Practical Assessment, Research, and Evaluation

Volume 12 *Volume 12, 2007* Article 16

2007

On the Performance of the l_Z Person-Fit Statistic

Ronald D. Armstrong

Zachary G. Stoumbos

Kung Mabel T.

Min Shi

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 16, December 2007 ISSN 1531-7714

On the Performance of the l_z Person-Fit Statistic

Ronald D. Armstrong and Zachary G. Stoumbos, *Rutgers University*

Mabel T. Kung, *California State University at Fullerton*

Min Shi, *California State Polytechnic University, Pomona*

Person-fit measurement refers to statistical methods used to detect improbable item-score patterns. This study investigates the detection effectiveness of the l_z statistic, which is one of the most popular and powerful person-fit statistics in the literature to date. The contributions of the present study are three-fold. First, the simulation results show that the detection power of the l_z statistic is largely hinged on test characteristics, particularly the test difficulty. Therefore, the l_z statistic should be used with caution in an operational testing environment. Second, this paper provides a clear explanation for the poor performance of the l_z statistic under certain situations. The third objective is to present a summary of the patterns and conditions for which the l_z statistic is not recommended for the detection of aberrancy. This can be used as a checklist for implementation purposes.

Person-fit or appropriateness measurement refers to statistical methods used to evaluate the fit of a response pattern to a particular test model. A number of person-fit statistics have been proposed to identify item-score patterns that are not in agreement with the expected response pattern based on item response theory (IRT) model (for example, Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1988; Molenaar & Hoijtink, 1990, 1996; Bracey & Rudner, 1992; Nering & Meijer, 1998; Rudner, 2001; and Childs, Elgie, Gadalla, Traub & Jaciw, 2004). Recently, Meijer & Sijtsma (2001) and Karabatsos (2003) presented extensive, excellent reviews of the methodological developments in evaluating person fit.

This article discusses the likelihood-based person-fit statistic l_z (Drasgow, Levine, & Williams, 1985) due to the following reasons. First, the l_z statistic has received a great deal of attention in more than forty person-fit statistics (Meijer & Sijtsma, 2001),and is one of the most popular person-fit statistics in educational measurement (see Drasgow, Levine & McLaughlin, 1991; Reise & Due, 1991; Reise, 1995; Nering 1995, 1997; and Nering & Meijer, 1998). Second, prior studies have found that the l_z statistic performed better than other person-fit statistics in many cases and have recommended the l_z statistic as one of the most powerful person-fit statistics for the detection of aberrant behavior. For example, Drasgow, Levine, & McLaughlin (1987) used the three-parameter logistic (3PL) model to compare the performance of the l_z, ZU, ZW, C, JK, O/E, $ECI2_z$ and $ECI4_z$ person fit statistics, and concluded that l_z is among the most capable statistics for identifying both spuriously high-scoring and low-scoring examinees. Li & Olejnik (1997) compared the distributions of l_z, ZU, ZW, $ECI2_z$, and $ECI4_z$ within the framework of the Rasch model (MacCann & Stanley, 2006) and found that the l_z statistic was the most powerful statistic, as it identified two-third of the misfitting

item-score patterns. Nering & Meijer (1998) compared the performance of the l_z statistic and the person response function (PRF) method, and found that the l_z statistic performed better than the PRF method in most cases.

In view of these studies on the power of the l_z statistic, its effectiveness under difference scenarios is still not clear. To our best knowledge, the study of examining the effects of item characteristics on the detection power of person-fit statistics is limited and incomplete. Using information-based methods, Reise & Due (1991) investigated the influence of the test length, the spread of the item difficulty parameter, and the value of the guessing parameter on the detection power of l_z statistic. It was shown that the l_z statistic was most efficient for a long test with items of varied difficulty levels and small guessing parameters. Meijer, Molenaar & Sijtsma (1994) extended Reise & Due (1991) to a nonparametric context. Specifically, they examined how the detection power of a nonparametric person-fit statistic, U3, was influenced by the test, person, and group characteristics. They suggested that the detection rate is a function of the item discrimination parameter (reliability), the test length, the percentage of non-fitting response vectors (NRVs) in the group, and the types of NRVs. In particular, different test parameters could work in a complimentary manner to achieve desirable rates of detection. In a personality assessment context, Reise (1995) considered a two-parameter logistic (2PL) model to investigate the power of the l_z statistic in detecting non-model-fitting responses and examined how the detection power was affected by different scoring strategies. It was found that the best detection rates were achieved when the difference between trait levels and item difficulty parameters was large. The findings of this paper indicate that this is not always the case.

The purpose of this paper is to further investigate the effects of item characteristics on the detection power of the l_z statistic in the parametric IRT context. First, we show through Monte Carlo simulations that the detection power of l_z statistic is a function of test characteristics. The detection rates could be low in various situations. Therefore, the l_z statistic should be used with caution in an operational testing environment. The simulation methods in this study are based on the literature (Levine & Rubin, 1979; Drasgow, 1982; Drasgow, Levine, & McLaughlin, 1987; Nering & Meijer, 1998). Second, the paper gives an explanation for the potentially poor performance of l_z statistic. The third objective is to present a summary of the patterns and conditions for which the l_z statistic is not recommended for the detection of aberrancy. This can be used as a checklist for application purposes. Lastly, the summary provides the discussions of the implications for practitioners and points out possible adjustments that can be used to improve detection effectiveness.

The l_Z Person-Fit Statistic

Suppose that an examinee with trait level θ is administered a test form of *n* items. Let the response of the i^{th} dichotomous (0 or 1) item score be represented by a Bernoulli random variable

$$X_i = \begin{cases} 1, & \text{if the response to item } i \text{ is correct,} \\ 0, & \text{otherwise,} \end{cases}$$

with probability density function

$$f(x_i; p_i(\theta)) = p_i(\theta)^{x_i}[1-p_i(\theta)]^{1-x_i}, \quad x_i = 0, 1,$$

where $p_i(\theta) = P(X_i = 1 | \theta)$, for i=1,2,…n.

That is, $p_i(\theta)$ denotes the probability of a correct response to the i^{th} item by an examinee with trait level θ. The value of θ is generally unknown in an operational testing environment, so an estimate of the θ value is often used in practice.

As one of the most popular person-fit statistics in the literature, the l_z statistic is the standardized version of the likelihood-based person-fit statistic l_0 (Levin & Rubin, 1979). Let

$$l_{0i} = X_i ln(p_i(\theta)) + (1 - X_i) ln(1 - p_i(\theta))$$

represent the natural logarithm of the probability for the observed response to item *i*. An observed response pattern, $(x_1, x_2, \ldots, x_n)$, is used to calculate a value for the statistic. The l_0 and l_z statistics are usually evaluated over all items on the test. The l_0 and l_z statistics can be respectively expressed by

$$l_0 = \sum_{i=1}^{n} l_{0i} \tag{1}$$

and

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}} = \sum_{i=1}^{n} \frac{l_{0i} - \bar{l}}{[Var(l_0)]^{1/2}} \tag{2}$$

where $\bar{l} = E(l_0)/n$.

The expectation and variance of the l_0 statistic can be written as

$$E(l_0) = \sum_{i=1}^{n} [p_i(\theta) ln(p_i(\theta)) + (1 - p_i(\theta)) ln(1 - p_i(\theta))] \tag{3}$$

and

$$Var(l_0) = \sum_{i=1}^{n} p_i(\theta)(1 - p_i(\theta)) \left\{ ln\left[\frac{p_i(\theta)}{1 - p_i(\theta)} \right] \right\}^2 \tag{4}$$

Based on a generalization of the Central Limit Theorem, the asymptotic distribution of l_0 is normal (Box, Hunter & Hunter, 1978; pp 87-90). A statistical test is associated with the l_z statistic and can be described by the hypotheses,

H_0: The response pattern is congruent with the specified response model,

versus

H_1: The response pattern is not congruent with the specified response model.

This hypothesis test is structurally conducted as a lower one-sided test, even though over-achieving and under-achieving performances are signaled and can be detected. Select situations, as will be shown, may benefit from a two-sided test. However, most aberrant situations are best detected with a lower one-sided test. Therefore, the null hypothesis will be rejected if the computed test statistic is less than a critical value specified so the test achieves a significance level α. The following example illustrates the reasoning for a lower one-sided test.

An Illustrative Example

Assume that an examinee with a known latent trait value of θ has taken a test and the average unconditional probability of a correct response to an item of the test is 0.60. Further, suppose that $\bar{l} = -0.67$ and $Var(l_0) = 0.525$ (these are some plausible values). Now consider a single item where $p_i(\theta) = 0.25$. This is the case of an examinee capable of answering correctly this test item 25% of the time. Aberrant behavior will be manifested by answering this item correctly.

In equation (2) where the contribution to the l_z statistic by a single item can be evaluated by considering the relationship between l_{0i} and $\bar{l}$. When $p_i(\theta)$ =0.25, $ln(0.25) \approx -1.39$ indicates that this response is aberrant (a correct response), since $ln(p_i(\theta)) < \bar{l}$ and the response further decreases the value of l_z. An incorrect response, in this case, has $ln(0.75) \approx -0.29$ (or $ln(1 - p_i(\theta)) > \bar{l}$), which increases the value of l_z. Likewise, when $p_i(\theta)$ =0.75, an incorrect response contributes negatively to the l_z statistic, while a correct response gives a positive contribution. In general, x_i =1 and $ln(p_i(\theta)) < \bar{l}$ indicate an aberrant behavior. Then again, x_i =0 and $ln(1 - p_i(\theta)) < \bar{l}$ also indicate possible aberrant behavior.

Caution is indeed required when utilizing the l_z statistic. Consider the same examinee and test described in the previous paragraph. Now note a single item where $p_i(\theta)$ =0.52. The contribution to l_z, given a correct response, is $(ln(p_i(\theta)) - \bar{l})/\sqrt{Var(l_0)} \approx +0.02218$. That is, responding correctly repeatedly to items with $p_i(\theta)$ around 0.5 would never suggest aberrant behavior.

To illustrate the performance of the l_z statistic under certain situations, a series of calculations show how the l_z statistic fluctuates when the item responses from the examinee are (a) all correct, and (b) alternate between correct and incorrect responses (see Figure 1(a), and 1(b)). For purposes of exposition, it is assumed that the latent trait value of the examinee is known and is not affected by the response patterns. The graphs are based on ten-item tests where the probability of a correct response increased by 0.01 for each item. Multiple tests were created with the lowest probability of a correct response ranging from 0.2 (a difficult test item) to 0.9 (an easy test item). The first graph plots average test difficulty versus the l_z statistic in the case where an examinee answers all items correctly regardless of the level of difficulty of the item. These swings indicate that a one-tailed test decision rule may not be realistic in all situations.

To further the discussion on the lower one-tail test for the l_z statistic, consider the situation when the examinee answers every other item correctly. The l_z value never becomes positive in this case. The reason for this behavior is that when the "all or nothing" responses occur, the l_z statistic follows a logarithmic curve. On the other hand, when alternating correct responses are recorded, the l_z statistic remains negative even when the probability of a correct response is around 0.5.

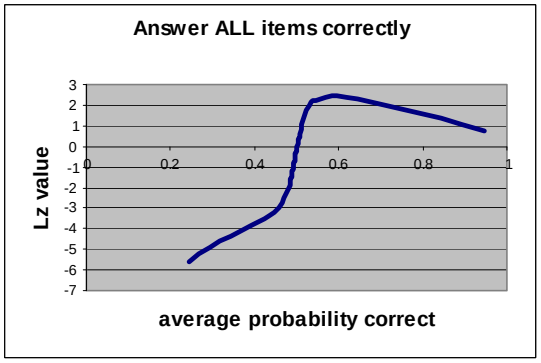

Figure 1(a) The l_z statistic when the examinee answers all items correctly.

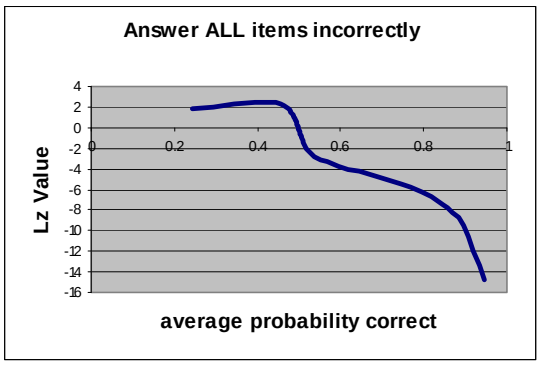

Figure 1(b). The l_z statistic when the examinee answers every other item correctly.

Simulation Method

Monte Carlo simulations were performed in the framework provided by a previous study of Nering & Meijer (1998). Their general purpose was to compare the detection rates of the l_z statistic against a person response function (PRF) method. The IRT model is used to calculate $p_i(\theta)$ as a function of an examinee's trait level and a set of item parameters, a, b, and c. The exact form of this model can be found in Lord (1980, page 12). The value of c is the probability of an infinitely unable examinee responding correctly to the item. The quality of incorrect choices affects this parameter. For example, a low-level examinee may eliminate incorrect choices, which will increase this value, or be drawn to attractive incorrect choices, which can lower the value. The value of b is a measure of the difficulty of the item and is on the same scale as θ. That is, easier items have a lower b value and more difficult items have a higher b value. When $\theta = b$, the probability of a correct response to any item by an examinee is $(1+c)/2$. The value of a measures the discrimination of the item, or how well the item can distinguish between lower-level and higher-level examinees. The larger a is, the steeper the curve is about $\theta = b$.

Benchmark

A variant of the method given by Nering & Meijer (1998) for a moderate length test was used as a benchmark. This study deviated from their approach only in the way that critical values were obtained. A 121 item test was created. While 121 items may be longer than most tests in practice, the use of a large number of items here leads to more stable estimates of l_z and the proportions that will be reported later. The in-control responses to all items followed a 3PL IRT model. The parameters were $a \sim$ N(1.0,0.1], b ~U(-2.7,2.7), and c ~N(0.20,0.10] (left bounded at 0.0). The a and c values were randomly drawn from the stated normal distributions. The distribution of the a parameters is tighter than observed in practice and leads to better estimates of l_z. The b values were *assigned* evenly spaced across the interval (–2.7,+2.7). Thus, the b values were –2.7, –2.655, –2.61, …, –0.09, –0.045, 0, +0.045, +0.09, …, +2.61, +2.655, +2.7. In general, we use the notation $b \sim \mathrm{U}(b_L, b_U)$ to indicate a discrete uniform assignment of the difficulty parameter over the closed interval (b_L, b_U).

Obtaining Critical Values

Let l_{zC} be a numerical value of the statistic l_z, where $\alpha = P(l_z \le l_{zC})$ conditional on a specified θ value and no aberrant behavior. The value l_{zC} is the probability of a Type I error, α, set at 0.05. Since the l_z statistic does not have tabulated critical values, Monte Carlo simulations

were employed to obtain them. These critical values were obtained by simulating 10,000 examinees on each of 61 equally spaced θ values ranging from −3.0 to +3.0 with an interval 0.1. The l_z was computed for each simulated examinee, and at each θ value, the 10,000 l_z values were sorted in ascending order and the critical value for that θ was taken to be the value in the 500th (or $(10{,}000*\alpha)^{th}$) position. This procedure created a critical value for each of the 61 points. Linear extrapolation was used when the (estimated) θ value of an examinee fell between those tabulated. Any estimated θ below -3.0 was brought back to -3.0 and any estimate above +3.0 was brought back to +3.0.

Obtaining Aberrant Responses

The responses to a single test, where the θ value and the described IRT parameters were used to simulate the administration of a test, are referred to as the Fitting Response Vectors (FRVs). In order to simulate non-FRVs (NRVs), Nering & Meijer (1998) applied a response manipulation method suggested by Levine & Rubin (1979) and Drasgow (1982). Two types of NRVs were simulated taking an FRV and then randomly selecting a proportion of the items within this response vector to manipulate. Every response on the FRV was given the same chance of being manipulated. For examinees with $\theta \geq 0$, a spuriously low (SL) NRV was simulated by taking each of the selected items and making the responses correct with probability 0.2 and incorrect with probability 0.8. A spuriously high (SH) NRV was simulated for examinees with $\theta \leq 0$ by rescoring all the selected items to be correct. All the manipulations for both SL and SH NRVs were made regardless of the responses in the FRV; thus, the number of responses changed from correct to incorrect (or vice versa) differed from examinee to examinee. The procedure was repeated as 18, 24 or 36 items were manipulated.

Table 1 presents the average number of responses changed during the simulations. The detection rates were defined as the proportion of NRVs that were correctly identified as aberrant by comparing the l_z values against the associated critical value l_{zC}.

Table 1. Average number of responses changed during the simulation of the benchmark test

# manipulated	Spuriously High Simulation θ Value						Spuriously Low Simulation θ Value					
	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.0	0.5	1.0	1.5	2.0	2.5
18	13.22	12.27	11.12	9.91	8.61	7.29	10.02	10.85	11.60	12.35	12.96	13.50
24	17.78	16.40	14.86	13.21	11.51	9.76	13.34	14.41	15.48	16.44	17.33	18.02
36	26.41	24.58	22.27	19.77	17.29	14.57	20.03	21.66	23.18	24.59	25.99	27.05

Simulation Results

θ Known

Tables 2 and 3 provide a summary of the simulation results when the known θ was used to calculate the l_z statistic. In both Tables 2 and 3, the benchmark results were close to the detection rates reported by Nering & Meijer (1998). The Group I columns report the detection rates under different assignments of the "level of difficulty" parameter b, while keeping the parameters a and c unchanged. As before, the b values were given a discrete uniform assignment across the interval. The effect of the b change from the benchmark in the SL case was to make to test easier, and in the SH case, was to make it more difficult. Thus, when simulating aberrant behavior, the responses of fewer items would be changed from the non-aberrant situation, and the detection of aberrant behavior should be reduced. This anticipated result was observed. More notable results occurred when the detection rate became biased. That is, the detection rates when aberrant behavior was present fell below the Type I error rate α (Lehmann, 1986 and Efron, 1986). The following describes why this occurred.

Table 2. Detection Rates for a 121-Item Test under SH with θ Known

θ	Items manipulated	Benchmark*	Group I			Group II		Group III	
			$b \sim$ U(-2.7,0)	$b \sim$ U(-2.7,-1)	$b \sim$ U(-2.7,-2)	$a \sim$ N(.5,.1)	$a \sim$ N(1.5,.1)	$c \sim$ N(.1,.1)	$c \sim$ N(.25,.1)
-2.5	18	95.21%	77.42%	45.29%	1.52%	80.66%	97.40%	99.91%	83.38%
	24	99.62%	93.86%	64.93%	0.84%	95.81%	99.94%	100.00%	97.29%
	36	100.00%	99.84%	92.53%	0.40%	100.00%	100.00%	100.00%	100.00%
	0	4.98%	5.67%	4.60%	5.14%	5.03%	4.80%	5.09%	4.74%
-2.0	18	86.02%	38.77%	5.81%	0.42%	65.41%	91.48%	99.39%	70.13%
	24	97.65%	56.84%	5.42%	0.10%	85.54%	98.89%	99.99%	89.29%
	36	99.97%	85.52%	5.44%	0.00%	99.19%	100.00%	100.00%	99.65%
	0	4.92%	5.10%	4.89%	4.60%	4.50%	4.65%	4.86%	4.75%
-1.5	18	71.98%	10.14%	0.92%	0.86%	45.96%	79.29%	96.33%	51.17%
	24	90.04%	12.57%	0.59%	0.28%	65.43%	94.09%	99.68%	72.16%
	36	99.55%	17.79%	0.11%	0.03%	92.15%	99.95%	100.00%	96.02%
	0	5.31%	5.16%	4.85%	4.84%	5.07%	4.29%	4.90%	4.75%
-1.0	18	52.39%	3.14%	0.91%	1.44%	27.09%	66.05%	86.65%	33.92%
	24	72.16%	2.15%	0.46%	0.94%	39.25%	85.55%	97.18%	50.48%
	36	95.88%	1.34%	0.07%	0.25%	65.33%	99.33%	99.98%	81.29%
	0	5.08%	5.25%	4.90%	5.27%	4.86%	4.80%	5.07%	4.68%
-0.5	18	34.77%	1.44%	1.47%	2.49%	12.40%	49.20%	68.02%	20.57%
	24	51.77%	0.90%	0.70%	1.49%	15.29%	69.38%	86.19%	29.00%
	36	81.55%	0.23%	0.27%	0.80%	25.04%	94.58%	98.92%	53.51%
	0	5.31%	4.60%	5.27%	5.48%	4.73%	4.67%	4.33%	4.71%
0.0	18	20.60%	1.18%	2.07%	2.55%	6.14%	35.47%	47.11%	12.91%
	24	29.69%	0.97%	1.46%	2.18%	6.64%	51.50%	66.19%	17.04%
	36	52.33%	0.16%	0.55%	1.28%	7.53%	81.87%	92.00%	27.97%
	0	5.28%	5.34%	5.28%	4.86%	4.84%	4.21%	4.92%	4.88%

* $b \sim$ U(-2.7,2.7) $a \sim$N(1.0, 0.1) $c \sim$N(0.2, 0.1)

Table 3. Detection Rates for a 121-Item Test under SL with θ known

θ	Items manipulated	Benchmark*	Group I			Group II		Group III	
			$b \sim$ U(-2.7,0)	$b \sim$ U(-2.7,-1)	$b \sim$ U(-2.7,-2)	$a \sim$ N(.5,.1)	$a \sim$ N(1.5,.1)	$c \sim$ N(.1,.1)	$c \sim$ N(.25,.1)
0.0	18	83.90%	5.62%	6.51%	8.48%	50.10%	95.52%	80.94%	85.90%
	24	94.96%	5.98%	7.45%	10.14%	69.31%	99.26%	94.04%	96.44%
	36	99.79%	6.39%	7.59%	11.90%	92.40%	99.99%	99.65%	99.89%
	0	5.08%	5.11%	5.13%	4.73%	4.90%	5.26%	5.33%	5.11%
0.5	18	95.45%	9.51%	4.58%	6.70%	69.17%	99.41%	94.03%	96.66%
	24	99.36%	10.77%	4.64%	7.22%	87.01%	99.98%	99.05%	99.57%
	36	100.00%	14.97%	4.29%	8.57%	98.96%	100.00%	99.99%	100.00%
	0	5.02%	6.01%	5.46%	5.39%	4.51%	4.99%	4.65%	4.79%
1.0	18	99.26%	25.19%	4.22%	4.84%	84.29%	99.91%	98.90%	99.47%
	24	99.96%	35.42%	4.19%	4.53%	95.78%	100.00%	99.90%	99.96%
	36	100.00%	58.74%	3.92%	4.08%	99.87%	100.00%	100.00%	100.00%
	0	4.95%	5.60%	5.04%	5.08%	4.85%	4.97%	4.57%	4.96%
1.5	18	99.93%	60.08%	13.32%	2.36%	93.83%	100.00%	99.92%	99.98%
	24	100.00%	79.80%	17.77%	1.89%	99.39%	100.00%	100.00%	100.00%
	36	100.00%	97.23%	29.10%	1.33%	100.00%	100.00%	100.00%	100.00%
	0	5.50%	5.08%	5.66%	5.22%	5.41%	4.96%	5.55%	5.78%
2.0	18	99.99%	91.81%	51.99%	5.66%	98.10%	100.00%	100.00%	100.00%
	24	100.00%	98.37%	71.33%	6.20%	99.94%	100.00%	100.00%	100.00%
	36	100.00%	99.98%	94.53%	6.20%	100.00%	100.00%	100.00%	100.00%
	0	4.61%	5.35%	5.59%	4.58%	5.11%	5.07%	5.22%	5.30%
2.5	18	100.00%	99.11%	85.92%	45.09%	99.46%	100.00%	100.00%	100.00%
	24	100.00%	100.00%	96.95%	63.71%	99.99%	100.00%	100.00%	100.00%
	36	100.00%	100.00%	100.00%	91.51%	100.00%	100.00%	100.00%	100.00%
	0	5.69%	5.23%	4.83%	4.72%	5.44%	4.84%	4.66%	4.94%

* $b \sim$ U(-2.7,2.7) $a \sim$N(1.0, 0.1) $c \sim$N(0.2, 0.1)

Again, consider equation (2) where the l_z statistic is defined. Based on a correct response to item i, the i^{th} term is negative when $ln(p_i(\theta)) < \overline{l}$ and positive when $ln(p_i(\theta)) > \overline{l}$. Figure 2 shows the plots of θ versus the average of $ln(p_i(\theta))$ minus $\overline{l}$ for a set of tests where the a and c parameters remained unchanged as the b parameter was changed. Comparing each of the plots against the corresponding column from Table 2 gives a pattern which relates to the power of the statistic. The values of θ where the curves are the most negative give the best detection rates. When the curve approaches the abscissa, the detection is approximately equal to the Type I error rate α. When the curve becomes positive, the hypothesis test becomes biased.

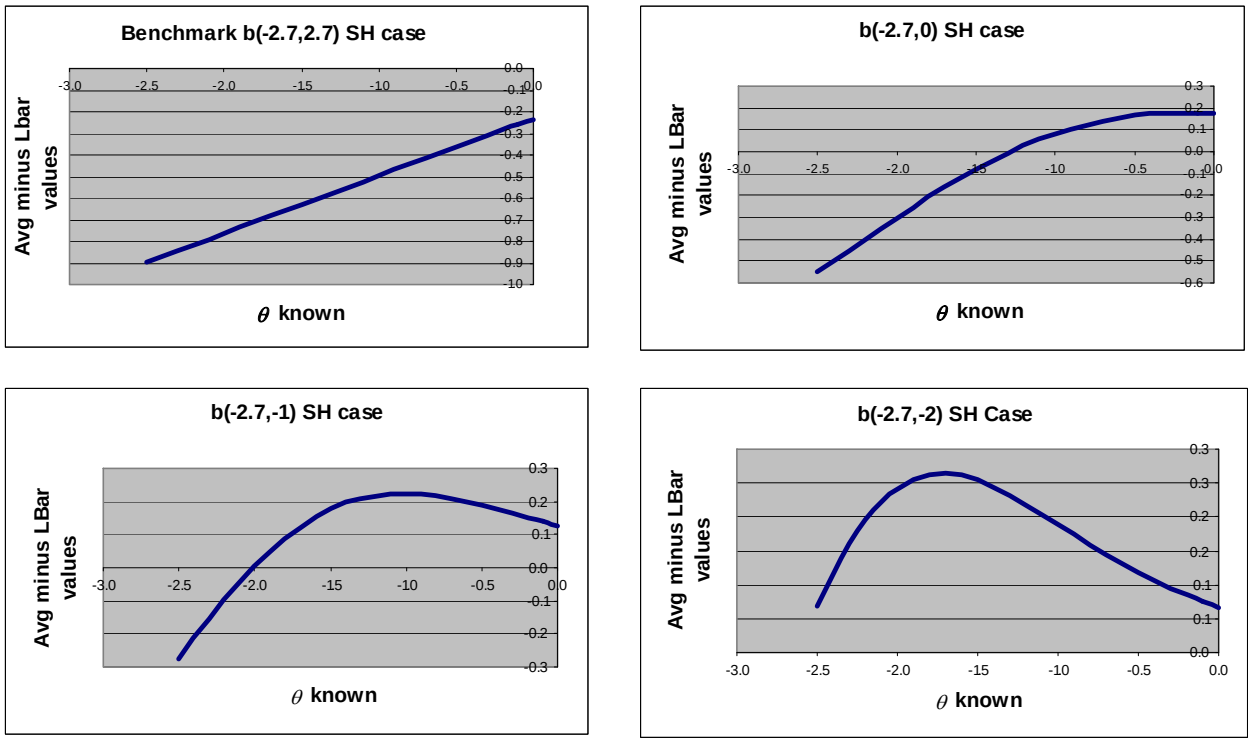

Figure 2. The θ versus the average of $ln(p_i(\theta))$ minus $\overline{l}$ (SH case).

Spuriously High (SH) Case

Of concern here is the consistency of the l_z statistic in detecting response aberrancy. As an example, consider the b ~U(−2.7,-1) column of Group I of Table 2 (SH case with θ known). The detection rate is summarized as follows:

I. θ = -2.5 the detection rate is well above the Type I error rate of 0.05

II. θ = -2.0, the detection rate is at approximately 0.05

III. -1.5 ≤ θ ≤ 0.0, the detection rate is generally below 0.05. The l_z statistic cannot identify the aberrant behavior.

The results can be compared with the plot of θ versus the average of $ln(p_i(\theta))$ minus $\overline{l}$ under b ~U(−2.7,-1) in Figure 2. The lack of detection occurs when the curve crosses the axis, from negative to positive. The detection rates are the poorest at θ = −1.0 (0.91%, 0.46%, 0.07%), where the curve (approximately) reaches a maximum. The detection increases slightly for θ = -0.5 and θ =0.0. The identification of aberrant behavior with rates less than 0.05 corresponds directly to θ values where the associated curve (Figure 2) is close to the axis or positive. In fact, when the curve is positive, detection power was found in the upper tail of the distribution. For example, with θ =-2.0 and b ~ U(-2.7,-2), an *upper* critical value was set by simulating 10,000 in-control administrations and taking the 95th percentile of the observed l_z values. This provided 64.38% of the observations *above* the upper critical value in the SH case when 36 items were manipulated.

Spuriously Low (SL) case

The pattern for the SH case is similar to that of the SL case. Figure 3 shows the plots of θ versus the average of $ln(1.0 - p_i(\theta))$ minus $\overline{l}$ where the a and c parameters remained unchanged as the b parameter was assigned uniformly over the different ranges. In this SL case, the $ln(1.0 - p_i(\theta))$ minus $\overline{l}$ is applied, rather than the $ln(p_i(\theta))$ minus $\overline{l}$ since the aberrant incorrect responses were to be detected. Hence, the plots are skewed on the reverse slopes of the SH case.

Nonetheless, the results coincide with the same analysis as in the SL case when these plots are compared to the Group I values in Table 3 (SL Case with θ known).

Out of the possible b parameters reported, the detection rate of the l_z statistic is at the poorest under b ~U(2,2.7) when θ =1.5 (2.36%, 1.89%, 1.33%). The highest ability, θ =2.5, shows reasonable detection power. The remaining values under b ~U(2,2.7) are around the Type I error rate of 0.05.

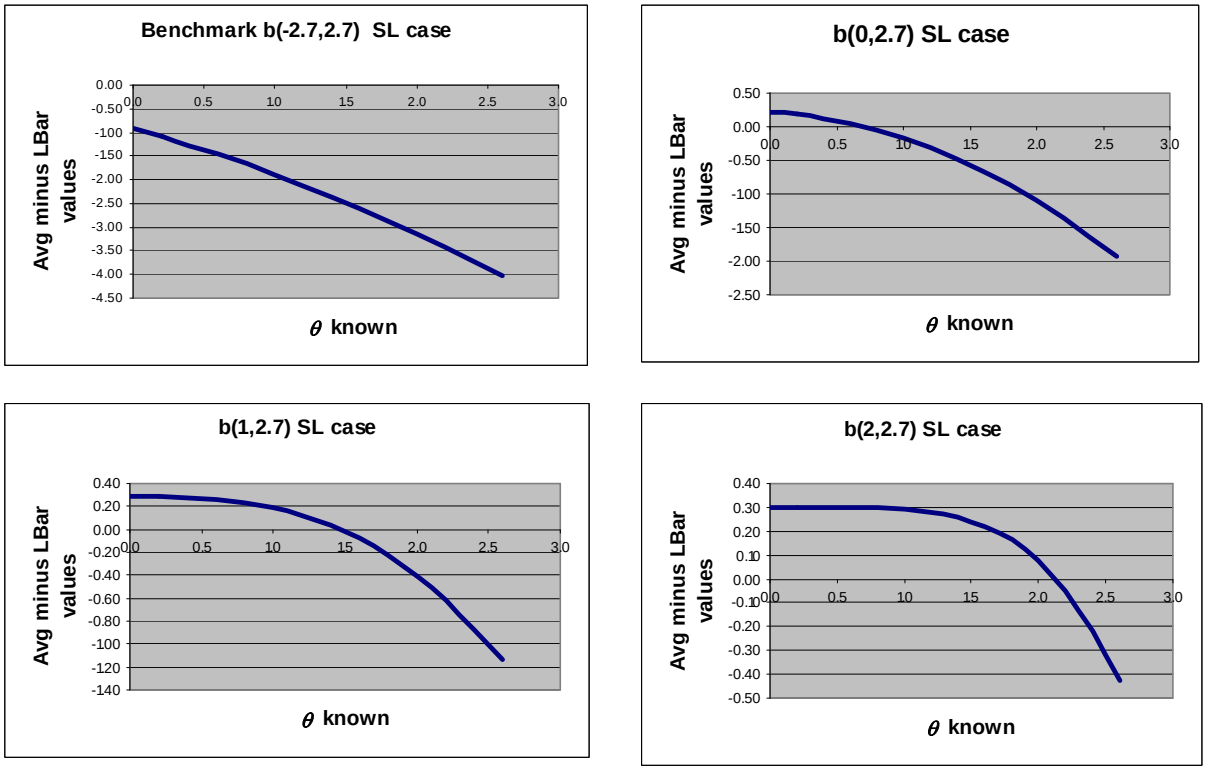

Figure 3. The θ versus the average of $ln(1.0-p_i(\theta))$ minus $\bar{l}$ (SL case).

Referring to Figure 3, the poorest detection can be seen when θ is between 0.0 and 2.0. The plot shows that the detection rate improves after θ =2.0 when it changes direction from positive to negative. Other similarities of the pattern can be matched between the detection rate and the change in direction of the plot under b ~U(1,2.7) when $\theta \leq 1$ for both situations. The study demonstrated that under certain values of θ, $0.0 \leq \theta \leq 2.0$, the l_z fails to give the signals for detection. The plots of Figure 3 do not follow the detection rates of Table 3 as closely as observed when comparing Figure 2 and Table 2. The reason has to do with the method of creating aberrant behavior in the simulation. A manipulated item was made correct with probability 0.2; thus, incorrect responses in the FRV may have changed to correct responses in the NRV. In some cases, this caused the value of l_z to be less than the critical value. If all manipulated responses were made incorrect, the detection results mirrored the results for the SL case.

Groups II and III (Different Normal Distributions)

The results reported for Groups II and III were anticipated. Different normal distributions N(0.5,0.1) and N(1.5,0.1) were examined for the discrimination parameter a, while keeping the parameters b and c unchanged from the benchmark. The detection rates increased when the distribution was shifted up and decreased when shifted down. In Group III, the normal distributions N(0.1,0.1), and N(0.25,0.1) were examined for c, while keeping the parameters a and b as in the benchmark. All the c values were left truncated at zero to ensure a non-negative value. The detection rates increased for lower c values and decreased for higher ones. These patterns stay consistent with the results reported by Reise & Due (1991) for the 2PL model where the parameter c equals 0.

θ Unknown and Estimated

Prior knowledge of an examinee may allow use of a θ value not derived from the current test. However, in many cases a θ estimate based on responses to the current test must be used to compute the statistic. Denote this estimate by $\hat{\theta}$. Several authors have observed the reduced detection power of the l_z statistic when $\hat{\theta}$ is used in place of θ. Some of the reduced power comes from the bias present in the estimator (Snijders (2001)), but most of the reduction comes from the shift coming from the aberrant responses. For example, if the θ known value is −1.0 and aberrant correct responses occur, the $\hat{\theta}$ may be around 0.0. This, in turn, shifts the probabilities used in calculating the statistic and makes the results appear less aberrant. The following analysis reveals additional problems with the use of $\hat{\theta}$ in certain situations.

The same random seeds were used for the simulations when using θ and $\hat{\theta}$ to calculate the l_z statistic. Thus, the sequence of in control responses and the location of items that were manipulated are directly comparable across corresponding cells of tables.

The simulations calculated the critical values as before using the true θ, but $\hat{\theta}$ was used in the calculation of the observed l_z statistic of equation (2). The maximum likelihood estimate (MLE) was taken as $\hat{\theta}$. A reduction in detection power was noted when comparing Tables 4 and 5 to Tables 2 and 3. The function maximized when calculating the MLE is the same as the function providing the l_0 statistic (Equation (1)). Since a lower one-sided hypothesis test was used, the reduction in the power of test when using the MLE is understandable. Most of the detection percentages were under 5.0% when no items were manipulated.

Spuriously High (SH) Case

A comparison of Tables 2 and 4 shows a general improvement in detecting aberrant behavior using θ instead of $\hat{\theta}$ when computing the l_z statistic. However, as θ became closer to 0.0, several values in Table 4 are larger than the corresponding values in Table 2. The improvement in detection when using $\hat{\theta}$ comes from the nature of the constructed test; that is, the range of the b values was spread evenly across the θ scale and the items had high discrimination. When responses were manipulated under the SH case, we had $\hat{\theta} > \theta$ but the true θ was used to determine the non-manipulated response pattern. Non-manipulated responses for items with low b were more likely to show aberrant behavior with $\hat{\theta}$ and the manipulated responses for items with high b still showed almost as much aberrancy because of the high discrimination. Incorrect responses to easy items reduce l_z more than correct responses to difficult items –because the 3PL allows an examinee at any ability level to guess a correct response with at least probability c. However, the probability of a correct response can be arbitrarily close to 1.

The detection results where the difficulty ranges were narrowed show even more loss of power than when θ was used to calculate l_z. The results for Groups II and III essentially mirror that discussed in the previous paragraph for the benchmark case.

Spuriously Low (SL) Case

A comparison of Tables 3 and 5 shows a consistent improvement in detecting aberrant behavior when using θ instead of $\hat{\theta}$ when computing the l_z statistic. This was particularly true at the lower θ where the detection power was sometimes more than doubled when using θ instead of $\hat{\theta}$.

The in-control (0 items manipulated) detection rates found in Table 4 and 5 were greater than the specified Type I error rate only for the SL case when b values were assigned between +2.0 and +2.7 and θ = 0.0, 0.5, and 1.0. Generally, the $\hat{\theta}$ during an in-control replication was close to θ. However, when b was between +2.0 and +2.7 and θ = 0.0, 0.5, and 1.0, the test was so difficult for this group that the average $\hat{\theta}$ fell almost one unit below θ. This caused the detection rates of the in-control case to be slightly greater than the Type I error.

Summary

The l_z person-fit statistic is regarded as one of the most powerful person-fit statistics in the psychological measurement literature. A number of previous studies have shown its effectiveness in detecting item-person misfitting patterns, which is partly manifested in our present study as well. However, our simulation results also showed that the l_z statistic is likely to produce a biased

hypothesis test under certain conditions. The purpose of this paper is not to destructively challenge the well established concept that the l_z statistic is one of the best person-fit statistics in the literature, but to point out when l_z provides unstable detection of aberrant behavior.

In the spuriously high (SH) case where responses were manipulated to make them with probability 0.8 and incorrect with probability 0.2, the l_z statistic shows a consistently poor performance when the trail level of θ is between −0.5 and 0.0. In addition, under the level of difficulty b ~U(−2.7,−1) and b ~U(−2,−1) with θ equal to −1.5, l_z fails to detect aberrant behavior. Likewise, this pattern is observed in the same range of $\hat{\theta}$, the maximum likelihood estimator (MLE) of θ .

Table 4. Detection Rates for a 121-Item Test under SH with θ Estimated

θ	Items manipulated	Benchmark*	Group I			Group II		Group III	
			b ~ U(-2.7,0)	b ~ U(-2.7,-1)	b ~ U(-2.7,-2)	a ~ N(.5,.1)	a ~ N(1.5,.1)	c ~ N(.1,.1)	c ~ N(.25,.1)
-2.5	18	87.87%	48.41%	20.56%	1.49%	51.37%	95.07%	98.89%	74.72%
	24	97.26%	66.24%	28.19%	0.84%	70.00%	99.44%	99.92%	91.79%
	36	99.98%	86.57%	35.97%	0.28%	89.40%	100%	100.00%	99.57%
	0	4.17%	3.37%	3.52%	3.11%	3.34%	4.46%	3.51%	4.30%
-2.0	18	80.89%	32.17%	7.18%	0.02%	44.55%	90.22%	96.94%	67.05%
	24	94.39%	43.32%	7.69%	0.00%	60.25%	98.46%	99.62%	86.07%
	36	99.70%	63.18%	8.44%	0.00%	81.57%	99.99%	100.00%	98.45%
	0	4.34%	4.27%	3.18%	0.03%	3.72%	4.13%	4.21%	4.40%
-1.5	18	69.47%	14.22%	0.73%	0.00%	34.99%	80.84%	92.14%	54.73%
	24	87.34%	19.39%	0.72%	0.00%	48.41%	94.81%	98.56%	74.89%
	36	98.76%	28.25%	0.68%	0.00%	67.81%	99.85%	99.97%	95.22%
	0	4.66%	3.55%	0.78%	0.00%	4.68%	4.29%	4.54%	4.05%
-1.0	18	56.05%	4.34%	0.02%	0.00%	25.98%	69.27%	83.21%	40.78%
	24	75.12%	5.45%	0.00%	0.00%	35.04%	87.52%	94.33%	60.85%
	36	94.99%	6.79%	0.00%	0.00%	51.34%	99.18%	99.65%	87.81%
	0	3.75%	1.63%	0.02%	0.00%	3.91%	3.79%	4.05%	3.38%
-0.5	18	41.59%	0.97%	0.00%	0.00%	16.88%	55.53%	67.34%	29.87%
	24	59.70%	0.82%	0.00%	0.00%	22.88%	76.63%	83.46%	44.16%
	36	85.43%	0.89%	0.00%	0.00%	32.99%	96.27%	97.10%	73.20%
	0	3.44%	0.76%	0.00%	0.00%	3.58%	3.36%	4.23%	3.20%
0.0	18	28.36%	0.12%	0.00%	0.00%	10.58%	41.67%	48.42%	19.56%
	24	41.63%	0.06%	0.00%	0.00%	13.12%	60.83%	65.20%	29.67%
	36	66.56%	0.09%	0.00%	0.00%	17.66%	88.23%	85.93%	51.42%
	0	3.69%	0.2%	0.00%	0.00%	3.12%	3.65%	4.34%	2.85%

* b ~ U(-2.7,2.7) a ~N(1.0, 0.1) c ~N(0.2, 0.1)

Table 5. Detection Rates for a 121-Item Test under SL with θ Estimated

θ	Items manipulated	Benchmark*	Group I			Group II		Group III	
			b ~ U(0,2.7)	b ~ U(1,2.7)	b ~ U(2,2.7)	a ~ N(.5,.1)	a ~ N(1.5,.1)	c ~ N(.1,.1)	c ~ N(.25,.1)
0.0	18	52.81%	4.72%	1.00%	5.85%	19.80%	81.70%	63.42%	47.37%
	24	66.72%	5.55%	1.03%	5.82%	26.82%	91.70%	78.94%	59.99%
	36	83.11%	7.13%	1.23%	5.61%	40.40%	97.75%	93.34%	75.23%
	0	3.51%	3.31%	0.59%	6.28%	2.80%	3.57%	4.14%	2.67%
0.5	18	74.79%	8.55%	1.67%	8.10%	25.74%	94.92%	81.83%	70.90%
	24	87.61%	9.78%	1.75%	7.54%	36.71%	98.76%	92.87%	83.78%
	36	96.06%	13.23%	2.24%	7.40%	55.57%	99.86%	98.82%	92.65%
	0	2.93%	4.77%	1.29%	9.69%	2.13%	3.12%	3.98%	2.36%
1.0	18	90.33%	14.19%	3.55%	7.89%	34.51%	98.89%	92.94%	88.33%
	24	96.86%	18.25%	4.21%	8.27%	49.02%	99.83%	98.24%	95.57%
	36	99.39%	26.66%	5.25%	7.01%	70.39%	99.99%	99.88%	98.84%
	0	2.72%	4.46%	2.56%	9.07%	1.55%	3.02%	3.65%	1.95%
1.5	18	97.26%	22.81%	9.87%	1.20%	43.91%	99.94%	98.26%	96.29%
	24	99.42%	31.88%	12.32%	1.76%	61.54%	99.99%	99.70%	99.17%
	36	99.96%	49.64%	15.30%	2.59%	81.78%	100.00%	99.99%	99.87%
	0	2.27%	2.63%	5.08%	0.72%	0.93%	2.62%	3.16%	1.82%
2.0	18	99.38%	37.39%	12.05%	5.24%	53.18%	99.98%	99.59%	99.18%
	24	99.95%	53.56%	18.99%	5.31%	72.62%	100.00%	99.96%	99.91%
	36	100.00%	75.64%	31.08%	4.51%	90.48%	100.00%	100.00%	99.98%
	0	1.42%	0.83%	1.52%	4.21%	0.45%	2.03%	2.36%	1.12%
2.5	18	99.92%	52.40%	9.91%	4.78%	61.45%	100.00%	99.97%	99.81%
	24	100.00%	73.83%	17.62%	8.13%	80.49%	100.00%	100.00%	99.99%
	36	100.00%	92.66%	44.62%	17.94%	95.25%	100.00%	100.00%	100.00%
	0	5.69%	2.20%	3.50%	0.00%	0.05%	0.80%	1.05%	0.51%

* b ~ U(-2.7,2.7) a ~N(1.0, 0.1) c ~N(0.2, 0.1)

Many of the tests where the l_z statistic performed poorly over a θ group are poorly designed tests for that group. Nevertheless, the results indicate potential problems when using l_z. The results presented here are consistent with the final conclusion of Riese & Due (1991). The practical use of l_z requires items on the test to have a difficulty range covering the trait levels of the population trait. In a further study of the l_z statistic, simulations were performed with tests assembled from an operational pool for the Law School Admission Test (LSAT) with all LSAT constraints satisfied. The results were consistent with the results reported here for the benchmark test. However, the reduction in detection power when using $\hat{\theta}$ was more pronounced because there were fewer items in the LSAT (101 vs. 121), the difficulty parameter values were not so evenly spread, and the discrimination parameters tended to be less. We report the results based on the simulation

approach of Nering and Meijer (1998) to allow replication by other researchers.

It is possible to improve the detection power of the l_z statistic when an estimate of θ has to be used. The MLE $\hat{\theta}$ should be replaced with another estimate because $\hat{\theta}$ maximizes the l_0 statistic. In selected cases, an overall improvement in detection power was obtained by using an expectation of a Bayesian posterior with a standard normal prior. More complicated procedures may also be considered. For example, removing suspicious responses before computing an estimate for θ or computing $\hat{\theta}$ over a certain portion of the test can be applied.

The l_z statistic is usually calculated using all the items on a test. As pointed by a referee, research on the l_z statistic toward the detection of test responses from examinees that have an unfair prior knowledge of the test content may lead to aberrance in response patterns to indicate cheating and item theft. The examinee can be expected to miss items where they did not cheat, but should have been easy items for the examinee given an artificially high $\hat{\theta}$. In this context, an upper-tailed test may be performed. High index values indicate aberrance rates for high-score tests above and beyond the expected levels. These high values indicate elevated levels of cheating. Brain dumps provide many but not all items on the web. Cheaters with access to these sites will have unfair and undetectable advantage on these items. There would be an inconsistent fit with the items they did not see in advance and (not surprisingly) responded incorrectly. Exploring the l_z statistic over incorrect items has potential to identify cheating behavior.

References

Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building.* New York. John Wiley & Sons, Inc.

Bracey, G. & Rudner, L.M. (1992). Person-fit statistics: high potential and many unanswered questions. *Practical Assessment, Research & Evaluation*, 3(7), [Available online: http://PAREonline.net/getvn.asp?v=3&n=7].

Childs, R.A., Elgie, S., Gadalla, T. Traub, R. & Jaciw, A.P. (2004). IRT-linked standard errors of weighted composites. *Practical Assessment, Research & Evaluation*, 9(13), [Available online: http://PAREonline.net/getvn.asp?v=9&n=13].

Drasgow, F. (1982). Choice of test models for appropriateness measurement. *Applied Psychological Measurement, 6,* 297-308.

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.

Drasgow, F., Levine, M.V. & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M.V. & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81,* 461-470.

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

Lehmann, E.L. (1986). *Testing statistical hypothesis*. New York, NY. John Wiley.

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.

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. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

MacCann, R.G. & Stanley, G. (2006). The use of Rasch Modeling to improve standard setting. *Practical Assessment, Research & Evaluation*, 11(2) [Available online: http://PAREonline.net/getvn.asp?v=11&n=2].

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

Meijer, R. R., Molenaar, I. W. & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111-120.

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

Molenaar, I.W. & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9,* 27-45.

Nering. M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

Nering. M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21, 2,* 115-127.

Nering, M. L. & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22, 1,* 53-69.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19, 3,* 213-229

Reise, S. P. & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217-226.

Rudner, L.M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20, 1,* 16-19.

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331-342.

Notes

This work was partly funded by the Law School Admission Council (LSAC). The opinions and conclusions contained in this publication are those of the authors and do not necessarily reflect the position or policy of LSAC.

The authors wish to thank Dr. Lawrence M. Rudner, editor, for his insights and invaluable suggestions.

Citation

Armstrong Ronald D., Stoumbos, Zachary G., Kung, Mabel T., & Shi, Min (2007). On the Performance of the l_Z Person-Fit Statistic *Practical Assessment Research & Evaluation*, 12(16). Available online: http://pareonline.net/getvn.asp?v=12&n=16

Authors

Ronald D. Armstrong and Zachary G. Stoumbos
Department of Management Science & Information Systems
Rutgers Business School
Rutgers University
Piscataway, NJ 08854-8054

Mabel T. Kung
Department of Information Systems & Decision Sciences
College of Business and Economics
California State University at Fullerton
Fullerton, CA 92834

Min Shi
Department of Finance, Real Estate and Law
College of Business Administration
California State Polytechnic University, Pomona
Pomona, CA 91768

Address correspondence to:

Professor Mabel T. Kung
Department of Information Systems & Decision Sciences
College of Business and Economics
California State University at Fullerton
800 North State College Blvd.
PO Box 6848, LH 540
Fullerton, CA 92834
mkung@fullerton.edu
(714) 278-2221
(714) 278-5940 (FAX)