

# Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics

George Karabatsos  
*University of Illinois–Chicago*

The accurate measurement of examinee test performance is critical to educational decision-making, and inaccurate measurement can lead to negative consequences for examinees. Person-fit statistics are important in a psychometric analysis for detecting examinees with aberrant response patterns that lead to inaccurate measurement. Unfortunately, although a large number of person-fit statistics is available, there is little consensus as to which ones are most useful. The purpose of this study was to compare 36 person-fit indices, under different testing conditions, to obtain a better consensus as to their relative merits. The results of these comparisons, and their implications, are discussed.

Sound decisions in educational settings hinge largely on accurate measurement of student characteristics. Such measurements can help identify those individuals who are qualified enough to enter a particular school, or receive a particular educational degree. Also, these measurements can be used to monitor students' learning progress. This may, for example, enable educators to productively tailor their curriculum, or help policy makers decide on important educational issues.

In contrast, the inaccurate measurement of test performance can lead to negative consequences. On the one hand, spuriously high test scores can lead to unqualified individuals being enrolled into an educational program (e.g., undergraduate, graduate, or professional), or being awarded an educational degree. On the other hand, qualified individuals with spuriously low test scores may be unfairly excluded from academic programs, or unfairly denied a degree. Furthermore, the inaccurate measurement of test performance undermines the assessment of students' learning progress, and curriculum planning efforts.

At least five factors can cause an examinee's score on a test to be spuriously high or spuriously low: cheating, careless responding, lucky guessing, creative responding, and random responding (Meijer, 1996a, 1996b). *Cheating* (e.g., copying from another examinee) refers to behavior where the examinee unfairly obtains the correct answers on test items that he/she is unable to answer correctly. *Careless responding* occurs when the examinee, in slipshod fashion, answers certain items incorrectly, for which he/she is able to answer correctly. *Lucky guessing* happens when the examinee guesses the correct answers to some test items, for which he/she does not know the correct answer. Some examinees with high ability engage in *creative responding*, where they obtain incorrect responses to certain easy items, because they interpret these items in a unique, creative manner. Finally, *random responding* refers to the situation where the examinee randomly chooses the multiple-choice option for each item on the test.

Considering the potential consequences that result from the inaccurate measurement of examinees, it is imperative to identify those individuals with inappropriate test responses. *Person-fit statistics* are designed to identify examinees with aberrant item response patterns (which lead to spuriously high or spuriously low test scores), and to distinguish those respondents from respondents with non-aberrant, *normal* item-response patterns.

Interest in person-fit analysis initiated in the early part of the century (e.g., Cronbach, 1946; Fowler, 1954; Glaser, 1949, 1950, 1951, 1952; Guttman 1944, 1950; Mosier, 1940; Sherif & Cantril, 1945, 1946; Spearman, 1910; Thurstone, 1927). This research intensified during the late 70s after item response theory (IRT) models were established as mainstream methods of psychometrics (Lord & Novick, 1968; Mokken, 1971; Rasch, 1960). In fact, *Applied Measurement in Education* (1996, 9(1)) recently devoted a special issue entitled "Person-fit research: Theory and applications." Also, Meijer and Sijtsma (2001), in their review, showed that there are over forty statistics available to test person-fit.

Researchers of fourteen previous studies compared the quality of person-fit statistics, either with simulated or real empirical data (Birenbaum, 1985, 1986; Drasgow, Levine, & McLaughlin, 1987; Harnisch & Linn, 1981; Harnisch & Tatsuoka, 1983; Kogut, 1987; Li & Olejnik, 1997; Meijer, 1998; Meijer, 1994; Meijer, Muijtjens, & van der Vleuten, 1996; Nering & Meijer, 1998; Noonan, Boss, & Gessaroli, 1992; Rogers & Hattie, 1987; Rudner, 1983). In ten of these studies, only 3 to 11 person-fit statistics were compared, and in the remaining four studies, only two were compared. This pattern characterizes the unsystematic nature of person-fit research (Meijer & Sijtsma, 1999, p. 13; Rudner, Bracey, & Skaggs, 1996; Rudner, Skaggs, Bracey, & Getson, 1995), and so it is difficult to directly compare the quality of all the person-fit statistics through the review of these studies. As a result, there is no universal agreement about which fit statistics are the most effective in detecting aberrant-responding examinees.

## A REVIEW OF PERSON-FIT STATISTICS

A person-fit statistic measures the degree of reasonableness of an examinee's answers to a set of test items. To formulate this idea mathematically, let  $X_{nj} = 1$  denote a correct response made by examinee  $n$  on test item  $j$ ,  $X_{nj} = 0$  an incorrect response, and  $\{n = 1, \dots, N\}$  denotes a random sample of examinees who responds to the set of test items  $\{j = 1, \dots, J\}$ . A basic standard of reasonableness is that, given an examinee  $n$ 's score over the test items,  $r_n = \sum_{j=1}^J X_{nj}$ , the examinee's item responses most probably contains the correct answers on the easiest  $r_n$  test items, and incorrect answers to the remaining  $J - r_n$  items. With respect to the parametric (dichotomous) models of item response theory (IRT), the easiness of an item is predicted by  $P_{nj1} \in [0, 1]$ , which refers to the probability of a correct response for examinee  $n$  on item  $j$ . These models estimate the probability  $P_{nj1}$  as a logistic function of a respondent's ability level  $\theta_n \in \text{Re}$ , an item difficulty parameter  $\delta_j \in \text{Re}$ , and possibly other item parameters (item discrimination, item lucky-guessing), depending on the particular model considered. For non-parametric IRT models, and classical test theory, the difficulty of an item  $j$  is based on  $\sum_{n=1}^N 1 - X_{nj}$ , the number of incorrect responses it contains over the sample of  $N$  examinees (see for e.g., van der Ark, 2001, p. 273, as how the item score stochastically orders examinee ability,  $\theta$ ).

As indicated in Meijer and Sijtsma's (2001) review, there is a large number of statistics invented for the purpose of identifying aberrant-responding examinees. Most of them are presented in Table 1. The table includes a total of 36 person-fit statistics, which are classified as either *non-parametric* or *parametric*. A non-parametric person-fit statistic is not based on estimated IRT model parameters, but is wholly calculated from the data set of  $N$  examinees' scored responses to  $J$  test items. In contrast, a parametric person-fit statistic measures the distance between the test data set and the estimated response predictions derived from the parameter estimates of an IRT model. Technical details about the statistics presented in Table 1 are provided in Appendix A and B; Meijer and Sijtsma (2001) give an excellent and thorough review of them.

The left side of Table 1 lists 11 *non-parametric* person-fit statistics. For a given examinee  $n$ , the statistic  $G$  counts the number of item response pairs that deviate from the "Guttman perfect pattern." Among  $J$  test items, such a pattern contains correct responses for only the easiest  $r_n$  items. The statistic  $G^*$  normalizes  $G$  to have the range  $[0, 1]$ . The Personal Point-Biserial statistic  $r_{pbis}$  measures the correlation between examinee  $n$ 's responses  $\mathbf{X}_n = (X_{n1}, \dots, X_{nj}, \dots, X_{nJ})'$  and the vector  $\mathbf{p} = (p_1, \dots, p_j, \dots, p_J)'$ , where  $p_j$  is the proportion of correct responses obtained by the  $N$  respondents on item  $j$ . The Caution Index  $C$ , with a covariance ratio, measures the degree to which an examinee's item-responses  $\mathbf{X}_n$  deviate from the perfect pattern.  $C$  has no upper bound, so the Modified Caution Index (MCI) is a modification of  $C$  to have a range  $[0, 1]$ . In particular, MCI is the ratio between two

TABLE 1  
Thirty-Six Person-Fit Statistics

| Non-Parametric<br>Person-Fit Statistics (11) |   | Parametric<br>Person-Fit Statistics (25)          |                                       |
|--|---|---|---------------------------------------|
| <i>G</i>                                     | (Guttman, 1944, 1950)                     | <i>U</i>  | (Wright & Stone, 1979)s               |
| <i>G*</i>                                    | (van der Flier, 1977)                     | <i>ZU</i>   | (Wright, 1980)                        |
| <i>r<sub>pbis</sub></i>                      | (Donlon & Fischer, 1968)                  | <i>lnU</i>  | (Wright & Stone, 1979)                |
| <i>C</i>                                     | (Sato, 1975)                              | <i>W</i>  | (Wright, 1980)                        |
| <i>MCI</i>                                   | (Harnisch & Linn, 1981)                   | <i>ZW</i>   | (Wright, 1980)                        |
| <i>U3</i>                                    | (van der Flier, 1980)                     | <i>lnW</i>  | (Wright & Stone, 1979)                |
| <i>ZU3</i>                                   | (van der Flier, 1982)                     | <i>EC11, EC12, EC13, EC14, EC15, EC16, EC11z,</i> |                                       |
| <i>H<sup>T</sup></i>                         | (Sijtsma, 1986;<br>Sijtsma & Mejer, 1992) | <i>EC12z, EC14z, EC16z</i>                        | (Tatsuoka, 1984)                      |
| <i>A, D, E<sub>i</sub></i>                   | (Kane & Brennan, 1980)                    | <i>l</i>  | (Levine & Rubin, 1979)                |
|  |   | <i>l<sub>z</sub></i>                              | (Drasgow, Levine, & Williams, 1985)   |
|  |   | <i>M</i>  | (Molenaar & Hojtkink, 1990)           |
|  |   | <i>M(p-value)</i>                                 | (Bedrick, 1997)                       |
|  |   | <i>Item-Grouping Person-fit Statistics</i>        |                                       |
|  |   | <i>D(θ)</i>                                       | (Trabin & Weiss, 1983)                |
|  |   | <i>l<sub>zm</sub></i>                             | (Drasgow, Levine, & McLaughlin, 1991) |
|  |   | <i>UB</i>   | (Smith, 1986)                         |
|  |   | <i>ZUB</i>  | (Smith, 1986)                         |
|  |   | <i>lnUB</i>                                       | (through<br>Wright & Stone, 1979)     |

covariances, namely, the covariance of  $\mathbf{X}_n$  with the perfect pattern, and the covariance of  $\mathbf{X}_n$  with the perfectly-inconsistent pattern. Among  $J$  test items, the perfectly-inconsistent pattern contains correct responses for only the most difficult  $r_n$  items. The statistic  $U3$  has the same form as  $MCI$ , replacing covariance with a log ratio.  $ZU3$  transforms  $U3$  to have a unit-normal distribution (i.e., with mean = 0 and s.d. = 1). The statistic  $H^T$  is a correlation index that measures the similarity between examinee  $n$ 's response vector  $\mathbf{X}_n$ , with the response vectors of the remaining  $N-1$  examinees. Finally, the Agreement Index ( $A$ ) and the Disagreement Index ( $D$ ) measure the agreement and disagreement between  $\mathbf{X}_n$  and  $\mathbf{p}$ , respectively, and Dependability ( $E_i$ ) is the ratio of  $A$  over the sum of the elements of  $\mathbf{p}$ .

Four non-parametric person-fit approaches mentioned in the literature are excluded from Table 1. The first is the Norm Conformity Index ( $NCI$ , Tatsuoka & Tatsuoka, 1983), which relates perfectly to  $G^*$ , with  $NCI = 1-2G^*$  (Meijer & Sijtsma, 2001). The second is the Individual Consistency Index ( $ICI$ ), which is the same as  $NCI$  but measures the agreement between examinee  $n$ 's responses and an a-priori defined item response pattern that reflects a particular cognitive skill.  $ICI$  was excluded because cognitive based psychometrics is not the focus in this study. The third is the Personal Biserial Correlation  $r_{bis}$  (Donlan & Fischer, 1968), which

is very similar to  $r_{pbis}$ , with the only exception that the former assumes a continuous normal distribution underlying an examined set of item responses. The last approach pertains to the very recent work of Sijtsma and Meijer (2001), who, within the context of non-parametric IRT, propose a new method in testing person-fit based on the person response function.

The right side of Table 1 lists 25 *parametric* person-fit statistics. The first set of parametric statistics is based on mean squares. The statistic  $U$ , for an examinee  $n$ , is the average of the squared response residuals  $(X_{nj} - P_{nj1})^2$  over the  $J$  item responses.  $W$  is the average of the item response residuals, weighted by the sum of the variances  $\sum_{j=1}^J P_{nj1}(1 - P_{nj1})$ . To interpret  $U$  and  $W$  on a unit-normal distribution, they may either have a  $Z$ -cubic root ( $Z$ ) or logarithmic ( $ln$ ) transformation. These transformations yield the statistics  $ZU$ ,  $lnU$ ,  $ZW$ , and  $lnW$ .

Several caution indices are based on IRT model parameters. Statistics  $EC11$ ,  $EC12$ , and  $EC14$  are based on covariances among  $\mathbf{X}_n$ ,  $\mathbf{p}$ ,  $\mathbf{P}_n = (P_{n11}, \dots, P_{nj1}, \dots, P_{Nj1})'$ , and  $\mathbf{G} = (G_1, \dots, G_j, \dots, G_J)'$ , where within item  $j$ ,  $G_j$  is the averages of the  $P_{nj1}$  over  $N$  examinees ( $\mathbf{G}$  is not in  $EC11$ ).  $EC13$  and  $EC15$  are analogs to  $EC12$  and  $EC14$ , respectively, which use correlation instead of covariance.  $EC16$  is based on a ratio, with the correlation between  $\mathbf{X}_n$  and  $\mathbf{P}_n$  in the numerator, and the variance of the  $\mathbf{P}_n$  elements in the denominator.  $EC11_z$ ,  $EC12_z$ ,  $EC14_z$ , and  $EC16_z$  are unit-normal transformations of  $EC11$ ,  $EC12$ ,  $EC14$ , and  $EC16$ , respectively.

The statistic  $l$  measures the log-likelihood fit of an examinee's responses  $\mathbf{X}_n$  with the predictions of an IRT model  $\mathbf{P}_n$ . The index  $l_z$  is the unit-normal transformation of  $l$ . The fit statistic  $M$ , for an examinee  $n$ , is the sum of the product  $X_{nj}\delta_j$  over the  $J$  items. This statistic is specifically applicable to the Rasch IRT model (Rasch, 1960), where the total score  $r_n$  can be used to condition out ability  $\theta_n$ . The  $p$ -value corresponding to the value of  $M$ , denoted by  $M(p\text{-value})$ , is estimated by the so-called Edgeworth tail approximation.

Five person-fit statistics, as listed in Table 1, measure the fit between a priori defined subsets of the  $J$  items. An item subset, for example, may refer to items of a particular subscale in a test, or to a group of items that have a common range of difficulty. The statistic  $D(\theta)$  is the sum, over the  $S$  subsets, of the average of the response residuals  $(X_{nj} - P_{nj1})$  within each item subset  $S$ .  $UB$  measures  $W$  for each subset  $s$ , then averages them over the  $S$  subsets.  $UB$  can be interpreted on a unit-normal distribution, through either the cubic-root transformation to obtain  $ZUB$ , or the logarithmic transformation to obtain  $lnUB$ . Finally, the log likelihood statistic  $l_{zm}$  is  $l_z$  summed over the  $S$  item subsets.

Seven parametric person-fit statistics identified in the literature are excluded from Table 1. The first two include the ratio of observed and expected Fisher information  $O/E$ , and the jackknife variance estimate  $JK$ , which are both sensitive to the flatness of the likelihood function. Drasgow et al. (1987) found that they are ineffective in detecting aberrant-responding examinees. The next two statistics are  $\lambda(\mathbf{X})$  (Levine & Drasgow, 1988) and  $T(\mathbf{x})$  (Klauer, 1991, 1995), and each are used

to test the null hypothesis that an examinee's responses are consistent with a parametric model assuming normal responses (e.g., Rasch model), against the alternative hypothesis that the responses are consistent with a parametric model assuming aberrant responses (e.g., a model for cheating behavior, local item dependence, careless responding). These statistics are optimal in the sense that no other fit statistic can achieve a higher rate of detection of aberrant-responding examinees. However, these approaches cannot be routinely implemented in a psychometric analysis, as they require a specification and estimation of the alternative parametric model. This study focuses on person-fit statistics that can be directly calculated and routinely implemented. Klauer and Rettig (1990) developed the three remaining fit statistics, which are related because they are a version of  $D(\theta)$ , standardized in order to asymptotically follow a  $\chi^2$  distribution for long tests. These statistics are  $\chi_{SC}^2$ , a Wald test, and a likelihood ratio test. Klauer and Rettig (1990) show that  $\chi_{SC}^2$  is distributed  $\chi^2$  only for very long tests ( $\geq 80$  items). Also, the Wald and likelihood ratio tests are not very practical, since for each one, the difference between the theoretical and empirical  $\chi^2$  distribution is quite large.

### Purpose of This Study

In this study, 36 person-fit statistics are compared in their ability to detect examinees with aberrant item-response patterns. The goal of this effort is to obtain a better consensus as to which person-fit statistics are best. The next section describes the methods used to compare them with simulated data, and the last two sections of the article details and discusses the results of these comparisons, respectively.

## METHODS

### Model Framework

The 36 person-fit statistics were compared within the context of the Rasch model (Rasch, 1960). This model is basic to the family of IRT models (van der Linden & Hambleton, 1997), and also, its parametric simplicity enables a straightforward investigation of the fit statistics. Furthermore, the Rasch model is frequently used in educational and psychological testing (e.g., Bond & Fox, 2001; Fisher & Wright, 1997; Wilson & Englehard, 2000).

The Rasch model postulates that when examinee  $n$  with ability  $\theta_n$  encounters item  $j$  with difficulty  $\delta_j$ , the probability of a correct response  $P_{nj1}$  depends on the logarithmic difference between  $\theta_n$  and  $\delta_j$ :

$$P_{nj1} = 1 - P_{nj0} = [1 + \exp(-(\theta_n - \delta_j))]^{-1}, \quad (1)$$

where  $P_{nj0}$  is the probability of an incorrect item response. It is well known that examinee  $n$ 's test score  $r_n = \sum_{j=1}^J X_{nj}$ , and item  $j$ 's score  $q_j = \sum_{n=1}^N X_{nj}$ , are sufficient statistics for  $\theta_n$  and  $\delta_j$ , respectively (e.g., Andersen, 1977). Therefore, one may view the ability and difficulty estimates, in an approximate sense, to satisfy  $\hat{\theta}_n \approx \ln(r_n/(L - r_n))$  and  $\hat{\delta}_j \approx \ln((N - q_j)/q_j)$ . A Newton-Raphson procedure, called the Unconditional Maximum Likelihood (Wright & Panchepakesan, 1969), is commonly used to estimate the parameters of the Rasch model. Plugging in parameter estimates  $\{\hat{\theta}_n, \hat{\delta}_j\}$  into (1) yields  $\hat{P}_{nj1}$ , the estimated (response) prediction of the model.

### Data Simulations

Sixty data sets were simulated according to a fully-crossed  $5 \times 4 \times 3$  design. The design included five types of aberrant responding examinees (cheaters, creative respondents, guessing, careless, and random respondents), four percentages of aberrant-responding examinees (5%, 10%, 25%, or 50%), and three test lengths (short—17 items, medium—33 items, long—65 items). Each data set consisted of 500 (simulated) normal and aberrant examinees.

A data set  $\mathbf{X} = (X_{nj} \mid n = 1, \dots, N; j = 1, \dots, J)$  was simulated by generating  $X_{nj} = 1$  with probability  $P_{nj1}$ , and  $X_{nj} = 0$  with probability  $P_{nj0} = 1 - P_{nj1}$ , independently for all  $n = 1, \dots, N$  and  $j = 1, \dots, J$ . These probabilities  $\mathbf{P} = (P_{nj1} \mid n = 1, \dots, N; j = 1, \dots, J)$  were obtained by inputting particular values of ability parameters  $\theta = (\theta_1, \dots, \theta_n, \dots, \theta_N)$  and item difficulty parameters  $\delta = (\delta_1, \dots, \delta_j, \dots, \delta_J)$  into the Rasch model (1).

Specifically, a data set  $\mathbf{X}$  was generated with difficulty parameters (the  $\delta_j$ ) specified to be equidistant with a mean of 0 and a range  $[-2, 2]$ . The  $\delta_j$  were spaced .25 logits for the 17-item test, .125 logits for the 33-item test, and .0625 logits for the 65-item test. The normal examinees of a data set had a uniform ability ( $\theta$ ) distribution with range  $[-2, 2]$ . Cheating examinees were simulated by generating item responses from a uniform low ability distribution with a  $\theta$  range of  $[-2, -.5]$ , and imputing correct responses for the 18% most difficult items on the test (items with  $\delta_j \geq 1.5$ ). Creative examinees were simulated by generating item responses from a uniform, high-ability distribution with a  $\theta$  range of  $[.5, 2]$ , and then imputing incorrect responses to the 18% easiest items (items with  $\delta_j \leq -1.5$ ). Examinees with lucky-correct guessing responses were generated from a uniform low ability distribution with a  $\theta$  range of  $[-2, -.5]$ , and each of the 41% most difficult items (items with  $\delta_j \geq .5$ ) was assigned a .25 probability of containing a correct response. The value .25 was chosen to simulate multiple-choice items, each having four answer choices. Careless examinees were simulated from a uniform high-ability distribution with a  $\theta$  range of  $[.5, 2]$ , and each of the 41% easiest items (with difficulty  $\delta_j \leq -.5$ ) was assigned a .5 probability of containing an incorrect response. To generate randomly responding examinees, each and every item was assigned a .25 probability of containing a correct response. This simulates

examinees who blindly respond to all multiple-choice test items, each item having four answer choices.

For each of the 60 simulated data sets, Rasch model parameters were estimated, and the 36 person-fit statistics were calculated for each (simulated) examinee. In particular, for each examinee  $n$ , 23 of the 25 parametric person-fit statistics were calculated from estimated predictions  $\hat{P}_n = (\hat{P}_{n11}, \dots, \hat{P}_{nj1}, \dots, \hat{P}_{nJ1})'$  derived from the Rasch model parameter estimates  $\{\hat{\theta}_n, \hat{\delta}_j | j = 1, \dots, J\}$ . The remaining two parametric person-fit statistics  $M$  and  $M(\text{p-value})$ , were calculated from the item difficulty estimates  $\hat{\delta} = (\hat{\delta}_1, \dots, \hat{\delta}_j, \dots, \hat{\delta}_J)'$ . Also, in calculating the five person-fit statistics that measure person-fit by grouping items ( $D(\theta)$ ,  $UB$ ,  $ZUB$ ,  $\ln UB$ ,  $l_{zm}$ ), three item groups were specified a priori. The easiest one-third of the items composed one group, the most difficult one-third of the items composed a second group, and the remaining "medium-difficulty" items were in a third group.

All the methods described in this subsection were executed by a written program in S-PLUS code (S-PLUS, 1995), and is available from the corresponding author upon request.

### Computing the Detection Rate of the Person-Fit Statistics

With the 60 simulated data sets, Receiver Operating Curve (ROC) analysis compared the 36 person-fit statistics in their ability to detect aberrant-responding examinees. All ROC analyses were conducted with the SPSS software, version 8.0 (SPSS, 1998).

For a given set of simulated data, the ROC analysis of a person-fit statistic estimated  $H(c)$ , the probability that an aberrant-responding examinee has a fit statistic value greater than  $c$ . The same analysis also estimated  $F(c)$ , the probability that a normal examinee has a fit statistic value greater than  $c$ . This interpretation of the probabilities applies to a person-fit statistic where a higher value implies a higher person misfit. They are easily re-interpreted for a person-fit statistic where a higher value implies better person-fit (i.e.,  $H(c)$  and  $F(c)$  refer to probabilities of a fit value less than  $c$ ).

The quantity  $H(c)$  is the sensitivity, or "hit rate," the probability that a person-fit statistic correctly identifies an aberrant-responding examinee, using  $c$  as the critical value of the fit statistic. So the "miss rate"  $1-H(c)$  is the probability that a fit statistic incorrectly classifies an aberrant-responding examinee as normal. The probability  $F(c)$  is the "false-alarm" rate of the person-fit statistic, conditional on  $c$ . This rate refers to the probability that a person-fit statistic (at critical value  $c$ ) incorrectly classifies a normal examinee as aberrant. The quantity  $1-F(c)$  refers to the specificity, the probability that a person-fit statistic correctly classifies a normal examinee as normal.

Now consider a two-dimensional graph, with  $F(c)$  on the  $x$ -axis,  $H(c)$  on the  $y$ -axis, and the coordinate  $\{\hat{F}(c), \hat{H}(c)\}$  containing the estimates of the false alarm

rate and the hit rate, conditional on the value  $c$ . The ROC curve, of a person-fit statistic, is represented by a line that connects a set of such coordinates, over all possible values of  $c$  (observed from the simulated data set). A person-fit statistic with high sensitivity and specificity will have a ROC line that curves close to the upper-left corner of the graph. Therefore, for a person-fit statistic, its sensitivity and specificity are together represented by area  $a \in [0,1]$  under the ROC curve. The value  $a = 1$  refers to a fit statistic having perfect sensitivity and perfect specificity.

The 36 person-fit statistics were compared on their ability to detect aberrant responding examinees on the basis of ROC area ( $a$ ) values, each value estimated for a person-fit statistic over several simulated data sets (each  $a$  estimate bracketed by a 95% confidence interval). Comparisons were made with respect to the five types of aberrant responding examinees (cheaters, creative respondents, guessing, careless, and random responders), the per cent of aberrant respondents contained in the data set (5%, 10%, 25%, or 50%), and the length of the test (17, 33, or 65 items).

## RESULTS

Figure 1 shows a comparison of the detection ability between 36 person-fit statistics, with respect to the type of aberrant-responding examinees (cheaters, creative respondents, lucky guessers, careless respondents, and random respondents). For each of these five groups, the ROC results are based on 6000 simulated examinees. The 6000 correspond to all the four conditions pertaining to the percentage of aberrant respondents in the sample (5%, 10%, 25%, and 50%), and all the three test-length conditions (17, 33, and 65 items). In the legend of the figure (and for the remaining figures in this article), “CI Lower Bound” and “CI Upper bound” refer to the lower and upper bounds of the 95% confidence interval of the ROC area estimate ( $a$ ).

The pattern of ROC area results in Figure 1 show that creative and cheating examinees are the most difficult to detect, lucky-guessers are slightly easier to detect, and careless-responding and random-responding examinees are the easiest to detect. By far,  $H^T$  and  $D(\theta)$  are best in detecting cheating and creative-responding examinees, whereas  $H^T$ ,  $D(\theta)$ , and  $E_i$  are most effective in identifying the lucky-guessing examinees, and also,  $D(\theta)$  and  $H^T$  are best in detecting careless respondents. Furthermore,  $r_{pbis}$ ,  $C$ ,  $MCI$ ,  $U3$ ,  $H^T$ ,  $E_i$ ,  $ECI3$ ,  $ECI5$ , and  $M$  are most effective in detecting random-responding examinees.

With respect to the four percentages of aberrant respondents in the sample (5%, 10%, 25%, 50%), Figure 2 shows a comparison of the detection ability between 36 person-fit statistics. Within each of these four groups, the ROC results were based on 7500 simulated examinees. The 7500 correspond to all the five conditions pertaining to the types of aberrant-responding examinees (cheaters,

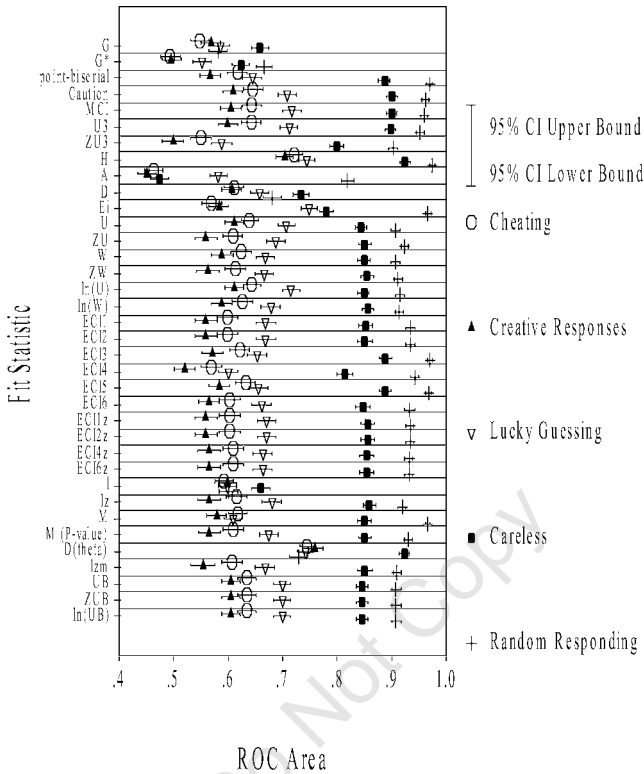


FIGURE 1 A comparison of 36 person-fit statistics, in their ability to detect cheating, creative, lucky-guessing, careless, and random-responding examinees.

creative respondents, lucky guessers, careless respondents, and random respondents), and all the three test-length conditions (17, 33, and 65 items).

The figure shows that, in general, detection rates decrease as the percentage of aberrant-responding examinees increase. By far, the 50% condition is the most difficult in which to detect aberrant respondents, and the 5%, 10%, and 25% conditions had approximately the same rates of misfit detection. In the 50% condition, there were negligible detection differences between the 36 person-fit statistics, and the most effective were *Ei*, *ECI1*, *ECI2*, and *ECI6*. Within each of the 5%, 10%, and 25% conditions, there were essentially no differences in detection rates between the 36 person-fit statistics.

Figure 3 shows a comparison of the detection ability between 36 person-fit statistics, with respect to test length (17, 33, or 65 items). Within each of these three groups, the ROC results are based on 10,000 simulated examinees. The 10,000 combine all the five conditions corresponding to the types of aberrant-responding

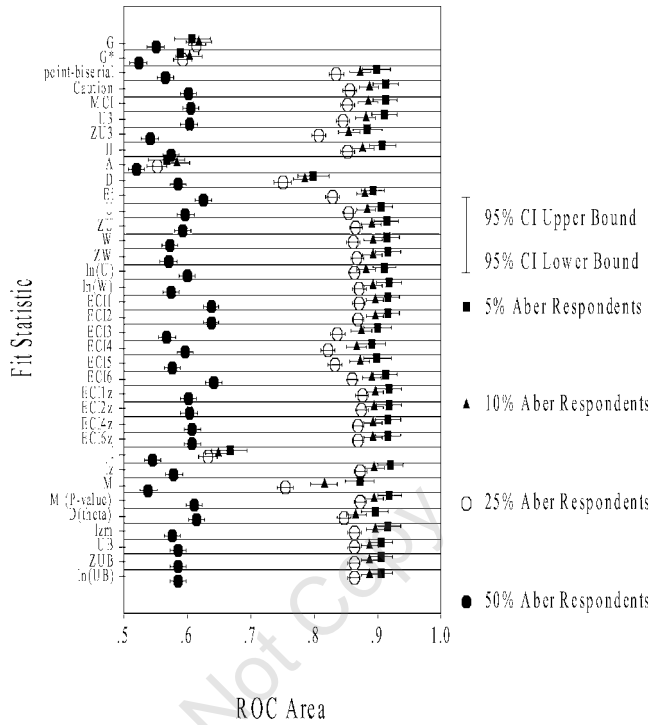


FIGURE 2 A comparison of 36 person-fit statistics, in their ability to detect aberrant-responding examinees under four different conditions, each condition representing the percentage of such examinees in the sample.

examinees (cheaters, creative respondents, lucky guessers, careless respondents, and random respondents), and all the four conditions pertaining to the percentage of aberrant respondents in the sample (5%, 10%, 25%, and 50%). The pattern of ROC area estimates in Figure 3 show that detection rates increase with test length. Also,  $H^T$ ,  $l$ , and  $D(\theta)$  are most effective in detecting aberrant respondents for short (17 item), medium (33-item), and long (65-item) tests.

Figure 4 presents a comparison of the overall detection ability between the person-fit statistics, based on the grand total of 40,000 simulated examinees. The 40000 combine all five groups of aberrant-responding examinees, all four conditions pertaining to the percentage of aberrant respondents in the sample (5%, 10%, 25%, and 50%), and all the three test-length conditions (17, 33, and 65 items). The figure reveals that the  $H^T$  statistic best detects aberrant-responding examinees, followed by  $D(\theta)$ ,  $C$ ,  $MCI$ , and  $U3$  which are tied for second. Twenty-four of the remaining 30 fit statistics are tied for third.

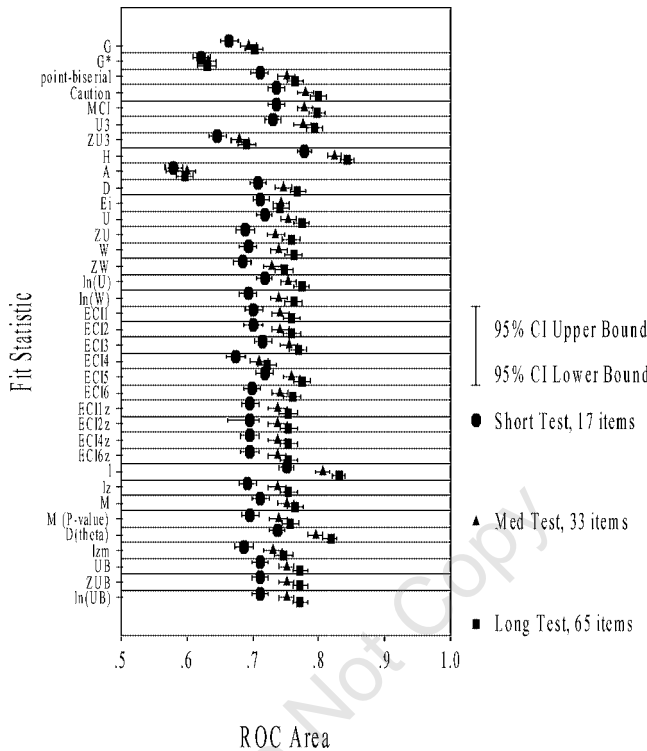


FIGURE 3 A comparison of 36 person-fit statistics, in their ability to detect aberrant-responding examinees under three different test-length conditions.

Given that overall,  $H^T$ ,  $D(\theta)$ ,  $C$ ,  $MCI$ , and  $U3$  performed best, it is useful to indicate how to best interpret them for the detection of aberrant-responding examinees. From the total of 40,000 simulated examinees, the estimated ROC curve of the best-performing,  $H^T$  statistic reveals that the critical values  $H^T \leq .22$  best identify aberrant-responding examinees. The .22 cutoff point corresponds to a value of  $c$  that, together, maximizes the sensitivity rate and specificity rate. Likewise, the estimated ROC curve for each of the second-best performing fit statistics determined that  $D(\theta) \geq .55$ ,  $C \geq .53$ ,  $MCI \geq .26$ , and  $U3 \geq .25$  are the critical values that optimally identify aberrant-responding examinees.

### DISCUSSION

Of the 36 person-fit statistics examined in this study,  $H^T$ ,  $D(\theta)$ ,  $C$ ,  $MCI$ , and  $U3$  are the best in identifying aberrant-responding examinees. Psychometric practitioners interested in applying any one of the five best-performing person-fit

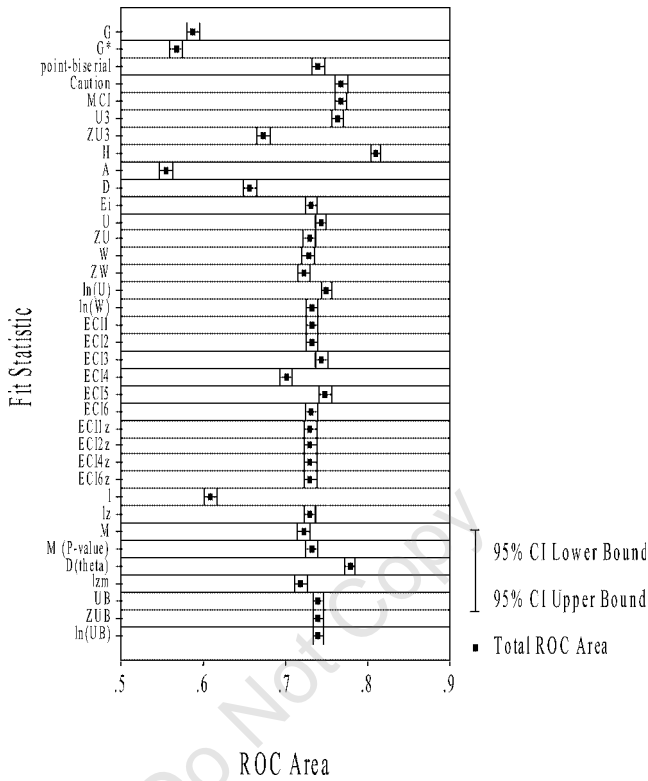


FIGURE 4 A comparison of 36 person-fit statistics, in their ability to detect aberrant-responding examinees, over all the simulate data generated in this study.

statistics are recommended to use critical values identified at the end of the previous section, obtained from the ROC analysis of all 40,000 simulated examinees. These critical values can be used to identify many types of aberrant-responding examinees, including cheaters, careless respondents, lucky-guessers, creative respondents, and random respondents.

Overall, the  $H^T$  statistic is best in identifying aberrant test respondents. It is also among the best in detecting each of the five different types of aberrant-responding examinees, and in detecting such examinees in each of the short, medium, and long test conditions. Three of the fit statistics that performed second-best, namely  $C$ ,  $MCI$ , and  $U3$ , have the same basic form (Meijer & Sijtsma, 2001, p. 109):

$$Q = \frac{\sum_{j=1}^{r_n} w_j - \sum_{j=1}^J X_{nj}w_j}{\sum_{j=1}^{r_n} w_j - \sum_{j=j-r_n+1}^J X_{nj}w_j} \tag{2}$$

The Caution index  $C$  is obtained from (2) by setting  $w_j = p_j$ , and multiplying  $\sum_{j=J-r_n+1}^J X_{nj} w_j$  by  $r_n$  and the three other terms by  $J$ . The statistic  $MCI$  is obtained by setting  $w_j = p_j$ , and  $U3$  instead uses the weight  $w_j = \ln[p_j / (1 - p_j)]$ . So it is not surprising that, over all the 60 simulated data sets,  $C$  and  $U3$  correlated .991,  $C$  and  $MCI$  correlated .994, and  $MCI$  and  $U3$  correlated .998.

There seems to be an important difference between  $H^T$  and the statistics  $C$ ,  $MCI$ , and  $U3$ . On the one hand,  $C$ ,  $MCI$ , and  $U3$  measure the (lack of) conformity between an examinee's item response pattern  $\mathbf{X}_n$  and the response pattern summarized over the sample of examinees, by the correct response proportions  $\mathbf{p} = (p_1, \dots, p_j, \dots, p_J)'$ . On the other hand, the  $H^T$  statistic measures the conformity between an examinee's response pattern  $\mathbf{X}_n$  and the response patterns of the remaining  $N - 1$  examinees. Thus, the  $H^T$  outperformed the other three, because it is more sensitive to all individual item response that patterns the data set. In fact,  $H^T$  is sensitive in detecting examinee item response patterns that violate the assumption of non-intersecting item characteristic curves (Sijtsma & Meijer, 1992). This assumption happens to characterize the Rasch model.

It is worth considering why, overall, the non-parametric  $H^T$  outperformed the parametric  $D(\theta)$ , and the non-parametric  $C$ ,  $MCI$ , and  $U3$  outperformed many well-known parametric fit statistics, including  $U$ ,  $W$ ,  $ZU$ ,  $ZW$ ,  $UB$ ,  $ZUB$ ,  $l$ ,  $l_z$ ,  $l_{zm}$ ,  $M$ , and  $M(\text{p-value})$ , and the family of  $ECI$  statistics. With respect to parametric statistical models, general speaking, Efron (1986) observed that the value of a parametric fit statistic (e.g., chi-square) is always biased to be overoptimistic, relative to the true value of that statistic. In terms of the current context, this is because a parametric fit statistic uses the data set  $\mathbf{X} = (X_{nj} | n = 1, \dots, N; j = 1, \dots, J)$  twice, once for the estimation of the model parameters to construct the estimated predictions  $\hat{\mathbf{P}} = (\hat{P}_{nj} | n = 1, \dots, N; j = 1, \dots, J)$ , and once again to measure its fit to the same predictions  $\hat{\mathbf{P}}$  (or similarly, in the case of  $M$  and  $M(\text{p-value})$ , to measure the fit of the data to the estimated item parameters  $(\hat{\delta}_j | j = 1, \dots, J)'$ ). It then follows that parametric person-fit statistics, based on IRT model parameters, do suffer from this dependence between data and parameter estimates (for a related comment, see Smith, 1988, p. 659). In contrast, non-parametric person-fit statistics circumvent this dependence, because they are not based on IRT model parameters. This difference may explain the superior performance of  $H^T$ ,  $C$ ,  $MCI$ , and  $U3$ , over the well-known parametric fit statistics.

This study provided a better consensus about which person-fit statistics, among the very many, are best in detecting aberrant-responding examinees. A future research study may focus on comparing the best performing  $H^T$ , with the more recent non-parametric person-fit methods of Sijtsma & Meijer (in press), and the optimal fit statistics  $\lambda(\mathbf{X})$  and  $T(\mathbf{x})$  (Klauer, 1991, 1995; Levine & Drasgow, 1988). In such a study, one may evaluate the tradeoff between the degree of superior detection ability of  $\lambda(\mathbf{X})$  and  $T(\mathbf{x})$ , relative to the ease of implementation of the non-parametric methods.

In light of the limitations of this study, it is also useful to suggest additional future research. In particular, this study focused on the Rasch model, only considered simulated data, and did not evaluate the effects of sample sizes on the person-fit statistics. The comparison of person-fit statistics, as conducted in this study, can be extended to contexts of other popular IRT models, thanks to the availability of easy-to-use software for IRT analysis (du Toit, 2003). Furthermore, real data sets enjoy the advantage over data sets simulated under artificial conditions, and also, small sample sizes may affect the detection ability of some person-fit statistics more than other statistics. Therefore, future research should compare the 36 (or more) person-fit statistics, by expanding focus to other models of IRT, such as the two- and three-parameter logistic models of IRT, while analyzing real data sets, and examining the effects of sample size. Though the results of this future research may generally show the non-parametric fit statistics to have superior ability to detect aberrant-responding examinees, as the parametric fit statistics tend to be overoptimistic because they use the data twice, as mentioned earlier.

### ACKNOWLEDGMENTS

This study was supported by Spencer Foundation grant SG200100020, and in part by National Science Foundation Grant SES-0242030, George Karabatsos, Principal Investigator. The author thanks the three anonymous referees for detailed comments and suggestions they made on an earlier version of this article.

### REFERENCES

- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81.
- Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika*, *62*, 191–199.
- Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual item response patterns. *Educational and Psychological Measurement*, *45*, 523–534.
- Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriate measures. *Applied Psychological Measurement*, *10*, 167–174.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475–494.
- Donlan, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement*, *28*, 105–113.
- Dragow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59–79.
- Dragow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171–191.

- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- du Toit, M. (Ed.). (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81, 461–470.
- Fisher, W. P., & Wright, B. D. (1997, Guest Eds.). Special issue on the applications of probabilistic conjoint measurement. *International Journal of Educational Research*, 21, 557–664.
- Fowler, H. M. (1954). An application of the Ferguson method of computing item conformity and person conformity. *Journal of Experimental Education*, 22, 237–245.
- Glaser, R. (1949). A methodological analysis of the inconsistency of responses to test items. *Educational and Psychological Measurement*, 9, 721–739.
- Glaser, R. (1950). Multiple operation measurement. *Psychological Review*, 57, 241–253.
- Glaser, R. (1951). The applications of the concepts of multiple operation measurement to the response patterns on psychological tests. *Educational and Psychological Measurement*, 11, 322–382.
- Glaser, R. (1952). The reliability of inconsistency. *Educational and Psychological Measurement*, 12, 60–64.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and Prediction* (pp.66–90). Princeton: Princeton University Press.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–46.
- Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver: Kluwer.
- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4, 105–126.
- Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, 56, 535–547.
- Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97–110). New York: Springer-Verlag.
- Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, 43, 193–206.
- Kogut, J. (1987). *Detecting aberrant item response patterns in the Rasch model* (Research Report 87-3). Enschede: University of Twente, Department of Education.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161–176.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269–290.
- Li, M. F., & Olejnik, S. (1997). The power of Rasch person fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21, 215–231.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18, 311–314.
- Meijer, R. R. (Guest Ed.). (1996a). Person fit research: Theory and applications [Special Issue]. *Applied Measurement In Education*, 9(1).
- Meijer, R. R. (1996b). Person-fit research: An introduction. *Applied Measurement in Education*, 9, 3–8.

- Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology*, 71, 147–160.
- Meijer, R. R., Muijtjens, A. M. M., van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education*, 9, 77–89.
- Meijer, R. R., & Sijtsma, K. (1999). *A review of methods for evaluating the fit of item score patterns on a test* (Research Report 99-01). The Faculty of Educational Science and Technology of the University of Twente.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107–135.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75–106.
- Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 47, 355–366.
- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the  $Iz$  person fit statistic. *Applied Psychological Measurement*, 22, 53–69.
- Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement*, 16, 345–352.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rogers, H. J., & Hattie, J. A. (1987). A Monte-Carlo investigation of several person and item person fit statistics for item response models. *Applied Psychological Measurement*, 11, 47–57.
- Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, 20, 207–219.
- Rudner, L. M., Bracey, G., & Skaggs, G. (1996). The use of a person fit statistic with one high quality achievement test. *Applied Measurement in Education*, 9, 91–109.
- Rudner, L. M., Skagg, G., Bracey, G., & Getson, P. R. (1995). *Use of person-fit statistics in reporting and analyzing National Assessment of Educational Progress Results*. NCES 95–713. Washington, DC: National Center for Educational Statistics.
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tokyo.
- Sherif, M., & Cantril, H. (1945). The psychology of attitudes. *Psychological Review*, 52, 259–319.
- Sherif, M., & Cantril, H. (1946). The psychology of attitudes. *Psychological Review*, 53, 1–24.
- Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden*, 7, 131–145.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, 16, 149–157.
- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191–207.
- Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, 359–370.
- Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement*, 48, 657–667.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- S-PLUS (1995). *S-PLUS documentation*. Seattle: Statistical Sciences.
- SPSS (1998). *SPSS Base 8.0 users guide*. Chicago: SPSS.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95–110.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20, 221–230.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34, 273–286.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, 25, 273–282.

van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y.P. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets & Zeitlinger.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]*. Lisse: Swets & Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13, 267–298.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wilson, M., & Engelhard, G. (Eds.). (2000). *Objective Measurement: Theory Into Practice, (Vol 5)*. Wesport, CN: Ablex.

Wright, B. D. (1980). Afterword. In G. Rasch (Ed.), *Probabilistic models for some intelligence and attainment tests: With foreword and afterword by Benjamin D. Wright*. Chicago: MESA Press.

Wright, B. D., & Panchepakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23–48.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

## APPENDIX A

### Eleven Non-Parametric Person-Fit Statistics

| <i>Number of<br/>response errors</i><br>(Guttman, 1944, 1950)   | <i>Normed G</i><br>(van der Flier, 1977)   | <i>Personal point-<br/>biserial correlation</i><br>(Donlan & Fischer, 1968) |
|---|--|---|
| $G = \sum_{h,e} X_{nh}(1 - X_{ne})$   | $G^* = G / r_n(L - r_n)$   | $r_{pbis} = Corr(\mathbf{X}_n, \mathbf{p})$                                 |
| <i>Caution Index (C)</i><br>(Sato, 1975)  | <i>Modified caution Index (MCI)</i><br>(Harnisch & Linn, 1981)   |   |
| $C = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{p})}{Cov(\mathbf{X}_n^*, \mathbf{p})}$   | $MCI = \frac{Cov(\mathbf{X}_n^*, \mathbf{p}) - Cov(\mathbf{X}, \mathbf{p})}{Cov(\mathbf{X}_n^*, \mathbf{p}) - Cov(\mathbf{X}'_n, \mathbf{p})}$ |   |
| <i>U3 and standardized U3</i> (van der Flier, 1980)   |  |   |
| $U3 = \frac{\ell n(\mathbf{X}_n^*) - \ell n(\mathbf{X}_n)}{\ell n(\mathbf{X}_n^*) - \ell n(\mathbf{X}'_n)}$   | $ZU3 = \frac{U3 - E(U3)}{\sqrt{Var(U3)}}$  |   |
| $H^T$ (Sijtsma, 1986): $H^T = \frac{\sum_{n \neq m} \beta_{nm} - \beta_n \beta_m}{\sum_{n \neq m} \min \{ \beta_n(1 - \beta_m), (1 - \beta_n)\beta_m \}}$ |  |   |

Agreement (*A*), Disagreement (*D*), Dependability (*E*)  
(Kane & Brennan, 1980)

$$A = \sum_{j=1}^J X_{nj} p_j \qquad D = \sum_{j=1}^J p_j - A \qquad E_i = A / \sum_{j=1}^J p_j$$

**Notation**

- J* number of items,  $\{j = 1, \dots, J\}$ .
- N* number of persons,  $\{n = 1, \dots, N\}$ .
- $X_{nj}$  person *n*'s scored response to test item *j*,  
where  $X_{nj} = 1$  is a correct response, and  $X_{nj} = 0$  is incorrect.
- $X_{nh}$  examinee *n*'s response to hardest item in item pair  $\{h, e\}$ .
- $X_{ne}$  examinee *n*'s response to easiest item in item pair  $\{h, e\}$ .
- $r_n$  number of correct responses for examinee *n* over the *J* test items.
- $p_j$  proportion correct by the *N* examinees on item *j*.
- p** item vector of proportion correct,  $\mathbf{p} = (p_1, \dots, p_j, \dots, p_J)'$ .
- $\mathbf{X}_n$  examinee *n*'s (scored) item response vector,  
with  $\mathbf{X}_n = (X_{n1}, \dots, X_{nj}, \dots, X_{nJ})'$ .
- $\ell n(\mathbf{X}_n)$  log-ratio correct for the *N* examinees, over the responses of  
examinee *n*, with  $\ln(\mathbf{X}_n) = \sum_{j=1}^J X_{nj} \ell n\left(\frac{p_j}{1-p_j}\right)$ .
- $\mathbf{X}_n^*$  examinee *n*'s response vector containing correct responses  
only for the easiest *r* items.
- $\ell n(\mathbf{X}_n^*)$  log-ratio correct for the *N* examinees, over these *r* items,  
with  $\ln(\mathbf{X}_n^*) = \sum_{j=1}^r \ell n\left(\frac{p_j}{1-p_j}\right)$ .
- $\mathbf{X}'_n$  examinee *n*'s response vector containing correct responses  
only for the most difficult  $J - r_n$  items.
- $\ell n(\mathbf{X}'_n)$  log-ratio correct for the *N* examinees, over these *r* items,  
with  $\ln(\mathbf{X}') = \prod_{j=J-r_n+1}^J \ell n\left(\frac{p_j}{1-p_j}\right)$ .
- $\beta_n$  proportion correct for examinee *n* over the *J* test items,  
 $\beta_n = J^{-1} r_n$ .
- $\beta_{nm}$  the covariance between the scored test responses of  
examinee *n* with examinee *m*, with  $\beta_{nm} = J^{-1} \sum_{j=1}^J X_{nj} X_{mj}$ .

APPENDIX B

Twenty-five Parametric Person-fit Statistics

*Mean Square Person-Fit Statistics*

*Unweighted mean square*  
(Wright & Stone, 1979)

$$U = J^{-1} \sum_{j=1}^J \nu_{nj}^{-1} (X_{nj} - P_{nj1})^2$$

*Weighted mean square*  
(Wright, 1980)

$$W = \left( \sum_{j=1}^J \nu_{nj} \right)^{-1} \sum_{j=1}^J (X_{nj} - P_{nj1})^2$$

*Z standardized mean square*  
(Wright, 1980)

$$Z(MS) = (MS^{.45} - 1)3 / \sqrt{\nu_{MS}} + (\sqrt{\nu_{MS}}/3)$$

*Log-standardized mean square*  
(Wright & Stone, 1979)

$$\ell n(MS) = (\ell n(MS) + MS + 1) (\sqrt{(L - 1)/8})$$

... where MS can either refer to U or W.

*Extended Caution Indices (ECI) (Tatsuoka, 1984)*

$$ECI1 = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{p})}{Cov(\mathbf{P}_n, \mathbf{p})} \quad ECI2 = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{G})}{Cov(\mathbf{P}_n, \mathbf{G})} \quad ECI3 = 1 - \frac{Corr(\mathbf{X}_n, \mathbf{G})}{Corr(\mathbf{P}_n, \mathbf{G})}$$

$$ECI4 = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{P}_n)}{Cov(\mathbf{P}_n, \mathbf{G})} \quad ECI5 = 1 - \frac{Corr(\mathbf{X}_n, \mathbf{P}_n)}{Corr(\mathbf{P}_n, \mathbf{G})} \quad ECI6 = 1 - \frac{Corr(\mathbf{X}_n, \mathbf{P}_n)}{\text{var}(\mathbf{P}_n)}$$

*Standardized ECI :* 
$$ECIb_z = ECIb - \frac{E(ECIb)}{SE(ECIb)}$$

... where b equals 1, 2, 4, or 6.

*Likelihood l*  
(Levine & Rubin, 1979)

$$l = \sum_{j=1}^J [X_{nj}(\ell n P_{nj1}) + (1 - X_{nj})(\ell n P_{nj0})]$$

*Standardized l*  
(Drasgow et al., 1985)

$$l_z = [l - E(l)] / \sqrt{\nu(l)}$$

*Molenaar's M*  
(Molenaar & Hoijtink, 1990)

$$M = - \sum_{j=1}^J X_{nj} \delta_j$$

*p-value of M*  
(Bedrick, 1997)

$$M(\text{p-value}) = [\phi(Z_M) - (Z_M^2 - 1)\gamma]^{-6}$$

*Item-Grouping Person-Fit Statistics*

$D(\theta)$  (Trabin & Weiss, 1983): 
$$D(\theta) = \sum_{s=1}^S \left( J_s^{-1} \sum_{j \in s} (X_{nj} - P_{nj1}) \right)$$

$l_{zm}$  (Drasgow et al., 1991): 
$$l_{zm} = \frac{\sum_{s=1}^S l_s - E(l_s)}{\sum_{s=1}^S \sqrt{v(l_s)}}$$

$UB$  (Smith, 1986): 
$$UB = (S - 1)^{-1} \frac{\sum_{s \in S} (X_{nj} - P_{nj1})^2}{\sum_{s \in S} P_{nj1} P_{nj0}}$$

$UB$  Z standardized statistic: 
$$Z(UB) = (UB^{1/3} - 1)(3/\sqrt{v_{UB}}) + (\sqrt{v_{UB}}/3)$$

$UB$  log-standardized statistic: 
$$\ell n(UB) = \left( \sqrt{\frac{S-1}{8}} \right) (\ell n UB + UB + 1)$$

**Notation**

- $J$  number of items,  $\{j = 1, \dots, J\}$ .
- $N$  number of persons,  $\{n = 1, \dots, N\}$ .
- $X_{nj}$  examinee  $n$ 's scored response to test item  $j$ , where  $X_{nj} = 1$  is a correct response, and  $X_{nj} = 0$  is incorrect.
- $P_{nj1}$  probability of a correct ( $X_{nj} = 1$ ) response, predicted by an IRT model.
- $P_{nj0}$  probability of an incorrect ( $X_{nj} = 0$ ) response, predicted by an IRT model, with  $P_{nj0} = (1 - P_{nj1})$ .
- $\mathbf{P}_n$  examinee  $n$ 's vector of  $P_{nj1}$  values across  $J$  items.
- $\mathbf{p}$  item vector of proportion correct,  $\mathbf{p} = (p_1, \dots, p_j, \dots, p_J)'$ .
- $\mathbf{X}_n$  examinee  $n$ 's (scored) item response vector, with  $\mathbf{X}_n = (X_{n1}, \dots, X_{nj}, \dots, X_{nJ})'$ .
- $\mathbf{G}$  vector containing mean  $P_{nj1}$  for each item, over the  $N$  examinees. Given by  $\mathbf{G} = (G_1, \dots, G_j, \dots, G_J)'$ , with  $G_j = N^{-1} \sum_{n=1}^N P_{nj}$ .
- $S$  number of non-overlapping item subsets, where  $s = 1, \dots, S$ .
- $L_s$  number of items in subset  $s$ .
- $l_s$  examinee  $n$ 's  $l$  value within item subset  $s$ .
- $v(l)$  examinee  $n$ 's  $v(l)$  value within item subset  $s$ .
- $\delta_j$  difficulty parameter of item  $j$ .
- $Z_M$  z-value corresponding to the value of  $M$ .
- $v_{nj}$  variance, where  $v_{nj} = P_{nj1} P_{nj0}$ .

$\nu_U$  variance, with  $\nu_U = J^{-1} \left[ \left( \sum_{j=1}^J \nu_{nj}^{-1} \right) - 4L \right]$ .

$\nu_W$  variance, with  $\nu_W = \nu_{nj}^{-1} \left( \sum_{j=1}^J \nu_{nj} - 4 \sum_{j=1}^J \nu_{nj}^2 \right)$ .

$\nu(l)$  variance, with  $\nu(l) = \sum_{j=1}^J (\nu_{nj}) \left( \ell n \frac{P_{nj1}}{P_{nj0}} \right)^2$ .

$\nu_{UB}$  variance, with  $\nu_{UB} = 2 / (S-1)$ .

$E(l)$  expectation of  $l$ , with  $E(l) = \sum_{j=1}^J (P_{nj1} \ell n[P_{nj1}]) + (P_{nj0} \ell n[P_{nj0}])$ .

Do Not Copy