Journal of Educational and Psychological Studies - Sultan Qaboos University (687-702) Vol.13 Issue 4, 2019
DOI: http://dx.doi.org/10.24200/jeps.vol13iss4pp687-702

# Residual-Based Person Fit Statistics over Test Sections

Rashid Almehrizi*
Sultan Qaboos University, Sultanate of Oman

Received: 10/4/2019 Accepted: 9/7/2019

**Abstract:** Most tests are composed of multiple sections (each section has group of items) such as different item formats, different content category, competencies, different difficulty levels, test dimensions, testlets, and interpretive exercise items. Students could show unexpected and unacceptable responses across these sections. Studying person fit over item level cannot detect aberrant response over test sections. The study proposes a residual-based person fit statistic over test sections with a dichotomous IRT model. The paper demonstrates the new section-level person fit statistic and investigates its distributional properties and power of detecting aberrance in person responses with comparison to Wright's between person fit statistic. The proposed section-level person fit statistic shows superior distributional properties with both true and real ability and item parameters. Moreover, the performance of the proposed person fit statistic is also examined with real data.

**Keywords:** Person fit, section-level, residual approach, dichotomous IRT models.

## مؤشرات ملاءمة الفرد القائمة على منهج البواقي عبر مستويات أقسام الاختبار

راشد المحرزي*
جامعة السلطان قابوس، سلطنة عمان

**مستخلص:** تتكون معظم الاختبارات من أقسام مختلفة (القسم الواحد يضم مجموعة من أسئلة الاختبار) وفقا لنوع الأسئلة أو موضوعات المحتوى أو القدرات المقاسة، أو مستوى صعوبة الأسئلة أو الأبعاد أو تجمعات الأسئلة كأسئلة القراءة والأسئلة التفسيرية المشتركة بمتن واحد. وقد يظهر بعض الطلبة استجابات غير متوقعة وغير مقبولة عبر هذه الأقسام والتي قد لا تستطيع مؤشرات ملاءمة الفرد على مستوى المفردة الواحدة أن تكشفها. وقدمت هذه الدراسة مؤشرين لملاءمة الفرد باستخدام البواقي على مستوى أقسام الاختبار باستخدام نماذج استجابة المفردة ثنائية الاستجابة. كما تحققت الدراسة من الخصائص الاحصائية لهذين المؤشرين وقدرتهما على كشف الاستجابات غير المطابقة للأفراد بالمقارنة مع مؤشرين آخرين يتم استخدامهما لنفس الغرض وهما مؤشرا ملاءمة الفرد البيني لرايت. أظهرت النتائج أن المؤشرين الجديدين يمتلكان خصائص إحصائية متميزه سواء باستخدام المعالم المولدة أو المقدرة للمفردات ولقدرات الأفراد. كما أظهرت نتائج التطبيق على بيانات حقيقية أداء جيدا للمؤشرين.

**الكلمات المفتاحية**: ملاءمة الفرد، أقسام الاختبار، منهج البواقي، نماذج استجابة المفردة ثنائية الاستجابة.

*mehrzi@squ.edu.om

Test developers desire to have ways and tools to detect examinees with aberrant responses and finding the reasoning behind their unexpected responses to increase test fairness and validity. Wright (1977) classified these types of responses as systematic aberrant responses. Such systematic aberrant responses might be a result of an interaction between the person and item difficulties, the person and different contents, the person and different sections measuring different dimensions, and others. Methods used for these purposes are known as person fit or appropriateness measurement.

Several advancements in person fit in item response theory era are found in the literature review for different uses. Felt, Castaneda, Tiemensma, and Depaoli (2017) proposed a person fit statistic to detect outliers in survey research using graded response model. Similarly, Fox. and Marianti, (2017) proposed person-fit statistics for joint IRT models to detect aberrant response accuracy and/or response time patterns. The person-fit tests take the correlation between ability and speed into account, as well as the correlation between item characteristics.

Meijer and Sijtsma (2001) provided an excellent methodology review for the person fit methods both based on group characteristics and item response theory (IRT). In dichotomous IRT context, there are number of person fit methods that have proposed and investigated. Among these are Wright's mean square statistics (Wright, 1977), Caution indices (Tatsouka & Linn, 1983), $l_z$ statistic (Drasgow et. al; 1985), Optimal person fit statistic (Levine, & Drasgow, 1988), and Person Response Function (Trabin & Weiss, 1983). All of these methods are used for general detection of person misfit to the expectation of the IRT models without providing test users with much about the causes of such person misfit (Meijer, & Sijtsma, 2001).

Although most of the existing person fit methods evaluate the person fit on item-level response patterns, there are circumstances where the interest of test developers focuses on person's response patterns over sections of items such as different test formats, test content topics, test dimensions, testlets, interpretive exercise items or any other characteristics of sections of items. For, example, an examinee not familiar with the test format could obtain a lower score than expected, or a person familiar with a task given in the interpretive-exercise might get a higher score than expected.

Students could show unexpected and unacceptable responses across these sections. Studying person fit over item level cannot detect aberrant responses over test sections. There are two statistics that exist on literature for investigating person fit on sections of items are the between standardized mean square statistics introduced by Smith (1985) and the $l_{zm}$ statistic introduced by Drasgow, Levine, and Mclaughlinm (1991) for multidimensional test batteries.

The between standardized mean square statistic is an extension of the total standardized mean square developed by Wright and Panchapakesan (1969) and Wright (1977). The statistic has unweighted and weighted versions that both examine the squared standardized residual difference of the person's scores over the sections. Both versions are evaluated as $\chi^2$ distribution and then transformed through certain transformation equation to follow a unit normal distribution although there is no sufficient theoretical justification for these transformations.

In the context of item-fit, Smith (1985; 1991; 1994; 1996) found that the between standardized mean square statistic showed poor distributional properties. Both the mean and standard deviation of this statistic deviated from their theoretical and expected values. Smith also found that the amount of the deviation of the mean and standard deviation increased as the number of the groups of examinees decreased. Given that the

between standardized mean square statistic is evaluated as a Pearson chi-square test, the number of groups of examinees in item fit has a large influence on the type I error rate and the power of the statistic (Smith, 1991; 1996).

In addition, previous research argued that the legitimacy of the use of the unit normal approximation for the transformed mean squares is heavily dependent on the degree to which the mean square is approximately distributed as a Pearson chi-square distribution. If the mean square statistic deviated from the chi-square distribution, it is no longer accurate to assume that the transformation is distributed as a unit normal distribution. For example, George (1979), Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978), and Reckase (1981) questioned the validity of the chi-square significant test for the mean square statistic through criticizing the use of a normal approximation to the binomial distribution of person responses to dichotomous items. They argued that as the probability of a correct response departs from 0.5, the distribution of observed frequencies becomes less and less normal; thus, the mean square statistics are no longer distributed as a Pearson chi-square. They argued that this issue becomes more pronounced when the test is composed of a small number of items. This chain-like dependence among the fit statistics is problematic. If a fit statistic does not meet its distributional assumptions for a particular test situation, then other statistics depending on this statistic will also not meet their distributional assumptions.

The other between person fit is the $l_{zm}$ statistic which is applied for multidimensional test batteries (Drasgow, Levine, & Mclaughlinm, 1991). This statistic assumes that different sections measure different dimensions; each section measures one dimension. These multidimensional sections could be independent or correlated. $l_{zm}$ statistic evaluates the person fit within each unidimensional section using the corresponding ability levels, and then accumulate that over these sections. Meijer and Sijtsma (2001) noted that this $l_{zm}$ is not very different from the $l_z$ statistic for long unidimensional test. They stated that "although was effective in detecting misfitting item-score patterns, detection rates were approximately equal to those for long, unidimensional tests with s number of items equaling the total number of items in the *S* sections". Thus the $l_{zm}$ statistic is equal to the $l_z$ statistic when all sections measure unique unidimension and the grouping of items into sections are based on other bases like item format, difficulty, reading passages, or others.

In summary, because of the poor distributional properties of the existing between standardized mean square statistic, the section-level person fit has received less attention. At the same time, the existing item-level person fit statistic could not be helpful where the aberrance is related to the characteristics of the section more than individual items.

In this paper, a new section-level (between) person fit statistic that employs residual difference approach is introduced. The statistical properties of the new section-level person fit statistic investigated and compared to Wright's between person fit statistic using simulated data.

**The Between Standardized Mean Square Statistic**

According to smith (1985), the between standardized mean square statistic standardizes person *a*'s scores on each section, *j*, that consists of $n_j$ items (*i*=1, 2, .., $n_j$),

$$z_{aj} = \frac{x_{aj} - \sum_{i=1}^{n_j} p_{ai}}{\sqrt{\sum_{i=1}^{n_j} p_{ai} q_{ai}}}, \quad j = 1,2,\dots,J \quad (1)$$

where $x_{aj} = \sum_{i=1}^{n_j} u_{ai}$, *J* is the number of sections, $u_{ai}$ is person's response on each item *i* on each set, *j*, $p_{ai}$ is the probability

that person with a certain level of ability ($\theta$) correctly answers each item, $i$, on a section $j$ based on any unidimensional IRT model, and $q_{ai} = 1 - p_{ai}$. Wright (1977) formed two versions to evaluate the square of $z_{aj}$: unweighted and weighted versions. Both versions can be evaluated as a chi-square test with a degree of freedom of one. The unweighted version is

$$UMS_a = \frac{1}{J-1}\sum_{j=1}^{J} z_{aj}^2 = \frac{1}{J-1}\sum_{j=1}^{J}\left(\frac{x_{aj}-\sum_{i=1}^{n_j} p_{ai}}{\sqrt{\sum_{i=1}^{n_j} p_{ai}q_{ai}}}\right)^2$$

$$= \frac{1}{J-1}\sum_{j=1}^{J}\frac{\left(x_{aj}-\sum_{i=1}^{n_j} p_{ai}\right)^2}{\sum_{i=1}^{n_j} p_{ai}q_{ai}}, \quad (2)$$

and the weighted version is,

$$WMS_a = \frac{J}{J-1}\frac{\sum_{j=1}^{J}\left(x_{aj}-\sum_{i=1}^{n_j} p_{ai}\right)^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} p_{ai}q_{ai}}, \quad (3)$$

Both versions are then transformed to standardize the mean square statistic to an approximate unit normal distribution. The transformation of the unweighted version is,

$$UBT_a = \left(\sqrt[3]{UMS_a} - 1\right)\left(\frac{3}{s}\right) + \left(\frac{s}{3}\right), \quad (4)$$

where $s = \sqrt{\frac{2}{J-1}}$.

The transformation of the weighted version is,

$$WBT_a = \left(\sqrt[3]{WMS_a} - 1\right)\left(\frac{3}{s}\right) + \left(\frac{s}{3}\right), \quad (5)$$

where $s = \sqrt{\frac{2}{J-1}}$.

**The New Section-Level Person Fit Statistic**

The new section-level person fit statistic is an extension of the approach developed by Al-Mehrzi (2004; 2010) for the item-level residual-based person fit statistic. The basic theme of the new section-level person fit statistic is based on estimating the amount of congruence on person's section score, and then examining whether this amount of congruence could be explained by the used IRT model or it might indicate aberrant section scores. The amount of congruence on person's score on each section score, $x_{aj}$, can be estimated by the squared residual difference, $BSR_{aj}$,

$$BSR_{aj} = \left(x_{aj} - \sum_{i=1}^{n_j} p_{ai}\right)^2. \quad (6)$$

$BSR_{aj}$ can take any values between zero and $n_j^2$. Across all ability levels, the $BSR_{aj}$ can take any continuous value that ranges between zero and $n_j^2$. This estimate of congruence cannot be used for classifying the persons's responses as model-fitting or misfitting because $BSR_{aj}$ depends on the person's ability levels. Then the distribution for this statistic under null distribution of fitting section scores is required. To overcome this issue, a standardized version of the $BSR_{aj}$ can be suggested through computing the expected score and variance of $BSR_{aj}$. Under the independent assumption of item responses with IRT, the expected score of the $BSR_{aj}$ is the known IRT variance of subtotal score for each section of items given the ability, i.e.,

$$E(BSR_{aj}) = \sum_{x_j=0}^{n_j}\left(x_j - \sum_{i=1}^{n_j} p_{ai}\right)^2 f(x_j|\theta_a)$$

$$= \sum_{i=1}^{n_j} p_{ai}q_{ai}, \quad (7)$$

Similarly, the variance of the squared residual difference, BSRaj can be obtained by the typical equation of variance, i.,e.,

$$Var(BSR_{aj}) = \sum_{x_j=0}^{n_j}\left[\left(x_j - \sum_{i=1}^{n_j} p_{ai}\right)^2 - \sum_{i=1}^{n_j} p_{ai}q_{ai}\right]^2 f(x_j|\theta_a)$$

$$= \sum_{x_j=0}^{n_j}\left(x_j - \sum_{i=1}^{n_j} p_{ai}\right)^4 f(x_j|\theta_a) - \left[\sum_{i=1}^{n_j} p_{ai}q_{ai}\right]^2 \quad (8)$$

where $x_j$ is all possible scores on each set of items, $j$, and $f(x_j \mid \theta_a)$ is the PDF of the $x_j$ scores given examinee $a$'s ability. This PDF can be obtained through the Lord and Wingersky (1984) recursion formula using $p_{ai}$ which was employed by Almehrizi (2013; 2016).

To get $f(x_j|\theta_a)$ through the recursion formula, define $x_j$ as the random variable of raw scores on the first $i$ items on the test ($x_j$ ranges between 0 and 1). Now, let $f(x_{ij} = x_{ij}|\theta_{ai})$ represent the probability mass function of $x_j$ when it is equal to $x_j$ on a test of $i$ items. For a

test of one item, $i = 1$ is entered into the formula,

$f(x_1 = 0|\theta_{a1}) = p_{a1}(0)$ and

$f(x_1 = 0|\theta_{a1}) = p_{a1}(0)$.

For the next $i > 1$, the recursion formula is as follows:

$f(x_i = x_i|\theta_{ai}) = f(X_{i-1} = x_i|\theta_{a(i-1)})p_{ai}(0) + f(X_{i-1} = x_i - 1|\theta_{a(i-1)})p_{ai}(1)$,

for $x_i = 0, 1, \ldots i$. (9)

To use this recursion formula, enter items into the recursion formula in any order, beginning with $i = 1$, and repeatedly apply the formula by increasing $i$ on each repetition. The process is stopped after $i = n_j$, which gives the required $f(x_j|\theta_a)$. That is, $f(x_j|\theta_a) = f(X_{n_j} = x_{n_j} = x|\theta_a)$.

Now, the standardization version of the squared residual difference across all sections can be obtained through two ways. The first ways require standardizing the $BSR_{aj}$ at each section level, then summing across all sections and finally dividing by the square root of the number of sections, $J$, as follows (it will be referred as unweighted version):

$$UBSR = \frac{1}{\sqrt{J}}\sum_{j=1}^{J}\frac{BSR_{aj} - E(BSR_{aj})}{\sqrt{Var(BSR_{aj})}}, \quad (10)$$

The other way of constituting the standardization version of the squared residual difference across all sections is through accumulating the amount of congruence in person's scores across all sets of items by summing the squared residual differences across test items, referred to as

$BSR_a = \sum_{j=1}^{J} BSR_{aj}$

and then standardizing it using its expected score, $E(BSR_a)$, and its variance, $Var(BSR_a)$. Under the independence assumption of the section scores in IRT,

$E(BSR_a) = \sum_{j=1}^{J} E(BSR_{aj}) = \sum_{i=1}^{n} p_{ai}q_{ai}$,

and $Var(BSR_a) = \sum_{j=1}^{J} Var(BSR_{aj})$.

i.e.,

$$WBSR = \frac{BSR_a - E(BSR_a)}{\sqrt{Var(BSR_a)}}, \quad (11)$$

where all terms are defined as before. If the used IRT model fits test data, it is hypothesized that both versions of the new section-level person fit statistic is distributed theoretically as a unit normal distribution. Extreme scores of the section-level person fit statistic in both tails of the unit normal distribution are considered as aberrant response patterns. Z-value of 1.96 represents 5% level of significance for thee hit rate for detecting aberrant response patterns.

Two studies were conducted to investigate the new section-level person fit statistic. The first study aimed to investigate the distributional properties (including mean, standard deviation, type I error, and power) of the new section-level person fit statistic and to compare them with the distributional properties of Wright's person fit mean square statistic with simulation data using true ability and item parameters. The effect of increasing either the number of sections, the number of items per section, or both on the distributional properties for these section-level person fit statistic also is investigated. The second study aimed to investigate the distributional properties (including mean, standard deviation, type I error, and power) of the new section-level person fit statistic and Wright's person fit mean square statistic with simulation data using real ability and item parameters. In addition, the performance of both statistics in detecting misfitted person responses was examined with real data.

**Study 1: Simulation with True Parameters**

## Method

Two types of simulated data were used to examine the distributional properties of the new section-level person fit statistic and to compare them with those of

Wright's between person fit mean square statistic. Data sets that fit the 3PL IRT model were simulated to examine the means, standard deviations, and type I error rates at a level of significance of 0.05 for both statistics. Then data sets that represent measurement disturbance were simulated to examine the power of the two statistics. The true item parameters for all data sets were generated as follows: $a_i$~lognormal (1.0, 0.04); $b_i$~Uniform (-2.5,2.5), $c_i$~Uniform (0.0,0.2). The generation of all data sets on this study was replicated 60 times.

For the ordinary samples, nine data sets were generated from varying two factors: the number of sections of items ($J$=2, 4 or 8), and number of items per section ($n_j$=5, 10, or 20). The total numbers of items in these nine data sets were: 10, 20, 40, 20, 40, 80, 40, 80, and 160. This arrangement of the nine data sets was employed to examine the effect of both increasing the number of sections of items and increasing the number of items within each section on the distributional properties of the two section-level person fit statistic.

The distributional properties of the two statistics were examined conditioned on seven ability levels within each of the nine data sets. They were -3, -2, -1, 0, 1, 2, and 3. For each data set, 1,000 response patterns were produced at each ability level using random number generators (with different seeds) from uniform distribution. Each random number was compared to the conditioned probability of a correct answer to each item. If the random number was larger than or equal to the conditioned probability of a correct answer to the item, the item response was set as 1, and vice versa.

For the aberrant samples, there are two arrangements. The first Arrangement consisted of thirty six data sets that were generated from varying four factors: the number of sections of items ($J$=2, 4 or 8), the number of items per section ($n_j$=5, 10, or 20), the type of aberrance (spuriously high or spuriously low), the severity of aberrance (mild or moderate). Eighteen of the aberrant samples had spuriously high response patterns, and the remaining samples had spuriously low response patterns. Spuriously high response patterns were created by generating ordinary response patterns as described for the ordinary sample, and then replacing randomly a given percentage of simulated responses with correct responses for each section on the data set separately. Spuriously low response patterns were created by generating ordinary response patterns and then replacing randomly a fixed percentage of items for each section with random responses (choosing randomly a correct option on multiple-choice item with four options). The mild aberrant response patterns were created using 20% of items per section (i.e., 1 of 5, 2 of 10 and 4 of 20). The moderate aberrant response patterns were created using 40% of items per section (i.e., 2 of 5, 4 of 10 and 8 of 20). Spuriously high response patterns were generated for low ability levels ($\theta$=-3, -2, -1), whereas spuriously low response patterns were generated for high ability levels ($\theta$=1, 2, 3).

The second arrangement of aberrant samples was generated to examine the effect of location of aberrant response patterns across sections of the test on the power of the section-level person indices. Eighteen data sets were generated from varying three factors: the number of sections of items ($J$=2, 4 or 8), the type of aberrance (spuriously high or spuriously low), the location of aberrant response patterns (within one section, evenly within half of the sections, and evenly within all sections). All of these eighteen data sets consisted of 40 total number of items and have 4 items with aberrant responses (the severity of aberrance is 10%). Spuriously high and spuriously low response patterns were created as described with the first arrangement of aberrant samples. There were three locations of aberrant response patterns across sections of the test. The first location assumes that the

four items with aberrant responses were observed on only one section, while other sections have no aberrant response patterns. The second location assumes that the four items with aberrant responses were divided evenly within half number of sections (e.g., with *J*=4, two sections, each with 2 items). The third location assumes that the four items with aberrant responses were divided evenly within all sections (e.g., with *J*=4, each section had one item with aberrant responses).

## Results

### Distributional characteristics

Table 1 presents the means of the four section-level person fit statistic for the nine ordinary samples. The results revealed that the means of both versions of Wright's statistic deviated to similar extent from zero at all ability levels within all data sets. The means of both Wight's statistics were around 0.7 for data sets with *J*=2, and were around 0.4 for data sets with *J*=4, and were around 0.2 for data sets with *J*=8. The means of Wright's statistics approached zero as number of sections increased. Results showed also that these mean values were not affected by increasing the number of items per section. For example, at *J*=2, the means of the unweighted version of Wright's statistic, *UBT*, at the seven ability levels were 0.765, 0.749, 0.746, 0.746, 0.743, 0.741, 0.200 when $n_j$=5, and they were 0.762, 0.745, 0.742, 0.741, 0.741, 0.761, 0.585 when $n_j$=10, and they were 0.742, 0.741, 0.741, 0.739, 0.742, 0.741, 0.766 when $n_j$=20. It can be noted that mean values of both Wright's statistics at ability level of 3 were different from the mean values at other ability levels with $n_j$=5 (the mean values for *UBT* were 0.200 with *J*=2, -0.282 with *J*=4, -0469 with *J*=8). However, they became similar to the mean values at other ability levels with $n_j$=20. This can be explained by the very low probability of a correct answer at this ability level given the range of true item difficulty parameter (ranged from -2.5 to 2.5).

**Table 1**
**Means of the four section-level person fit indices for various test lengths**

| | | *UBT* | | | *WBT* | | | *UBSR* | | | *WBSR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *J* | $\theta$ | $n_j$=5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 |
| | -3 | 0.765 | 0.762 | 0.742 | 0.760 | 0.761 | 0.742 | 0.000 | 0.004 | 0.002 | -0.001 | 0.004 | 0.002 |
| | -2 | 0.749 | 0.745 | 0.741 | 0.748 | 0.745 | 0.741 | -0.001 | -0.002 | -0.002 | -0.001 | -0.002 | -0.002 |
| | -1 | 0.746 | 0.742 | 0.741 | 0.745 | 0.741 | 0.741 | -0.004 | 0.000 | 0.001 | -0.004 | 0.000 | 0.001 |
| 2 | 0 | 0.746 | 0.741 | 0.739 | 0.745 | 0.740 | 0.738 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 1 | 0.743 | 0.741 | 0.742 | 0.738 | 0.738 | 0.741 | -0.003 | 0.000 | 0.000 | -0.003 | 0.001 | 0.000 |
| | 2 | 0.741 | 0.761 | 0.741 | 0.751 | 0.750 | 0.736 | 0.002 | 0.002 | -0.004 | 0.002 | 0.002 | -0.005 |
| | 3 | 0.200 | 0.585 | 0.776 | 0.286 | 0.630 | 0.777 | -0.005 | 0.000 | -0.002 | -0.006 | 0.000 | -0.003 |
| | -3 | 0.442 | 0.424 | 0.419 | 0.438 | 0.423 | 0.419 | 0.004 | -0.005 | -0.003 | 0.004 | -0.005 | -0.003 |
| | -2 | 0.434 | 0.425 | 0.428 | 0.433 | 0.424 | 0.427 | 0.000 | -0.003 | 0.003 | 0.000 | -0.004 | 0.003 |
| | -1 | 0.428 | 0.425 | 0.431 | 0.427 | 0.425 | 0.431 | -0.001 | -0.002 | 0.008 | -0.001 | -0.002 | 0.008 |
| 4 | 0 | 0.425 | 0.424 | 0.430 | 0.424 | 0.423 | 0.429 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 | 0.009 |
| | 1 | 0.426 | 0.423 | 0.423 | 0.422 | 0.419 | 0.421 | 0.002 | -0.002 | 0.001 | 0.002 | -0.002 | 0.001 |
| | 2 | 0.389 | 0.428 | 0.426 | 0.415 | 0.416 | 0.421 | 0.002 | -0.002 | -0.002 | 0.000 | -0.001 | -0.001 |
| | 3 | -0.282 | 0.203 | 0.393 | -0.091 | 0.281 | 0.398 | -0.002 | -0.001 | -0.002 | -0.001 | -0.003 | -0.001 |
| | -3 | 0.270 | 0.272 | 0.276 | 0.269 | 0.271 | 0.276 | -0.004 | -0.001 | 0.005 | -0.004 | -0.001 | 0.005 |
| | -2 | 0.278 | 0.280 | 0.278 | 0.276 | 0.279 | 0.278 | -0.003 | 0.005 | 0.004 | -0.004 | 0.005 | 0.004 |
| | -1 | 0.278 | 0.276 | 0.275 | 0.278 | 0.276 | 0.275 | -0.003 | 0.002 | 0.002 | -0.002 | 0.002 | 0.002 |
| 8 | 0 | 0.271 | 0.279 | 0.275 | 0.270 | 0.278 | 0.274 | -0.005 | 0.006 | 0.002 | -0.006 | 0.005 | 0.002 |
| | 1 | 0.271 | 0.275 | 0.274 | 0.268 | 0.272 | 0.273 | -0.001 | 0.001 | 0.002 | -0.001 | 0.001 | 0.003 |
| | 2 | 0.221 | 0.273 | 0.273 | 0.257 | 0.267 | 0.266 | 0.002 | -0.001 | -0.002 | 0.003 | -0.001 | -0.003 |
| | 3 | -0.469 | 0.038 | 0.214 | -0.174 | 0.121 | 0.220 | -0.003 | -0.003 | 0.003 | -0.001 | -0.001 | 0.003 |

Unlike Wright's statistic, both versions of the proposed statistic had mean values very close to zero at all ability levels within all nine data sets as showed in Table 1. This was evident even within data set with the smallest number of sections and the smallest number of items per section ($J$=2, $n_j$=5). The mean values were 0.000, -0.001, -0.004, 0.000, -0.003, 0.002, -0.005 for the unweighted version, and -0.001, -0.001, -0.004, 0.000, -0.003, 0.002, -0.006 for the weighted version at the seven ability levels, respectively. The same patterns of the mean values of the proposed statistic were observed with different number of sections and number of items per section.

Table 2 presents the standard deviations of the four section-level person fit statistic for the nine ordinary data sets. For all data sets, the standard deviations of both versions of Wright's statistic were not equal to one at almost all ability levels. The standard deviations were more deviated from one with data set with small number of sections ($J$=2) and became less deviated as number of sections increased ($J$=4, $J$=8). For example, for data set with $J$=2 and $n_j$=5, the standard deviations of *WBT* were 0.825, 0.854, 0.852, 0.856, 0.859, 0.830, and 1.202 for the seven ability levels, respectively. Although the standard deviations of Wright's statistic improved with tests with more number of sections, they were still deviated form one at some ability levels especially those that were distant from average item difficulty. For example, at ability levels of 2 and 3 for the data set with $J$=8 and $n_j$=5, the standard deviations of *UBT* were 1.099 and 2.157, respectively.

Moreover, increasing the number of items within each section had almost little or no effect on improving the deviation of the standard deviations for both versions of Wright's statistic.

**Table 2**
**Standard deviations of the four section-level person fit indices for various test lengths**

| | | *UBT* | | | *WBT* | | | *UBSR* | | | *WBSR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $J$ | $\theta$ | $n_j$=5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 |
| | -3 | 0.817 | 0.822 | 0.862 | 0.825 | 0.824 | 0.862 | 1.004 | 1.014 | 1.003 | 1.002 | 1.014 | 1.003 |
| | -2 | 0.854 | 0.852 | 0.859 | 0.854 | 0.853 | 0.859 | 0.998 | 1.000 | 0.993 | 0.997 | 1.001 | 0.993 |
| | -1 | 0.851 | 0.860 | 0.865 | 0.852 | 0.860 | 0.865 | 0.996 | 1.003 | 0.999 | 0.997 | 1.004 | 0.999 |
| 2 | 0 | 0.855 | 0.861 | 0.866 | 0.856 | 0.862 | 0.866 | 0.990 | 0.998 | 0.998 | 0.989 | 0.999 | 0.998 |
| | 1 | 0.852 | 0.862 | 0.861 | 0.859 | 0.866 | 0.863 | 0.994 | 1.003 | 0.992 | 0.993 | 1.002 | 0.993 |
| | 2 | 0.831 | 0.827 | 0.853 | 0.830 | 0.846 | 0.858 | 1.005 | 1.006 | 0.999 | 1.002 | 1.007 | 0.998 |
| | 3 | 1.238 | 0.976 | 0.742 | 1.202 | 0.933 | 0.739 | 0.980 | 1.007 | 0.980 | 0.974 | 1.002 | 0.973 |
| | -3 | 0.892 | 0.919 | 0.936 | 0.901 | 0.921 | 0.937 | 1.012 | 0.999 | 1.002 | 1.011 | 0.999 | 1.002 |
| | -2 | 0.915 | 0.927 | 0.936 | 0.918 | 0.929 | 0.937 | 1.006 | 0.994 | 0.997 | 1.006 | 0.994 | 0.997 |
| | -1 | 0.928 | 0.931 | 0.940 | 0.929 | 0.932 | 0.940 | 1.006 | 0.995 | 1.004 | 1.006 | 0.994 | 1.003 |
| 4 | 0 | 0.935 | 0.939 | 0.946 | 0.937 | 0.941 | 0.947 | 1.002 | 0.997 | 1.004 | 1.001 | 0.997 | 1.004 |
| | 1 | 0.933 | 0.934 | 0.941 | 0.944 | 0.943 | 0.945 | 1.000 | 0.993 | 1.001 | 1.001 | 0.995 | 1.001 |
| | 2 | 0.981 | 0.913 | 0.927 | 0.938 | 0.944 | 0.940 | 1.003 | 1.007 | 0.992 | 0.998 | 1.008 | 0.994 |
| | 3 | 1.733 | 1.255 | 0.945 | 1.604 | 1.136 | 0.939 | 1.008 | 0.991 | 0.978 | 1.009 | 0.982 | 0.981 |
| | -3 | 0.953 | 0.965 | 0.973 | 0.957 | 0.967 | 0.974 | 0.998 | 0.999 | 1.001 | 0.997 | 0.999 | 1.001 |
| | -2 | 0.941 | 0.963 | 0.967 | 0.946 | 0.965 | 0.969 | 1.000 | 1.009 | 1.000 | 1.000 | 1.008 | 1.001 |
| | -1 | 0.946 | 0.966 | 0.970 | 0.948 | 0.967 | 0.971 | 0.997 | 1.004 | 1.003 | 0.997 | 1.004 | 1.002 |
| 8 | 0 | 0.962 | 0.969 | 0.971 | 0.963 | 0.971 | 0.972 | 0.997 | 1.002 | 1.003 | 0.996 | 1.002 | 1.003 |
| | 1 | 0.971 | 0.967 | 0.972 | 0.982 | 0.977 | 0.978 | 0.998 | 1.000 | 1.004 | 0.998 | 1.001 | 1.004 |
| | 2 | 1.099 | 0.956 | 0.963 | 1.015 | 0.982 | 0.981 | 1.002 | 0.991 | 0.998 | 1.006 | 0.993 | 0.997 |
| | 3 | 2.157 | 1.464 | 1.109 | 1.840 | 1.305 | 1.097 | 0.989 | 0.994 | 1.007 | 0.990 | 0.997 | 1.008 |

Hence, the standard deviation values for both versions of Wright's statistics deviated from the theoretical value of one at all data sets, and were more influenced by the number of sections than by the number of items per section.

On the other side, the standard deviations of both versions of the new statistic were very close to one at all ability levels for all ordinary data sets. For example, for data sets with J=2 and *nj*=5, the standard deviations for *UBSR* were 1.004, 0.998, 0.996, 0.990, 0.994, 1.005, and 0.980 at the seven ability levels, respectively. This was evident even at ability levels that were distant from the average item difficulty for all data sets.

Table 3 presents type I error rates of the four section-level person fit statistic at a level of significance of 0.05 ($z$=1.96). Results showed that both versions of Wright's had inflated type I error rates at all ability levels within all data sets. These deviations were larger at ability levels that were distant from the average item difficulty ($\theta$=2, 3). In addition, the type I error rates for both versions of Wright's statistic were influenced by the number of sections on the test. They approached the theoretical hit rates of 0.05 as the number of sections increased.

This indicated that increasing the number of subsets improved the ability of both versions of Wright's statistic to better control type I error rates although they remained different from 0.05. Results showed that there was an effect (although it was modest) of increasing the number of items per section on approaching the type I error rates of both versions of Wright's statistic to the level of significance of 0.05. Hence, both versions of Wright's statistic were not able to control type I error rate at its expected value, and increasing the number of sections did improve the ability of Wright's statistic to control the type I error rates.

As showed in Table 3, both versions of the new statistic had type I error rates close to 0.05 at all ability levels within all data sets. Across all data sets, the type I error rates ranged between 0.040 and 0.056 for the weighted version of the new statistic, and they ranged between 0.043 and 0.055 for the unweighted version.

**Power of detecting aberrance**

Table 4 presents the power of the four section-level person fit statistic to detect both spuriously high and spuriously low aberrant responses using the hypothetical cutting score of 1.96 for the thirty six data sets that have 20% of items with aberrant response patterns. The power of the new statistic showed a similar pattern to Wright's statistic. However, Wright's statistic had higher power rates than the new statistic at all conditions. This might be caused by the inflated type I error rates showed by Wright's statistic. Moreover, the new statistic was able to detect aberrant response to the same rate as Wright's statistic even with those ability levels where Wright's statistic had very inflated type I error rates.

Similar patterns of detection of both types of aberrant response patterns (spuriously high and spuriously low) achieved by the new statistic as compared to Wright's statistic were found for those data sets that have more severe aberrant response patterns (40% of items) as shown in Table 5. Moreover, as previous studies showed (e.g, Drasgow, Levine, & Mclaughlin, 1991), the power of the four section-level person fit statistics were higher to detect more severe aberrant response patterns (40% of items) than less severe aberrant response patterns (20% of items) at all ability levels across all aberrance data sets. For example, for the data set with $J$=2, $n_j$=20, the power rates of *UBSR* to detect 20% of items with spuriously high aberrant response patterns were 0.406, 0.578 & 0.722 at $\theta$=-1, -2 & -3, respectively; whereas its power to detect

**Table 3**
**Type I error rates of the four section-level person fit indices at a significance level of 0.05 for various test lengths**

| | | *UBT* | | | *WBT* | | | *UBSR* | | | *WBSR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *J* | $\theta$ | $n_j$=5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 |
| 2 | -3 | 0.073 | 0.067 | 0.074 | 0.074 | 0.067 | 0.076 | 0.052 | 0.045 | 0.046 | 0.054 | 0.044 | 0.045 |
| | -2 | 0.069 | 0.074 | 0.082 | 0.071 | 0.076 | 0.082 | 0.047 | 0.048 | 0.049 | 0.046 | 0.049 | 0.049 |
| | -1 | 0.080 | 0.082 | 0.083 | 0.080 | 0.082 | 0.083 | 0.049 | 0.053 | 0.051 | 0.050 | 0.053 | 0.051 |
| | 0 | 0.082 | 0.090 | 0.087 | 0.082 | 0.089 | 0.087 | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 |
| | 1 | 0.077 | 0.081 | 0.084 | 0.083 | 0.082 | 0.084 | 0.047 | 0.050 | 0.049 | 0.047 | 0.051 | 0.049 |
| | 2 | 0.081 | 0.073 | 0.078 | 0.075 | 0.075 | 0.080 | 0.044 | 0.044 | 0.047 | 0.045 | 0.045 | 0.047 |
| | 3 | 0.110 | 0.100 | 0.093 | 0.136 | 0.089 | 0.099 | 0.045 | 0.043 | 0.048 | 0.040 | 0.043 | 0.045 |
| 4 | -3 | 0.068 | 0.051 | 0.057 | 0.070 | 0.052 | 0.058 | 0.055 | 0.047 | 0.046 | 0.055 | 0.047 | 0.047 |
| | -2 | 0.059 | 0.056 | 0.057 | 0.058 | 0.056 | 0.057 | 0.049 | 0.046 | 0.048 | 0.049 | 0.046 | 0.049 |
| | -1 | 0.050 | 0.052 | 0.056 | 0.051 | 0.053 | 0.056 | 0.048 | 0.048 | 0.049 | 0.048 | 0.048 | 0.049 |
| | 0 | 0.061 | 0.060 | 0.057 | 0.059 | 0.060 | 0.057 | 0.049 | 0.050 | 0.049 | 0.049 | 0.049 | 0.049 |
| | 1 | 0.054 | 0.054 | 0.054 | 0.057 | 0.059 | 0.054 | 0.047 | 0.047 | 0.049 | 0.048 | 0.048 | 0.049 |
| | 2 | 0.075 | 0.061 | 0.055 | 0.063 | 0.062 | 0.058 | 0.049 | 0.048 | 0.045 | 0.048 | 0.047 | 0.046 |
| | 3 | 0.145 | 0.104 | 0.079 | 0.141 | 0.095 | 0.074 | 0.049 | 0.050 | 0.050 | 0.038 | 0.056 | 0.048 |
| 8 | -3 | 0.055 | 0.055 | 0.050 | 0.055 | 0.055 | 0.051 | 0.048 | 0.048 | 0.045 | 0.048 | 0.047 | 0.045 |
| | -2 | 0.056 | 0.050 | 0.054 | 0.055 | 0.051 | 0.054 | 0.045 | 0.046 | 0.045 | 0.046 | 0.047 | 0.045 |
| | -1 | 0.044 | 0.049 | 0.049 | 0.044 | 0.049 | 0.049 | 0.044 | 0.045 | 0.045 | 0.044 | 0.045 | 0.045 |
| | 0 | 0.052 | 0.050 | 0.052 | 0.053 | 0.051 | 0.052 | 0.044 | 0.045 | 0.045 | 0.044 | 0.045 | 0.045 |
| | 1 | 0.051 | 0.054 | 0.053 | 0.053 | 0.055 | 0.053 | 0.044 | 0.044 | 0.046 | 0.044 | 0.045 | 0.045 |
| | 2 | 0.075 | 0.055 | 0.053 | 0.066 | 0.058 | 0.056 | 0.050 | 0.045 | 0.044 | 0.049 | 0.045 | 0.044 |
| | 3 | 0.528 | 0.138 | 0.077 | 0.399 | 0.113 | 0.079 | 0.048 | 0.052 | 0.052 | 0.048 | 0.051 | 0.053 |

**Table 4**
**Power values of the four section-level person fit indices at a significance level of 0.05 for 20% aberrant response patterns**

| | | *UBT* | | | *WBT* | | | *UBSR* | | | *WBSR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *J* | $\theta$ | $n_j$=5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 |
| Spuriously high response patterns | | | | | | | | | | | | | |
| 2 | -3 | 0.398 | 0.562 | 0.782 | 0.398 | 0.563 | 0.782 | 0.342 | 0.480 | 0.722 | 0.345 | 0.476 | 0.721 |
| | -2 | 0.314 | 0.448 | 0.650 | 0.314 | 0.448 | 0.648 | 0.255 | 0.375 | 0.578 | 0.254 | 0.374 | 0.576 |
| | -1 | 0.234 | 0.321 | 0.484 | 0.233 | 0.319 | 0.483 | 0.176 | 0.250 | 0.406 | 0.176 | 0.249 | 0.405 |
| 4 | -3 | 0.509 | 0.695 | 0.914 | 0.507 | 0.695 | 0.913 | 0.474 | 0.678 | 0.907 | 0.474 | 0.677 | 0.906 |
| | -2 | 0.367 | 0.541 | 0.790 | 0.359 | 0.534 | 0.788 | 0.352 | 0.528 | 0.781 | 0.347 | 0.523 | 0.779 |
| | -1 | 0.236 | 0.360 | 0.580 | 0.232 | 0.356 | 0.579 | 0.230 | 0.351 | 0.567 | 0.227 | 0.347 | 0.565 |
| 8 | -3 | 0.671 | 0.878 | 0.989 | 0.658 | 0.877 | 0.989 | 0.653 | 0.876 | 0.989 | 0.645 | 0.875 | 0.989 |
| | -2 | 0.495 | 0.712 | 0.938 | 0.481 | 0.706 | 0.937 | 0.504 | 0.721 | 0.942 | 0.492 | 0.717 | 0.939 |
| | -1 | 0.295 | 0.484 | 0.783 | 0.287 | 0.477 | 0.779 | 0.309 | 0.498 | 0.791 | 0.305 | 0.493 | 0.789 |
| Spuriously low response patterns | | | | | | | | | | | | | |
| 2 | 1 | 0.270 | 0.331 | 0.453 | 0.264 | 0.327 | 0.451 | 0.215 | 0.272 | 0.389 | 0.209 | 0.264 | 0.381 |
| | 2 | 0.483 | 0.586 | 0.763 | 0.445 | 0.573 | 0.754 | 0.392 | 0.522 | 0.711 | 0.371 | 0.506 | 0.700 |
| | 3 | 0.741 | 0.878 | 0.972 | 0.765 | 0.861 | 0.971 | 0.623 | 0.804 | 0.953 | 0.586 | 0.796 | 0.948 |
| 4 | 1 | 0.310 | 0.398 | 0.553 | 0.281 | 0.379 | 0.544 | 0.294 | 0.386 | 0.544 | 0.267 | 0.367 | 0.533 |
| | 2 | 0.616 | 0.733 | 0.900 | 0.554 | 0.710 | 0.892 | 0.557 | 0.716 | 0.893 | 0.523 | 0.691 | 0.884 |
| | 3 | 0.906 | 0.974 | 0.998 | 0.900 | 0.966 | 0.998 | 0.834 | 0.949 | 0.997 | 0.799 | 0.947 | 0.997 |
| 8 | 1 | 0.428 | 0.542 | 0.740 | 0.373 | 0.512 | 0.728 | 0.421 | 0.548 | 0.747 | 0.378 | 0.519 | 0.732 |
| | 2 | 0.801 | 0.907 | 0.987 | 0.727 | 0.884 | 0.984 | 0.763 | 0.904 | 0.988 | 0.712 | 0.880 | 0.984 |
| | 3 | 0.986 | 0.999 | 1.000 | 0.982 | 0.998 | 1.000 | 0.965 | 0.997 | 1.000 | 0.951 | 0.997 | 1.000 |

40% of items with spuriously high aberrant response patterns were 0.910, 0.980 & 0.997 at $\theta$=-1, -2 & -3, respectively. For the data set with $J$=4, $n_j$=5, the power rates of *WBSR* to detect 20% of items with spuriously low aberrant response patterns were 0.267, 0.523 & 0.799 at $\theta$=1, 2 & 3, respectively; whereas its power to detect 40% of items with spuriously low aberrant response patterns were 0.787, 0.978 & 0.999 at $\theta$=1, 2 & 3, respectively.

Moreover, results at both levels of severity of aberrance showed that the four section-level person fit statistic detected spuriously low aberrant response patterns better than spuriously high response patterns at corresponding ability levels across all aberrant data sets. This result also was found in Drasgow, Levine, and Mclaughlin (1991).

The power rates of the four section-level person fit statistic to detect different aberrant response patterns were increasing by the test length that resulted from increasing the number of sections or/and increasing the number of items per section. The lowest power rates were for the data set with $J$=2 & $n_j$=5 (total $n$=10), whereas the highest power rates were for the data set with $J$=8 & $n_j$=20 (total $n$=160). This pattern can also be observed by comparing the power rates of each person fit statistic horizontally (increasing the number of items per section, $n_j$, for each number of sections, $J$) and vertically (increasing the number of sections, $J$, for each number of items per section, $n_j$).

The results in Tables 4 and 5 allowed for answering the question: Which one has more effect on increasing the power of

**Table 5**
**Power values of the four section-level person fit indices at practical cutting scores (corresponding to a hit-rate of 0.05) for 40% aberrant response patterns**

| | | *UBT** | | | *WBT** | | | *UBSR* | | | *WBSR* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $J$ | $\theta$ | $n_j$=5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 | $n_j$ =5 | $n_j$ =10 | $n_j$ =20 |
| Spuriously high response patterns | | | | | | | | | | | | | |
| | -3 | 0.779 | 0.948 | 0.999 | 0.779 | 0.947 | 0.998 | 0.726 | 0.921 | 0.997 | 0.729 | 0.921 | 0.997 |
| 2 | -2 | 0.669 | 0.877 | 0.989 | 0.664 | 0.875 | 0.988 | 0.601 | 0.834 | 0.980 | 0.598 | 0.834 | 0.980 |
| | -1 | 0.511 | 0.740 | 0.935 | 0.508 | 0.737 | 0.935 | 0.435 | 0.673 | 0.910 | 0.435 | 0.671 | 0.911 |
| | -3 | 0.994 | 0.994 | 1.000 | 0.994 | 0.994 | 1.000 | 0.994 | 0.994 | 1.000 | 0.994 | 0.994 | 1.000 |
| 4 | -2 | 0.968 | 0.968 | 1.000 | 0.967 | 0.967 | 1.000 | 0.966 | 0.966 | 1.000 | 0.965 | 0.965 | 1.000 |
| | -1 | 0.867 | 0.867 | 0.992 | 0.865 | 0.865 | 0.992 | 0.863 | 0.863 | 0.991 | 0.860 | 0.860 | 0.991 |
| | -3 | 0.990 | 1.000 | 1.000 | 0.990 | 1.000 | 1.000 | 0.989 | 1.000 | 1.000 | 0.989 | 1.000 | 1.000 |
| 8 | -2 | 0.947 | 0.999 | 1.000 | 0.943 | 0.999 | 1.000 | 0.949 | 0.999 | 1.000 | 0.946 | 0.999 | 1.000 |
| | -1 | 0.799 | 0.978 | 1.000 | 0.792 | 0.978 | 1.000 | 0.814 | 0.980 | 1.000 | 0.808 | 0.979 | 1.000 |
| Spuriously low response patterns | | | | | | | | | | | | | |
| | 1 | 0.505 | 0.658 | 0.854 | 0.495 | 0.647 | 0.850 | 0.443 | 0.597 | 0.816 | 0.429 | 0.590 | 0.811 |
| 2 | 2 | 0.769 | 0.910 | 0.989 | 0.748 | 0.904 | 0.988 | 0.711 | 0.882 | 0.984 | 0.693 | 0.875 | 0.983 |
| | 3 | 0.943 | 0.994 | 1.000 | 0.951 | 0.993 | 1.000 | 0.898 | 0.986 | 1.000 | 0.883 | 0.986 | 1.000 |
| | 1 | 0.807 | 0.807 | 0.959 | 0.796 | 0.796 | 0.958 | 0.801 | 0.801 | 0.957 | 0.787 | 0.787 | 0.955 |
| 4 | 2 | 0.984 | 0.984 | 1.000 | 0.980 | 0.980 | 1.000 | 0.984 | 0.984 | 1.000 | 0.978 | 0.978 | 1.000 |
| | 3 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.999 | 1.000 |
| | 1 | 0.815 | 0.951 | 0.998 | 0.779 | 0.942 | 0.997 | 0.814 | 0.953 | 0.998 | 0.783 | 0.944 | 0.998 |
| 8 | 2 | 0.989 | 1.000 | 1.000 | 0.980 | 0.999 | 1.000 | 0.986 | 0.999 | 1.000 | 0.980 | 0.999 | 1.000 |
| | 3 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

* $z$=2.15 for $J$=2, $z$=2.0 for $J$=4.

the section-level person fit statistic in detecting different types of aberrant response patterns:increasing the number of sections or increasing the number of items per section? This can be answered by comparing the power rates for the first data set ($J$=2, $n_j$=5) with its right next neighbor data set ($J$=2, $n_j$=10) and bottom next neighbor data set ($J$=4, $n_j$=5); and by comparing the power rates for the first data set ($J$=2, $n_j$=5) with its right second neighbor data set ($J$=2, $n_j$=20) and bottom second neighbor data set ($J$=8, $n_j$=5). For spuriously high response patterns, the more effect on the power rates for the person fit statistic was for increasing the number of items per section, whereas the more effect on the power rates for the person fit statistic was for increasing the number of sections for spuriously low response patterns.

In addition, a consistent pattern can be noted in Table 4 and Table 5 through comparing the data sets that have similar test length with different combinations of number of sections and number of items per section (e.g., $J$=2, $n_j$=20; $J$=4, $n_j$=10; $J$=8, $n_j$=5). The power rates of the four section-level person fit statistics in detecting spuriously high response patterns were higher for test with fewer number of sections and large number of items per section ($J$=2, $n_j$=20) than other combinations. However, their power rates in detecting spuriously low response patterns were similar regardless the combinations of number of sections and number of items per section.

**Study 2: Simulation with Real Parameters**

The second study aimed to compare between the proposed section-level person fit statistic and Wright's person fit mean square statistic in terms of their distributional characteristics (mean, standard deviation, and type I error at two levels of significance) and their power of detecting person misfit when both real ability and item parameters were used.

## Method

The real data used in this study is the verbal ability test that is administered by the National Center for Assessment in Higher Education (NCAHE) in Saudi Arabia. The verbal ability test contains 66 multiple-choice items that measure verbal reasoning ability in four areas: verbal meaning 13 items, verbal analog 16 items, sentence completion 17 items, and verbal comprehension 20 items. The test is administered in the Arabic language as an admission test for students in grade 11 and 12 who want to apply to higher education institutions.

A random sample of 7000 examinees were selected from the 2008/2009 administration of one test form and then were used to estimate both the ability and item parameters with the 3PL model using BILOG program. The unidimensionality of the test was examined by fitting one second-order common factor through confirmatory factor analysis with Lisrel 8.52 using tetrachoric correlations among test items. Results showed high model-fit indices (RMSEA=0.037, GFI=0.93, CFI=0.98). Moreover, Yens (1981) $Q_1$ statistic for overall 3PL model fit was not significant. The estimated parameters were then used to simulate responses to the 66 test items using pseudo random numbers. If the generated random numbers are larger than or equal to the calculated probabilities of a correct answer to items (using 3PLM with the estimated ability and item parameters), the item responses are set to 1, and vice versa. This process was replicated 60 times.

The distributional properties of the two person fit statistic were also examined with two partitions of the verbal ability test for research purpose. The first one had the original four sections (13, 16, 17, 20 items) and the second one had two combined sections resulted from combining the first two section (verbal meaning & verbal analog, 29 items) and the last two sections (sentence comple-

tion & verbal comprehension, 37 items). Moreover, the real data for the 7000 examinees in this verbal ability test with the four original sections was analyzed using the four section-level person fit statistic to compare the ability of these statistics to detect examinees with aberrant responses in real data.

## Results

Table 6 and 7 present the means, standard deviations, and type I error rates for the four person fit statistic for all 7000 examinees at five ability intervals with about equal sample size. Results showed that all statistics had patterns for the three distributional properties similar to what have been found earlier in the paper when using the true ability and item parameters. Table 6 showed that both Wright's statistics had deviated means from zero at all ability intervals with both two sections and four sections. This deviation was larger with two sections. The means for *UBT* ranged between 0.738 and 0.749 with two sections and 0.420 and 0.433 with four sections. Moreover, the standard deviations of both Wright's statistics were deviated from one with both four sections and two sections at all ability intervals (*WBT* ranged between 0.858 and 0.870 for two sections; and between 0.932 and 0.945 for four sections). As a result of the deviations of both means and standard deviations, both versions of Wright's statistic showed higher type I error rates than 0.05 for both numbers of sections. For example, type I error rates for *UBT* ranged between 0.082 and 0.086 with two sections, and between 0.0052 and 0.055 with four sections. When the practical cut scores for Wright's statistics found in study 1 were used, the type I error rates of *UBT*, for example, reduced to 0.052 and 0.057 with two sections and to 0.054 and 0.057 with four sections.

Table 6 showed that the proposed person fit statistic showed superior distributional properties (means and standard deviations) when using estimated ability and item parameters at all ability intervals. The means and standard deviations of both versions of the proposed statistic were at their hypothetical values for four sections and two sections. In addition, Table 7 showed that the proposed statistic controlled type I error rates around 0.05. For example, the type I error rates for the *WBSR* ranged between 0.044 and 0.053 with two sections, and between 0.0046 and 0.049 with four sections.

**Table 6**
**Means for the person fit statistic using estimated ability and item parameter with 4 and 2 sections**

| | | | *UBT* | | *WBT* | | *UBSR* | | *WBSR* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 |
| *Mean* | *All data* | 7000 | 0.742 | 0.426 | 0.741 | 0.423 | 0.002 | 0.000 | 0.002 | 0.000 |
| | $p_1$ - $p_{19}$ | 1399 | 0.743 | 0.427 | 0.740 | 0.423 | 0.005 | 0.002 | 0.006 | 0.003 |
| | $p_{20}$ - $p_{39}$ | 1401 | 0.743 | 0.430 | 0.740 | 0.426 | 0.004 | 0.005 | 0.005 | 0.005 |
| | $p_{40}$ - $p_{59}$ | 1395 | 0.738 | 0.423 | 0.737 | 0.420 | -0.002 | -0.003 | -0.002 | -0.004 |
| | $p_{60}$ - $p_{79}$ | 1405 | 0.738 | 0.420 | 0.737 | 0.417 | -0.006 | -0.010 | -0.006 | -0.010 |
| | $p_{80}$ - $p_{99}$ | 1400 | 0.749 | 0.433 | 0.749 | 0.430 | 0.007 | 0.005 | 0.008 | 0.005 |
| *SD* | *All data* | 7000 | 0.862 | 0.932 | 0.864 | 0.938 | 1.005 | 1.000 | 1.005 | 1.000 |
| | $p_1$ - $p_{19}$ | 1399 | 0.866 | 0.937 | 0.870 | 0.945 | 1.012 | 1.007 | 1.014 | 1.009 |
| | $p_{20}$ - $p_{39}$ | 1401 | 0.866 | 0.935 | 0.869 | 0.942 | 1.007 | 1.003 | 1.007 | 1.004 |
| | $p_{40}$ - $p_{59}$ | 1395 | 0.862 | 0.933 | 0.864 | 0.938 | 1.006 | 0.999 | 1.006 | 0.998 |
| | $p_{60}$ - $p_{79}$ | 1405 | 0.858 | 0.927 | 0.858 | 0.932 | 0.988 | 0.988 | 0.989 | 0.989 |
| | $p_{80}$ - $p_{99}$ | 1400 | 0.859 | 0.927 | 0.859 | 0.932 | 1.009 | 1.001 | 1.010 | 1.002 |

**Table 7**
**Type I error rates for the person fit statistic using estimated ability and item parameters with 4 and 2 sections**

| | | UBT | | WBT | | UBT* | | WBT* | | UBSR | | WBSR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 | *j*=2 | *j*=4 |
| *All data* | 7000 | 0.085 | 0.054 | 0.085 | 0.056 | 0.055 | 0.050 | 0.056 | 0.051 | 0.052 | 0.048 | 0.051 | 0.048 |
| $p_1 - p_{19}$ | 1399 | 0.086 | 0.055 | 0.087 | 0.057 | 0.057 | 0.050 | 0.058 | 0.052 | 0.052 | 0.049 | 0.053 | 0.049 |
| $p_{20} - p_{39}$ | 1401 | 0.085 | 0.055 | 0.086 | 0.057 | 0.057 | 0.051 | 0.056 | 0.052 | 0.053 | 0.049 | 0.052 | 0.049 |
| $p_{40} - p_{59}$ | 1395 | 0.085 | 0.054 | 0.084 | 0.055 | 0.056 | 0.050 | 0.057 | 0.050 | 0.052 | 0.049 | 0.052 | 0.049 |
| $p_{60} - p_{79}$ | 1405 | 0.086 | 0.052 | 0.088 | 0.054 | 0.052 | 0.048 | 0.051 | 0.049 | 0.050 | 0.047 | 0.049 | 0.046 |
| $p_{80} - p_{99}$ | 1400 | 0.082 | 0.054 | 0.082 | 0.056 | 0.055 | 0.050 | 0.055 | 0.051 | 0.051 | 0.047 | 0.051 | 0.048 |

* $z$= 2.15 for $J$=2, $z$ = 2.0 for $J$=4.

## Discussion and Conclusions

Similar to the mainstream findings of many previous studies (Hambleton, et. al., 1978; Smith, 1994; 1996), this study showed that Wright's between person fit statistic had poor distributional properties. The means, standard deviations, and type I error rates for both versions of Wright's statistic deviated from their presumed values with both true and real ability and item parameters. In addition, the degree of deviation of these distributional properties was influenced by the number of sections. A small number of sections resulted in more deviations of these distributional properties. The increase of the number of items per section did not improve the proximity of these means, standard deviations, and type I error rates of both versions of Wright's statistic. Similar to previous studies (Divgi, 1986; George, 1979; Hambleton, 1978), the results of this study suggested that the applied transformations to Wright's statistic did not remove the effect of the number of sections within the test on the statistic. In addition, different empirical cut scores were found for the Wright's statistic to control the hit-rate at a level of significance of 0.05 with different number of sections of the test. Having more sections on the test leads to empirical cut scores similar to the hypothetical cut score of 1.96.

Using the hypothetical cut scores with all data sets, results showed that the power of Wright's between person fit statistic was relatively high since it was contaminated by the inflated type I error rates. In addition, the power of Wright's between person fit statistic has no meaning provided that the statistic was shown to not follow the unit normal distribution. Hence, when using a single cut score for classifying misfit person response patterns, the power of Wright's statistic should be interpreted with caution.

On the other hand, results showed that both versions of the new section-level person fit statistic had superior distributional properties with both true and estimated ability and item parameters. The means, standard deviations, and type I error rates of both versions of the new section-level person fit statistic were approximate to their expected values under the assumption of unit normal distribution within all data sets including the smallest data set ($J$=2, $n_j$=5). In addition, the distributional properties of the new section-level person fit statistic were unaffected by either small number of sections, small number of items per section, or both.

Unlike Wright's statistic, the new section-level person fit statistic was not influenced by the number of sections. The adequacy of the new statistic was parallel within tests with a large number of sections or with a small number of sections. This suggests that, when using the new section-level person fit statistic, the sections can be formed without worrying about the effect of the number of sections on the distributional properties of the person fit statistic.

The new section-level person fit statistic also showed larger power rates for data sets with a large number of sections and a large number of items per section than

data sets with a small number of sections and a small number of items per section. In addition, increasing the number of items per section improved the power of the new section-level person fit statistic more than increasing the number of sections. This suggests that keeping a large number of items per section is preferred when forming the sections in order to achieve more power to detect person misfit with the new section-level person fit statistic.

Moreover, with the support of the result of this study, the section-level person fit statistic are more sensitive to aberrant response patterns that are existing heavily within one or few sections as compared to those aberrant response patterns that are distributed evenly across all sections of the test.

The results of outperforming of the new section-level statistic over Wright's statistic (found in this study) are supported by several justifications. Unlike Wright's statistic, the new statistic evaluates the squared residual differences of person's scores on each section (equation 6), which is a continuous quantity that might take any values between 0 and $n_j^2$. This fact might suggest that standardizing this quantity is possible and leads to be distributed as unit a normal distribution. Second, the new section-level person fit statistic uses a normal distribution to evaluate the fit of person responses. The number of sections used to calculate the statistic has less effect on the new statistic than its effect on the Pearson chi-square test used with Wright's statistic.

Third, the new section-level person fit statistic involves only a straightforward standardization of the squared residual differences, whereas Wright's between person fit mean square statistic involves three steps: standardizing person's section scores, squaring it (evaluated as chi-square test), and normalizing it (evaluated as normal distribution test). These chain-like steps on Wright's statistic have been found to cause dependency and difficulty in explaining any poor distributional properties of the statistic (see Smith, 1991). Finally, the mean and standard deviation of the squared residual difference used in the new section-level person fit statistic are computed based on the prediction of the IRT model. However, the standard deviation of the person fit mean square used to normalize Wright's between person fit mean square statistics (in equation (4) and (5) is data-driven (Smith, 1986).

In summary, the results of this study showed that the proposed section-level person fit statistic performed well with person fit analysis even with the commonly-used small number of sections associated with person fit analysis. The proposed section-level person fit statistic deserves more investigation under different tests conditions and different test applications.

## References

Almehrizi, R. (2013). Coefficient alpha and reliability of scale scores. *Applied psychological Measurement, 37*(6), 438-459.

Almehrizi, R. (2016). Normalization of Mean Squared Differences to Measure Agreement for Continuous Data. *Statistical Methods in Medical Research*, 25, 1975-1990.

Al-Mahrazi, R. (2004). *Investigating a new modification of the residual-based person fit index and its relationship with other indices in dichotomous item response theory*. Unpublished Ph.D Dissertation, University of Iowa.

Al-Mehrzi, R. (2010). Comparison among new residual-based person fit indices and Wright's indices for dichotomous three-parameter IRT model with standardized tests. *Journal of Educational and Psychological Studies, Sultan Qaboos University, 4*(2), 14-26.

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

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Felt,J. M., Castaneda, R., Tiemensma, J.,& Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Front Psychol., 8*, 1-9.

Fox. J. P. & Marianti, S. (2017). Person fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*, 243-262.

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

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.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 453-461.

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. & Sijtisma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 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 the American Educational Research Association, Los Angeles, CA.

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

Smith, R. M. (1988). The distribution properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

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

Smith, R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbance. *Educational and Psychological Measurement, 54*, 42-55.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between person fit methods of detecting item bias. *Educational and Psychological Measurement, 56*, 403-418.

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.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.