Comparison Among New Residual-based Person-Fit Indices and Wright's Indices for Dichotomous Three-Parameter IRT Model with Standardized Tests

Rashid Al-Mehrzi
Sultan Qaboos University, Oman
mehrzi @squ.edu.om

Wright's residual-based person fit indices were the first person fit indices with dichotomous IRT model and commonly used with Rasch model software. Although there were number of studies which suggested modifications to improve the statistical properties of the Wright's indices, they remained to lack good statistical properties.The study presented a new person fit index and how it can be interpreted and applied for detecting person misfit. Moreover, through a simulated data, the study investigated the statistical properties and the power rates of the new index and compared it with Wright's indices. Results showed that the new index had superior statistical properties under different test conditions and overcome the Wright's index.

Keywords: Item response theory, Person fit, Residual difference, standardized tests.

مقارنة معاملات مقترحة لملائمة استجابة الفرد المعتمدة على البواقي مع معاملات رايت في النموذج الثلاثي لنظرية الاستجابة الثنائية للمفردة في الاختبارات المقننة

راشد المحرزي
جامعة السلطان قابوس

تعد معاملات ملائمة استجابة الفرد المعتمدة على البواقي للعالم رايت من أوائل معاملات ملائمة استجابة الفرد في نماذج نظرية الاستجابة الثنائية للمفردة والتي يشيع استخدامها في برامج تحليل الاختبارات حسب نموذج راش. وبرغم من ظهور العديد من الدراسات التي قدمت تعديلات لتحسين الخصائص الإحصائية لمعاملات رايت، إلا أنها ما زالت تفتقر إلى الخصائص الإحصائية المرغوبة. قدمت الدراسة الحالية معاملا جديدا لملائمة الفرد وكيفية تفسيره واستخدامه لكشف الأفراد ذوي الاستجابات غير الملائمة. كما قارنت الدراسة - من خلال بيانات محاكاة- الخصائص الإحصائية للمعامل المقترح ومعدلات قوته للكشف عن الأفراد ذوي الاستجابات غير الملائمة مع معاملات رايت. أظهرت النتائج أن المعامل المقترح حصل على خصائص إحصائية أفضل تحت ظروف اختباريه مختلفة بدرجة أعلى عن معاملات رايت.

الكلمات المفتاحية: نظرية الاستجابة الثنائية للمفردة، ملائمة استجابة الفرد، الفرق الباقي، الاختبارات المقننة.

.

There are many testing circumstances in which both developers and users of standardized tests might question an examinee's test score. For example, an examinee who is unfamiliar with a new item format might do badly on these items. In addition, students who perform well on multiple-choice items and simultaneously perform badly on constructed response items might raise the question of whether these students are using test-taking strategies or even cheating on the multiple-choice items. Students who have reading difficulty might do badly on a group of items measuring language ability besides arithmetic ability on an arithmetic test. On a reading test, an examinee might do badly on some reading passages because he/she is unfamiliar with the topics of the reading passages. In all of these circumstances and others, test developers consider such responses to be unacceptable and raise concerns about the validity of the students' scores (Meijer, Muijtjens, & Van der Vleuten, 1996).

Many methods have been proposed to obtain information from an examinee's response pattern across test items (Al-Mahrazi, 2003; Meijer & Sijtsma, 2001). The methods used for understanding response patterns, both expected and unexpected, are known as person fit indices or appropriateness measurement indices. In an IRT context, these methods focus on investigating whether the item responses of an examinee are congruent with the expectation of performance ascribed to the model used for calibrating test data. The response patterns for the majority of examinees tend to conform to expectations based on overall test performance and item interrelationships. However, unexpected response patterns do occur and must be examined and understood if the examinees' scores are to be maximally useful.

Wright's (1977) mean square index is one of such person fit indices and has been the focus of a fair number of research studies designed to both understand its utility as well as enhance its applicability (George, 1979; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Reckase, 1981; Smith, 1991, 2000; Smith, Schumacker, & Bush, 1998). Within the framework of Rasch measurement, where the index was initially proposed, this mean square index was proposed as the central method for assessing data fit to the Rasch model. Wright (1977) proposed two versions of the mean square index: an unweighted and a weighted total-fit mean square. Harnisch and Tatsuoka (1983) applied these mean square indices to a three-parameter logistic model and showed how these indices could be adopted for any dichotomous IRT model.

The interest in Wright's (1977) index is understandable given the popularity of the Rasch model and the usefulness of a residual approach in assessing data fit to a given model in the measurement field and other fields. Almost all available software packages for Rasch model calibration (WINSTEPS, BIGSTEPS, FACETS, QUEST, and RUMM) utilize these mean square indices for assessing both model fit and person fit. However, many researchers raised a number of issues with the use of Wright's indices examining fit. Some later researchers (Hambleton et al., 1978; Smith, 1982; Waller, 1981) found both mean square indices are influenced by test lengths and sample size. Waller (1981) argued that Wright and Panchapakesan's index required a large sample size in order to provide precise results. Hambleton et al. (1978) argued against using a large sample size with the mean square index. They said that when the sample size is large, the chi-square test would always show a rejection of the null hypothesis of model fit. Smith (1998) stated that the total-fit mean square is sensitive to sample size and reliance on a single critical value for the mean square can result in an under-detection of misfit.

Many psychometricians continued to raise a number of criticisms associated with the use of these mean square indices (Andersen, 1973; Hambleton et. al., 1978; George, 1979; Gustafsson, 1980; Reckase, 1981; Van den Wollenberg, 1980, 1982; & Wainer, Morgan, & Gustafsson, 1980). For example, George (1979), Hambleton et. al.

(1978), and Reckase (1981) criticized the use of a normal approximation to the binomial distribution of examinee's response to an item. Smith(1998) showed that the empirical distribution was far off the expected theoretical distribution, as a result, using critical values based on the theoretical distribution, the mean square index was insensitive to aberrant item response patterns. Karabatsos (2000) summarized this by noting "this chain-like dependence among the fit indices is problematic: if a fit index does not meet its distributional assumptions for a particular test situation, then other indices dependent on this index will also not meet their distributional assumptions" (p. 162). In spite of the numerous modifications and versions of Wright's index, the concerns regarding its appropriateness continue. Karabatsos (2000) argued "but the fact that these indices need correction indicates that they are flawed to begin with. Therefore, it seems necessary to suggest a few alternatives…" (p. 171).

The focus of this investigation is to develop a modification of Wright's person fit index. This modification is a major one with relative to previous modifications to Wright's person fit index. The study is devoted to deriving and describing a new residual-based person fit index that stems from the total-fit mean square suggested by Wright (1977). The study outlines the derivation and interpretation of two versions of the new residual-based person fit index. The statistical properties of the new person fit index are examined and compared to Wright's mean square indices with simulated data.

Wright's Mean Square Index

Wright and Panchapakesan's (1969) mean square index standardizes the person's observed item score, y_{ij}, which is considered as the variable of interest. They called it as the standardized residual difference for person j's observed score on item i,

$$z_{ij} = \frac{\left(y_{ij} - p_{ij}\right)}{\sqrt{p_{ij} q_{ij}}}, i = 1, 2, \ldots, n. \quad (1)$$

where p_{ij} is the probability of obtaining a correct score on item i by a person j with a given ability value, θ_j, using any IRT model given, and $q_{ij} = 1-p_{ij}$. The z_{ij} score is used as an indication of unexpected responses. Wright and Panchapakesan (1969) argued that these standardized residual difference scores are distributed as standard normal with a mean of zero and a standard deviation of one if the data fit the specified IRT model. Wright (1977) used this z_{ij} score to propose two versions of the mean square index: an unweighted and a weighted total-fit mean square. The unweighted total-fit mean square is,

$$UMS = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{ij}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \frac{\left(y_{ij} - p_{ij}\right)^2}{p_{ij} q_{ij}}. \quad (2)$$

The weighted total-fit mean square is,

$$WMS = \frac{\sum_{i=1}^{n} \left(p_{ij} q_{ij}\right) z_{ij}^2}{\sum_{i=1}^{n} p_{ij} q_{ij}} = \frac{\sum_{i=1}^{n} \left(y_{ij} - p_{ij}\right)^2}{\sum_{i=1}^{n} p_{ij} q_{ij}}. \quad (3)$$

Wright (1977) believed that both mean square indices are useful and needed. Wright and Stone (1979) suggested transformations to both unweighted and weighted mean square indices to remove the effect of sample size. The unweighted mean square index is transformed by a log transformation to as follows,

$$UT = \left[\ln UMS + UMS - 1\right] \sqrt{\frac{n-1}{8}}. \quad (4)$$

The weighted mean square index is transformed by a cube-root transformation,

$$WT = \left(\sqrt[3]{WMS} - 1\right)(3/r) + (r/3), \quad (5)$$

Where $r = \frac{\sqrt{\sum_{i=1}^{n} p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}}{\sum_{i=1}^{n} p_{ij} q_{ij}}$

is the standard deviation of the *WMS* index. Wright and Stone (1979) argued that both *UT* and *WT* scores are distributed as a unit normal with a mean of zero and a standard deviation of one. Large positive values of both *UT* and *WT* indicate aberrant response patterns.

The New Modified Residual-Based Person Fit Indices

The proposed person fit index is similar to Wright's (1977) mean square index for the purpose of person fit analysis in that it employs the residual approach to assess the fit of a person's response pattern. However, the new index uses a simple function of the residual difference between the person's observed response and the probability of correctly answering the item as a measure of the degree of aberrance in a person's response pattern. The square of the residual difference is used as a core for this person fit index. Two versions of the proposed residual-based person fit index are formalized: Unweighted and Weighted.

In the unweighted version, the squared residual difference between the person's observed response and the probability of correctly answering each individual item, SRij, is computed as,

$$SR_{ij} = \left(y_{ij} - p_{ij} \right)^2, \quad = 1, 2, \ldots, n. \quad (6)$$

The values for SRij could take any value that ranges from 0 to 1. The closer the value of SRij is to 1, the less is the correspondence between the person's response and the IRT model prediction and, hence, the more aberrant is the person's response. However, this SRij is not sufficient to detect misfitting person responses because there is no identified value of SRij that can be used to determine whether the person's response is aberrant at any ability value. This squared residual difference can be standardized at any ability value by subtracting from it its expected value and then dividing by its variance. The expected value of SRij, is:

$$E\left(SR_{ij} \right) = p_{ij} q_{ij} = Var\left(y_i \middle| \theta_j \right), \quad i = 1, 2, \ldots, n. \quad (7)$$

and the variance of SRij scores at each ability level is,

$$Var\left(SR_{ij} \right) = p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2, \quad i = 1, 2, \ldots, n. \quad (8)$$

Then, the standardized squared residual index, USRij, for a person's response to an individual item is defined as,

$$USR_{ij} = \frac{SR_{ij} - E\left(SR_{ij} \right)}{\sqrt{Var\left(SR_{ij} \right)}} = \frac{\left(y_{ij} - p_{ij} \right)^2 - p_{ij} q_{ij}}{\sqrt{p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}}$$

$, i = 1, 2, \ldots, n. \quad (9)$

However, Equation 9 can have undefined values in three cases in which the denominator has a value of zero: 1) pij = 0.0, 2) pij = 1.0, 3) pij = 0.5. The first two cases do not typically occur with the logistic IRT models. The last case might exist and, hence, USRij is set to be zero. This fixed value will not affect the performance of the index, because any response (0 or 1) is acceptable by the IRT model with this probability value.

Then, an overall unweighted person fit index across all n items, referred to here as USR, is computed as,

$$USR = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} USR_{ij} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\left(y_{ij} - p_{ij} \right)^2 - p_{ij} q_{ij}}{\sqrt{p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}} \quad (10)$$

Equation (10) can be further simplified as demonstrated in Al-Mahrazi (2003) to,

$$USR = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\left(p_{ij} - y_{ij} \right)\left(p_{ij} - q_{ij} \right)}{\sqrt{p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}} \quad (11)$$

The weighted version of the new person fit index is based on computing the sum of the squared residual differences across all test items, SRj,

$$SR_j = \sum_{i=1}^{n} SR_{ij} = \sum_{i=1}^{n} \left(y_{ij} - p_{ij} \right)^2 \quad . (12)$$

SRj provides test users with a simple measure of the degree of correspondence between the observed person's responses to test items and the prediction of the IRT model. SRj takes on values that range from zero to n. The closer the value of SRj to n, the more aberrant is the person's response pattern. Similarly, this SRj is not sufficient to detect misfitting person responses because there is no identified unique value of SRj that can be used to determine whether the person's response is misfitting at any ability value. This squared residual difference is then standardized at any ability value by subtracting from it its expected value and dividing by its variance. So, the standardized score of the SRj is the weighted version of the new index, and it is referred to here as WSR:

$$WSR = \frac{SR_j - E\left(SR_j\right)}{\sqrt{Var\left(SR_j\right)}} = \frac{\sum_{i=1}^{n} y_{ij} - p_{ij}^{\wedge 2} - \sum_{i=1}^{n} p_{ij} q_{ij}}{\sqrt{\sum_{i=1}^{n} p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}} \quad .(13)$$

Equation (13) can be further simplified to,

$$WSR = \frac{\sum_{i=1}^{n} \left(p_{ij} - y_{ij} \right)\left(p_{ij} - q_{ij} \right)}{\sqrt{\sum_{i=1}^{n} p_{ij} q_{ij} \left(p_{ij} - q_{ij} \right)^2}} \quad (14)$$

If the data fit the IRT model, both USR and WSR are likely to follow a unit normal distribution with a mean of zero and a standard deviation of one. Hence, the values of the USR and WSR indices would be large and positive to indicate that the person is more likely to have an aberrant response pattern. This suggests that a one-tailed significance test (right tail) should be used to evaluate both the USR and WSR indices for the person.

Method

The study examined Wright's index and the new index for three-parameter logistic IRT model with respect to two criteria that are essential for any effective person fit index. These two criteria are: 1) the empirical null distribution of the index matches its hypothetical null distribution and this null distribution is invariant across different test conditions including the ability levels, and 2) the index reliably detects aberrant responses of various types. The properties of the four residual-based person fit indices were examined at seven ability values.

The analyses of the properties of these indices were conducted within each of twelve data sets that resulted from the combinations of the following test conditions: two test lengths, (n = 15 & n = 50), three levels of item difficulty, bi (less difficult, medium difficult, more difficult), and two levels of item discrimination (low ai, high ai). The first level of item difficulty represents tests with easy items, and it is generated from a uniform distribution in the interval U(-3.0, 0.0), the second represents tests with medium difficult items that are generated from a uniform distribution in the interval U(-1.5, 1.5), and the last represents tests with difficult items that are generated from a uniform distribution in the interval U(0.0, 3.0). All intervals of the three levels of difficulty parameters have the same moderate spread. The first level of item discrimination represents tests with items having low discrimination and generated from lognormal distribution (0.6, 0.02), while the second level of item discrimination represents tests with items of high discrimination which is generated using lognormal (1.4, 0.06). The guessing parameters , ci, for all data sets were generated using uniform distribution (0.0, 0.2).

The aberrant responses are simulated using an information-based approach suggested by Reise and Due (1991). The information-based approach involves simulating aberrant responses according to a model in which items are

differentially discriminating for different individuals. One such model is (Reise & Due, 1991) applied to the 3PL IRT model,

$$p_{ij} = Prob\ y_{ij} = 1 \mid \theta_j = c_i + \frac{1 - c_i}{1 + exp[-1.7\, a_p a_i (\theta_j - b_i)]}, \quad i = 1, 2, \ldots, n. \quad (15)$$

where all terms are similar to the terms on the 3PL IRT model, except for the term ap. The ap parameter is the aberrancy level parameter. The ap parameter is different across individuals and might be different or constant across items for a particular individual. In this study, the ap parameter is treated as constant across items for each particular individual.

Two values of ap is used here to simulate aberrant responses. The ap = 1.0 condition represents the case when equation 15 is identical to the 3PL IRT model. The data sets generated using this level of aberrancy represents data sets that fit the 3PL IRT model. At the other aberrancy condition, ap = 0.5, the item responses using equation 15 differ from the responses generated by the 3PL IRT model. This condition represents existence of aberrant responses.

Each simulated data set follows a similar procedure. First, the n, ai, bi, ci, and ap for each data set are specified for the 3PL IRT model. Then, for each data set, using the specified test length and item parameters, 1000 response vectors are generated at each of the seven points on θ scale using the model in Equation 17. The seven true θ values range from –3.0 to 3.0 with an increment of 1.0. This procedure is replicated 50 times. At ap = 1.0, the means, standard deviations, and type I error for all indices were examined. At the other aberrancy condition (ap = 0.5), the power rates of the indices were examined. For each simulated data set, all person fit indices are computed for each simulated response vector using the generated item responses specified in equation 15, while the predictions were based on the 3PL IRT model (Equation 15 when ap = 1.0).

Results

Table 1 through Table 3 present the statistical properties (means, standard deviations, type I error rates) of the four indices; UTa, WTa, USRa and WSRa, for the twelve data sets when data sets fit the 3PL IRT model (ap=1.0). Table 1 revealed that both means and standard deviations of the UT index deviated from their theoretical values at almost all theta values within all data sets. These deviations were larger at theta values that were farther from the difficulty level of each data set. For example, the mean values of the UT index were 0.106, 0.152, 0.118, -0.057, -0.596, -1.762, & -3.464; and the standard deviations were 0.941, 0.604, 0.866, 1.704, 3.067, 4.958, & 7.832, at ability levels of -3, -2, -1, 0, 1, 2, &3 for data set with n = 15, less difficult and low discriminating items. Results showed also that increasing items' discrimination worsens the deviation of the mean and standard deviation values of the UT index at ability levels that are farther from the difficulty level of items. For the same data set, the mean values became -0.096, 0.034, -0.162, -1.176, -3.177, -5.851, & 185.303; and the standard deviation became 1.841, 1.337, 2.456, 7.278, 54.921, 93.320, & 6220.932. Moreover, Lengthening test didn't improve the closeness of the mean and standard deviation of the UT index to zero.

Table 1.
The means and standard deviations of *UT* and *WT* indices under different test conditions.

		UT								*WT*							
		Mean				Standard deviation				Mean				Standard deviation			
		n=15		*n*=50		*n*=15		*n*=50		*n*=15		*n*=50		*n*=15		*n*=50	
b	*q*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*
More difficult	-3	-0.251	-0.483	-0.115	-0.285	2.337	2.883	2.342	3.055	-0.028	-0.050	-0.003	-0.008	1.106	1.125	1.038	1.045
	-2	-0.130	-0.442	-0.086	-0.290	1.964	2.890	1.952	2.887	-0.020	-0.051	-0.005	-0.011	1.073	1.112	1.024	1.045
	-1	0.005	-0.319	-0.007	-0.190	1.444	2.531	1.465	2.524	-0.001	-0.042	-0.003	-0.003	1.029	1.115	1.006	1.038
	0	0.106	-0.077	0.077	-0.081	0.941	1.749	0.967	1.935	0.002	-0.006	0.001	-0.007	1.002	1.018	0.999	1.008
	1	0.153	0.036	0.023	0.004	0.585	1.320	0.604	1.431	-0.001	0.005	-0.002	-0.005	1.005	1.001	1.003	1.006
	2	0.115	-0.247	-0.189	-0.231	0.841	2.491	0.874	2.815	-0.006	-0.002	0.004	0.007	0.993	0.989	0.994	1.003
	3	-0.051	-1.162	-0.055	-1.230	-1.660	7.169	1.798	9.161	0.000	0.005	-0.002	-0.007	1.006	0.981	1.009	1.014
Medium difficult	-3	-0.073	-0.394	-0.046	-0.275	1.710	2.781	1.714	2.816	-0.017	-0.043	-0.011	-0.011	1.058	1.121	1.017	1.045
	-2	0.063	-0.207	0.037	-0.150	1.118	2.237	1.216	2.354	0.003	-0.022	0.005	-0.004	1.008	1.079	1.005	1.021
	-1	0.139	-0.006	0.074	-0.017	0.709	1.483	0.736	1.619	0.002	-0.003	0.005	-0.001	0.991	1.006	0.994	1.003
	0	0.147	-0.014	0.077	-0.029	0.625	1.592	0.656	1.711	0.002	0.005	0.000	0.006	0.999	1.002	1.001	0.997
	1	0.052	-0.520	0.023	-0.587	1.209	3.682	1.301	4.553	0.000	0.003	0.003	0.004	1.000	0.974	0.999	0.994
	2	-0.259	-2.233	-0.189	-2.610	2.316	15.061	2.497	10.747	-0.020	-0.025	-0.002	-0.040	1.052	0.995	1.015	1.123
	3	-1.077	-6.106	-0.750	-8.019	4.070	24.666	4.464	83.827	-0.082	0.832	-0.022	0.037	1.129	0.446	1.086	0.849
Less difficult	-3	0.106	-0.096	0.054	-0.063	0.941	1.841	0.978	2.015	0.002	-0.012	0.006	0.000	0.996	1.020	1.000	1.005
	-2	0.152	0.034	0.082	0.001	0.604	1.337	0.601	1.403	0.002	0.006	0.004	-0.001	0.995	1.004	0.996	1.000
	-1	0.118	-0.162	0.061	-0.219	0.866	2.465	0.879	2.837	0.002	0.006	0.000	0.006	0.992	0.995	1.002	0.996
	0	-0.057	-1.176	-0.062	-1.311	1.704	7.278	1.783	8.089	-0.004	-0.003	-0.003	-0.007	1.008	0.983	1.002	1.015
	1	-0.596	-3.177	-0.391	-5.682	3.067	54.921	3.319	24.426	-0.050	0.274	-0.007	-0.138	1.108	0.706	1.040	1.145
	2	-1.762	-5.851	-1.406	7.708	4.958	93.320	5.957	560.860	-0.091	2.101	-0.066	0.625	1.100	0.223	1.173	0.497
	3	-3.464	185.303	-3.778	188.491	7.832	6220.932	10.251	6102.853	0.023	6.787	-0.169	2.915	0.875	0.065	1.251	0.148

The WT index performed better than the UT index. Both mean and standard deviation values of the WT index were better approximating to their theoretical values at all θ values with comparison to the UT index within all data sets. However, the mean values of the WT index showed deviated mean and standard deviation values at θ values farther from the average difficulty level of test items, and they deviated farther within data sets having high discriminating items. For example, the mean values were -0.091 and -0.023 at θ=2 and 3 for data set with n=50, less difficult and low discriminating items, whereas they were 2.101 and 6.787 at θ=2 and 3 for data set with n=15, less difficult and high discriminating items. The corresponding values for the standard deviation of the WT index were 1.100 and 0.875 for the low discriminating items and 0.223 and 0.065 for the high discriminating items. Moreover, increasing test length to n=50, reduced the deviation of the mean and standard deviation values of the WT index at all θ values within all data sets.

Moreover, Table 1 showed that the existence of guessing in test items improved the performance of both UT and WT indices at low ability values with comparison to high ability values within all data sets. The means and standard deviations were less deviated from their expected values at the low ability values within all data sets. This was especially evident with data sets of more difficult items (where low ability values are farther from the difficulty level of the data set).

On the other hand, results as shown in Table 2 revealed that the USR and WSR indices performed well at almost all ability values within all data sets. The means and the standard deviations of both indices were approximately equal to their theoretical values. The weighted version of the new person fit indices (WSR index) showed extremely closeness of means and standard deviations to their theoretical values at all ability values, even with data set with extreme characteristics. For example, the

WSR index had mean values of -0.006, 0.007, 0.006, -0.002, -0.002, -0.002, & 0.012

and standard deviations of 0.999, 0.999, 1.004, 0.992, 0.998, 0.997, & 1.031 at the ability values of -3, -2, -1, 0, 1, 2, & 3, respectively, for the most extreme data sets (where n=15, less difficult, high discriminating items).

Table 2.
The means and standard deviations of *USR*, and *WSR* indices under different test conditions

		USR								*WSR*							
		Mean				Standard deviation				Mean				Standard deviation			
		n=15		*n*=50		*n*=15		*n*=50		*n*=15		*n*=50		*n*=15		*n*=50	
b	*q*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*
More difficult	-3	0.007	0.000	0.004	0.001	1.007	0.997	1.006	0.999	0.007	0.001	0.004	0.001	1.006	1.001	1.005	0.001
	-2	0.002	-0.001	-0.001	-0.004	1.000	0.998	1.001	0.997	0.001	-0.002	-0.001	-0.002	0.999	0.998	1.001	-0.002
	-1	0.006	0.002	-0.004	0.004	1.002	1.007	0.996	1.002	0.007	0.000	-0.002	0.004	1.002	1.005	0.998	0.004
	0	0.002	-0.001	0.001	-0.007	1.004	1.001	1.000	1.000	0.003	-0.001	0.001	-0.006	1.006	1.001	0.999	-0.006
	1	-0.002	0.006	-0.003	-0.004	1.009	0.995	1.003	1.005	0.000	0.006	-0.002	-0.005	1.007	0.997	1.004	-0.005
	2	-0.006	-0.004	0.004	0.007	1.001	0.994	0.995	1.000	-0.007	-0.004	0.004	0.007	1.001	0.996	0.997	0.007
	3	-0.003	0.003	0.001	0.000	0.995	1.001	1.006	1.012	0.001	0.003	0.001	-0.003	0.994	0.995	1.003	-0.003
Medium difficult	-3	-0.001	0.005	-0.008	-0.003	1.002	1.004	0.997	0.998	-0.001	0.004	-0.009	-0.002	1.001	1.003	0.996	-0.002
	-2	0.004	0.004	0.006	-0.001	0.998	1.004	1.000	0.995	0.005	0.004	0.006	0.000	0.998	1.004	1.001	0.000
	-1	0.003	-0.003	0.004	0.000	1.000	0.996	0.995	0.999	0.001	-0.002	0.005	-0.001	0.996	0.998	0.995	-0.001
	0	0.003	0.006	-0.002	0.005	1.001	1.006	1.002	0.998	0.002	0.006	0.000	0.006	1.002	1.006	1.003	0.006
	1	0.000	-0.001	0.002	0.004	1.004	0.993	1.000	0.996	0.001	-0.001	0.003	0.003	1.006	0.995	1.002	0.003
	2	-0.003	-0.004	0.000	0.007	1.004	1.010	0.997	0.989	-0.003	0.007	0.000	0.007	1.002	0.992	0.997	0.007
	3	-0.004	0.002	0.002	0.013	0.999	0.939	0.993	1.148	-0.004	0.006	0.001	0.002	1.002	1.003	0.994	0.002
Less difficult	-3	0.001	-0.005	0.007	0.002	0.998	1.000	1.002	1.000	0.002	-0.006	0.006	0.001	1.001	0.999	1.001	0.001
	-2	-0.001	0.003	0.004	-0.003	0.998	0.998	0.999	0.997	0.002	0.007	0.004	-0.001	0.997	0.999	0.997	-0.001
	-1	0.003	0.005	-0.002	0.004	1.002	1.000	1.004	1.000	0.002	0.006	0.000	0.005	1.000	1.004	1.004	0.005
	0	-0.002	0.000	-0.004	-0.004	1.001	0.998	0.996	1.002	-0.001	-0.002	-0.003	-0.004	0.999	0.992	0.997	-0.004
	1	-0.004	0.009	0.001	0.000	1.002	1.158	0.994	1.004	-0.004	-0.002	0.001	0.002	1.003	0.998	0.994	0.002
	2	0.006	0.011	-0.002	0.035	0.995	1.178	1.001	1.614	0.006	-0.002	-0.003	0.000	0.998	0.997	0.997	0.000
	3	0.001	0.109	0.005	0.147	1.004	3.884	1.006	4.594	0.001	0.012	0.004	0.002	1.004	1.031	1.004	0.002

However, for the same data set, the USR index had mean values of -0.005, 0.003, 0.005, 0.000, 0.009, 0.011, & 0.109, and standard deviations of 1.000, 0.998m 1.000, 0.998, 1.158, 1.178, & 3.884 at the ability values of -3, -2, -1, 0, 1, 2, & 3, respectively. The USR index had acceptable mean values at all ability values but deviated standard deviations at ability values farther from the difficulty level within those data sets of less difficult and high discriminating items (both n=15 and n=50).

Table 3 present the type I error rates of the four indices for the twelve data sets at α=0.5. As expected from its extremely deviated means and standard deviations, the UT index had high type I error rates at most ability levels for all data sets; especially those with high discriminating items. This high type I error rates of the UT index did not improve as test length increased.

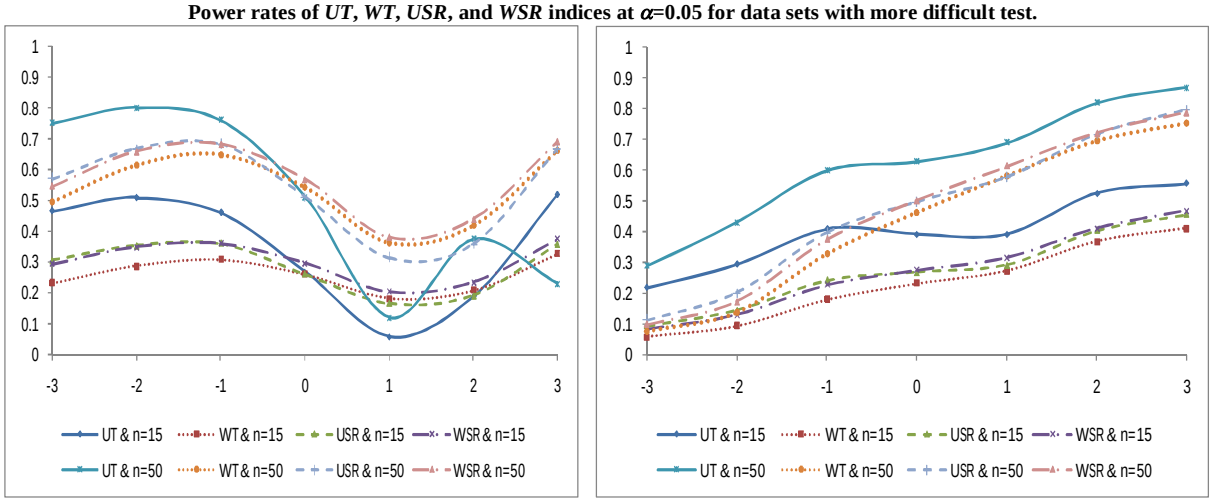
The WT index showed acceptable type I error rates at most ability levels for all data sets. However, it was unable to control type I error at ability values farther from the difficulty level for the high discriminating data sets. For example, it had type I error rates of 0.575, and 1.000 at θ=2 and 3 for the data set of n=15, less difficult and high discriminating items, and 0.081, and 0.961 for the data set of n=50, less difficult and high discriminating items.

Table 3.
Type I error rates of *UT*, *WT*, *USR*, and *WSR* indices at α=0.05 under different test conditions.

		UT				*WT*				*USR*				*WSR*			
		n =15		*n* =50		*n* =15		*n* =50		*n* =15		*n* =50		*n* =15		*n* =50	
b	*q*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*	Low *a*	High *a*
More difficult	-3	0.169	0.170	0.193	0.193	0.043	0.043	0.045	0.045	0.067	0.065	0.060	0.062	0.067	0.065	0.060	0.060
	-2	0.157	0.176	0.171	0.194	0.044	0.043	0.045	0.043	0.065	0.066	0.058	0.061	0.065	0.063	0.058	0.059
	-1	0.123	0.176	0.125	0.190	0.047	0.046	0.047	0.045	0.063	0.066	0.056	0.061	0.064	0.065	0.056	0.060
	0	0.063	0.138	0.057	0.150	0.053	0.047	0.051	0.047	0.059	0.061	0.055	0.056	0.064	0.063	0.057	0.058
	1	0.011	0.109	0.008	0.119	0.053	0.049	0.051	0.049	0.056	0.060	0.053	0.056	0.062	0.064	0.057	0.058
	2	0.048	0.125	0.044	0.169	0.051	0.050	0.050	0.051	0.054	0.060	0.054	0.059	0.062	0.065	0.057	0.061
	3	0.140	0.106	0.155	0.158	0.048	0.049	0.048	0.048	0.060	0.056	0.058	0.059	0.064	0.071	0.058	0.062
Medium difficult	-3	0.143	0.167	0.149	0.192	0.044	0.043	0.043	0.043	0.063	0.068	0.056	0.059	0.063	0.069	0.055	0.057
	-2	0.093	0.157	0.093	0.174	0.480	0.046	0.048	0.044	0.059	0.066	0.055	0.059	0.062	0.066	0.056	0.058
	-1	0.030	0.114	0.023	0.131	0.520	0.049	0.050	0.050	0.055	0.060	0.052	0.056	0.062	0.064	0.056	0.058
	0	0.015	0.127	0.013	0.147	0.052	0.052	0.051	0.049	0.058	0.062	0.053	0.056	0.062	0.066	0.056	0.058
	1	0.100	0.114	0.109	0.171	0.051	0.051	0.051	0.052	0.060	0.059	0.055	0.061	0.065	0.067	0.059	0.062
	2	0.165	0.100	0.194	0.141	0.046	0.053	0.045	0.044	0.065	0.051	0.059	0.057	0.064	0.077	0.058	0.067
	3	0.187	0.064	0.224	0.099	0.048	0.131	0.042	0.069	0.068	0.040	0.064	0.045	0.070	0.072	0.061	0.080
Less difficult	-3	0.063	0.133	0.058	0.153	0.051	0.047	0.051	0.049	0.056	0.062	0.055	0.059	0.063	0.063	0.058	0.059
	-2	0.013	0.114	0.009	0.112	0.049	0.049	0.051	0.049	0.055	0.061	0.054	0.055	0.059	0.064	0.056	0.058
	-1	0.054	0.115	0.046	0.171	0.052	0.052	0.052	0.050	0.055	0.062	0.054	0.061	0.063	0.067	0.059	0.059
	0	0.150	0.116	0.154	0.151	0.048	0.048	0.046	0.048	0.062	0.057	0.055	0.059	0.064	0.068	0.056	0.063
	1	0.182	0.079	0.217	0.120	0.044	0.082	0.043	0.047	0.069	0.044	0.062	0.049	0.068	0.088	0.060	0.077
	2	0.183	0.023	0.231	0.066	0.048	0.575	0.041	0.081	0.070	0.022	0.066	0.036	0.083	0.023	0.064	0.081
	3	0.155	0.004	0.212	0.012	0.060	1.000	0.047	0.961	0.066	0.004	0.069	0.011	0.070	0.004	0.061	0.012

Table 3 showed that both USR and WSR indices were able to control type I error rates closer to its nominal value at all ability values within all data sets. Similar to the WT index, both USR and WSR indices were unable to control type I error at the two most extreme high ability levels (θ=2 & 3) for the less difficult and high discriminating data set. However, the type I error rates for these two indices were deflated as opposite to the case with the WT index of which the type I error rates were inflated. The type I error rates were 0.022 & 0.004 for the USR index and 0.023 & 0.004 for the WSR index at θ=2 & 3 for the data set with n=15, less difficult and high discriminating items. The type I error rates were 0.036 & 0.011 for the USR index and 0.081 & 0.012 for the WSR index for the data set with n=50, less difficult and high discriminating items.

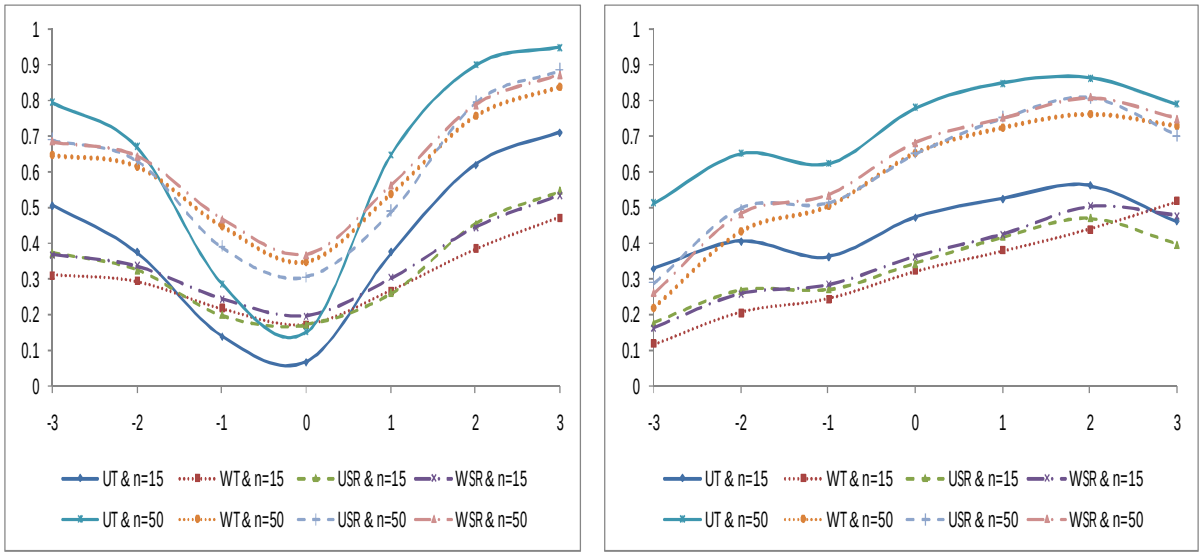
Figure 1.
Power rates of *UT*, *WT*, *USR*, and *WSR* indices at α=0.05 for data sets with more difficult test.



Low *a* **High *a***

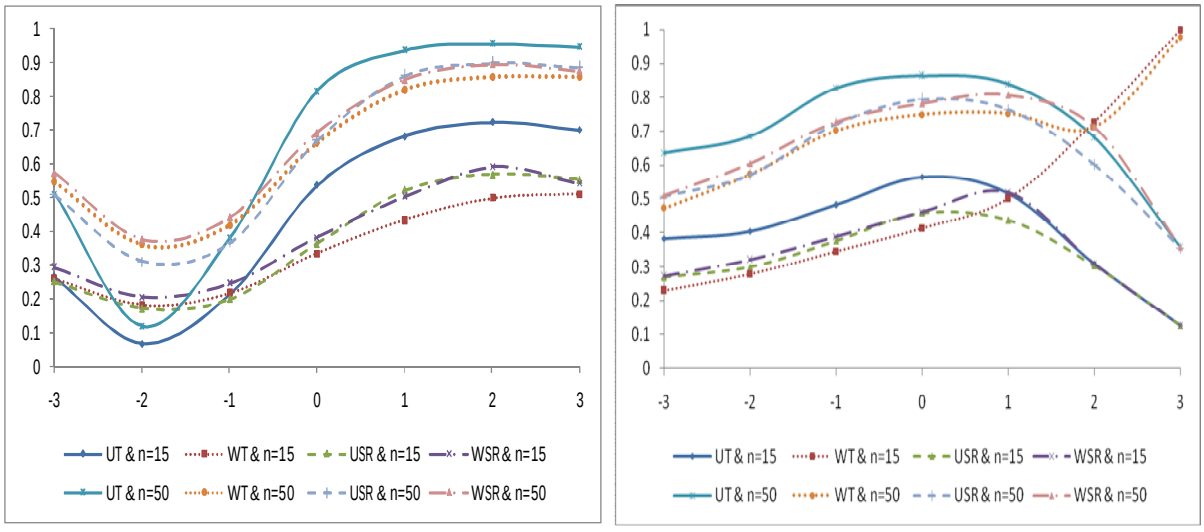
Figure 1 through figure 3 present the power of the four indices for the twelve data sets with the aberrancy level of ap=0.5. The power rates of the UT index were higher than all other indices at almost all test conditions as a result of the inflated type I error rates, and hence, it is not considered in further discussion. The WT index and the new indices (USR and WSR) showed similar power rates at most ability levels within all data sets. The WSR index had a slightly higher power rates than the WT index at most ability levels. The USR index had less power rates than WSR index at ability levels closer to the difficulty level of the data set with low discriminating items. For the high discriminating data set, the power rates of the USR index were closer to the WSR index.

Figure 2.
Power rates of *UT*, *WT*, *USR*, and *WSR* at α=0.05 for data sets with medium difficult test



Low *a* **High *a***

Figure 3.
Power rates of UT, WT, USR, and WSR at α=0.05 for data sets with less difficult test.



All indices had different power rates pattern with different levels of item discrimination. For the data sets with low discriminating items, the power rates of all person fit indices were higher at ability levels that were farther from the difficulty levels than at ability levels that were closer t to the difficulty level of the data set (see

the left graph on the three figures). However, for the data sets with high discriminating items, the power rates of the four indices were low at low ability levels (power was less than 0.15) and they were increasing as the ability levels increased from -3 to 3 for the more difficult and medium difficult data sets. However, for the high discriminating and less difficult data sets, the power rates of the four indices were high at low ability levels and it was slightly increased as the ability levels approached medium ability. At the high ability values, the power rates of the WT index jumped higher and reached 1.000. However, for the other indices, the power rates dropped down and became low at high ability values. In addition, results shown in the three figures revealed that the power rates of the four indices were higher within data sets with longer test (n=50) than within data sets with shorter test (n=15).

Discussion and Conclusions

The results of the study revealed that the new indices (USR and WSR) performed well in terms of means and standard deviations at all ability values within all data sets. The USR index had acceptable mean values at all ability values but deviated standard deviations at only those ability values which were farther from the difficulty level within those extreme data sets (less difficult and high discriminating items, both n=15 and n=50). In addition both USR and WSR indices were able to control type I error rates closer to its nominal value at all ability values within all data sets. Exceptions of that were at the two most extreme high ability levels (θ=2 & 3) for the less difficult and high discriminating data set. The type I error rates for these USR and WSR indices were deflated (less than 0.05). on the other hand, the WT index had inflated type I error rates. Although the USR and WSR indices were unable to control type I error rates at these conditions, the small type I error rates shown by them have less consequences on the detection of aberrant responses in real testing as compared to the high type I error rates with the WT index. This is because the positive error decision resulted from the USR and WSR indices (person fits while he/she is not) has less price with comparison with the negative error decision resulted from the WT index (person misfits while he/she fits).

Moreover, the two new indices showed good power rates. The WSR index had similar and even slightly higher power rates than the WT index at most ability levels within all data sets. The USR index had less power rates than WSR index at ability levels closer to the difficulty level of the data set with low discriminating items. This indicates that the USR index is less sensitive to person misfit at ability levels that are closer to the difficulty level of the data set. However, for the high discriminating data set, the power rates of the USR index were closer to the WSR index. This could lead to say that having high discriminating items on the test increases the power rates of the person fit indices and make them give similar results.

Although the two versions of the new person fit index proposed here and Wright's person fit indices are similar in that they both use the residual difference between the observed and expected person's item responses, the new person fit index differs procedurally from Wright's person fit index in several ways. First, the proposed person fit index standardizes a squared transformation of the residual difference between a person's observed response and the expected probability for each individual item. This squared transformation can assume any value between zero and one. On the other hand, Wright's index standardizes the person's response. The person's item response is dichotomous, i.e., either zero or one. Hence, the variable of interest that is standardized in the new indices can be considered as a continuous variable, whereas the variable of interest that is standardized in Wright's index is a discrete variable. This structure of the new person fit indices suggests that these new person fit indices could address the criticisms that are raised with Wright's indices regarding the use of the normal approximation to the binomial

distribution of a person's responses to dichotomous items and the use of the Pearson chi-square as a distribution of Wright's mean square indices. In addition it is expected that the new person fit indices should be less influenced by the typical problems associated with sample size as supported by the results of this study. Both USR and WSR showed superior statistical properties when data fits the IRT model and similar or even better power of detecting aberrant responses as demonstrated by this simulation study.

Moreover, both USR and WSR indices are straightforward indices which require only standardizing the squared residual difference which is a stable quantification of aberrant responses. On the other hand, the UT and WT are chain-like dependence indices since they require standardizing the person's response to item, squaring it, summed it, and transforming it to follow a unit normal distribution. This simplicity of the new indices is an advantage and provides test users with easy interpretation and understanding of the causes of aberrance in person's responses.

References

AL-Mahrazi, R. S. (2003). *Investigating a new modification of the residual-based person fit index and its relationship with other indices in dichotomous Item Response Theory*. Unpublished doctoral dissertation. University of Iowa.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

George, A. A. (1979). *Theoretical and practical consequences of the use of standardized residual as Rasch model fir statistics*. Paper presented at the annual meeting of American Educational Research Association, San Francisco, CA.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*. *33*, 205-233.

Hambleton,R. K., Swaminathan, H., Cook,L. L., Eignor, D. R. & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research, 48*, 467-510.

Harnisch, D. L. & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory (pp 104-122). In R.K Hambelton (ED.), *Applications of item response theory*. Vancouver, Canada: Kluwer

Karabatsos, G (2000). A Critique of Rasch Residual Fit Statistics. *Journal of Applied Measurement*; 1 1, 152-178.

Meijer, R. R., Muijtjens, A. M., & Van der Vleuten, C. P. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9*, 77-90.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new development. *Applied Measurement in Education, 8*, 261-272

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating person fit. *Applied Measurement in Education, 25*, 107-135

Reckase, M. D. (1981). *The validity of latent trait models through the analysis of fit and invariance*. Paper presented at the annual meeting of American Educational Research Association, Loss Angeles, CA.

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

Smith, R. M. (1982). *Detecting measurement disturbances with the Rasch model*. Unpublished Doctoral Dissertation. University of Chicago.

Smith, R. M. (1991). The distribution properties of Rasch standardized residuals. *Educational and Psychological Measurement, 51*, 541-565.

Smith, R. M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Smith, R. M. (2000).Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1,* 199-218.

Smith, R. M. Schumacker, R.E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Van den Wollenberg, A. L. (1980). *On the Wright-Panchapakesan goodness of fit test for the Rasch model*. Intern Rapport 80-MA-02, K.U. Nijmegen.

Van den Wollenberg , A. L. (1982). Two new test statistics for the Rasch model, *Psychometrika, 47,* 123-140.

Wainer, H., Morgan, A., & Gustafsson, J. E. (1980). A review of estimation procedures for the Rasch model with a view towards longer test. *Journal of Educational Statistics, 1*, 35-69.

Waller, M. I. (1981). A procedure for comparing logistic latent trait models. *Journal of Educational Measurement, 18*, 119-125.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-115.

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

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