Reading List (available online):
Meijer, R. R. (1996). Person-fit research: an introduction. *Applied Measurement in Education*, 9(1), 3-8.

Aberrant Response Patterns and Person-Fit Statistics

The purpose of educational assessment is to provide information in order to make sound decisions. Teachers use test scores to identify students, assign grades, and modify instruction. Administrators use assessment results to evaluate teachers, schools, and instructional methods. The federal government now uses assessment results to identify schools and school districts in need of improvement (LeTendre, 2000). As long as the information provided by these assessments is accurate, these decisions can lead to educational improvements.

Inaccurate assessment results, in contrast, can lead to poor decisions. In licensure decisions, inaccurate test scores could result in unqualified individuals obtaining licenses. In making admissions decisions, erroneous test results could lead to deserving individuals being denied entry into an educational program. In any setting, inaccurate test results undermine efforts to evaluate and improve educational programs. It is therefore important to identify factors that lead to inaccurate test results.

Traditionally, research into measurement error and inaccurate assessment results has focused on characteristics of the tests (Bracey, 1992). Reliability studies have identified test administration characteristics (unclear directions, scoring errors) and item characteristics (misworded, ambiguous, multidimensional items) that lead to an examinee's test score misrepresenting the amount of ability the examinee possesses. A great deal of effort is placed into identifying and correcting items that do not "fit" and could lead to inaccurate test results.

Relatively less attention has been placed on identifying examinee characteristics that lead to inaccurate results (examinees who do not "fit"). Consider the following eight types of examinees whose unusual patterns of responses (aberrant response patterns) could yield test scores that would not provide an accurate estimate of their underlying ability level: (Meijer, 1996; Wright & Stone, 1979)

Aberrant response patterns that artificially *increase* ability estimates:

1. Cheaters: Examinees who unexpectedly answer many difficult items correctly.
2. Lucky Guessers: Examinees who unexpectedly provide a few correct responses to difficult items

Aberrant response patterns that artificially *decrease* ability estimates:

3. Fumblers: Examinees who are confused at the beginning of the test (unexpected string of incorrect answers to the first several items.)
4. Plodders: Examinees who never get to the end of the test (unexpected string of incorrect answers to the final several items)
5. Language Deficiency: Examinees whose language deficiencies artificially decrease their ability estimates (unexpected incorrect answers to easy items).

Aberrant response patterns that may artificially *increase or decrease* ability estimates:

6. Random Guessers or Constant Responders: Examinees who respond with no thought to the item content (unexpected correct and incorrect answers).
7. Creatives: Examinees who interpret items in a creative way (unexpected correct and incorrect answers).
8. Palm Beachers: Examinees who misalign their answer sheets (unexpected correct and incorrect answers).

Considering the consequences that could result from the inaccurate assessment of examinees, it is important to identify examinees with unusual patterns of test results (aberrant responders). Attempts to systematically identify these aberrant responders have led to the development of over 40 person-fit indices. These indices measure the extent to which an examinee's item responses deviate from an expected pattern of responses, and distinguish these examinees from examinees with normal response patterns. In this project, six classes of person-fit indices will be discussed:

Nonparametric Indices (not based on IRT): (1) Guttman-based, (2) Agreement-based, (3) Correlation-based indices
Parametric Indices (based on IRT models): (4) Deviation-based, (5) Correlation-based, (6) Likelihood-based indices

After a brief description of these three classes of person-fit indices, several simulations will be conducted to evaluate the usefulness of each index in correctly identifying aberrant responders.

Nonparametric Person-Fit Indices (listed in Appendix A)

- Guttman-based Indices

A person-fit statistic measures "the degree of reasonableness of an examinee's answers to a set of test items." (Karabatsos, 2003). Thus, if we know examinee i earned a test score of x_i on a test with n items, it is reasonable to expect the examinee answered the easiest x items correctly and the most difficult $n-x$ items incorrectly. Thus, if we are able to order test items by increasing difficulty, we could easily measure the reasonableness of any examinee's response pattern. Nonparametric person-fit indices estimate the difficulty of each of the items from the set of N examinees' scored responses to the n test items.

The first nonparametric person-fit index developed in 1944 was the G statistic (Guttman, 1944). Consider Table 1, which displays the responses from nine examinees on a 4-item test. The items have been sorted by their difficulty estimates (calculated as the proportion of correct responses from the sample of examinees).

Table 1

	Items						
Examinee	1 (Difficult)	2	3	4 (Easiest)	Response Pairs	**G**	**G***
A	1	1	1	1	(11) (11) (11) (11) (11) (11)	0 + 0 + 0 + 0 + 0 + 0 = 0	0.00
B	0	1	1	1	(01) (01) (01) (11) (11) (11)	0 + 0 + 0 + 0 + 0 + 0 = 0	0.00
C	0	0	1	1	(00) (01) (01) (01) (01) (11)	0 + 0 + 0 + 0 + 0 + 0 = 0	0.00
D	0	0	0	1	(00) (00) (01) (00) (01) (01)	0 + 0 + 0 + 0 + 0 + 0 = 0	0.00
E	0	0	0	0	(00) (00) (00) (00) (00) (00)	0 + 0 + 0 + 0 + 0 + 0 = 0	0.00
F*	0	0	1	0	(00) (01) (00) (01) (00) (10)	0 + 0 + 0 + 0 + 0 + 1 = 1	0.33
G*	0	1	0	1	(01) (00) (01) (10) (11) (01)	0 + 0 + 0 + 1 + 0 + 0 = 1	0.33
H*	1	0	0	1	(10) (10) (11) (00) (01) (01)	1 + 1 + 0 + 0 + 0 + 0 = 2	0.50
I*	1	1	1	0	(11) (11) (10) (11) (10) (10)	0 + 0 + 1 + 0 + 1 + 1 = 3	1.00
Difficulty	p = 0.33	0.44	0.55	0.67	Item Pairs: (12) (13) (14) (23) (24) (34)	$G = \sum_{h,e} u_{nh}(1-u_{ne})$	$G^* = \frac{G}{x_i(n-x_i)}$
* denotes aberrant response pattern							

An examinee's response pattern is said to be a "Guttman Perfect Pattern" if the examinee responds correctly to only the easiest items. Examinees A-E display this pattern. If we examine each possible pair of item responses for an examinee, the G statistic counts the number of item response pairs that deviate from the Guttman Perfect Pattern. For example, Examinee H answered the only the easiest and most difficult items correctly. Since the examinee answered the most difficult item correctly, we may reasonable assume the examinee should have also answered items 2 and 3 correctly. Thus, the examinee deviates from the Guttman Perfect Pattern by 2 units (G=2). Examinees yielding a low value of G have response patterns close to what we would expect. High values of G indicate a greater deviation from the Guttman Perfect Pattern. Since the value of G tends to increase as the test length increases, the statistic G* was developed. G* normalizes the range of G to [0,1] by taking the test length into account.

If we are willing to assume that test items all differ in difficulty and the item difficulties are calculated from a representative sample, the G statistic appears to be an easily-interpreted person-fit index. The main drawbacks to using G or G* as person-fit indices are (1) no agreed-upon critical value exists to identify aberrant responders and (2) G* does not take into account the magnitude of each Guttman error. (Karabatsos, 2000). Despite those drawbacks and its simplicity, Meijer (1994) states that G* is a powerful fit statistic that performs as well or better than fit statistics which have more complex formulas.

Two additional person-fit indices that can be calculated directly from the G statistic are the *Norm Conformity Index* (Tatsuoka, 1983) and the *Individual Consistency Index* (Karabatsos, 2003). See Appendix A for information about these indices.

- Agreement-based Indices

A second subset of nonparametric person-fit indices developed by Kane and Brennan (1980) are based on A, the Agreement Index. The Agreement Index measures the agreement between an examinee's response to item j, u_j, and the item difficulty, p_j (the proportion of examinees answering the item correctly). Table 2 displays these agreement-based indices calculated on our 4-item test:

Table 2

	Items						
Examinee	1 (Most Difficult)	2	3	4 (Easiest)	**A** (Agreement)	**D** (Disagreement)	**E** (Dependability)
A	1	1	1	1	.33 + .44 + .55 + .67 = **2.00**	2.00 – 2.00 = **0.00**	1.00
B	0	1	1	1	.44 + .55 + .67 = **1.67**	1.67 – 1.67 = **0.00**	1.00
C	0	0	1	1	.55 + .67 = **1.23**	1.23 – 1.23 = **0.00**	1.00
D	0	0	0	1	**0.67**	0.67 – 0.67 = **0.00**	1.00
E	0	0	0	0	**0.00**	0.00 – 0.00 = **0.00**	--
F*	0	0	1	0	**0.55**	0.67 – 0.55 = **0.12**	0.82
G*	0	1	0	1	.44 + .67 = **1.11**	1.23 – 1.11 = **0.12**	0.90
H*	1	0	0	1	.33 + .67 = **1.00**	1.23 – 1.00 = **0.23**	0.81
I*	1	1	1	0	.33 + .44 + .55 = **1.33**	1.67 – 1.33 = **0.33**	0.80
Difficulty	p = 0.33	0.44	0.55	0.67	$A = \sum_{j=1}^{J} u_j p_j$	$D = A_{\max} - A$	$E = \frac{A}{A_{\max}}$
* denotes aberrant response pattern							

To calculate an examinee's Agreement Index, we simply sum the p-values for every item the examinee answers correctly. We then compare the value A_i for an examinee to $A_{\max}$, the highest possible value of A_i for an examinee with total score x_i. The difference between A_i and $A_{\max}$ for an examinee, called the Disagreement Index (D), measures the amount of aberrance in an examinee's response pattern (the disagreement between an examinee's response to the items and the relative difficulty of each item).

As an example, consider examinees C and H, each answering two items correctly. We would expect an examinee with a total correct score of 2 to answer the two easiest items correctly. Therefore, $A_{\max}$ would be the sum of the two highest p-values ($0.67 + 0.55 = 1.23$, in this example). When we calculate A_i for any examinee with a total score of 2, any deviation from $A_i = 1.23$ would indicate that the examinee exhibited some aberrance in his response pattern (the examinee answered relatively easy items incorrectly and relatively difficult items correctly.)

Summing the p-values for the items answered correctly, we find $A_C = 1.23$ and $A_H = 1.00$. Thus, examinee C's Disagreement Index is $D_C = 1.23 - 1.23 = 0.00$, which corresponds to the expected response pattern. For examinee H, $D_H = 1.23 - 1.00 = 0.23$, which tells us examinee H's response pattern deviates from what we would expect.

The maximum value of D will increase as the length of the test increases. To normalize the range to [0, 1], the Dependability Index (E_i) calculates the ratio of an examinee's observed Agreement Index with $A_{\max}$. A value of $E_i = 1.00$ represents examinees whose responses fit the Guttman Perfect Pattern. As with most person-fit indices, these statistics are descriptive -- no widely accepted critical value has been established to identify examinees with significantly aberrant response patterns.

Comparing the Disagreement Index from Table 2 to either G or G* from Table 1, one can see that the examinees are given the same aberrant response rank ordering. This is generally true of all person-fit indices (both parametric and nonparametric). In fact, Harnisch and Linn (1981) examined the correlations among eight nonparametric person-fit indices and found correlations ranging from 0.65 to 0.90.

• Correlation/Covariance-based Indices

A third group of nonparametric (group-based) person-fit indices are based on the relationship between an examinee's vector of item responses, $\mathbf{X}_i$ and the vector $\mathbf{P}$ containing the proportion of correct responses obtained by the N examinees on each item.

Table 3

	Items						
Examinee	1 (Difficult)	2	3	4 (Easy)	r_{pbis}	MCI	H^T
A	1	1	1	1	--	0	---
B	0	1	1	1	0.765	0	.273
C	0	0	1	1	0.890	0	.250
D	0	0	0	1	0.788	0	.273
E	0	0	0	0	--	0	---
F*	0	0	1	0	0.240	0.36*	-.091*
G*	0	1	0	1	0.455	0.50*	.000*
H*	1	0	0	1	0.020	0.50*	-.250*
I*	1	1	1	0	-0.788	0.82*	-.818*
Difficulty	p = 0.33	0.44	0.55	0.67	$r_{pbis} = Corr(\mathbf{X}_n, \mathbf{p})$	$MCI = \frac{Cov(\mathbf{X}_n^*, \mathbf{p}) - Cov(\mathbf{X}, \mathbf{p})}{Cov(\mathbf{X}_n^*, \mathbf{p}) - Cov(\mathbf{X}_n^{'}, \mathbf{p})}$	*See Appendix A*
* denotes aberrant response pattern							

Donlan and Fischer (1968) created the personal point-biserial correlation, r_{pbis}, and the personal biserial correlation, r_{bis}, as person-fit indices. The personal point-biserial is calculated as the correlation across all items between an examinee's item scores and the item difficulty p-values. The biserial measures that same correlation under the assumption that a continuous distributed variable underlies an examinee's item responses (Meijer & Sijtsma, 2001). Values of r_{pbis} for the nine examinees on our our simple 4-item test can be found in Table 3.

Sato (1975) proposed the Caution Index (C) to identify aberrant responders. The Caution Index, like other nonparametric person-fit indices, measures the degree to which an examinee's item responses deviate from the Guttman Perfect Pattern. To calculate C, one first measures the covariance between the vector of an examinee's item responses with the vector of item difficulties. The same covariance is then calculated using only the x easiest items (where x represents the examinees total correct score). The Caution Index is calculated as one minus the ratio of these covariances.

Harnisch and Linn (1981) developed the Modified Caution Index (MCI) which normalizes Sato's Caution Index. MCI represents the ratio between the covariance of the vector of Guttman Perfect Model responses and the covariance of a response vector that is perfectly opposite the Guttman pattern for the most difficult x items. Table 3 displays MCI values for the nine examinees. Karabatsos (2003) recommends using a critical value of MCI > 0.26 to identify aberrant-responders. Table 3 shows that all the aberrant responders in the 4-item test had MCI values above this critical value.

Van der Flier (1980) proposed two transformations of the Caution Index. The U3 statistic has the same formula as MCI, but replaces the covariance with a ratio of logarithms. Subtracting the mean of U3 and dividing by its standard error, van der Flier developed the standardized U3 statistic (ZU3), which can be interpreted as a standard normal variable.

Another correlation-based person-fit index developed by Sijtsma (1986) is the H^T statistic. This person-fit index measures the agreement between an examinee's vector of item responses and the response vectors of the remaining examinees. According to a simulation conducted by Karabatsos (2003), "… the critical values of $H^T \leq 0.22$ best identify aberrant-responding examinees." Applying this critical value to the simple 4-item test displayed in Table 3, examinees F, G, H, and I would all be correctly identified as aberrant responders.

• Research on Nonparametric Person-fit Indices

A number of simulation and empirical studies (Harnisch & Linn, 1981; Rudner, 1983; Meijer, 1996; Karabatsos, 2003) have been conducted to compare nonparametric person-fit studies. With the exception of A_i, Harnisch and Linn found that correlations among nonparametric person-fit indices ranged from .65 to .90. Rudner and Karabatsos obtained similar high positive correlations among the indices.

Parametric (IRT-based) Person-Fit Indices (Appendix B)

In item response theory, the probability of correctly answering item i is a function of an examinee's latent ability, θ, and the characteristics of the item. For the three-parameter logistic model, the probability of examinee a correctly answering item i can be expressed as:

$$P_{ia} = P_{ia}(\theta_a) = P_{ia}(u_i = 1 \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_a - b_i)}}$$, where

$a_i =$ the item discrimination parameter (proportional to the slope of the function at its inflection point)

$b_i =$ the item difficulty (location) parameter

$c_i =$ the guessing (lower asymptote) parameter.

$\theta_a =$ the latent ability of examinee a

$u_i =$ the examinee's scored response to the item

Due to the assumption of local independence, the probability of obtaining a string of n responses from an examinee with ability θ_a is equal to the product of the probabilities of the individual item responses:

$$P(u_1, u_2, ..., u_n \mid \theta) = \prod_{i=1}^{n} P_i^{u_i} (1 - P_i)^{(1-u_i)}$$

Parametric person-fit indices measure the extent to which an examinee's observed pattern of responses deviates from the response pattern expected from an examinee with ability θ_a.

• Deviation-based Indices

The discrepancy between an examinee's scored response to item i and the examinee's expected probability of a correct response to that item is given by

$$Y_{ia} = (u_{ia} - P_{ia})$$

A large value of Y_{ia} would occur if an examinee gives an aberrant response to item i (a low ability examinee answers a difficult item correctly or a high-ability examinee answers an easy item incorrectly). A low value of Y_{ia} occurs when an examinee answers relatively easy items correctly and relatively difficult items incorrectly. Since each item is an independent Bernoulli trial, the residuals may be standardized as follows:

$$z_{ia} = \frac{u_{ia} - E[u_{ia}]}{\sqrt{Var[u_{ia}]}} = \frac{u_{ia} - P_{ia}}{\sqrt{P_{ia}(1 - P_{ia})}}$$

Squaring both sides yields a statistic z_{ia}^2 that may be assumed to have an approximate χ^2 distribution (Lamprianou et al, 2001).

Wright and Stone (1979) proposed a person-fit index that measures the average value of z_{ia}^2 over the n test items. This index, called the *Outfit Mean Square* (short for "outlier sensitive mean square residual goodness-of-fit statistic") is calculated

$$U = OMS = \sum_{i=1}^{n} \frac{z_{ia}^2}{n}$$

A close examination of this formula shows the size of each squared standardized residual in the OMS index depends on the variance of the individual item scores, $P_{ia}(1-P_{ia})$. These variances, in turn, depend on the discrepancy between an examinee's ability and the item difficulty. Therefore, items with extreme difficulty estimates will contribute more to the OMS than items with difficulty estimates near the examinee's ability level. Wright (1980) proposed another residual-based index, called the *Infit Mean Square* (IMS), that is more sensitive to items with moderate difficulty estimates. IMS, short for "information weighted mean square residual goodness-of-fit statistic" (Lamprianou et al, 2001), takes the variance-weighted average of the squared residuals

$$W = IMS = \frac{\sum (u_{ia} - P_{ia})^2}{\sum P_{ia}(1-P_{ia})}$$

Several studies have attempted to discover a critical IMS/OMS value for identifying aberrant responders. According to Karabatsos (2000), "Convention suggests that 1.3 defines the minimum critical value for *OMS* or *IMS* for classifying a person or item as misfitting the model" (Karabatsos, 2000, p.155). Furthermore, Wright (1995) suggested that a lower limit for both IMS and OMS should be set at 0.8 to identify response patterns that fit the model too well.

To further aid in interpretation, the mean-square indices can be transformed into the standard normal distribution through a cube-root or logarithmic transformation (Wright, 1980). These transformations, listed in Appendix B, lead to a critical value of approximately 2.0 if the Type I error rate is $\alpha = .05$ (Karabatsos, 2000).

In general, these parametric deviation-based indices share much in common with the nonparametric deviation-based indices. The nonparametric indices described earlier increase as an examinee's response pattern deviates from the Guttman Perfect Model. Likewise, the parametric indices increase as an examinee's response pattern deviates from the pattern expected from a probabilistic model. The values of these indices are at a minimum when only the examinees with the highest ability levels answer the most difficult items correctly.

As a simple computational example, Table 4 displays the responses for six examinees on a 5-item test. Each item on the test was assumed to have fixed discrimination $(a = 0.8)$ and guessing $(c = 0.2)$ parameters. The difficulties of the items range from $b_1 = -2$ to $b_5 = 2$. Maximum likelihood estimates of examinee ability were calculated from the item responses. Notice that the first three examinees have the same relatively low level of ability $(\theta = -0.515)$, while the last three examinees have $\theta = +0.515$. The probability of a correct response was calculated from the three-parameter logistic model.

Table 4

		$P(\theta = -.515)$	Examinee 1	Examinee 2	Examinee 3	$P(\theta = 0.515)$	Examinee 4	Examinee 5	Examinee 6
Item #1	b = -2	0.906	1	1	0	0.975	1	1	0
Item #2	b = -1	0.727	1	0	0	0.910	1	0	0
Item #3	b = 0	0.465	0	0	0	0.735	1	1	1
Item #4	b = +1	0.290	0	1	1	0.473	0	0	1
Item #5	b = +2	0.225	0	0	1	0.294	0	1	1
		OMS	0.410*	1.275	3.818*		0.360*	2.750*	10.483*
		Z(OMS)	-2.64	0.99	6.08		-1.88	2.95	8.48
		IMS	0.476*	1.434*	2.926*		0.513*	2.136*	3.460*
		Z(IMS)	-2.24	1.45	4.67		-1.26	2.16	3.74

** MS value is above or below the critical value suggested by Wright (1995)*

Examinees 1 and 4 obtained MS values that fall below the 0.8 threshold suggested by Wright, indicating that their responses fit the model too well. This is as it should be, since both examinees responses fit the Guttman Perfect Model. Examinee 2's response pattern was not aberrant enough to yield an OMS or IMS value above 1.3. Examinees 3, 5, and 6 were identified as aberrant responders by both the IMS and OMS statistics.

Notice the discrepancies between the OMS and IMS values for the final three examinees (especially examinee 6). The value of OMS has been inflated by the extreme probabilities of a correct response to items #1 and #2. For these examinees, the OMS is a better measure of aberrance. The IMS is a better index to use for examinees 1-3 due to the moderate probabilities for correct responses.

While OMS and IMS are commonly reported person-fit indices (IMS is part of the standard BICAL output), they are not without problems. In a study of several person-fit indices, Karabatsos discovered that in addition to a the level of aberrance in an examinee's response pattern, the value of OMS or IMS depends on (1) how well the data fit the model, (2) sample size, (3) test length, (4) the distribution of examinee ability, (5) the distribution of item difficulty parameters, (6) the amount of lucky guessing (Karabatsos, 2000, p.164). These factors make it difficult to compare mean-square values across tests or different groups of examinees.

- Correlation/Covariance-based Indices

Tatsuoka (1984) developed six person-fit statistics that are parametric extensions of the nonparametric Caution Index. Recall the Caution Index measures the degree to which an examinee's item response vector deviates from the Guttman Perfect Pattern:

$$C = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{p})}{Cov(\mathbf{X}_n^*, \mathbf{p})}$$

The first *Extended Caution Index* (ECI_1) simply modifies the denominator of the Caution Index to norm against the covariance between the expected probabilities of correct responses under the IRT model and the vector of item p-values.

$$ECI1 = 1 - \frac{Cov(\mathbf{X}_n, \mathbf{p})}{Cov(\mathbf{P}_n, \mathbf{p})}$$

Tatsuoka proposed four additional extended caution indices all based on the same logic. They all measure the ratio between two covariances or correlations involving (1) the examinee's response vector $(\mathbf{X}_n)$, (2) the vector of item p-values $(\mathbf{P}_n)$, or (3) a vector containing the average probability of a correct response across all examinees $(\mathbf{G})$.

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

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

The sixth extended caution index developed by Tatsuoka norms against the variance of the probabilities of correct responses for an examinee.

$$ECI6 = 1 - \frac{Corr(\mathbf{X}_n, \mathbf{P}_n)}{\mathrm{var}(\mathbf{P}_n)}$$

All six Extended Caution Indices can be transformed to the standard normal by subtracting the corresponding expected value and dividing by the standard error.

$$ECIb_z = \frac{ECI_b - E(ECI_b)}{SE(ECI_b)}$$

These standardized extended caution indices can then be interpreted against standard critical values. The expected values and standard errors corresponding with each ECI can be found in Tatsuoka, 1984.\

Table 4 shows the values of ECI2 and ECI6 for six examinees on a five-item exam (see Table 3 for the raw data). As expected, the values of both indices increase as the examinee's response pattern becomes more aberrant.

Table 4

	Examinee 1	Examinee 2	Examinee 3		Examinee 4	Examinee 5	Examinee 6
ECI2	-0.90	0.47	2.90		-0.92	1.20	2.60
ECI6	-0.93	0.34	2.60		-0.92	1.20	2.60

- Likelihood-based Indices

Given an examinee's responses to a set of test items with known item parameters, the ability level of the examinee (θ_a) can be estimated by maximizing the log of the likelihood function

$$\ell_0 = \ln\left[L(\mathbf{u} \mid \theta; \mathbf{a}, \mathbf{b}, \mathbf{c})\right] = \ln\left[\prod_{i=1}^{n} P_i^{u_i} Q_i^{(1-u_i)}\right] = \sum_{i=1}^{n}\left[u_i \ln(P_i) + (1-u_i)\ln(Q_i)\right]$$

The value of θ_a that maximizes this function represents the examinee's most likely level of ability given the examinee's item responses and the item parameters. Examinees whose item responses conform to the IRT model produce likelihood functions with relatively high maximum values. Examinees whose responses deviate from what is predicted by the IRT model produce low maximum values of the likelihood function (Davey, et. al, 2003, p.6). Thus, the relative magnitude of ℓ_0 can be used to identify examinees whose responses are aberrant.

Levine & Rubin (1979) first proposed using the log-likelihood function ℓ_0 as a person-fit index. Recognizing the fact that the shape of the likelihood function changes across examinees, tests, and the chosen ability scale, Drasgow, Levine, and Williams (1985) proposed a standardized form of the ℓ_0 statistic that is approximately asymptotically standard normal

$$\ell_z = \frac{\ell_0 - E[\ell_0]}{(\mathrm{var}[\ell_0])^{1/2}} \quad \text{, where} \quad E[\ell_0] = \sum_{i=1}^{n}\left[P_i \ln(P_i) + (1-P_i)\ln(1-P_i)\right]$$

$$\text{and} \quad \mathrm{var}(\ell_0) = \sum_{i=1}^{n} P_i(1-P_i)\left[\ln\frac{P_i}{1-P_i}\right]^2$$

Examinee response patterns that yield large negative values of ℓ_z (below –2.0) are identified as being aberrant (they do not conform to the pattern predicted by the IRT model). Large positive values of ℓ_z (above +2.0) indicate an examinee's response pattern may conform too well to the IRT model.

The ℓ_z statistic has been demonstrated to be superior to many other person-fit indices. In a discussion of several person-fit indices, Schmitt states, "Although researchers call for the development of even more powerful person-fit indices, there is some indication the ℓ_z index is the most accurate available for the two- and three-parameter models" (Schmitt et. al, 1999, p.42). Li & Olejnik go one step further, concluding, "Practitioners need no longer be confused by the large number of possible person-fit indexes available to detect non-fitting examinees. The ℓ_z index will provide as reliable and accurate identification of unusual responding as other person fit statistics" (Li & Olejnik, 1997, p.229).

Though supported by many, the ℓ_z index is not without limitations. In particular, the index is highly dependent upon the length of the test. As the test length gets shorter, the accuracy of detecting aberrance decreases (Schmitt et al, 1999). Moelnaar & Hoijtink also argue that the distribution of the index is only standard normal if true θ_a values are used. They found that when estimated ability levels are used, the variance of ℓ_0 becomes smaller than expected (Molenaar & Hoijtink, 1990).

For illustration purposes, Table 5 shows the values of the ℓ_z index calculated for six examinees on the five-item test (using the Rasch model). First, maximum likelihood estimates of examinee ability were calculated. Plugging those ability estimates into the log-likelihood function yielded the ℓ_0 values. Finally, the standardized ℓ_z values were calculated. Based on the critical value of -2.0, examinees 2, 3, 5, and 6 would have been identified as aberrant responders on this exam.

Table 5

			Examinee 1	Examinee 2	Examinee 3		Examinee 4	Examinee 5	Examinee 6
Item #1	b = -2		1	1	0		1	1	0
Item #2	b = -1		1	0	0		1	0	0
Item #3	b = 0		0	0	0		1	1	1
Item #4	b = +1		0	1	1		0	0	1
Item #5	b = +2		0	0	1		0	1	1
		MLE of _	-0.515	-0.515	-0.515		0.515	0.515	0.515
		ℓ_0	-0.876	-4.276	-11.076		-0.876	-4.276	-11.076
		$E[\ell_0]$	-1.813	-1.813	-1.813		-1.813	-1.813	-1.813
		$\mathrm{var}(\ell_0)$	1.424	1.424	1.424		1.424	1.424	1.424
		ℓ_z	0.785	-2.064*	-7.762*		0.785	-2.064*	-7.762*

** Flagged as significant*

Molenaar & Hoijtink (1990, p.96) showed that under the Rasch model, ℓ_0 can be rewritten as the sum of

$$\ell_0 = \left[-\sum \ln(1+e^{\theta-b}) + \theta\sum u_i\right] + \left[-\sum b_i u_i\right] = [d_0] + [M]$$

After demonstrating d_0 is independent of an examinee's response vector (and M is dependent), Molenaar & Hoijtink recommend using the simple M rather than ℓ_0 as a person-fit index. The distribution of M is unknown, but an approximate p-value of M can be calculated from

$$M(p-value) = \frac{1}{\left[\Phi(Z_M) - \left(X_M^2 - 1\right)\gamma\right]}.$$

- Other Parametric Indices

Appendix B lists several person-fit indices that measure the fit between predefined subsets of items on a particular test. The statistic $D(\theta)$ is similar to the OMS index in that it calculates the sum of the average response deviations within each subset of items. The UB index is equivalent to the IMS index – it calculates the weighted average of response deviations within each item subset. Likewise, the ℓ_{zm} statistic is equal to ℓ_z summed over the item subsets. All three indices can be transformed to an approximate standard normal distribution.

Two indices, $\lambda(x)$ and $T(x)$, called *optimal person-fit statistics* test the null hypothesis of normal examinee responses against the alternative hypothesis that an examinee's responses are consistent with a model assuming aberrant responses. If a researcher is able to propose specific alternative hypothesis models (to model cheating, sleeping, fumbling, etc. responses) then, according to Karabatsos (2000, p.282), "… no other fit statistic can achieve a higher rate of detection of aberrant-responding examinees."

Several other person-fit indices that will not be described in this paper are included at the end of Appendix B. For a description of these lesser-known indices, the reader is referred to Meijer & Sijtsma (2001).

Simulation Study

To evaluate the usefulness of several person-fit indices, a dataset was simulated under the following conditions:

Test Length: $n = 50$ items

Item Discrimination Parameters: Randomly generated from a normal distribution: $a_i \sim ND(1, 0.3)$. The values of these parameters fell in the range (0.39, 1.52).

Item Difficulty Parameters: Generated to be equidistant over the range [-2.00, 2.00].

Item Pseudo-Guessing Parameters: Randomly generated from a uniform distribution: $c_i \sim Uni(0, 0.25)$. The values of these parameters fell in the range (0.01, 0.24)

Examinee Ability Estimates: $N = 500$ ability estimates were randomly generated from a standard normal distribution: $\theta_a \sim ND(0,1)$

Examinee Response Probabilities: The three-parameter logistic model was used to calculate the probabilities of correct responses for each examinee-item combination (P_{ia})

Examinee Responses: Random numbers ranging from [0, 1] were generated. These random numbers (R_{ia}) were then compared to the response probabilities generated earlier. The following rule was used to generate examinee responses:

$$u_{ia} = \begin{cases} 0 & \text{if } P_{ia} < R_{ia} \\ 1 & \text{if } P_{ia} > R_{ia} \end{cases}$$

The following table displays the item parameters simulated under these conditions.

Item	a_i	b_i	c_i	p-value
1	1.26	-2.00	0.03	0.96
2	0.86	-1.92	0.11	0.92
3	0.73	-1.84	0.15	0.86
4	0.40	-1.76	0.07	0.79
5	1.23	-1.67	0.04	0.89
6	0.88	-1.59	0.17	0.90
7	0.75	-1.51	0.09	0.86
8	0.74	-1.43	0.11	0.82
9	0.39	-1.35	0.01	0.70
10	0.86	-1.27	0.03	0.79
11	0.41	-1.18	0.04	0.69
12	0.48	-1.10	0.01	0.71
13	1.11	-1.02	0.16	0.80
14	1.21	-0.94	0.04	0.79
15	0.83	-0.86	0.18	0.79
16	1.11	-0.78	0.23	0.80
17	1.51	-0.69	0.14	0.79
18	1.02	-0.61	0.07	0.73
19	1.22	-0.53	0.18	0.70
20	0.60	-0.45	0.01	0.59
21	0.69	-0.37	0.03	0.61
22	0.69	-0.29	0.18	0.63
23	0.93	-0.20	0.06	0.60
24	1.08	-0.12	0.06	0.57
25	0.62	-0.04	0.15	0.61
26	1.52	0.04	0.24	0.61
27	0.79	0.12	0.06	0.53
28	0.62	0.20	0.08	0.52
29	1.13	0.29	0.01	0.40
30	1.07	0.37	0.18	0.50
31	0.70	0.45	0.03	0.41
32	1.32	0.53	0.01	0.35
33	0.95	0.61	0.05	0.36
34	0.83	0.69	0.02	0.35
35	1.44	0.78	0.17	0.40
36	1.21	0.86	0.17	0.41
37	1.01	0.94	0.22	0.41
38	1.25	1.02	0.09	0.27
39	1.46	1.10	0.12	0.29
40	0.93	1.18	0.18	0.34
41	0.51	1.27	0.05	0.30
42	0.83	1.35	0.19	0.35
43	1.24	1.43	0.02	0.18
44	0.85	1.51	0.13	0.28
45	0.54	1.59	0.17	0.32
46	0.86	1.67	0.23	0.33
47	1.07	1.76	0.08	0.18
48	0.62	1.84	0.13	0.22
49	0.85	1.92	0.19	0.26
50	1.13	2.00	0.03	0.10

p-value represents the proportion of examinees answering the item correctly.
The correlation between difficulty parameters and p-values is -0.97

The test simulated by this method had the following characteristics:

Total Number Correct Score: $\mu = 27.58$
$\sigma = 9.076$
Minimum: 5
Maximum: 49

Distribution of Scores:

Frequency	Stem &	Leaf
.00	0 .	
7.00	0 .	9&
26.00	1 .	012233444
77.00	1 .	55555666667777788888999999
77.00	2 .	00000011111222233333444444
106.00	2 .	55555666666677777777778888889999999
94.00	3 .	00000111111112222222233333444444
55.00	3 .	5555566677778889999
44.00	4 .	00011112233444
14.00	4 .	566789

Internal Reliability Estimate: Coefficient Alpha = 0.9007

Distribution of Item a_i Parameters:

Stem	&	Leaf
0.3		9
0.4		018
0.5		14
0.6		022299
0.7		03459
0.8		333556668
0.9		335
10		12778
11		1133
12		1123456
13		2
14		46
15		12

Distribution of Item b_i Parameters:

Stem	&	Leaf
-2	.	0
-1	.	556789
-1	.	011234
-0	.	566789
-0	.	012234
0	.	012234
0	.	566789
1	.	011234
1	.	556789
2	.	0

Appendix C displays the simulated item responses for the sixty examinees with the highest and lowest abilities.

• Simulation #1

One nonparametric (MCI) and three parametric person-fit indices (OMS, IMS, and ℓ_z) were then computed for the simulated data. These specific indices were selected because they have agreed-upon critical values for identifying aberrant responders:

Person-Fit Index	Critical Value	Source
(O/I)MS	MS > 1.3	Meijer, & Sijtsma (2001)
ℓ_z	ℓ_z < -2.0	Reise (1996)
MCI	MCI > 1.3	Meijer, & Sijtsma (2001)

Of particular interest was the number of aberrant responders flagged by each index. In this data set, no examinees were intentionally simulated to be aberrant (no guessers, plodders, or cheaters were generated). Even so, some aberrance is expected due to the guessing parameters estimated for each item. The following table displays the number of examinees flagged as being aberrant by each index.

	OMS	IMS	ℓ_z	MCI
Aberrant Responders	64	13	9	53
# of Examinees	500	500	500	500
% Aberrant	12.8%	2.6%	1.8%	10.6%

As the table shows, the indices identified between 1.8% and 12.8% of the examinees as aberrant. The differences in the number of examinees flagged may be due to two reasons: (1) the effectiveness of each index in identifying aberrant responders and (2) the different (mostly unknown) Type I error rates of the agreed-upon critical values. For this data set, it isn't possible to determine which index is most effective, since the actual number of aberrant responders is unknown.

Perhaps of more interest is to determine the level of agreement among the indices for this particular data set. If the indices only differed in the number of examinees identified (and had no disagreement on the aberrance of any particular examinee), we could reasonably believe the differences among the indices is due to the Type I error rates due to their unique critical values. The following contingency tables display the number of examinees identified as being aberrant by each index.

		IMS	
		Not Aberrant	Aberrant
OMS	Not Aberrant	435	1
	Aberrant	52	12

		IMS	
		Not Aberrant	Aberrant
ℓ_z	Not Aberrant	485	6
	Aberrant	2	7

		IMS	
		Not Aberrant	Aberrant
MCI	Not Aberrant	441	6
	Aberrant	46	7

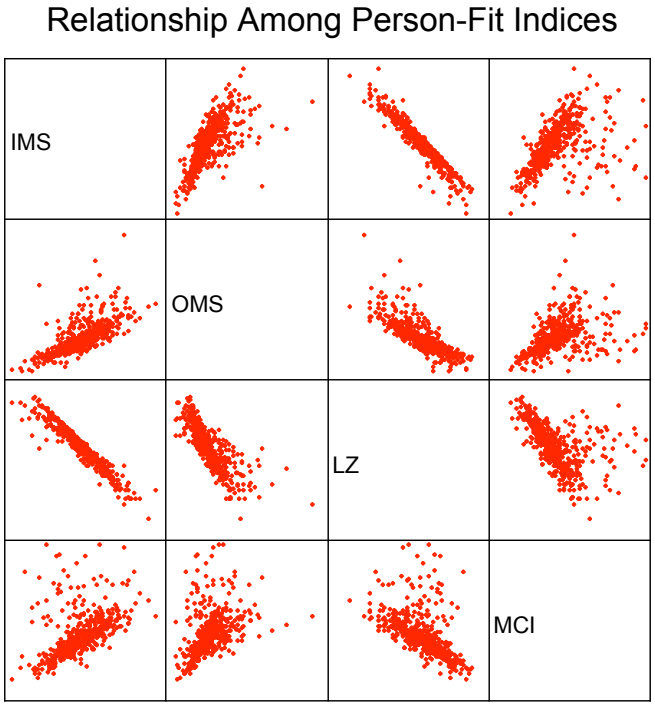
		ℓ_z	
		Not Aberrant	Aberrant
MCI	Not Aberrant	442	5
	Aberrant	4	9

The first table displays the level of agreement between the two deviation-based indices: *Outfit MS* and *Infit MS*. Most of the difference between the two indices was due to OMS simply flagging more examinees as being aberrant. This is expected, since the OMS inflates when item difficulties and examinee abilities widely differ. Note only one examinee flagged by the IMS index was not flagged by the OMS index.

The second table displays the agreement between IMS and the ℓ_z index. Of the 15 examinees flagged by at least one of these indices, only 7 were flagged by both indices. This same level of disagreement exists between ℓ_z and the MCI, with only 9 of the 18 potentially aberrant examinees being flagged by both indices. The following matrix displays the correlations among the indices for this data set:

	IMS	OMS	ℓ_z	MCI
IMS		.704	-.957	.482
OMS			-.803	.440
ℓ_z				-.539

Note that the three parametric indices had high intercorrelations and lower correlations with the nonparametric MCI statistic. The following scatterplots display the nature of the relationships among the indices. It is apparent that the MCI differs the most from the other indices.



• Simulation #2: Detecting Cheaters

In order to determine which index does the best job of flagging aberrant responders, the data set was modified. Ten percent (N = 50) of the examinees with below-average ability estimates were randomly selected to become "cheaters." The responses from these examinees were modified so that they all answered the most difficult 10 items (20% of the test) correctly. This modification ensures these fifty examinees will have aberrant response patterns. After obtaining maximum likelihood estimates of the examinee abilities, the person-fit indices were computed. The following table displays the detection rates for each index.

		Number flagged as aberrant			
	Simulated	IMS	OMS	ℓ_z	MCI
Cheaters	50	33	37	40	43
Non-cheaters	450	12	47	7	35
Total	500	45	84	47	78
% of cheaters identified as aberrant		66%	74%	80%	86%
% of non-cheaters identified as aberrant		2.7%	10.4%	1.6%	7.8%

Once again, the OMS index flagged the highest number of potentially aberrant responders, while the IMS and ℓ_z indices flagged the fewest examinees. Since fifty examinees were simulated to be "cheaters" on this exam, we can evaluate the effectiveness of each index in detecting this type of aberrant responding. The nonparametric MCI statistic was able to detect 86% of the simulated cheaters. The parametric indices had somewhat less success, properly identifying between 66-80% of the cheaters.

The last row of the table can loosely be interpreted as the false alarm rate. Non-cheating examinees may still have aberrance in their response patterns, so the indices will flag these unintentionally aberrant responders. If we're interested in detecting cheating and not other types of aberrant response methods, it appears as though the ℓ_z index provides the fewest false alarms with a relatively high detection rate. If we're interested in simply identifying the highest percentage of cheaters (without worrying about false alarms), we might choose to compute the MCI.

All four indices had a relatively high level of agreement with each other. All 37 cheaters correctly identified by the IMS index were also flagged by the other three indices. The three cheaters not identified by the MCI index were also not detected by the other three indices. These undetected cheaters had the highest estimated ability levels, which apparently made it difficult for the indices to determine if the examinees cheated or if they simply had enough ability to answer the items correctly.

The detection rates (and false alarm rates) for all four indices can be increased by changing the critical values used to flag aberrant responders. This is not useful in practice, since we would have no reason to change the critical values (we wouldn't be able to check detection rates on actual data).

It is unclear how well these results generalize for detecting cheaters from actual data. Karabatsos (2003) demonstrated that the percentage of simulated aberrant responders impacts the detection rates of each person-fit index. Data simulated to have 5% or 10% aberrant responders yielded higher detection rates than data that was simulated to have 25% or 50% aberrant responders. (Karabatsos, 2003)

• Simulation #3: Detecting Random Guessers

The original data set was once again modified to include 50 examinees who randomly guessed the answers to each item (assuming the test was composed of 4-option multiple-choice items). Forty lower-than-average ability examinees were randomly selected to become "random guessers." The simulated responses from these examinees were modified so that each examinee had a 25% chance of correctly answering each item. The other ten "random guessers," whose responses were similarly modified, were selected from the high-end of the ability scale. These examinees were thus selected under the assumption that most random guessers are of lower ability.

After obtaining new maximum likelihood ability estimates under the three-parameter logistic model, the four person-fit indices were computed from this data set. The following table displays the detection rates for flagging random guessers.

		Number flagged as aberrant			
	Simulated	IMS	OMS	ℓ_z	MCI
Random Guessers	50	44	46	46	48
Normal Responders	450	12	38	9	24
Total	500	56	84	55	72
% of random guessers identified as aberrant		88%	92%	92%	96%
% of normal responders identified as aberrant		2.7%	8.4%	2.0%	5.3%

We again see the pattern of the OMS and MCI indices flagging the highest number of examinees, the MCI having the highest detection rate, and the ℓ_z index having the lowest percentage of false alarms. All four indices were noticeably more effective in detecting random guessers than they were in detecting cheaters in the previous data set. The false alarm rate also was lower for all four indices for this data set than they were in the cheaters data set.

The indices had a slight disagreement in which random guessers were flagged as being aberrant. Both random guessers who were not detected by the MCI index were correctly flagged by the three parametric indices. The nonparametric MCI also correctly flagged two low-ability examinees that were not detected by the other indices. Most of the random guessers not detected by the parametric indices were of low ability. Once again, this inability to detect aberrance is most likely due to the fact that random guessers have ability levels near what we would expect from low-ability examinees.

- Simulation #4: Detecting Plodders

Modifications were made to the original simulated data in order to generate plodders – examinees who do not complete the final items on an exam. Since plodders might be of high or low ability, fifty examinees were selected at random to become plodders. The responses to the final 10 test items from these fifty examinees were all generated to be incorrect. It should be noted that the final 10 items on this simulated test were also the ten most difficult items.

Maximum likelihood estimates were obtained and the four person-fit indices were once again computed.

		Number flagged as aberrant			
	Simulated	IMS	OMS	ℓ_z	MCI
Plodders	50	21	23	39	34
Normal Responders	450	21	43	19	29
Total	500	42	66	58	63
% of plodders identified as aberrant		42%	46%	78%	68%
% of normal responders identified as aberrant		4.7%	9.6%	4.2%	6.4%

All four indices had more trouble detecting plodders than they did in detecting cheaters or random guessers. It appears as though the ℓ_z index was best in detecting plodders, followed by the nonparametric MCI and the two mean-square indices. The false alarm rates were relatively low, with between 4-10% of normal responders identified as being aberrant by the four indices.

The indices had very little agreement over which plodders were, in fact, aberrant. Only 15 plodders were correctly flagged by all four indices. Of the 11 plodders not detected by the ℓ_z index, nine were correctly identified by the MCI.

Low-ability plodders were least likely to be detected. This is not surprising, since low-ability examinees had a high probability of missing the final ten items anyway.

- Simulation #5: Detecting Multiple Types of Aberrance

As a final test for the four person-fit indices, the original simulated data was once again modified. Sixty examinees were randomly selected to become aberrant. Twenty of these examinees had responses modified so that they became "cheaters." Another twenty had responses modified to become "random guessers." The final twenty examinees became "plodders" through appropriate modifications to their response strings. After computing maximum likelihood estimates of ability, the four person-fit indices were computed.

		Number flagged as aberrant			
	Simulated	IMS	OMS	ℓ_z	MCI
Cheaters	20	7	9	14	16
Random Guessers	20	14	15	15	18
Plodders	20	11	14	14	9
Total Aberrant	60	32	38	43	43
Normal Responders	440	24	44	14	28
Total	500	56	82	57	71
% of aberrant responders identified as aberrant		53%	63%	72%	72%
% of normal responders identified as aberrant		5.5%	10%	3.2%	6.4%

Overall, the person-fit indices had lower detection rates when multiple types of aberrant responses were simulated. The false alarm rates remained similar to the rates obtained in the previous simulations. Based on these five simulations, it appears as though the MCI and ℓ_z indices have the highest detection rates and lowest false alarm rates.

- Discussion

The objective of person-fit measurement is, "… to detect item-score patterns that are improbable given an IRT model or given other patterns in a sample." (Meijer & Sijtsma, 2001) These aberrant response patterns can have a negative impact on the validity of test scores. Examinees with aberrant response patterns can obtain ability estimates significantly higher or lower than their actual level of ability, depending on the type of aberrance they exhibit. This can cause serious problems in admissions testing or licensure testing situations. Furthermore, a close examination of the five simulations in this study demonstrated that item parameter estimates are impacted by the presence of aberrant responders.

The particular cause of an examinee's aberrant response pattern is difficult to determine. Aberrance, in the form of cheating, plodding, or random guessing, can be due to poorly designed tests, poor test administrations procedures, or examinee psychological idiosyncrasies. Some studies (Scmitt, Chan, & Sacco, 1999) have attempted to find correlates of aberrant response patterns.

Over 40 person-fit indices have been developed to detect aberrant response patterns from a set of test data. While the indices were developed from a variety of approaches (classical test theory vs. item response theory; deviation-, correlation-, and likelihood-based), they all measure the degree to which an examinee's response pattern matches the response pattern we expect to find. Since all the indices measure the same thing, only two questions remain:

(1) Which index should be used to detect aberrance?
(2) What should be done with aberrant responders?

An answer to the first question has been the focus of a number of studies. Li & Olejnik (1997), in a study of several indices, conclude the ℓ_z index provides as reliable and accurate detection as other person fit statistics (p.229). Meijer & Sijtsma (2001) conducted a literature review and concluded Moelnaar's M statistic is usually the best choice for detecting aberrance. Karabatsos (2003) evaluated 36 person-fit indices and found the nonparametric indices outperform the IRT-based indices. This is generally due to the fact that IRT-based indices are calculated from the estimated $\hat{\theta}$ instead of the actual θ. Karabatsos goes on to conclude that the H^T index is the best (followed by $D(\theta)$, C, MCI, and $U3$).

Regardless of the particular index chosen, all person-fit measures have limitations. First off, these indices are difficult to calculate for large sample sizes or test lengths. Secondly, the indices can only determine that aberrance exists – they do not indicate which type of aberrance is detected. This limits the usefulness of the indices as anything other than a filter for identifying potential aberrance.

Once a potentially aberrant examinee has been identified, what should be done? Examinees who cheat, randomly guess, or plod will not receive accurate ability estimates. Meijer & Sijtsma (2001) provide three options when an aberrant responder has been detected:

(1) Eliminate the aberrance (either the examinees entire response pattern or the portion of responses that show the greatest aberrance) and reestimate item parameters and ability estimates.
(2) Do not report an ability estimate for aberrant responders and retest
(3) Ignore the aberrance in estimating ability and item parameters.

References Cited:

Bell, R.C. (1982). Person fit and person reliability. *Education Research and Perspectives*, 9:1, 1982, 105-113.

Bracey, G. (1992). Person-fit statistics: high potential and many unanswered questions. *ERIC/TM Digest*.

Davey T., E. Stone, & R. McKinley (2003). Robust estimation of IRT parameters. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 21-25.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1 (2)*, 152-176.

Karabatsos, G. (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education 16:4, 277-298.*

Lamprianou, I., N. Nelson, & B. Boyle (2001). Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils that speak English as an additional language. *European Congress on Educational Research*, September 2001.

Li, M.F. & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement* 21:3, p.215-231.

Meijer, R.R. & K. Sitsma, Person-fit statistic – what is their purpose? *Rasch Measurement Transactions*, Fall 2001, 15:2 p.823

Meijer, R.R. (1996). Person-fit research: an introduction. *Applied Measurement in Education*, 9(1), 3-8

Meijer R.R., & Sijtsma K. (2001) Methodology review: evaluating person fit. *Applied Psychological Measurement* 25(2), 107-135

Reise, S. P. (1990). A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. *Applied Psychological Measurement*, 14(2), 127-137.

Reise, S. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9(1), 9-26

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, Fall.

Rudner, L. M. et. al. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education*, 9(1), 91-109

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied psychological measurement*, 23 (1), 41-54.

Smith, R.M. (1991) The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*. 51(3) 541-565.

LeTendre, M.J. (2000). Identifying schools and school districts in need of improvement or corrective action when moving from transitional to final assessments. Website: http://www.ed.gov/policy/elsec/guid/assessmemo.html

Van den Wittenboer, G., J. Hox, & E. Leeuw. Aberrant response patterns in elderly respondents: latent class analysis of respondent scalability. Chapter 14.

Van Krimpen-Stoop, E. M. L., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23 (4),* 327-346.

Wright, B.D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions*, 9(2), p.430-1.

References for Formulas:

Bedrick, E.J. (1997). Approximating the conditional distribution of person fit indices for checking the Rasch model. *Psychometrika*, 62, 191-199.

Donlan, T.F. & Fischer, F.E. (1968). An index of an individual's agreement with groups determined item difficulties. *Educational and Psychological Measurement*, 28, 105-113.

Drasgow, 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 polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*. 38, 67-86.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.

Harnisch, D. L., & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.

Kane, M.T. & Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4, 105-126.

Karabatsos, G. (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education 16:4, 277-298.*

Klauer, K.C. (1990). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, 56, 535-547.

Levine, M.V. & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics,* 4, 269-290.

Levine, M.V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika,* 53, 161-176.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices*. Psychometrika, 55(1)*, 75-106.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tokyo.

Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden*, 7, 131-145.

Smith, R.M. (1986) Person fit in the Rasch model. *Educational and Psychological Measurement*. 46, 359-370.

Tatsuoka, K.K. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20, 221-230.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.

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 flier (1977). Environmental factors and deviant response patterns. *Basic problems in cross-cultural psychology*. Amsterdam: Swets & Zeitlinger.

van der flier (1980). *Comparability of individual test performance*. Lisse: Swets & Zeitlinger.

Wright, B.D. (1980). *Probabilitistic Models for some intelligence and attainment tests: With foreword and afterword by Benjamin D. Wright*. Chicago: MESA Press.

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