) for the 10-item and 20-item data sets and, to keep about the same distances between adjacent item parameters, within the interval \([-2.75, 2.75]\) for the 30-item data set. The ability scores used in the data simulations were randomly generated within the standard normal distribution, \( \theta \sim N(0,1) \). The examinees with an ability score below \(-0.61\) (27th percentile) were identified as low ability examinees.

Random guessing was simulated by assigning a probability of \( .25 \) for correct response on 20% (or 40%) of the most difficult items for low ability examinees (i.e., simulating random guessing on multiple-choice items with four response options of which one is correct). Cheating was
simulated by assigning a probability of .90 for correct response on 20% (or 40%) of the most difficult items for low ability examinees. With this, 90% chances for cheating were used because a 100% successful cheating is unlikely to occur in real testing conditions. The term “cheating” is used here as a working label for consistency in terminology with previous research on person fit, but it may also be that some “special knowledge” leads to response patterns in which low ability examinees answer correctly some of the most difficult items.

The procedures used in four replications of simulated data sets, estimation of item and person parameters, and calculation of the compared person-fit statistics \((t, t^*, Z_3^*, Z_3, H^F)\) were facilitated by the use of the computer programs SIMITEM (Smith, 2000), SAS (SAS Inc., 1990), WINSTEPS (Linacre and Wright, 2001), and a C++ computer program developed for the purpose of this study.

**Evaluating the Detection Efficiency of the Person-Fit Statistics**

The overall efficiency of the person-fit statistics compared in this study was evaluated with the area under the ROC curve produced by a person-fit statistic. An ROC curve is formed by plotting dots with coordinates \(F(c)\) and \(H(c)\) for all (observed) values \(c\) of a person-fit statistic, where \(F(c)\), on the \(x\)-axis, is the proportion of normal response patterns incorrectly identified as aberrant, \(H(c)\), on the \(y\)-axis, is the proportion of aberrant response patterns correctly identified as aberrant, and \(c\) is the cutting score to which the person-fit statistic is compared to make classification decisions for \(F(c)\) and \(H(c)\). Thus, \(F(c)\) is the false-alarm rate (FAR) and \(H(c)\) is the hit rate (HR) of a person-fit statistic at the cutting score \(c\) (see, e.g., Drasgow et al., 1987). The area under the ROC curve measures the degree to which a person-fit statistic minimizes the Type I error rate \((FAR)\) and the Type II error rate \((1 - HR)\). Theoretically, the ROC area can vary from 0 to 1, with 1 being its “ideal” value indicating perfect hit-rate \((HR = 1)\) and no false-alarm \((FAR = 0)\) for the person-fit statistic (see also, Karabatsos, 2003). In addition, the highest \(HR\) for each person-fit statistic \((t, t^*, Z_3^*, Z_3, H^F)\) within two specific levels of Type I error rate, .10 and .05 (in percentages, \(FAR = 10\%\) and 5\%, respectively), is provided for practical references.

**Results**

The results in Table 1 show that there are differences between the estimates of probability for correct item response as represented by \(P_t(\hat{\theta}_n)\), with Equation 1, and \(\hat{\gamma}_{t^*}\), with Equation 2, for any item at all ability levels. Therefore, it is logical to expect differences between the values of the person-fit indices \(t\) and \(t^*\) (or \(Z_3\) and \(Z_3^*\)) as they are calculated through the use of \(P_t(\hat{\theta}_n)\) and \(\hat{\gamma}_{t^*}\), respectively.

The estimates of the ROC areas (with a 95% confidence interval across four simulations) for the person-fit statistics \(t, t^*, Z_3, Z_3^*, H^F\) are represented graphically in Figures 1 and 2 for simulations of guessing and cheating (or “special knowledge”), respectively. In Table 2, the person-fit statistics are given in decreasing order (from left to right) of their ROC area values (i.e., in decreasing order of their observed efficiency in detecting aberrant response patterns). The ROC curves for some scenarios are also provided (see, Figures 3, 4, 5, and 6) to visually illustrate the comparison of \(t, t^*, Z_3, Z_3^*, \) and \(H^F\). A brief summary of the results follows.

- The non-parametric person-fit statistic, \(H^F\), outperforms the parametric statistics, \(t, t^*, Z_3, \) and \(Z_3^*\) in detecting aberrancy with 30 items and 20 items (e.g., see Figures 4, 5, and 6), but this is not the case for very short tests of 10 items (e.g., see Figure 3). Also, for each condition, the superiority of \(H^F\) with tests of 30 and 20 items is higher in detecting cheating than guessing—e.g., compare Figures 5 and 6 for the case of 30 items, with 40% guessing and cheating (or “special knowledge”), respectively.
- The adjusted parametric fit statistics, \(t^*\) and \(Z_3^*\) outperform (slightly, yet consistently) their original counterparts, \(t\) and \(Z_3\).
Comparing the two adjusted parametric statistics, $t^*$ does somewhat better than $Z_3^*$ in detecting guessing and about the same as $Z_3^*$ in detecting cheating (or "special knowledge").

For practical references with using the person-fit statistics $t$, $t^*$, $Z_3$, $Z_3^*$, and $H^r$, Tables 3 and 4 provide the highest hit rate and its cutting score (given in parentheses) obtained for these

### Table 1

**Probability for Correct Item Response Estimated as $P(\hat{\theta}_i)$ and $\hat{\pi}_r$ (in Parentheses) Through the Use of Equations 1 and 2, Respectively, With Simulated Data ($N = 9,000$) on 10 Binary Items**

<table>
<thead>
<tr>
<th>Raw score</th>
<th>Item difficulty ($\hat{\delta}_i$)</th>
<th>$\hat{\theta}_i$</th>
<th>$\hat{\pi}_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$-$2.006</td>
<td>$-$1.390</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.2936</td>
<td>0.1833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.3714)</td>
<td>(0.2006)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.5222</td>
<td>0.3712</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.5991)</td>
<td>(0.4115)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.6887</td>
<td>0.5444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.7453)</td>
<td>(0.5910)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.8039</td>
<td>0.6889</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.8406)</td>
<td>(0.7292)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.8807</td>
<td>0.7995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9023)</td>
<td>(0.8280)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.9302</td>
<td>0.8780</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9417)</td>
<td>(0.8951)</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.9613</td>
<td>0.9307</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9669)</td>
<td>(0.9396)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.9807</td>
<td>0.9648</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9830)</td>
<td>(0.9687)</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.9926</td>
<td>0.9864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9934)</td>
<td>(0.9877)</td>
</tr>
</tbody>
</table>

Note: At each raw score level ($r = 1, ..., 9$), $P(\hat{\theta}_i)$ depends on both $\hat{\theta}_i$ and $\hat{\delta}_i$, whereas $\hat{\pi}_r$ depends only on $\hat{\delta}_i$.

### Table 2

**The Person-Fit Statistics $t$, $t^*$, $Z_3^*$, $Z_3$, and $H^r$, in Decreasing Order (From Left to Right) of their ROC Area Values**

<table>
<thead>
<tr>
<th>Test length/Level of aberrancy</th>
<th>Type of aberrant response behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 items 40%</td>
<td>$H^r$, $t^<em>$, $t$, $Z_3^</em>$, $Z_3$</td>
</tr>
<tr>
<td>30 items 20%</td>
<td>$H^r$, $t^<em>$, $t$, $Z_3^</em>$, $Z_3$</td>
</tr>
<tr>
<td>20 items 40%</td>
<td>$H^r$, $t^<em>$, $t$, $Z_3^</em>$, $Z_3$</td>
</tr>
<tr>
<td>20 items 20%</td>
<td>$H^r$, $t^<em>$, $t$, $Z_3^</em>$, $Z_3$</td>
</tr>
<tr>
<td>10 items 40%</td>
<td>$t^<em>$, $Z_3^</em>$, $t$, $Z_3$, $H^r$</td>
</tr>
<tr>
<td>10 items 20%</td>
<td>$t^<em>$, $Z_3^</em>$, $t$, $Z_3$, $H^r$</td>
</tr>
</tbody>
</table>

Note: Each person-fit statistic outperforms the statistics that follow (from left to right) in detecting aberrant response patterns.
Figure 1. ROC area for the person-fit statistics $t, t^*, Z_{3*}, Z_3$, and $HT$ when low ability examinees guess on 20% or 40% of the most difficult items in tests of 10, 20, and 30 items.

Figure 2. ROC area for the person-fit statistics $t, t^*, Z_{3*}, Z_3$, and $HT$ when low ability examinees cheat (or use some "special knowledge") on 20% or 40% of the most difficult items in tests of 10, 20, and 30 items.
Figure 3. Receiver Operating Curves (ROC) for the person-fit statistics $t$, $t^*$, $Z^*_3$, $Z_3$, and $H^T$ when low ability examinees guess on 20% of the most difficult items for tests of 10 items.

Figure 4. Receiver Operating Curves (ROC) for the person-fit statistics $t$, $t^*$, $Z^*_3$, $Z_3$, and $H^T$ when low ability examinees guess on 20% of the most difficult items for tests of 20 items.
Figure 5. Receiver Operating Curves (ROC) for the person-fit statistics $t$, $t^*$, $Z_3^*$, $Z_3$, and $H^T$ when low ability examinees guess on 40% of the most difficult items for tests of 30 items.

Figure 6. Receiver Operating Curves (ROC) for the person-fit statistics $t$, $t^*$, $Z_3^*$, $Z_3$, and $H^T$ when low ability examinees cheat (or use some “special knowledge”) on 40% of the most difficult items for tests of 30 items.
statistics at two specific levels of Type I error (false alarm) rate: .10 and .05; (in fact, the actual FAR values at the .10 and .05 levels are about .098 and .048, respectively, across all scenarios). For example, the highest hit rate for $H^c$ in detecting guessing on 40% of the most difficult items (out of 30 items) at the .10 level is .61 obtained at a cutting score of .31 (see Table 3). That is, using $H^c < .31$ to detect aberrant response patterns in this case leads to $HR = .61$ and $FAR = .098$ (i.e., 61% correct detection at about 9.8% false alarm rate). As another example, the highest hit rate for $H_T$ in detecting guessing on 40% of the most difficult items (out of 30 items) at the .10 level is .61 obtained at a cutting score of .31 (see Table 3). That is, using $H_T < .31$ to detect aberrant response patterns in this case leads to $HR = .61$ and $FAR = .098$ (i.e., 61% correct detection at about 9.8% false alarm rate). The results in Tables 3 and 4 also show that, at the .01 and .05 false alarm rates, (a) the non-parametric statistic, $H^c$, substantially outperforms the parametric statistics $t$, $t^*$, $Z_3$, and $Z_3^*$ with tests of 30 and 20 items, but the four parametric statistics perform better than $H^c$ with tests of 10 items, and (b) $t^*$ and $Z_3^*$ perform better than $t$ and $Z_3$, respectively, on tests of 30 items, somewhat better on tests of 20 items, and about the same on tests of 10 items.

For practical references to Tables 3 and 4 (or other person-fit outcomes), it would be useful to first obtain some information about the level of aberrancy in response patterns (e.g., 20% or 40%) by examining the Rasch model item fit with the data at hand. Then, an examination of the response patterns may provide indication about the type of aberrancy (e.g., guessing or cheating). For example, if aberrancy is suggested for 40% of 10 binary items, given in the increasing order (from left to right) of their difficulty, one can expect guessing on the four most difficult items

### Table 3

<table>
<thead>
<tr>
<th>False Alarm Rate</th>
<th>Test length/Level of Aberrancy</th>
<th>t</th>
<th>$t^*$</th>
<th>$Z_3^*$</th>
<th>$Z_3$</th>
<th>$H^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 items</td>
<td>40%</td>
<td>.41 (1.56)</td>
<td>.46 (1.53)</td>
<td>.39 (1.56)</td>
<td>.34 (1.60)</td>
<td>.61 (.31)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.41 (1.53)</td>
<td>.46 (1.52)</td>
<td>.34 (1.54)</td>
<td>.29 (1.56)</td>
<td>.52 (.32)</td>
</tr>
<tr>
<td></td>
<td>&lt;.10</td>
<td>.38 (1.40)</td>
<td>.41 (1.41)</td>
<td>.33 (1.36)</td>
<td>.28 (1.37)</td>
<td>.46 (.15)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>.42 (1.39)</td>
<td>.46 (1.42)</td>
<td>.32 (1.37)</td>
<td>.28 (1.36)</td>
<td>.43 (.15)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.33 (2.15)</td>
<td>.34 (2.30)</td>
<td>.39 (2.61)</td>
<td>.36 (2.24)</td>
<td>.33 (.03)</td>
</tr>
<tr>
<td></td>
<td>10 items</td>
<td>.40 (2.20)</td>
<td>.41 (2.30)</td>
<td>.41 (2.38)</td>
<td>.40 (2.10)</td>
<td>.30 (0.04)</td>
</tr>
<tr>
<td>30 items</td>
<td>40%</td>
<td>.33 (1.83)</td>
<td>.37 (1.80)</td>
<td>.32 (1.79)</td>
<td>.28 (1.82)</td>
<td>.51 (.27)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.32 (1.80)</td>
<td>.36 (1.79)</td>
<td>.27 (1.76)</td>
<td>.22 (1.79)</td>
<td>.38 (.28)</td>
</tr>
<tr>
<td></td>
<td>&lt;.05</td>
<td>.28 (1.68)</td>
<td>.30 (1.69)</td>
<td>.24 (1.64)</td>
<td>.21 (1.61)</td>
<td>.32 (.10)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>.29 (1.69)</td>
<td>.32 (1.75)</td>
<td>.25 (1.65)</td>
<td>.22 (1.61)</td>
<td>.33 (.10)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.20 (2.45)</td>
<td>.19 (2.60)</td>
<td>.19 (2.81)</td>
<td>.19 (2.44)</td>
<td>.19 (.05)</td>
</tr>
<tr>
<td></td>
<td>10 items</td>
<td>.23 (2.48)</td>
<td>.24 (2.62)</td>
<td>.25 (2.85)</td>
<td>.24 (2.56)</td>
<td>.19 (.06)</td>
</tr>
</tbody>
</table>

Note: The best hit rate (in bold) is obtained for (a) $t$, $t^*$, $Z_3^*$, and $Z_3$ when they exceed in absolute value the cutting score (in parentheses), and (b) $H^c$, when its algebraic value is smaller than the cutting score (in parentheses).
in the presence of numerous response patterns of the type (1000001010), (0100001000), and (1100000101), whereas numerous response patterns such as (1000001111), (1100001111), and (0100001111) would indicate cheating (or “special knowledge”).

**Conclusion**

Recent research on Rasch model measurement has sharpened its focus on person-fit analysis given the importance of such analysis for the accuracy of Rasch data and validity of their interpretations (e.g., Karabatsos, 2003; Li and Olejnik, 1997; Smith, 1991a, 1991b; Smith, Schumacker, and Bush, 1998). This paper provides an adjustment for two person-fit statistics relatively often discussed in previous research and used with some computerized applications— the cube root transformation of the unweighted total person fit statistic, $t$ (Smith, 1982, 1991b; Wright and Stone, 1979) and the standardized likelihood-based person fit statistic, $Z_3$ (Hulin, Drasgow, and Parsons, 1983). With Rasch data, the calculation of $t$ and $Z_3$ is based on the Rasch estimate of probability for correct item response, $P(\hat{\theta}_r)$, through the use of Equation 1. The value of $P(\hat{\theta}_r)$ depends, in turn, on the estimates of item difficulty, $\hat{\delta}_i$, and person’s ability, $\hat{\theta}_r$; (with the Rasch model, the person’s raw score, $r$, is a sufficient statistics for determining $\hat{\theta}_r$.) However, Van den Wollenberg (1982) argues that $\hat{\pi}_{ri}$ (with Equation 2) is more accurate estimate of the expected value of correct responses for an item than $P(\hat{\theta}_r)$ with Equation 1.

This study investigates an adjustment of the person-fit indices $Z_3$ and $t$ by replacing $P(\hat{\theta}_r)$

<table>
<thead>
<tr>
<th>False Alarm Rate</th>
<th>Test length/Level of Aberrancy</th>
<th>$t$</th>
<th>$t^*$</th>
<th>$Z_3^*$</th>
<th>$Z_3$</th>
<th>$H^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 items</td>
<td>40%</td>
<td>.60 (1.80)</td>
<td>.64 (1.76)</td>
<td>.68 (1.87)</td>
<td>.64 (1.90)</td>
<td>.99 (21)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.41 (1.62)</td>
<td>.46 (1.58)</td>
<td>.49 (1.66)</td>
<td>.43 (1.70)</td>
<td>.88 (24)</td>
</tr>
<tr>
<td>20 items</td>
<td>&lt; .10</td>
<td>.65 (1.38)</td>
<td>.67 (1.40)</td>
<td>.55 (1.46)</td>
<td>.50 (1.48)</td>
<td>.92 (28)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>.34 (1.38)</td>
<td>.37 (1.40)</td>
<td>.35 (1.40)</td>
<td>.30 (1.40)</td>
<td>.52 (29)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.57 (1.76)</td>
<td>.58 (1.83)</td>
<td>.62 (1.38)</td>
<td>.59 (1.26)</td>
<td>.22 (–.04)</td>
</tr>
<tr>
<td>10 items</td>
<td>40%</td>
<td>.30 (1.95)</td>
<td>.31 (2.06)</td>
<td>.36 (1.81)</td>
<td>.34 (1.61)</td>
<td>.11 (–.04)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.51 (2.06)</td>
<td>.55 (2.01)</td>
<td>.63 (2.06)</td>
<td>.59 (2.10)</td>
<td>.99 (19)</td>
</tr>
<tr>
<td></td>
<td>&lt; .05</td>
<td>.50 (2.07)</td>
<td>.54 (2.01)</td>
<td>.63 (2.06)</td>
<td>.59 (2.10)</td>
<td>.99 (19)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>.30 (1.92)</td>
<td>.35 (1.87)</td>
<td>.41 (1.89)</td>
<td>.35 (1.94)</td>
<td>.78 (22)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.23 (1.63)</td>
<td>.25 (1.66)</td>
<td>.25 (1.62)</td>
<td>.21 (1.63)</td>
<td>.36 (06)</td>
</tr>
<tr>
<td>10 items</td>
<td>40%</td>
<td>.40 (2.04)</td>
<td>.41 (2.16)</td>
<td>.45 (1.81)</td>
<td>.44 (1.64)</td>
<td>.12 (–.05)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>.21 (2.21)</td>
<td>.22 (2.33)</td>
<td>.26 (2.14)</td>
<td>.24 (1.95)</td>
<td>.06 (–.07)</td>
</tr>
</tbody>
</table>

Note. The best hit rate (in bold) is obtained for (a) $t$, $t^*$, $Z_3^*$, and $Z_3$ when they exceed in absolute value the cutting score (in parentheses), and (b) $H^T$, when its algebraic value is smaller than the cutting score (in parentheses).
with \( \hat{\pi}_n \) in calculating the terms on the right-hand side of Equations 4 and 8, respectively. Then, using Rasch data, the person-fit statistics \( t \) and \( Z_3 \) are compared to their adjusted counterparts, \( t^* \) and \( Z_3^* \), in power on detecting aberrant response patterns. The parametric statistics \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \) are compared also with the non-parametric fit statistic \( H^F \) (Sijtsma, 1986; Sijtsma and Mejer, 1992) which has been found to perform better than many other (parametric and non-parametric) person-fit statistics (Karabatsos, 2003).

With the conditions of data simulations in this study, the responses of low ability examinees are based on random guessing or cheating (“special knowledge”) on 20% or 40% of the most difficult items in short tests (10, 20, and 30 items). The results show that:

- The adjusted person-fit statistics \( t^* \) and \( Z_3^* \) outperform (slightly, yet consistently) their original counterparts, \( t \) and \( Z_3 \), respectively. Also, \( t^* \) does somewhat better than \( Z_3^* \) in detecting guessing and about the same as \( Z_3^* \) in detecting cheating (or “special knowledge”).

- The non-parametric person-fit statistic, \( H^F \), substantially outperforms the parametric statistics, \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \) in detecting aberrancy with 30 items and 20 items, but this is not the case for very short tests of 10 items. Also, the superiority of \( H^F \) with tests of 30 and 20 items is much higher in detecting cheating (or “special knowledge”) than guessing.

- The power of \( H^F \) in detecting aberrant response patterns, for both guessing and cheating (or “special knowledge”), increases with the increase of (a) test length and (b) level of aberrancy (40% versus 20%) for both guessing and cheating (or “special knowledge”). With \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \), this is true for cheating (or “special knowledge”), but not necessarily for guessing.

- The relative performance of the person-fit statistics \( t \), \( t^* \), \( Z_3 \), \( Z_3^* \), and \( H^F \), described in the above findings, remains about the same when the Type I error rate is held at the (frequently used in practice) levels of .10 and .05 (false alarm rate closely below 10% and 5%, respectively). For practical references, Tables 3 and 4 provide also the cutting score associated with the highest hit rate observed within the .10 and .05 levels of Type error rate.

In summary, the adjusted parametric person-fit statistics, \( t^* \) and \( Z_3^* \), tend to outperform their counterparts, \( t \) and \( Z_3 \), respectively, in detecting aberrant response patterns for low ability examinees who guess or cheat (use some “special knowledge”) in their responses on 20% or 40% of the most difficult items in short tests (10, 20, and 30 items). In this context, the non-parametric person-fit statistic \( H^F \) substantially outperforms \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \) on tests of 20 and 30 items, but not with very short tests of 10 items. This seems consistent with results reported by Karabatsos (2003) that \( H^F \) outperforms \( t \) and \( Z_3 \) (and numerous other parametric and non-parametric person-fit statistics) for tests of 17, 33, and 65 items.

The fact that \( H^F \) substantially outperforms the Rasch person fit statistics, \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \), makes it possible to assess person fit to the Rasch model, without having to fit data to a Rasch model. Indeed, \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \) are each calculated under the null hypothesis \( H_0 \): “Person \( n \) fits the Rasch model, which implies that the item characteristic curves (ICCs) are parallel and strictly increasing (over 0),” against the alternative \( H_1 \): “Not \( H_0 \).” On the other hand, \( H^F \) is calculated under the (weaker) null hypothesis \( H_0 \): “Person \( n \) fits a psychometric model which implies that the ICCs are non-intersecting and non-decreasing (over 0),” against the alternative \( H_1 \): “Not \( H_0 \).” Thus, it is interesting that while \( t \), \( t^* \), \( Z_3 \), and \( Z_3^* \) are each calculated under a “precise” null hypothesis of the Rasch model (the study concerns person fit to the Rasch model), \( H^F \) performs substantially better for non-short tests, even though the null hypothesis that is implied by \( H^F \) (i.e., \( H_0 \)) does not precisely represent the Rasch model. Of course, the Rasch model must be tested for data fit when the results and their interpretations (beyond the purpose of assessing person fit) are based on properties of the Rasch model.
It is important also to emphasize that the use of the adjusted person-fit statistics $t^*$ and $Z^*_3$ (if preferred to $t$ and $Z_3$, respectively) must be limited to short tests. This is because $t^*$ and $Z^*_3$ are based on symmetric functions, the estimation of which is computationally very expensive with, say, more than 30 items. In fact, given that $t^*$ and $Z^*_3$ provide advantage over the other person fit statistics only for very short tests (e.g., of 10 items), then the “computational expense” with the symmetric functions is not an issue with these adjusted person fit statistics for the Rasch model. In conclusion, the hope is that the results provided in this article may have useful methodological and practical implications for Rasch analysis of person-fit.

References


