DETECTION OF UNMOTIVATED TEST TAKERS THROUGH AN ANALYSIS OF RESPONSE PATTERNS:

BEYOND PERSON-FIT STATISTICS

by

TARA L. TWISTE

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

David Rindskopf

Date — Chair of Examining Committee

Mario Antonio Kelly

Date — Executive Officer

Jay Verkuilen
Keith Markus
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

DETECTION OF UNMOTIVATED TEST TAKERS THROUGH AN ANALYSIS OF RESPONSE PATTERNS:

BEYOND PERSON-FIT STATISTICS

by

Tara L. Twiste

Adviser: Professor David Rindskopf

The identification of patterned responding in unmotivated test takers was investigated through the formation of a novel method. The proposed method relied on marginal proportions of answer choice options as well as the transitional proportions between responses on item pairs. A chi square analysis was used to determine the degree of significance of each participant's patterned responding. The method was compared to the existing person-fit statistic l_z (Drasgow, Levine & McLaughlin, 1987). Three publically available data sets - including a political science survey, an elementary school arithmetic scale and a general college course final exam - were used to test the occurrence of patterned responding and the ability of the proposed method to identify such unmotivated behavior.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Detection of Unmotivated Test Takers through an Analysis of Response Patterns: Beyond Person-fit Statistics

Accuracy of measurement is crucial in all areas of research and data collection. In education we need to be confident that test results are appropriate and useful. This need has sparked many areas of research including reliability and validity studies, as well as attempts to detect bias in questioning and differential scoring for subgroups. Literature has also addressed the effects of cheating and guessing, but still a largely unexplored area involves the consequences of a lack of motivation and attention to the task as a potentially serious area of inaccurate measurement. Motivation for a task can wean from simple distraction, boredom, inherent disinterest in the topic, or a lack of consequences or feedback for performance.

Small scale testing, surveys and questionnaires are heavily relied upon in social science and education research. We use these sources to gather data from students, teachers, parents and the community, to inform our policies, to evaluate programs and staff, and to understand experiences. Educational policy in particular repeated uses tests and surveys to support or change the direction of the field. The results of these measures could be distorted rendering the policies faulty and ineffective. There are obvious and extensive consequences to this practice. The marketing field relies on consumer data driven by self report. Job placement and screening rests on personality and character assessments. Even clinical psychology and diagnostics are often forced to consider, if not utterly depend upon, an individual's self-report. One can easily imagine how widespread the consequences of a lack of motivation may be in cases where the

participant has no vested interest. In fact, research suggests that there exists a high degree of unmotivated behavior in survey settings and in non-consequential testing situations. Barnette (1996) found that 14% of his population provided responses consistent with non-attending patterns. Additionally, not addressing the occurrences of these aberrant response patterns has been shown to directly affect the measurement statistics used in analysis. In 1999, Barnette demonstrated the effects on internal consistency of non-attending response patterns. With a replacement of as little as 5% of the response patterns, a significant change in Cronbach's alpha was found (Barnette, 1999). There has been no research into the effect on other statistics such as correlations and t-tests. But without further investigation we cannot assume that these statistics are safe from the consequences of bad data.

In light of this past research, it is apparent that we may be facing a serious lack of confidence in much of our data. It is of great importance that the data of unmotivated participants be identified, the effect of their inclusion investigated, and their removal considered in order to ensure accurate results and measurements.

A possible method for identification of these individuals is to consider the pattern of their responses. A study participant or test taker not focused on the task may randomly select answer options, repeatedly choose the same response, create a pattern from the response scale, or respond in some combination of these behaviors. The notion of patterned responses has previously been addressed in the field of person-fit statistics. This area of research relies on statistical methods to evaluate the fit of an individual's response pattern to either those of the population, as in group-based methods, or to the

test model, as in item response theory (IRT) based statistics. The category of group-based person-fit statistics is generally defined by the focus on comparing an individual's item score pattern with the item score patterns of all other individuals within the sample (Meijer & Sijtsma, 1999). Though flourishing in the late 1970s and early 1980s, group-based methods were largely replaced by the IRT-based statistics due to flaws in interpretation and standardization as well as the growing acceptance and use of item response theory.

Also known as latent trait theory, item response theory (IRT) is a model-based measurement theory in which trait level estimates depend on both persons' responses and on the properties of the items that are administered (Embretson & Reise, 2000). Depending on the model used, the response pattern of individuals may be predicted. Of course, it is rarely the case that the predicted answers correspond exactly with actual answers. Various IRT person-fit statistics seek to quantify and determine the meaningfulness of these discrepancies.

The major statistics in both of these areas will be discussed in the following literature review. This discussion will address when the methods are appropriately used and what may be gained from their use. It will also consider the inherent shortcomings and difficulties in applying these methods to the identification of non-attending participants.

One of the earliest person-fit statistics, proposed by Donlon & Fischer (1968), rested on an application of the biserial correlation coefficient (r_{bis}). In item analysis the r_{bis} has been used to measure the extent to which success on an individual item reflects

success on the test. It is calculated as the correlation between a column of item responses and a column of criterion scores, in this case the total test score. Donlon and Fischer, however, were more interested in determining the correlation between a person's distribution of item difficulties on a specific test and the distribution of item difficulties in some reference population. To accomplish this they proposed the personal biserial coefficient (r_{perbis}), a correlation between a row of responses by a person and a row of item difficulty indices:

$$r_{perbis} = \frac{\bar{\Delta}_R - \bar{\Delta}_1}{S_{\Delta R}} \times \frac{p_R'}{u'} \quad (1)$$

where $\bar{\Delta}_1$ is the mean item difficulty for items marked correctly, $\bar{\Delta}_R$ is the mean item difficulty for items reached, $S_{\Delta R}$ is the standard deviation of $\bar{\Delta}_R$, p_R' is the number of items marked correctly divided by the number of items reached and u' is the ordinate in the unit normal distribution which divides the area under the curve into the proportions p_R' and 1 - p_R'. This index will therefore measure the relationship between the difficulty of items for an individual test taker and the difficulties of items for the entire group.

Consider an individual response pattern of 11100, where 1 indicates a correct response and 0 indicates an incorrect response, on a 5 item test with corresponding item difficulties of 0.1, 0.3, 0.5, 0.7, 0.9. Note that Donlon & Fischer (1968) reversed the typical item difficulty scale so that higher values indicate more difficult items. In this example, the mean difficulty for items marked correctly ($\bar{\Delta}_1$) is 0.3 and assuming the test was completed, the mean item difficulty of items reached ($\bar{\Delta}_R$) is 0.5. The standard deviation of the mean difficulty of items reached ($S_{\Delta R} = \sqrt{\frac{\sum(x-M)^2}{N-1}}$) is 0.32. The number

of items marked correctly divided by the number of items reached, or p_R', is 0.6 which makes u' equal to 0.3863. For this individual, $r_{perbis} = 0.97$.

The index can be interpreted such that a positive coefficient will result when agreement is found between the individual and the group. A negative coefficient will occur in situations of disagreement such as when an individual responds to difficult items correctly and easy items incorrectly. A near zero coefficient indicates chance responses by the test taker given the assumption that these responses would not correlate with difficulty at all. Therefore in the above example where $r_{perbis} = 0.97$, we can see that our hypothetical individual's response pattern is similar to that of the group and is therefore not aberrant. The individual answered the easy items correctly and the difficult items incorrectly.

The goal of Donlon & Fischer (1968) was to provide information in addition to total test score without requiring additional testing time. Though research is quite limited on this statistic, Donlon and Fischer performed their own analysis where r_{perbis} did provide valuable information above and beyond the test scores. They sampled 614 PSAT test takers whose scores were considered in the chance range. This population was specifically chosen due to criticisms at the time of test publishers and users reporting scores that fell within the range of chance. The belief was that these scores cannot, and should not, be considered valid. Donlon & Fischer's analysis, however, showed that despite the low scores the personal biserial coefficients were significantly different from chance, i.e. they were not close to zero. This result indicated that in fact the scores could be considered valid and indicated low levels of competence.

The statistics that followed, however, typically relied on a comparison of score patterns for item pairs and the expected count under the Guttman model. The Guttman scale is based on a matrix of zeros and ones, incorrect and correct responses respectively, in which each row represents a single examinee's response pattern. Each column is created by the responses of examinees to the corresponding individual item. This matrix is formed so that the items are arranged from left to right in ascending order of difficulty and the examinees are arranged in descending order of total test score. For example, consider a 5 item test administered to 5 students, where the response patterns for Students A through E are as follows: 10000, 11100, 11000, 11110, 11111. The example can easily be formed into the Guttman model seen in Table 1.

Table 1

Dichotomous Responses of 5 Students to a 5 Item Test Arranged in Guttman Order

	Item (*i*)					
Student (j)	1	2	3	4	5	n_j
E	1	1	1	1	1	5
D	1	1	1	1	0	4
B	1	1	1	0	0	3
C	1	1	0	0	0	2
A	1	0	0	0	0	1
n_i	5	4	3	2	1	

Guttman proposed that no individual should have a correct answer following an incorrect answer, therefore making item score combinations of (0,1) impossible. For Guttman, any deviation from this pattern would indicate an aberrant response pattern.

In practice however, a perfect Guttman scale is not expected. From this understanding, Sato (1975, in Japanese), as discussed in English in Tatsuoka & Tatsuoka (1978) and Harnisch & Linn (1981), constructed the Student-Problem Table. The format and structure was identical to the Guttman matrix; however this table is characterized as having *mostly* 1s in the upper left hand corner and *mostly* 0s in the lower right hand corner. Sato used this table to construct two curves which were then used to determine the extent of response pattern homogeneity. The S-curve is created by drawing a vertical line in each row that falls to the left of the individual examinee's total number correct score, while the P-curve is a horizontal line drawn in each column indicating how many examinees answered the item correctly. Look at another, slightly more complicated example of a 5 item test this time given to 8 students where the following response patterns for Students A through H are 11111, 11110, 11011, 11010, 10101, 11100, 10100, 11000. Using the same table format as in the previous example, the students would fall into the rows based on descending total score and the items arranged most to least difficult would make up the columns as seen in Table 2.

Table 2

Dichotomous Responses of 8 Students on a 5 Item Test

	Item (*i*)					
Student (j)	1	2	3	4	5	n_j
A	1	1	1	1	1	5
B	1	1	1	1	0	4
C	1	1	0	1	1	4
D	1	1	0	1	0	3
E	1	0	1	0	1	3
F	1	1	1	0	0	3
G	1	0	1	0	0	2
H	1	1	0	0	0	2
n_i	8	6	5	4	3	

Note. Students arranged by descending order of total score and items in ascending order of difficulty.

In drawing the S- and P-curve we ignore what response falls into each box while considering what a perfect Guttman based on the student's number correct score and the item difficulty would look like. In Table 3, Student A has a total score of 5 so the vertical S-curve is dropped to the right of item 5. The same process holds for Student B with a score of 4 – the vertical line will be placed to the right of item 4. Student C also has a total score of 4, and despite the fact that the incorrect answer is item 3 the vertical S-curve line would fall to the right of item 4. With a total score of 3, the S-curve would fall

to the right of item 3 for Students D, E and F, and so on down through Student G and H. The P-curve can be placed by considering each item rather than each person. All students responded correctly to item 1 so the horizontal P-curve line would fall below Student H in the item 1 column. However, since only 6 students answered item 2 correctly, the P-curve would shift up to fall below Student F in the item 2 column. This would again continue through all the items. In ideal situations the S- and P-curves would overlap and a single step down line would separate the 1s from 0s. This would indicate that there were no aberrant response patterns – each individual with the same total score had the same pattern of responses. However, we see that this is not the case in Table 3. The area between the S- and P-curve was the basis of Sato's homogeneity index. The more divergent the two curves the smaller the index.

Table 3

S-P Curves for 5 Item Test Given to 8 Students

	Item				
Student	1	2	3	4	5
A	1	1	1	1	1
B	1	1	1	1	0
C	1	1	0	1	1
D	1	1	0	1	0
E	1	0	1	0	1
F	1	1	1	0	0
G	1	0	1	0	0
H	1	1	0	0	0

Note. P curve represented by the dotted line, S curve represented by the solid line.

Sato developed another index for identifying aberrant response patterns which he referred to as the caution index (*Ci*). As the name suggests the value of this index indicates that caution may be needed in interpreting the total score. This index is defined in Equation 2.

$$C_i = \frac{\sum_{j=1}^{n_{i.}}(1-u_{ij})n_{.j} - \sum_{j=n_{i.}}^{J} u_{ij}n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - n_{i.}\left(\frac{\sum_{j=1}^{J} n_{.j}}{J}\right)} \quad (2)$$

Where *i* represents the examinee and *j* refers to an individual item, u_{ij} is 1 if the examinee answers the question correctly and 0 if the answer is incorrect, $n_{i\cdot}$ is the total correct score

for the examinee and $n_{.j}$ is the total number of examinees who answered the item correctly.

This index provides potentially valuable information about a test taker that is not contained in the total score. For example, in Table 4 Students D, E and F all have a test score of 3, but their caution indices vary from 0 to 0.88 indicating that the scores are not equally meaningful for each student. However, since there is no predetermined range for caution index values, interpretation of the index is difficult, if not impossible. Although, it can easily be seen that an index of 0 occurs when a perfect Guttman response pattern is present, there is not much that can be said about a student with any other caution index value except that it is large or small in comparison to the sample of students.

Harnisch and Linn (1981) proposed a modified caution index (C_i*) to correct for this problem. The formula in Equation 3 gave Sato's C_i a lower bound limit of 0 and an upper bound limit of 1:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}}(1-u_{ij})n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij}n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J+1-n_{.j}}^{J} n_{.j}} \quad (3)$$

As seen in Table 4, the modified caution index lends itself to greater interpretability. The caution index of Students D and G are similar (0.29 and 0.27 respectively), but their modified indices are exactly the same (0.14). Though these students do not have the same test score, nor did they answer similar questions correctly, they both have response patterns that are one step away from a Guttman pattern.

Though the modified caution index offers some improvement over Sato's original C_i we encounter a very serious flaw common to most group-based statistics. The null

distribution is not known. There is no defined critical value with which to classify response patterns as aberrant. The values used in the research have largely been arbitrarily determined based on some characteristics of the data set in question. The critical values used have ranged from 0.3 (Harnish & Linn, 1981) to 0.6 (Harnish, 1983). This practice leads to inconsistency and prevents the evaluation and comparison of results.

Table 4

Caution and Modified Caution Indices of 8 Students on a 5 Item Test

	Item (*i*)							
Student (*j*)	1	2	3	4	5	n_j	C_i	C_i^*
A	1	1	1	1	1	5	0	0
B	1	1	1	1	0	4	0	0
C	1	1	0	1	1	4	.9	.4
D	1	1	0	1	0	3	.29	.14
E	1	0	1	0	1	3	.88	.42
F	1	1	1	0	0	3	0	0
G	1	0	1	0	0	2	.27	.14
H	1	1	0	0	0	2	0	0
n_i	8	6	5	4	3			

Despite this flaw, the advent of group-based statistics continued. Like Harnisch and Linn, van der Flier (1982) also borrowed Sato's concepts. Van der Flier's greatest

interest was in the effects of cultural background differences on the interpretation and meaning of test scores. His statistic was intended to determine the deviance of score patterns that would then indicate the appropriateness of comparing an individual's test score to the scores of a specific group. To do this van der Flier's method compared a test taker's score pattern on the individual items to the expected score pattern of a specified reference group. The expected score pattern in this situation was based on the mean item scores of a group. Therefore the deviance score was considered a reflection of the extent to which items have the same order of difficulty for an individual test taker as for the group. Van der Flier called this index U''' and it is defined in Equation 4.

$$U'''(X) = \frac{\log P_{max} - \log P(X)}{\log P_{max} - \log P_{min}} \quad (4)$$

where $P(X)$ is the probability of the score pattern X, and P_{max} and P_{min} are the probabilities of the least and most deviant patterns yielding the same test score. For example, on a 5 item test with a total score of 3 a possible response pattern could be (11011) as in the case of Student C. However, the least deviant response pattern would follow the Guttman pattern exactly meaning that all 3 correct responses would precede the 2 incorrect answers. The responses of 11100 have this pattern. On the other hand, the most deviant and unexpected response is the opposite where all the incorrect responses come before the correct responses (00111). Our eight student, five item example, however, cannot be used to demonstrate U''' because the probability of answering item 1 correctly is 1. The reverse Guttman (i.e. P_{min} - the most deviant response pattern) will result in a probability of 0. The log of P_{min} would then be log(0) which is undefined. The example has been modified to include only items 2-5 and for

additional simplicity only patterns resulting in a total score of 3 will be used. *P(max)* is calculated by multiplying the probabilities of answering the item correctly when the response is correct and multiplying by *1-P(i)* if the item was incorrect based on the perfect Guttman response pattern of 1110. For example, *P(max)* = (.75)(.625)(.5)(1-.375) = .146. *P(min)* is calculated similarly, but based on the reverse Guttman pattern of 0111. Then, *P(min)* = (1-.75)(.625)(.5)(.375) = .029. Table 5 shows the item patterns, probabilities and *U‴*.

Table 5

Probability of Response Patterns and van der Flier's Deviance Index

	Item (*i*)					
Student (j)	2	3	4	5	*P(X)*	*U‴*
B	1	1	1	0	.146	0
C	1	0	1	1	.053	.635
P(i=1)	.75	.625	.50	.375		

Note. P(i=1) from 8 student, 5 item example.

As expected, the value of *U‴* for Student B is 0 due to the Guttman pattern of responses. A reverse Guttman, not present in our example, would result in *U‴* = 1.

Van der Flier found that the deviance score provided valuable information above and beyond the test score alone. Using two populations that differed in the extent to which skills required to answer questions were over-learned, he generated two samples of score patterns. From one sample, van der Flier then estimated the p_i values of the items in test score brackets for both populations. He used the second sample to find two

deviance scores for all individuals based on the p_i values of each population. Finally, individuals were assigned to either population based on their two deviance scores. Depending on the process by which the cumulative probabilities were found, either empirically or theoretical, van der Flier was able to correctly allocate 73.1% and 72.3% of the individuals to the appropriate population. Further, he found that this information could not be derived from test scores alone; allocation based solely on the test score was not as accurate and was only slightly correlated with allocation (r = .069) based on deviance scores. Deviance scores provided additional information.

Van der Flier's U''' statistic can also be conceptualized as a measure of proximity of the response vector S to a reverse Guttman vector. Considering the dominance matrix of a response pattern matrix ($\bar{S}'S$), $U = U_a + U_b$, where U is the sum of all elements of N, U_a is the sum of all above-diagonal elements of N and U_b is the sum of all elements in the lower triangle of the matrix $\bar{S}'S$. U_a/U is therefore the proportion of (0,1) pairs among all possible ordered pairs of (s_i, s_j), where [$i > j$] of unlike elements. Oppositely, U_b/U is the proportion of (1,0) pairs.

The norm conformity index (*NCI*) proposed in 1983 by Tatsuoka & Tatsuoka is a transformation of van der Flier's U'''. See Equation 5. This index measures the proximity of the response pattern to a baseline pattern in which all 0s precede all 1s when items are arranged in descending order of difficulty. Following from the preceding language, *NCI* is the proportion of (1,0) pairs among ordered pair of (s_i, s_j) where [$i > j$], or U_b/U. However, Tatsuoka & Tatsuoka's index includes a rescaling so that the range of

possible results is -1 to 1 where 1 represents a perfect Guttman vector and -1 a reverse Guttman vector.

$$NCI = \left(\frac{2U_a}{U}\right) - 1 \quad (5)$$

Tatsuoka & Tatsuoka discuss the *NCI* as a measure of the proximity of *S* to a reverse Guttman vector where all 0's precede all 1's. However, the calculations call for arranging the items in descending order of difficulty. In the previous example of a 5 item test administered to 8 students, the items would now be ordered from 5 to 1 as in Table 6. Recalling that U_a and U are taken from the dominance matrix N, for Student A the value is

$$N_A = [00000]\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$$

Therefore $U_a = 0$, $U = 0$ and $NCI = 0$. Students B, F and H all have perfect reversed Guttman patterns so their *NCI* will be 1.

Table 6

Dominance Matrix Counts and the Norm Conformity Index

	Item							
Student	5	4	3	2	1	U_a	*U*	*NCI*
A	1	1	1	1	1	0	0	0
B	0	1	1	1	1	4	4	1
C	1	1	0	1	1	2	4	0
D	0	1	0	1	1	5	6	.67
E	1	0	1	0	1	3	6	0
F	0	0	1	1	1	6	6	1
G	0	0	1	0	1	5	6	.67
H	0	0	0	1	1	6	6	1

From this simple example the usefulness and ability to interpret *NCI* beyond the perfect reverse Guttman is questionable. Students D and G both have an *NCI* of 0.67 and their response patterns are one transition from a reverse Guttman. However, Students A, C and E have very different response patterns and their *NCI*'s all equal 0. Tatsuoka & Tatsuoka do not provide much guidance into the interpretation of these values.

NCI can also be used to determine the extent to which an individual response pattern conforms to that of a defined reference group. The special case as discussed previously occurs when items are arranged in descending order of difficulty, but the items may be ordered in any meaningful way.

In the same 1983 paper, Tatsuoka & Tatsuoka presented the individual consistency index (*ICI*): Equation 6. This index was designed to measure the extent to which an individual's response pattern remains consistent over time. Based on the assumption that as students learn their responses on similar items will move from highly variable to highly consistent, *ICI* was intended to be used for identifying a student's level of mastery. This index may have been a novel concept as Tatsuoka & Tatsuoka profess, but its calculation is based on previously mentioned statistics and can be considered a weighted average of these. Much like calculating the norm conformity index, the individual consistency index is partially based on the dominance matrix values *U* and U_a. In fact, the index rests in the *NCI* value itself in conjunction with the number of transformations necessary to convert a response pattern to a reverse Guttman and an assigned weight.

$$ICI = \sum_{j=1}^{J} w_j C(0')_j \tag{6}$$

where $w_j = \frac{U_p}{\sum U_p}$; $C(0')_j = C(0)_j + \frac{2t_j}{U_j}$ and $C(0)_j = NCI$. Consider a single student who has taken a 20 item test consisting of 4 parallel 5 item subtests. Suppose the test items are administered in no particular order of difficulty and that his/her responses for each subtest are as follows 10100, 11010, 00111, 10010. We begin by computing the *U* and U_a values for each subtest which will be used to find the $C(0)_j$ and the associated weight found in Table 7.

Table 7

Initial Statistics for Calculating the Internal Consistency Index

	Item								
Subtest	1	2	3	4	5	U_j	U_{ja}	$C(0)_j$	w_j
1	1	0	1	0	0	6	1	-.67	.25
2	1	1	0	1	0	6	1	-.67	.25
3	0	0	1	1	1	6	6	1.0	.25
4	1	0	0	1	0	6	2	-.33	.25

Note. Responses of 1 student on a 20 item test with 4, 5 item subtests.

The items must now be rearranged to fall in descending order of difficulty. U_j remains the same regardless of item order, as does the associated weight. But, U_{ja}, now referred to as U'_{ja}, must be recalculated based on $\bar{S}'S$. The new values will form $C_p(0')_j$. As shown in Table 8, the internal consistency index for this individual is then 0.998.

Table 8

Change in Statistics for ICI

	Item								
Subtest	2	5	3	4	1	t_j	U_j	U'_{ja}	$C_p(0')_j$	w_j
1	0	0	1	1	1	4	6	5	.663	.25
2	1	0	0	1	1	3	6	4	.33	.25
3	0	1	1	0	0	3	6	3	2	.25
4	0	0	0	1	1	4	6	6	1	.25

Note. Items reordered in ascending order of difficulty.

The meaning of the *ICI* value must be interpreted with a consideration of total test score. A high *ICI*, as in our example, indicates that the student is very consistently applying rules of operation to the test questions. If a student also has a high test score then they are applying rules correctly. However, it is possible that a student is consistently applying misconceptions to solve problems. This situation would result in a high *ICI* and a low total test score. These students are convinced of inaccurate information and may be resistant to remediation. On the other hand, students with low test scores and low *ICIs* have no rules for problem solving and are randomly applying methods of solution. *ICI* proved quite useful for the purposes of discriminating between these types of students. However, it has the real life disadvantage of requiring multiple, often 3 or 4, parallel subtests.

A thorough evaluation of the usefulness of group-based person-fit statistics discussed is difficult. Meijer & Sijtsma (1999) claimed that the research is incomplete or inconsistent and characteristics of the data sets are often unclear leaving others at a loss for comparisons and replications. However, a number of issues can be determined; the drawbacks are obvious. As discussed, if the null distribution is not known for the statistics, it can then not be decided on the basis of significance probabilities when a score pattern is unlikely given Type 1 error rates. However, the most serious shortcoming is that the statistics depend on the total score. This means that when one critical value is used across all test scores, the probability of classifying a score pattern as aberrant will vary depending on where in the score range the individual's total score lies.

With these problems, the use of group-based statistics can be relied on for little more than exploratory and perhaps descriptive purposes.

With a growing awareness of item response theory in the late 1970s and early 1980s, new developments in person-fit methodology attempted to correct for the problems inherent in group-based statistics. The class of IRT based person-fit statistics was born. The advantage of these methods is the opportunity to evaluate the fit of item score patterns to an IRT model. Most IRT person-fit statistics share the common form found in Equation 7.

$$V = \sum_{g=1}^{k} [X_g - P_g(\theta)] \; w_g(\theta) \tag{7}$$

Where $P_g(\theta)$ is the probability of answering item g out of k total items correctly given a specific ability, X_g is the item response as coded 1,0 and $w_g(\theta)$ is an assigned weight. Due to the nature of binary scoring required in tests based on typical IRT models, $X_g^2 = X_g$ so for a suitable new weight of $v_g(\theta)$, the general form of the person-fit statistic can be expressed as Equation 8.

$$V^* = \sum_{g=1}^{k} [X_g - P_g(\theta)]^2 \; v_g(\theta) \tag{8}$$

In item response theory, $P_g(\theta)$ is a function of both person and item characteristics. The specific number of item characteristics depends on the model chosen. The most common models are the one-, two- and three-parameter logistic models. As their names suggest, they each incorporate 1, 2 or 3 item parameters or characteristics into the calculation of $P_g(\theta)$.

The one-parameter logistic model (1PLM), also known as the Rasch model, assumes that individual test performance is affected by person ability and item difficulty.

In Equation 9 the probability an individual of ability θ answering an item correctly is based on a constant, e = 2.718 and the difficulty parameter for item *i,* b_i.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \quad (9)$$

Item difficulty can be understood as the point on the ability scale where the probability of a correct response is 0.5. Therefore the greater the value of b_i the greater the ability needed to have a 50% chance of answering the item correctly.

The two-parameter logistic model (2PLM) in Equation 10 takes the same form, however with the addition of an item discrimination parameter, a_i, and a scaling factor, *D*. When *D*=1.7, it has been shown that $P_i(\theta)$ for the two-parameter normal ogive and logistic models differ by less than 0.01 (Hambleton, Swaminathan & Rogers, 1991).

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad (10)$$

Item discrimination is proportional to the slope of the item characteristic curve at b_i and is an indication of how well an item separates ability levels.

Both the previous models assume that there is a near zero probability of a low ability test taker answering an item correctly. Obviously, this excludes the possibility of correctly guessing item answers. The form of the three-parameter (3PLM) found in Equation 11 allows for this with the addition of the pseudo-chance-level parameter, c_i. However, this should not be thought of precisely as a guessing parameter due to the typical empirical finding that c_i can be lower than a value resulting from random guessing.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad (11)$$

IRT based person-fit statistics can be divided into two general classes, residual-based and likelihood-based, depending on the assumptions and model used to determine the weight in the general person-fit equation. Much of the work in the area of residual-based statistics was pioneered by Wright and Stone (1979) and Wright and Masters (1982). They proposed two mean-squared residual-based statistics called *U* and *W*. The weight of *U* in Equation 12 is based on the conditional variances of the item mean scores.

$$v_g(\theta) = \frac{1}{kP_g(\theta)[1-P_g(\theta)]} \tag{12}$$

Where *k* is the total number of items. Placed into the general form of the person-fit statistic,

$$U = \sum_{g=1}^{k} \frac{[X_g - P_g(\theta)]^2}{kP_g(\theta)[1-P_g(\theta)]} \tag{13}$$

This would result in a low value if a student answered an expected item incorrectly when they had a low probability of answering that item correctly, and a high value if they answered an item incorrectly for which they had a high probability of answering correctly. The opposite would be true if they answered unexpected items correctly. The individual item values are summed across the entire test.

Since the formula of *U* is based on a specific number of items, *k*, and the discrepancy between a student's response and the probability of a correct response, unexpected responses to items far from an individual's ability level, perhaps from guessing, cheating or carelessness, would strongly skew its value. The statistic *W* was proposed to be less sensitive to this effect by removing the factor of *k* in Equation 14.

$$W = \frac{\sum_{g=1}^{k}[X_g - P_g(\theta)]^2}{\sum_{g=1}^{k} P_g(\theta)[1-P_g(\theta)]} \tag{14}$$

Using the R-2.8.0 program add-on ltm (Rizopoulos, 2006) and an existing dataset consisting of 1000 student's responses to 5 items on the LSAT exam as an example, the item parameters can easily be calculated. Table 9 contains the difficulty and discrimination coefficients resulting from a 2PL model.

Table 9

Item Parameters of 5 Items on the LSAT Exam

	Item Parameters	
Item	*a* - discrimination	*b* – difficulty
1	0.825	-3.360
2	0.723	-1.370
3	0.890	-0.280
4	0.689	-1.866
5	0.657	-3.124

With the basic R program it is also possible to estimate the ability parameter (θ) for each individual. For simplicity, this example will focus on only 5 possible response patterns and individual parameters seen in Table 10.

Table 10

Dichotomous Response Patterns, Total Score and Ability Estimates of 5 Students on the 5 Item Subtest of the LSAT Exam

	Response (X) to Item (i)						
Student	X_1	X_5	X_4	X_2	X_3	n_j	θ
1	0	0	0	0	1	1	-3.78
2	0	1	1	0	0	2	-2.87
3	0	1	1	0	1	3	-1.31
4	1	1	1	1	0	4	-0.07
5	1	1	0	1	1	4	0.39

Note. Items ordered in ascending order of difficulty.

The 2PLM $P_g(\theta)$ formula can be used with a, b, and θ to determine the probability that a student with a given ability level would answer the item correctly. With the probability of a correct response on each item, both the U and W statistics can be calculated.

Table 11

Probability of a Correct Response, U and W Person-Fit Values

	$P_g(\theta)$						
Student	1	5	4	2	3	*U*	*W*
1	0.36	0.32	0.10	0.05	0.00	40.3	2.1
2	0.67	0.57	0.24	0.14	0.02	1.23	1.57
3	0.95	0.88	0.66	0.52	0.17	4.86	2.56
4	0.99	0.97	0.89	0.83	0.58	0.35	0.72
5	0.99	0.98	0.93	0.90	0.73	2.91	2.55

U can be understood as the mean of the squared standardized residuals. However, the interpretation of the resulting values is not explicit. Intuitively, high results would suggest aberrant response patterns. But there are no guidelines for critical values available. We can only discuss values in comparison to one another.

Knowing how *U* is calculated we can at least discuss why some of the resulting values in Table 11 are high. Student 1 is of very low ability (θ = -3.78) and therefore has a very low probability of answering any of the 5 items correctly. It is not surprising that this student's total score on the test is 1. However, the item that this individual answered correctly is the most difficult (b = -0.28). With a low ability the probability of answering this item correctly is very low, $P_g(\theta)$ = 0.005. The unexpected response to this single item drastically affects the total *U* value contributing to *U* = 40.3 for this individual. In fact, it is possible that this item likely impacts the resulting fit statistic for Student 3.

Student 3 answered item 3 correctly despite an average ability, θ = -1.31, and a low probability of answering the item correctly ($P_g(\theta)$ = 0.17). Additionally, Student 3 had an unexpected response to item 1. This student answered the item incorrectly despite a very high probability, $P_g(\theta)$ = 0.95, of answering this easy item, b = -3.360, correctly. Collectively the result is U = 4.86. The U value of Student 5 (U = 2.91), though not nearly as extreme as that of Students 1 and 3, is much higher than both remaining students. Perhaps Student 5's response pattern should also be considered aberrant. In this case, the high value can potentially be explained by the unexpected response on item 4. This student is of the highest ability (θ = 0.39) and had a probability equal to 0.93 of answering this item correctly, but item 4 was the only item Student 5 answered incorrectly. Through this discussion, it is both obvious how one unexpected response to an item far from a student's ability can affect the resulting value of U, and how difficult interpretation can be without the convenience of known critical values.

In general, the comparative results of the W statistic, between 0.72 and 2.56, demonstrate the decreased sensitivity to unexpected responses on items far from the student's ability level. The fit statistic of Students 1, 3 and 5 are now within range of the other students. But, again we have the problem of interpreting the significance of this statistic due to a lack of critical values. Is it that the response patterns of Students 1, 3 and 5 are to now be considered normal, or that all the values at or near 2 are aberrant?

Wright and Masters (1982) proposed a transformation of each statistic which they assumed were asymptotically standard normally distributed, Equations 15 and 16. With k-1 degrees of freedom,

$$ZU = [lnU + U + 1]\left(\frac{df}{8}\right)^{\frac{1}{2}} \quad (15)$$

and where q is the variance of W,

$$ZW = \frac{3(W^{\frac{1}{3}} - 1)}{q + \left(\frac{q}{3}\right)} \quad (16)$$

Considering the general critical values of +/- 1.96, *ZU* in Table 12 confirms that the response patterns of Students 1, 3 and 5 are significantly aberrant. On the other hand, the outcomes of ZW indicate that none of the response patterns of these students is of concern. It is possible that the unexpected responses to a few items are continuing to skew even the standard form of *U* and that the decreased sensitivity of *W* is more apparent in its standard form.

Table 12

Standardized Versions of U and W Statistics

Student	*U*	*ZU*	*W*	*ZW*
1	40.3	30.4*	2.10	1.05
2	1.23	0.3	1.57	0.6
3	4.86	3.8*	2.56	1.35
4	0.35	-1.2	0.72	-0.38
5	2.91	2.1*	2.55	1.37

* $p < .05$.

Around the same time, Levine and Rubin (1979) were taking a very different route to person-fit statistics one based on the log-likelihood function:

$$l = \sum_i \{X_i ln P_i(\theta) + (1 - X_i) \ln[1 - P_i(\theta)]\} \quad (17)$$

where *l* is the logarithm of the likelihood function evaluated at the maximum likelihood estimation of θ. Given Equation 17, the value of *l* would be close to zero for students with expected responses, i.e. students who answered items correctly when they had a high probability of doing so and answered items incorrectly when they had a low probability of answering the item correctly. The value of *l* would grow more negative as the students' responses grew more unexpected based on their probability of answering the item correctly, resulting in a skewed negative distribution. For example, we can apply Equation 17 to the responses of 5 students on a 5 item subtest of the LSAT and calculate *l*. In Table 13, we find that Students 1 and 3 not only have the most extreme index values, but that the values are well beyond that of all the other students. Though to a lesser degree, Students 2 and 5 have extreme *l* values compared to Student 4. Perhaps only Students 1 and 3 have aberrant response patterns, or perhaps all four of these students should be considered beyond the normal range. Levine and Rubin were not clear about interpreting the index values.

Table 13

Probability of Answering Item Correctly and the Resulting l Statistic

	$P_g(\theta)$					
Student	1	5	4	2	3	*l*
1	0.36	0.32	0.10	0.05	0.00	-6.29
2	0.67	0.57	0.24	0.14	0.02	-3.27
3	0.95	0.88	0.66	0.52	0.17	-5.95
4	0.99	0.97	0.89	0.83	0.58	-1.21
5	0.99	0.98	0.93	0.90	0.73	-3.16

Note. Items arranged in order of ascending difficulty.

As with many of the proposed statistics before it, the problems applying *l* are obvious. *l* is strongly negatively skewed and since it is not standardized, the classification of any response pattern as aberrant depends on θ and there is no known null distribution on which to judge the resulting value. With no knowledge of the distribution and appropriate critical values we cannot accurately interpret these results.

In 1987, Drasgow, Levine and McLaughlin attempted to standardize the statistic, correct for the dependence on trait level and create an asymptotically standard normal distribution. The statistic took the mathematical form found in Equation 18.

$$l_z = \frac{l - E(l)}{Var(l)^{1/2}} \tag{18}$$

$$E(l) = \sum \{P_i(\theta)[lnP_i(\theta)] + [1 - P_i(\theta)]\ln[1 - P_i(\theta)]\}$$

$$Var(l) = \sum P_i(\theta)[1 - P_i(\theta)][\ln(\frac{P_i(\theta)}{1 - P_i(\theta)})^2]$$

where l_z is the standardized form of the person-fit statistic l. $E(l)$ is the expected value of l and $Var(l)$ is the variance of l. X_i is the examinee's response to item i, coded 1 for correct and 0 for incorrect. $P_i(\theta)$ is the probability that a randomly selected examinee with ability equal to θ will answer item i correctly based on the before mentioned *ICC* equation. *ln* is the natural logarithm of that function.

Continuing with our example, the expected value and variance of l along with the l_z statistic are in Table 14.

Table 14

Expected Values and Variance of l and the Resulting l_z Values

Student	l	$E(l)$	$Var(l)$	l_z
1	-6.29	-1.82	1.18	-4.1*
2	-3.27	-2.36	1.07	-0.87
3	-5.95	-2.36	1.29	-3.17*
4	-1.21	0	1.38	0.4
5	-3.16	-1.28	1.5	-1.53

* $p < .05$.

l_z identifies the patterns of only Students 1 and 3 as significantly aberrant. However, the assumption of normality has been called into question by several researchers. Standardization has shifted the distribution, but l_z is still negatively skewed. Molenaar & Hoijtink (1996) found that l_z can only be considered standard normally distributed when true θ values are used. When the maximum likelihood estimates are used instead, the distribution of the statistic changes and the variance of l_z is smaller than

expected. They claimed that this was especially true in tests of only moderate length. Given our very short five question example, relying on the critical values of a standardized normal l_z is perhaps inaccurate.

Several other IRT-based statistics have been proposed to compete with l_z. Two of these, JK and O/E, stem from the work of Drasgow, Levine and McLaughlin (1987). These statistics considered the flatness of the likelihood function and assumed that the likelihood function would be flat where there was no single value of theta providing good fit for an item score pattern. JK and O/E, though, are very difficult to work with due to the complication of the calculations and because multiple θ estimates are required. This means that θ must be recalculated as potentially problematic items are removed. In addition to the degree of difficulty in computation, Drasgow, Levine and McLaughlin also found that JK and O/E were not effective tools in identifying aberrant response patterns. Though they seemed to be well standardized, detection rates were very low and inconsistent. In all cases, they found that l_z performed at least as well or outperformed its competitors. Due to these findings, JK and O/E will not be discussed in detail; they are mentioned as examples of different methodologies proposed to classify response patterns.

With the types of IRT-based person-fit statistics discussed, the question of when and why to use any one in particular remains. Research into this question has focused on comparing the statistics on ability to detect aberrant response patterns and the influence of person and test characteristics on the statistics. Overall, the detection rates are low for all statistics, but they improve when the test is long, the spread of item discrimination is wide, and there is a large number of aberrant responses present (Meijer & Sijtsma, 1999).

Generally the favorite throughout the literature, l_z, historically outperforms the competitors. Li and Olejnik (1997) found l_z to perform at least as well as all other common statistics, and to be the most powerful in identifying aberrant score patterns. Despite this, remaining problems stem from the assumption of normality previously discussed by Molenaar and Hoijtink (1986). Additional discrepancies were found depending on the IRT model used. Noonan, Boss and Gessaroli (1992) found that l_z approximated the normal distribution but that it was negatively skewed for the two and three parameter logistic model. Using the Rasch model, however, l_z was found to be slightly positively skewed (Li & Olejnik, 1997).

The existing statistics, though seemingly related to the current question of non-attending respondents through the concept of patterned responses, cannot be used to adequately address this task. In addition to the shortcomings outlined in the previous review, IRT-based person fit statistics exclude small-scale tests without underlying IRT modeling, and all tests in which dichotomous scoring is not feasible. Also problematic in using IRT-based statistics when our assumption is that bad data is present is that item parameters depend on the data. If the data is bad then so might be the parameters. Group-based statistics appear more applicable because they compare individual score patterns to those of a group regardless of the scoring model and type of test. However, as discussed these statistics have serious shortcomings and cannot be reliably used to identify aberrant response patterns.

Perhaps, relying on any person-fit statistic – with their assumptions and model-driven mathematical base – is more problematic in this situation than beneficial. Why

not begin with identifying the potential aberrant individuals based on the pattern of response only? Once the researcher is aware of possibly problematic information in the data set, he/she may look more closely at the responses and make an educated decision to keep or remove the individual's record.

In a relatively recent advance, Barnette (1999) has identified 8 possible patterns of non-attending behavior – mono-extreme, mono-middle, big-stepper, small-stepper, checker-extreme, checker-middle, random and random mixture. The first two relate to the repetition of a single response across all items, as indicated by the prefix 'mono'. In a mono-extreme pattern the response repeated is in stark contrast to the group's mean. For example, a pattern of all 1s or all 7s on a 7-point Likert scale would be considered a mono-extreme pattern if the population's mean was close to the midpoint of the scale. In the case of mono-middle patterns, all responses would fall in the middle of the scale. All 3s or all 4s might fit this description. The two 'stepper' classes are called so because the pattern of responses starts at 1 and increases at one point increments up the scale and then decreases at the same rate across all items. A big-stepper continues the pattern over the entire scale – 1234567654321 in the 7-point Likert example, whereas a small-stepper restricts the response range to the extreme ends of the scale only, such as 1232112321. 'Checker' respondents are exemplified by the repeating pattern of only two responses. In the case of a checker-extreme these responses will stretch the entire scale as in the pattern of 1717171717. A checker-middle will use only responses at the middle of the scale, 3535353535 for example. Unlike the strict formation of all previous examples, the random category allows for a completely unsystematic pattern of all possible responses.

The random mixture which includes a combination of all possible patterned response types is perhaps most applicable to actual data situations.

Though helpful in discussing many possible patterns of responding, Barnette did not provide a method for detecting these patterns, but rather based his research on generated data sets forced to include them. In considering a detection technique that could be used, each pattern would require separate methods. Thus, creating an identifier for Barnette's patterned responses would be a tedious and inefficient process.

Proposed New Method

The purpose of this investigation was to identify patterned responding through an analysis of the marginal proportion of each response category and the conditional proportion of a particular response to item *i+1* given the response to item *i*, hereafter referred to as the transitional proportion. The marginal proportions were calculated by totaling each response option for each test taker and dividing by the number of items. A student with disproportionate responses in one category should raise a flag of concern. Then a single matrix of the transitional frequencies was created for each student representing all responses on the test. These frequencies were divided by the number of possible item transitions – the total number of items minus 1 – to get the transitional proportion. A closer look at the matrix revealed an individual's specific responding pattern. For example, a transitional proportion of 1 between response option A on item *i* and response option B on item *i+1* would indicate that the individual followed every answer A with answer B. Though it is unlikely such an extreme case would occur in real test data, high proportion would suggest potentially problematic patterned responses.

In an advantage over several previous methods discussed, the marginal and transitional frequencies can be compared to expected values using a chi square test of significance. However, there are three reasonable options for determining and using expected responses. As is often the case, we might assume that on a typical test correct responses are equally distributed between the answer choices. This is not always realized in actual testing conditions and even slight variation from an equal distribution could cause potential problems. If a test does not follow this strict division of answer options then even students who provide correct answers to every test item could potentially be considered aberrant responders. In this case, we might use the answer key for expected values and as a basis for determining disproportionate response patterns in individuals. Even students with extremely low scores on the test, if attending to the task and addressing individual questions to the best of their ability, are expected to select answer options in a relatively equal proportion. If they fall upon a habit of selecting a single response option, C for example, when they do not know the correct answer then their responses can be identified as patterned and their test results questioned. However, it may be argued that students who do not do well on a particular exam violate the assumptions of fitting expected responses based on the answer key. In a final approach, the way in which the entire group of test takers responded to the items formed the base of the expected values. The average frequencies of the overall sample were used to judge the individual's response patterns. Using observed responses in the sample as expected responses, though avoiding the pitfall just discussed, is problematic in that the determination of aberrance would depend on the performance of the group. With these

limitations, and without prior knowledge of the specific testing conditions and test purposes, we cannot make educated choices about the assumptions underlying each expected response method. In this project all three methods were used.

The chi square results of this new method were compared to those of Drasgow, Levine and McLaughlin's (1987) l_z, an established person-fit statistic. This helped to establish usability and interpretability of the proposed method. Usability here can be thought of as the convenience and practicality of the methods. How difficult is the application of each method and how meaningful are the results? Interpretability, on the other hand, can be conceptualized as the extent to which the results make sense. Can the resulting coefficients be understood and easily described? Though one of the most serious limitations to all person-fit statistics has been the lack of criteria in this regard, this study attempted to provide evidence, or lack of evidence, of these concepts through real world testing data.

In comparison to l_z, there are both advantages and disadvantages to using these chi squares to determine response aberrance. Among the advantages are that the use of IRT is not required and no assumptions of any psychometric model apply. Therefore, a large sample size is not necessary. Additionally, in theory this method will discriminate between students of low ability from those purely lacking motivation. Despite responding incorrectly to many items, a low ability student attending to the test should not be responding in an identifiable patterned way. Finally, an advantage to this new method is automation. The program created here could be modified for generalized use. The most serious disadvantage, however, of marginal and transitional chi squares is that a

large number of items is required. In a typical situation of 4-option answers, a short test of 12 items would leave only 3 items per cell in a chi square analysis of the marginal frequencies. If we wanted the same number of expected items within each cell of a transitional frequencies chi square test we would need 48 items. This is because the chi square of the marginal frequencies is composed of only 4 cells while that of the transitional frequencies has 16 cells. Tests of various lengths were used in this study to explore the impact of the number of items. An additional limitation is embedded in the concept that the new method is model-free. Moving away from models ensures that assumptions do not interfere with application, but there is often a sacrifice in power. With so many alternatives, model-free methods have low power.

METHOD

Participants

The participants in this investigation came from three existing test data discussed separately below.

Psychology Exam

This dataset consists of responses to a 100 item, four-option multiple choice final exam administered to 379 introductory psychology students enrolled at McGill University in Montreal, Quebec in 1989. The data was made public and available by James Ramsey through the TestGraf program.

Beginning Teacher Evaluation Study (BTES)

The data consist of responses to a 60 item, four-option multiple choice mathematics achievement test administered to 118 fifth graders in the San Francisco Bay area in the 1975-1976 academic year as part of the larger multiphase Beginning Teacher Evaluation Study. The data were made available by the Far West Laboratory for Educational Research and Development through the Inter-University Consortium for Political and Social Research (ICPSR). The data were originally collected by the Far West Laboratory while under contract to the California Commission for Teacher Preparation and Licensing. The purpose of the original investigation was to generate and explore effective teaching behaviors. It utilized many sources of data beyond achievement tests, including teacher logs and surveys.

Political Science Survey

The original data set consisted of responses to a 19 item multiple choice survey administered to 1,419 undergraduate students enrolled in an introductory political science course at the University of Illinois at Urbana-Champaign between the fall of 2003 and the spring of 2005. The questionnaire focused on current political knowledge including, "Who is Vicente Fox?", and general governmental process questions such as, "What is the length of a U.S. senator's term?" Only 10 items were kept for this analysis due to differences in response scale that could not be accommodated by the new method. The selected 10 items were all four-option multiple choice questions while the other items were based on a varying number of response options. The students were given extra credit in their courses for completing the survey.

Materials

The program for calculating and testing all statistics was created using SPSS 15.0. R was used to apply Item Response Theory to the samples and establish the parameters needed for the basis of Drasgow, Levine and McLaughlin's (1987) statistic. R was also used for graphing functions.

Design and Procedure

Based on the proposed method of examining marginal and transitional proportions, the current project focused on the creation of a computer based screening tool to identify individuals with highly patterned responses. The previously described three data sets were used to understand the existence and occurrence of response patterns in various testing and survey situations. Two of the data sets, the Psychology Exam and

the BTES mathematics achievement test, can be considered traditional but very different tests: one is strictly a performance test administered in a college course, while the other is an achievement inventory administered to elementary school children. The third data set, the Political Science Survey, is a questionnaire taken by college undergraduates. However, despite the survey nature of the administration, items easily lend themselves to dichotomized correct/incorrect coding due to the four-option response format rather than traditional opinion or attitudinal Likert scales found in surveys.

To begin the analysis, descriptive statistics including the mean and distribution of total scores for each sample were determined to better understand the traditional test results. All data sets were checked for potentially confounding variables such as missing data and unusual response options.

The analysis proceeded through the calculation of marginal and transitional response frequencies and proportions in SPSS. Marginal frequencies were computed by determining the total of each response option instance and then divided by the total number of items to get the proportions. Transitional frequencies were computed by totaling the number of responses within each pair of items, i and $i+1$. Transitional proportions were then computed by dividing by the total number of items minus one – the total number of transitions. The largest marginal and transitional proportions for each individual were selected for calculating descriptive statistics. Stem and leaf plots determined the distribution of those proportions and box plots initially identified potential outliers. However, box plots provide a highly conservative estimate of outliers and are traditionally meant for normal distributions. According to the default SPSS algorithm,

outliers in box plots are points found beyond the distance of 1.5 times the interquartile range below quartile 1 and above quartile 3. This calculation leads to approximately 4 outliers in 1,000 cases in a normal distribution. In order to capture a greater number of outliers, an adjusted box plot was manually calculated based on identifying 10 out of 1,000 cases or 1% of the population. This adjustment multiplied the interquartile range by a factor of 1.22 instead of 1.5.

The marginal and transitional frequencies for each test and each individual were used in the determination of aberrance through a chi square test of significance. All three sets of expected values discussed previously were used. The first set was calculated simply by assuming equal frequencies of all answer options. For example, the expected marginal frequency in a 100 item, 4-option test was 25 for answer option A, B, C and D. With 16 possible item pairs, assuming equal frequencies resulted in the expected transitional frequencies of approximately 6 (99/16) within each pair. Secondly, the expected responses were determined using the answer key in each test. The frequencies resulting from a perfect score on the test served as the expected values. The answer key was incorporated as an additional subject in the test samples so that the initial determination of the marginal and transitional frequencies created our expected values without the need for separate analysis. To test the last expected response method, the marginal and transitional frequencies of all test takers were calculated. With the three sets of expected values, chi square tests were performed on the marginal and transitional frequencies for each individual in each testing condition. There were a maximum of six separate chi squares per individual.

The chi square values were initially investigated through descriptive statistics. The relationship between chi square and score was determined through correlations and plotting. The correlation between score and the chi square based on the answer key was expected to be the strongest and to be negative. As score decreases the number of deviations from expected based on the answer key increase and therefore the chi square will also increase. It is unclear what kind of relationship score might have with the chi squares based on an equal distribution of responses and the answer key.

Critical values for the significance testing were assumed to rest on three and 12 degrees of freedom in the marginal and transitional conditions respectively. Responses are conditional on previous items, leaving three degrees of freedom for each of the four rows in the transitional frequencies. In practice, however, distributions will look different than expected based only on the degrees of freedom employed. Additionally, though the data tend to resemble chi squares, the true distribution is unknown. The chi square was forced and is not expected to fit exactly. Part of the goal of this study was to explore potential cut points for the marginal and transitional chi squares created.

Probability-probability (P-P) plots were used to investigate the distribution of the marginal and transitional chi squares as well as help select the appropriate degrees of freedom to use in significance testing. Multiple plots were run with various degrees of freedom because we began with only a theoretical idea of which are appropriate for these chi squares. Distribution fit was determined by the deviation found in the detrended chi square plots. In these plots, the vertical axis shows the numeric difference between the existing data and what would be expected in a true chi square distribution. The absolute

value of the maximum deviation point was used here because we are interested in the degree and not the direction of discrepancy. For consistent comparison, the degrees of freedom with the best overall fit in each condition and for each data set were used in the testing of significance despite the variation in fit for each type of expected value source. Box plots were additionally employed as a method of identifying outliers in the chi squares. However, due to the before mentioned conservatism and assumptions of box plots, it is assumed that they would result in fewer outliers. The adjusted box plots would identify more outliers, but it is unclear how that would compare to outliers identified in significance testing. The item level responses of individuals deemed aberrant were also visually inspected for emerging patterns in their responses.

In order to compare the results of the new method to that of the l_z person-fit statistic, the raw responses on all tests were converted to a standard 1/0 or correct/incorrect dichotomized scale. R was then used to calculate the item (b) and person parameters (θ) of each test with an underlying IRT Rasch model. The probability of answering item i correctly given the ability θ, $P_i(\theta)$, for each item for each individual was determined using the one-parameter logistic IRT model in Equation 9.

With this information, the l_z person-fit statistic was computed in SPSS for each individual in each testing condition by applying Drasgow, Levine and McLaughlin's (1987) formula in Equation 18.

Descriptive statistics on l_z such as the mean and standard deviation determined the underlying distribution. Correlations between l_z, score and all chi squares served to investigate the relationship between the new method and the existing person fit statistic.

The response patterns deemed significantly aberrant through the application of the new method and through the traditional person-fit statistic was compared and discussed. Due to the nature of the statistics, l_z is attempting to determine which response patterns differ from the IRT model and the chi squares are simply looking for patterns in responding, the results are not expected to overlap heavily. Cohen's Kappa was used as a test of agreement between the two.

The practical implication of identifying and removing highly patterned test results was determined through an analysis of reliability. Cronbach's alpha was used to assess the extent to which the total test score variance was due to true score variance for all items and then for subsets of items with patterned results removed based on significance in chi square conditions and l_z. Alpha is based on an assumption of a single true score and equal covariance between all items. However, these assumptions may be violated here if students are conceptualized as having at least two true scores – one based on taking the test while motivated and another while taking the test without motivation. Guttman's lambda 2 was used as an additional source of reliability.

RESULTS

Psych 101

The mean score on the Psych 101 exam was 64.2 with a standard deviation of 12.3 and a range between 24 and 91. All students completed the exam and the vast majority, 94%, provided an answer for every item. Of the 23 students to leave at least one question blank, only two failed to respond to more than one item. With a mean score of 58.3, these students seemed to do worse overall than the entire class, but this mean was not significantly different from the rest of the population; with a score range of 33 to 78, they were not the worst performers on the exam. The missing data was ignored and all students' responses were used for analysis.

Response patterns were initially investigated through the calculation of marginal and transitional proportions. Descriptive statistics, stem and leaf plots, and box and whisker plots on the largest proportions for each student were used to determine the distribution and identify potential response pattern outliers. Maximum marginal proportions had a minimum value of .26 and a maximum value of .40. With a mean of .30 and a standard deviation of .03, many students favored at least one answer option. The transitional proportions had a mean of .10, a standard deviation of .01, a minimum value of .08 and a maximum of .02. As seen in Figures 1 and 3, the distributions of both proportions had a positive skew, but the divergence from normal was more evident in the marginal proportion. With a skewness value of 0.9, 11 outliers with values greater than .36 were identified by the stem and leaf plot of the largest marginal proportions. The same 11 outliers were identified by the corresponding box plot. The transitional

proportion skewness value was 0.5 and with a lower cut point of .141, only 7 outliers were found in the stem and leaf plot. The corresponding box plot identified only three of these students as outliers due to a slightly elevated cut point of .143. The box plot distributions can be found in Figures 2 and 4.

```
 Frequency    Stem & Leaf

   25.00      26 .  000000000000
   54.00      27 .  000000000000000000000000000000
   64.00      28 .  0000000000000000000000000000000000
   53.00      29 .  00000000000000000000000000
   62.00      30 .  00000000000000000000000000000000
   40.00      31 .  00000000000000000000
   32.00      32 .  00000000000000000
   21.00      33 .  0000000000
   16.00      34 .  00000000
    2.00      35 .  0
   11.00 Extremes    (>=.360)

 Stem width:      .01
 Each leaf:       2 case(s)
```

Figure 1. Stem and leaf plot of the largest marginal proportions for each individual.

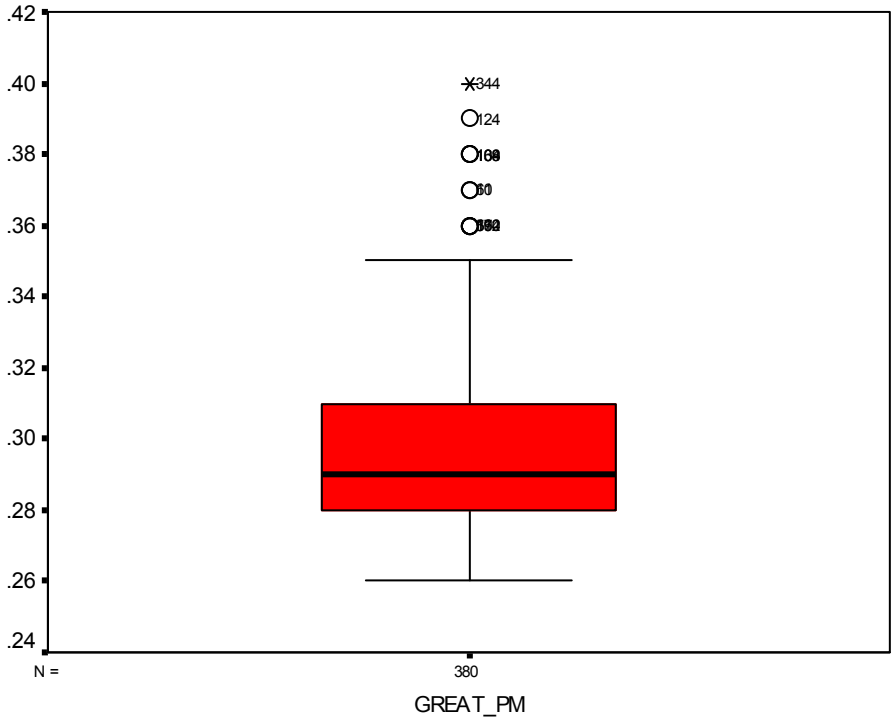

Figure 2. Box and whisker plot of the largest marginal proportions for each individual.

Frequency	Stem & Leaf
32.00	8 . 00000000000
78.00	9 . 000000000000000000000000
112.00	10 . 1111111111111111111111111111111111
92.00	11 . 11111111111111111111111111
42.00	12 . 11111111111111
17.00	13 . 111111
7.00	Extremes (>=.141)

Stem width: .01
Each leaf: 3 case(s)

Figure 3. Stem and leaf plot of the largest transitional proportions for each individual.

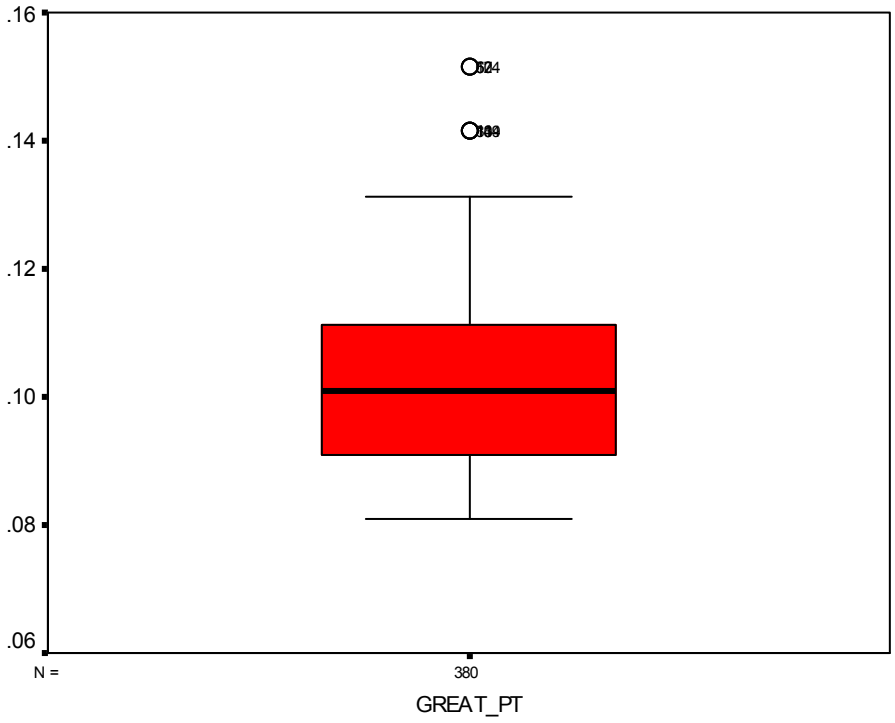

Figure 4. Box and whisker plot of the largest transitional proportions for each individual.

The adjusted box plots based on the distance of 1.22 times the interquartile range below quartile 1 and above quartile 3, resulted in a new cut point of .348 for the largest marginal proportion. Thirteen students were identified as outliers in this condition. The adjusted formula for the transitional proportions only slightly decreased the cut score to .138, but no additional students were identified.

Three separate chi square analyses were performed on the marginal and transitional response patterns depending on the source of the expected values – an equal distribution between response options, the answer key, and the average of all students' responses. The expected values for the analysis based on an equal distribution of responses across the item answer options was 25 for all four marginal frequencies and

6.2875 for all 16 possible transitions. The expected values for the marginal chi square analysis on the average and key responses are indicated in Table 15. Table 16 includes the expected values for the transitional chi squares. Due to some missing item level data, the expected values in the average condition do not sum to the total number of items or transitions.

Table 15

Expected Values for Chi Square Analyses of Marginal Frequencies

	Expected Value Source	
Response	Average	Key
A	23.61	25
B	25.98	28
C	24.02	24
D	26.32	23

Table 16

Expected Values Based on the Answer Key and the Average Response for Transitional Chi Square Analyses

	Response (*i*+1)							
	Answer Key				Average			
Response (*i*)	A	B	C	D	A	B	C	D
A	4	8	6	7	4.87	6.25	5.58	6.81
B	7	6	6	8	6.48	5.82	6.37	6.88
C	7	8	4	5	5.81	6.59	4.76	6.73
D	6	6	8	3	5.68	7.19	7.25	5.79

Significant negative correlations were found between total test score and all six chi squares (Table 17). As expected, the chi squares based on the keyed responses had the strongest correlation with score, -0.47 and -0.54 for marginal and transitional frequencies respectively: as students did better on the test, there was less deviation from expected responses and their chi square values decreased.

Table 17

Spearman's Rho Correlations between Marginal and Transitional Chi Squares and Total Score

Measure	1	2	3	4	5	6	7
				Marginal			
1.Equal	---						
2.Key	0.78**	---					
3.Average	0.90**	0.71**	---				
				Transitional			
4.Equal	0.67**	0.53**	0.58**	---			
5.Key	0.56**	0.75**	0.49**	0.60**	---		
6.Average	0.64**	0.53**	0.66**	0.90**	0.70**	---	
7.Score	-0.31**	-0.47**	-0.25**	-0.31**	-0.54**	-0.33**	---

Note. ** $p < .01$.

Plotting the chi squares against total score provided a closer look at the relationship between each of the chi squares and score. Despite the negative correlation and general trend, Figure 5 demonstrates that the range in marginal chi square statistics varied from a minimum mean of less than one to a high of 11 in students with the lowest total test scores. For those students, the marginal chi square based on the answer key was generally larger than that of the equal and average responses. The average responses resulted in the lowest chi square values in those students.

At the upper end of the score distribution the variation in the marginal chi squares regardless of the source of the expected value decreased, but the mean still fluctuated between zero and four. However, the relationship between the source of expected value within student changed. The chi square based on the key now produced the lowest values – evidence of the negative correlation between test score and the keyed response chi square. Chi squares based on the equal and average responses overlap and result in similar values for these high performing students.

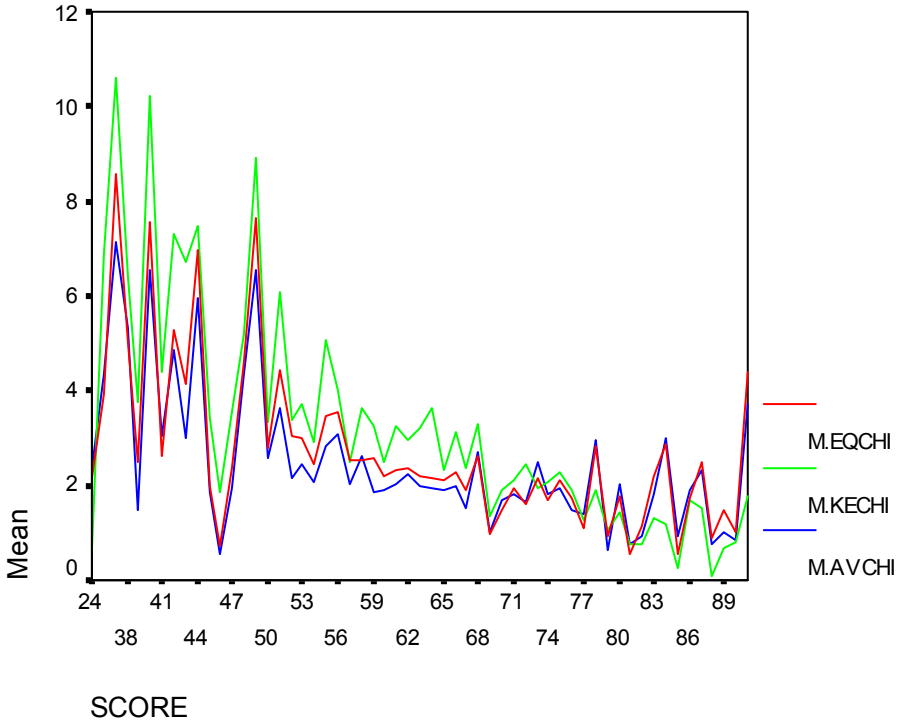

Figure 5. Mean marginal chi squares plotted against total score.

As shown in Figure 6, the pattern appears again in the plot of the transitional chi squares against total test score, but the values resulting from the keyed responses are greater than all other sources. The chi squares based on an equal distribution and on the answer key overlap considerably. At a test score of approximately 75, all three chi squares appear to result in similar values within students. For the highest performing students, the answer key produced the lowest chi square values, followed by the average

responses and an equal distribution of responses.

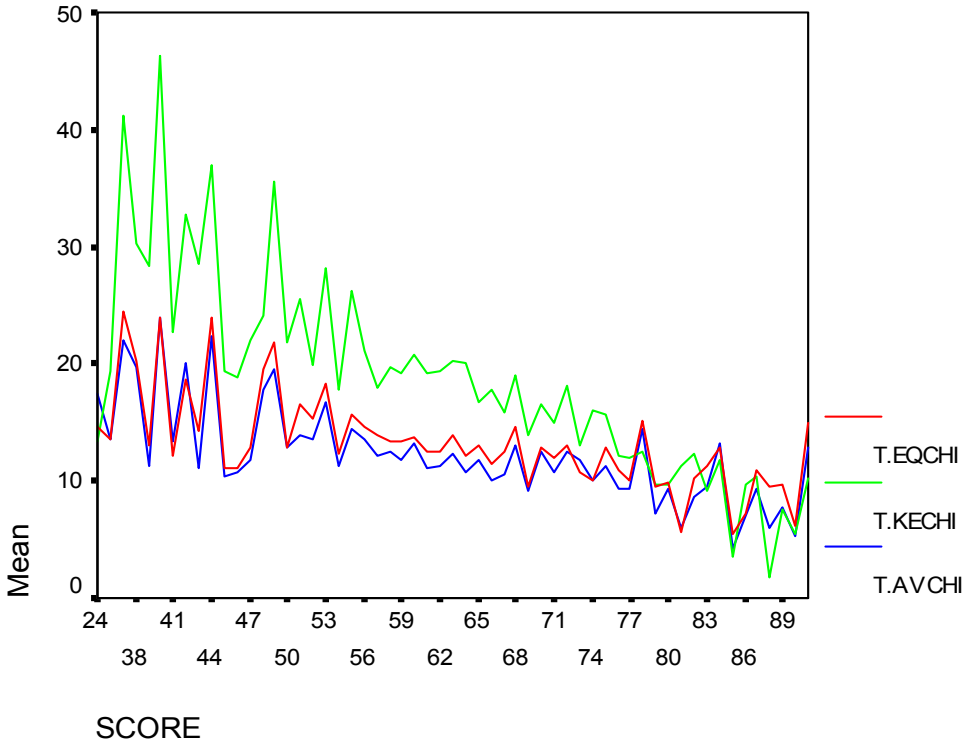

Figure 6. Mean transitional chi squares plotted against total score.

Smoothed line graphs demonstrating the relationship between score and chi square were created in R. As the significant negative correlations suggest, a steep decline occurs in chi squares based on the answer key in both the transitional and marginal conditions as score increases. The chi squares based on the average response and on an equal distribution in the marginal condition had similar patterns of steady decline followed by a slight increase at a score of approximately 70. This score is close to the mean of 64.2 for the entire population. The transitional chi squares of the average score and equal distribution are also similar to each other, but they cross at very low scores and

the chi square based on the average response decreases more quickly as score increases. Figures 7 and 8 show the smoothed line graph.

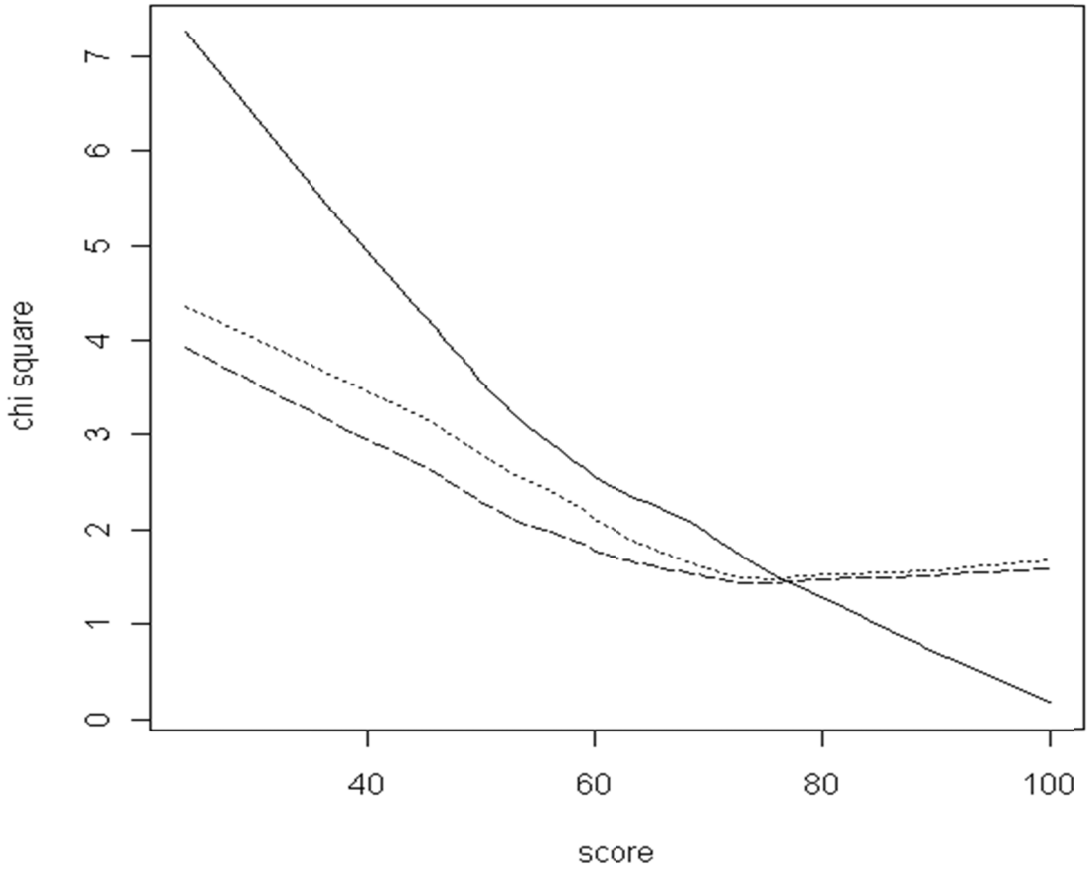

Figure 7. Smoothed line plot of marginal chi square against score.
Figure Key: solid line = answer key, dashed line = average, dotted line = equal

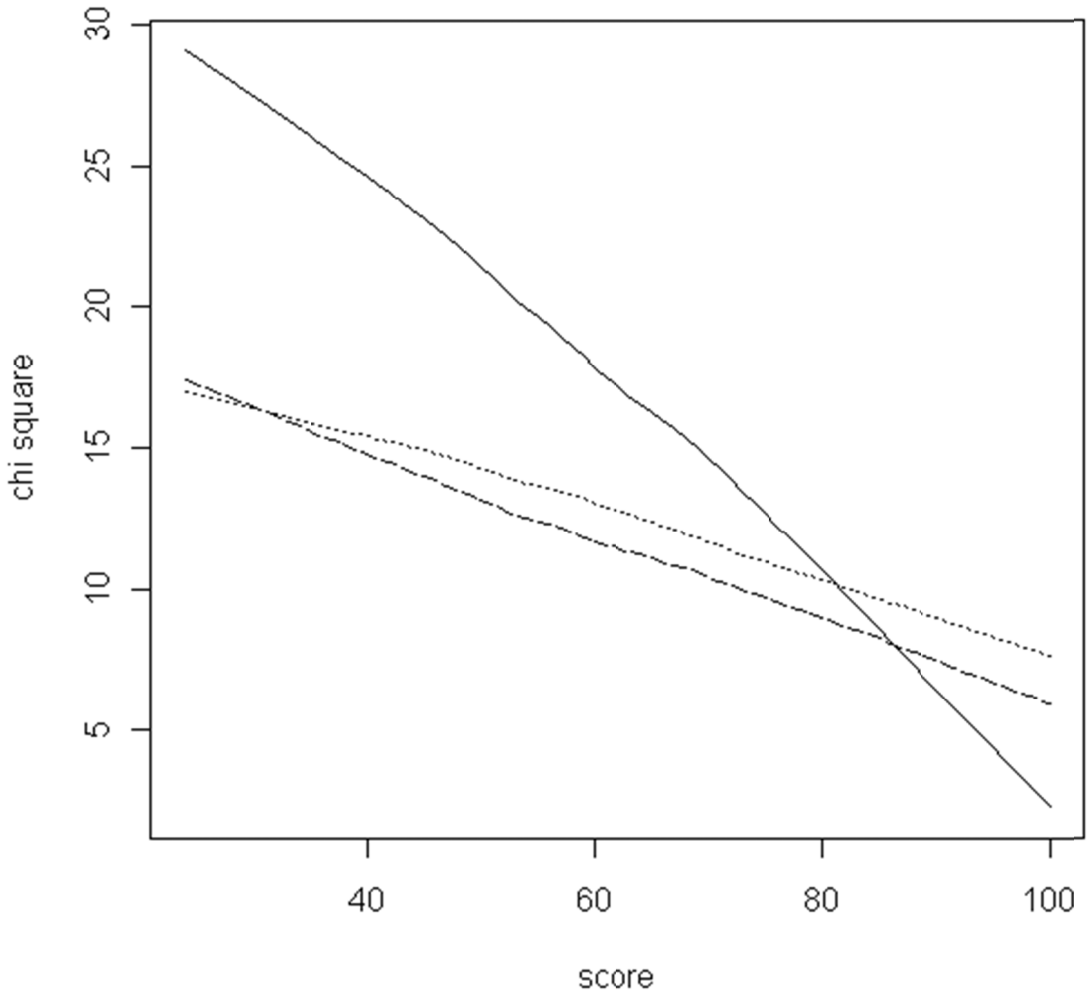

Figure 8. Smoothed line plot of transitional chi square against score.
Figure Key: solid line = answer key, dashed line = average, dotted line = equal

Probability-probability (P-P) plots were used to check the distribution of the marginal and transitional frequencies. Three degrees of freedom in the case of marginal chi squares and 12 degrees of freedom for the transitional chi squares were assumed to be correct based on the conditional responses within the four rows. However, to check these

assumptions several P-P plots were run with varying degrees of freedom. Distribution fit was determined by the deviation from chi square values in the detrended chi square plots. Absolute values were taken of the maximum deviation for each P-P plot. Two degrees of freedom had the best overall fit for marginal frequencies despite the fact that three degrees of freedom was theoretically correct. Three degrees of freedom did, however, work best for the distribution based on the answer key.

The P-P plots for the transitional chi squares revealed a particularly poor fit for the expected values based on the key responses regardless of the degrees of freedom. The maximum deviation was 0.17. It was lowest for 15 degrees of freedom and highest at 12 degrees of freedom. Despite this slight advantage in the case of the keyed responses, 15 degrees of freedom provided the worst overall fit for the transitional chi squares. Thirteen degrees of freedom provided a slightly better fit in the case of the equal distribution of expected value, 0.07 as compared to 0.09, but overall 12 degrees of freedom provided the best fit. The absolute values of the maximum deviations found in the detrended P-P plot for the marginal and transitional chi squares are shown in Table 18.

Table 18

Absolute Maximum Value of Deviation in Detrended Chi Square P-P Plot for Marginal and Transitional Frequencies

	Chi-square Expected Value		
df	Equal	Key	Average
	Marginal		
2	.11	.15	.09
3	.16	.09	.17
	Transitional		
12	.09	.34	.08
13	.07	.27	.15
14	.18	.17	.26

For consistent comparisons, all marginal chi squares were tested for significance based on two degrees of freedom with an alpha level of 0.05 and a critical value of 5.99. All three transitional chi squares were tested for significance based on 12 degrees of freedom with an alpha level of 0.05 and a critical value of 21.03. However, due to the range in accuracy of identifiable degrees of freedom for the marginal and transitional chi squares, box plots were used along with typical significance testing to investigate students' responses that fell far from normal within the total population. The results are individually discussed below followed by the summary table of outliers in each condition in Table 19.

The results of the marginal chi squares indicate that the observed answer choice frequencies of 29 students were significantly different than expected based on an equal distribution between all four response options. The corresponding box plot identified 18 outlying students. According to the box plot, an extreme chi square value was equal to or greater than 7.7. This value is much higher than the critical value used for a chi square with two degrees of freedom. Twenty-eight students had an observed frequency of responses significantly different than expected by the average of all students' responses. Again, the box plot identified fewer outliers; only 22 students had a chi square greater than or equal to 6.8. This is somewhat larger than the critical value of 5.99 used with two degrees of freedom. The response patterns of 52 students were found to be significantly different than the answer key. Seventeen of those students were also identified as outliers in the box plot. An outlying value according to the box plot was 9.0.

Of the total population, 322 students did not have marginal chi squares that were significantly different than expected regardless of the method by which the expected values were created for the marginal chi square. However, all three chi squares were significantly different than expected for 22 students. The total score mean of these students was 52.8 with a standard deviation of 12.4. The mean score for this group was significantly lower than the rest of the population, $F(1,378)=21.0$, $p = .000$. But with a range from 36 to 78, these scores fell well within the normal score range in the population and did not include the lowest scores.

Transitional chi squares identified 29 students whose response patterns were significantly different than expected when the average response was used. When an

equal distribution of response options was used for the expected value 41 students had significant chi squares. For chi squares with expected values based on the answer key, 114 students were found to have patterns significantly different than expected. The majority of students were not found to have significantly different response patterns regardless of the method used, N=256. Twenty-one students, however, had response patterns significantly different than expected in all three transitional chi square cases. The mean test score of these students was 52.0 with a standard deviation of 11.7. The scores varied from 36 to 74. This mean was significantly lower than the rest of the population of test takers, $F(1,378)=23.1$, $p = .001$.

The box plots identified fewer outliers in all three categories of transitional chi squares. Only eight students were classified as outliers in the box plot of chi squares based on an equal distribution of expected values. Accounting for the decrease, chi square values in the box plots equal to or greater than 28.3 were classified as outliers. This value is much higher than the 21.0 cut off used in a test of significance with 12 degrees of freedom. There were 12 students identified as outliers by the box plot of the chi square with expected values based on the average response of the population. These students had chi squares equal to or greater than 26.6 - a value much higher than the critical value of the significance test. In the case where the expected value was based on the answer key, 17 students were identified as outliers by the box plot. These students had chi squares equal to or greater than 41, almost double the critical value used in significance testing.

As expected, the revised box plots with a lower cut point identified more outliers in each marginal chi square condition. Twenty-two students were identified in the chi squares with expected values based on an equal distribution of response and based on the answer key. Twenty-seven students were identified as outliers in the chi square with the expected value based on the average response. The revised box plots on the transitional chi square with an equal distribution of expected values revealed 14 outliers based on the cut point of 26.1. With expected values based on the average response and a new cut point of 24.5, 16 students were identified as outliers. Twenty-one students were identified as outliers in the transitional chi square based on the answer key and a revised cut point of 37.8.

Sixteen students had significant values on all six chi squares. The mean score of these students, at only 50 points, was significantly lower than the rest of the population, $F(1,378)=21.1$, $p = .000$. Looking at the individual responses, repetitive selection of response options is apparent, but otherwise no distinct patterns can be detected. The individual item level data for these 16 students is presented in Appendix A.

Table 19

Summary of Outliers for all Chi Squares by Method of Identification

	Method for identifying outliers			
Expected Value	Critical Value	Stem & Leaf	Box Plot	Adjusted Box plot
		Marginal		
Equal	29	22(≥7.3)	18(≥7.7)	22(≥6.95)
Average	28	24(≥6.4)	22(≥6.8)	27(≥6.2)
Key	52	21(≥8.7)	17(≥9)	22(≥8.2)
		Transitional		
Equal	41	9(≥29)	8(≥28.3)	14(≥26.1)
Average	29	12(≥26)	12(≥26.6)	16(≥24.5)
Key	114	19(≥40)	17(≥41)	21(≥37.8)

The response patterns were also analyzed using the person-fit statistic, l_z. As shown in Table 20, significant correlations exist between l_z and test score, as well as between l_z and the six chi squares based on marginal and transitional frequencies. l_z is positively correlated with test score while negatively correlated with all chi squares.

Table 20

Spearman's Rho Correlations between l_z, Chi Squares and Test Score

	Marginal			Transitional			
	Equal	Key	Average	Equal	Key	Average	Score
l_z	-0.22**	-0.23**	-0.29**	-0.20**	-0.23**	-0.31**	0.34**

Note. $^{**}p < .01$.

Values of l_z had a mean of 0.02, a median of 0.16 and a standard deviation of 1.03, and ranged from -4.86 to 2.21. Of the total population, 21 students were found to have significantly aberrant response patterns according to the results of the l_z analysis. The test scores for these students ranged from 24 to 61 with a mean score of 48.8 and a standard deviation of 9.5. This mean was significantly lower than the rest of the population, $F(1,378)=38.23$, $p = .000$. Only three of these students also had significant chi squares for all three marginal chi square cases, while 10 of them had no significant marginal chi squares at all. With eight students having both, the marginal chi square based on the answer key provided the greatest overlap in significance. Six students with a significant l_z also had significant marginal chi squares based on the average response, and four had significant chi squares based on an equal distribution of responses. Three students had a significant l_z plus three significant transitional chi squares. Seven students had no significant transitional chi squares at all. The pattern of minimal overlap between the chi square results and l_z appeared again in the transitional condition where 14 students also had significant chi squares based on the answer key, six had significant chi squares based on the average response and five had significant chi squares based on the equal distribution of responses. Two students presented with response patterns significantly

different than expected in all six chi squares and l_z. However, the majority of students were not found to have significantly different response patterns by any of the seven methods. Table 21 summarizes the significant outcomes in chi squares compared to l_z.

Table 21

Marginal and Transitional Chi Squares Compared to Significance of l_z

		l_z			
		Not Significant	Significant	Not Significant	Significant
Expected Value	Significance	Marginal		Transitional	
Equal	Not Significant	334	17	323	16
	Significant	25	4	36	5
Average	Not Significant	337	15	336	15
	Significant	22	6	23	6
Key	Not Significant	315	13	259	7
	Significant	44	8	100	14
All	Not Significant	340	18	341	18
	Significant	19	3	18	3

Cohen's Kappa was used to quantify the degree of agreement between l_z and the chi squares. Though there are no standard critical values, Landis and Koch (1977) provided some benchmarks for discussing the strength of agreement. No kappa values

between l_z and any of the chi squares demonstrated a strong relationship. At .194, the strongest Kappa value was found between the marginal chi square based on the average response and l_z. The weakest relationship, .095, was between l_z and the transitional chi square based on an equal distribution of responses. All of the kappa values fell in the slight agreement category set forth by Landis and Koch.

The reliability of the test based on the entire population was .885 with a lower bound confidence interval of .868 and an upper limit of .900. Removing test takers with significant indications of patterned responding resulted in slightly lowered reliability estimates in all conditions, however all were within the 95% confidence interval. The greatest reduction in reliability was found when students with any significant chi squares, with any significant marginal chi squares or with any transitional chi squares were removed, $\alpha = .868$, $\alpha = .872$, $\alpha = .871$ respectively. Removing these students also resulted in the greatest decrease in population, $n = 252$, $n = 322$ and $n = 256$ respectively. Little difference was found in the reliability estimates resulting from removing students with only one significant chi square. The marginal and transitional chi squares based on the expected value of equally distributed response patterns, with $\alpha = .881$ and $\alpha = .882$, had the highest reliability among the chi square conditions. Chi squares based on the average response had reliability estimates of $\alpha = .877$ and $\alpha = .880$ for the marginal and transitional conditions. Removing the responses of students with significant marginal or transitional chi squares based on the answer key both resulted in a reliability estimate of $\alpha = .874$. The reliability estimate after removing students with a significant l_z was $\alpha = .878$ which fell in the middle of all the estimates. The method of removing students with

significant chi squares did as well as removing students with significant person-fit statistics, both of which however did not improve the test reliability.

Due to the fact that students in the data set can be conceptualized as having two true scores – one while motivated and one while unmotivated – the assumptions of Cronbach's alpha may be violated. Guttman's lambda estimate was included as another source of reliability. Guttman's lambda 2 was slightly larger than the corresponding alpha reliability estimate in all conditions, but the same pattern of decreasing reliability occurred. Lambda 2 was .889 for the entire population.

Beginning Teacher Evaluation Study (BTES)

In the original population of 118 students, eight were found to have multiple responses on at least one item and six had no valid responses at all. The remaining 104 tests were used in this analysis. Out of 60 total items, the mean test score for these students was 23.6 with a standard deviation of 9.8 and a range between 6 and 50. The test score distribution was positively skewed. Eighty-one students had at least one missing item. The mean for this group, M = 21.99, was significantly lower than those without any missing items, M=29.1, $F(1,103)=10.2$, $p = .000$. These students were kept in the analysis despite the missing data, but an adjusted chi square was proposed to minimize the impact.

Marginal and transitional frequencies and proportions were calculated for each student. Descriptive statistics, stem and leaf plots, and box and whisker plots were run on the largest proportion for each individual to determine the distribution and identify potential outliers. The mean of the maximum marginal proportion was .26, but the

individual values varied considerably between students. The standard deviation was .09 and proportions ranged from .03 to .40. A skewness of -0.58 resulted from the students at the low end of the range. No outliers were identified in either the stem and leaf or box plots found in Figures 9 and 10. The transitional proportions, ranging from .02 to .31, had a mean of .10 and a standard deviation of .04. A positive and larger skewness of 1.03 was found here. One student was identified as an outlier based on a cut point of .31 in the stem and leaf (Figure 11). With a lower cut point of .20, the box plot also identified one outlier (Figures 12).

```
Frequency    Stem & Leaf

   1.00      0 .  3
   3.00      0 .  568
   9.00      1 .  013333333
  12.00      1 .  566666688888
  12.00      2 .  000000113333
  26.00      2 .  55566666666888888888888888
  25.00      3 .  0000111111111111133333333
  14.00      3 .  55555666668888
   3.00      4 .  000

 Stem width:      .10
 Each leaf:      1 case(s)
```

Figure 9. Stem and leaf plot of the largest marginal proportions for each individual.

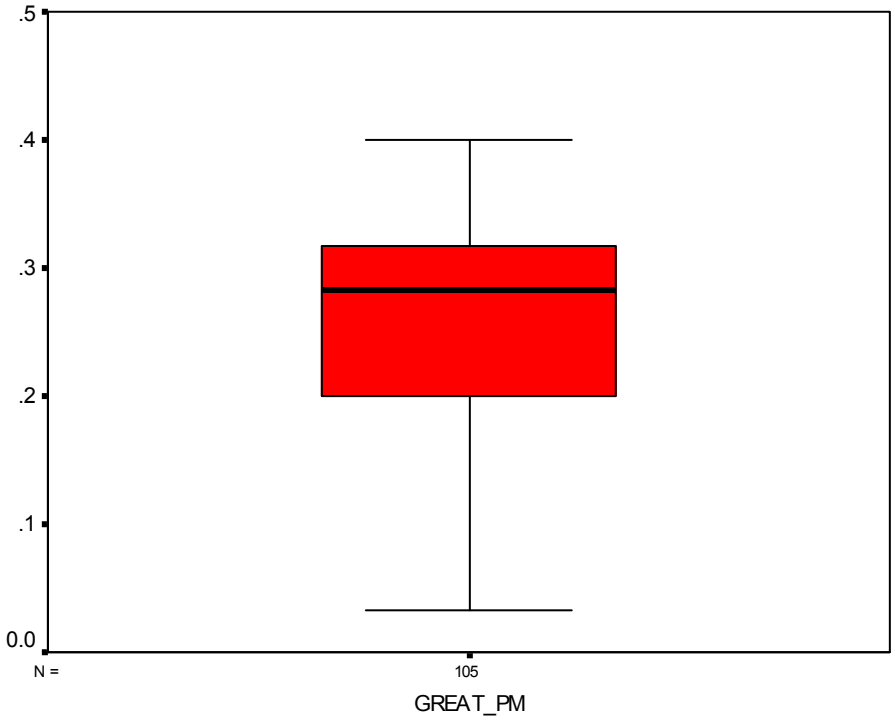

Figure 10. Box and whisker plot of the largest marginal proportions for each individual.

Frequency Stem & Leaf

3.00 1 . 666
3.00 3 . 333
10.00 5 . 0000000000
15.00 6 . 777777777777777
13.00 8 . 4444444444444
21.00 10 . 111111111111111111111
17.00 11 . 88888888888888888
9.00 13 . 555555555
10.00 15 . 2222222222
1.00 16 . 9
2.00 18 . 66
1.00 Extremes (>=.305)

Stem width: .01
Each leaf: 1 case(s)

Figure 11. Stem and leaf plot of the largest transitional proportions for each individual.

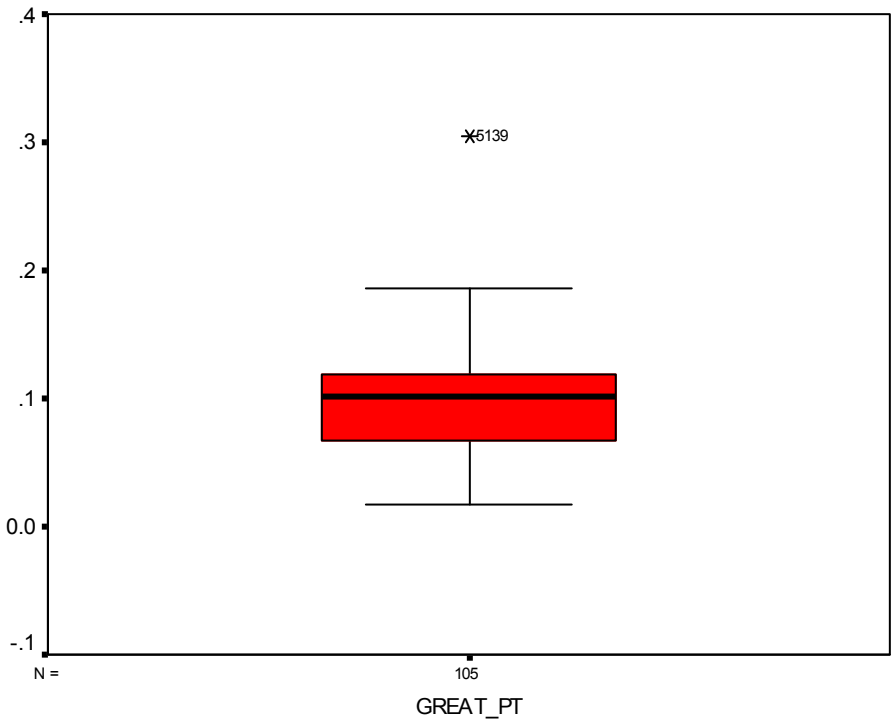

Figure 12. Box and whisker plot of the largest transitional proportions for each individual.

The adjusted box plots resulted in a slight decrease in cut point in both conditions, .52 and .19 for the marginal and transitional proportions respectively. This drop did not affect the number of outliers.

Chi square analyses were performed on the marginal and transitional response patterns according to the source of the expected values – an equal distribution between response options, the answer key, and the average of all students' responses. The equal distribution of response options resulted in an expected value of 15 in all four marginal frequencies and 3.69 for each of the 16 transitional frequencies. The expected values for

the average responses and the answer key are listed in Table 22 for marginal chi squares and Table 23 for transitional chi squares.

Table 22

Expected Values for Chi Square Analysis of Marginal Frequencies

	Expected Value Source	
Response	Average	Key
A	10.38	17
B	10.31	12
C	14.76	20
D	8.16	11

Table 23

Expected Values Based on the Answer Key and the Average Response for Transitional Chi Square Analyses

	Response (*i*+1)							
	Answer Key				Average			
Response (*i*)	A	B	C	D	A	B	C	D
A	3	6	4	3	2.30	2.41	3.38	1.53
B	2	2	6	2	2.23	2.37	3.56	1.35
C	9	2	5	4	3.61	2.95	3.55	3.43
D	3	2	4	2	1.61	1.84	2.84	1.43

Due to the large extent of missing data, expected values based on the average response do not sum to the total number of items or to the total number of transitions. Only 43 items were represented in the marginal average and only 40 in the transitional average. An adjusted expected value was created and used in calculating the individual students' chi squares. This adjustment considered the number of valid responses for each student. Equation 19 represents the revised chi square.

$$X^2 = \sum_i^k \frac{[O - (E\, {}^{g}/_{k})]^2}{(E\, {}^{g}/_{k})} \quad (19)$$

where *k* is the total number of items on the test and *g* is the number of valid items for each individual.

Only the marginal and transitional chi squares based on the answer key were significantly correlated with total test score, r = -.287, p=.00 and r = -.425, p=.00 respectively. The negative correlations suggest that as score increased the chi squares in both cases decreased. The correlations are summarized in Table 24.

Table 24

Spearman's Rho Correlations between Marginal and Transitional Chi Squares and Total Score

Measure	1	2	3	4	5	6	7
				Marginal			
1.Equal	---						
2.Key	.468**	---					
3.Average	.499**	.745**	---				
				Transitional			
4.Equal	.656**	.463**	.547**	---			
5.Key	.206*	.726**	.627**	.491**	---		
6.Average	.206*	.629**	.743**	.644**	.795**	---	
7.Score	.060	-.287**	.096	-.090	-.425**	-.186	---

Note. ** $p < .01$. * $p < .05$.

Plotting the individual chi squares by score demonstrated a consistent pattern not revealed by the correlations alone. At the very low end of the score band, chi square values for all conditions and expected value sources peaked dramatically. At a score of 18 and 14 in the marginal and transitional chi squares respectively, the means for all expected value sources leveled out and there was only a small fluctuation by score. The chi squares based on the average response resulted in higher values than the other chi square sources in both the marginal and transitional conditions. Figures 13 and 14 plot the mean chi squares against total score.

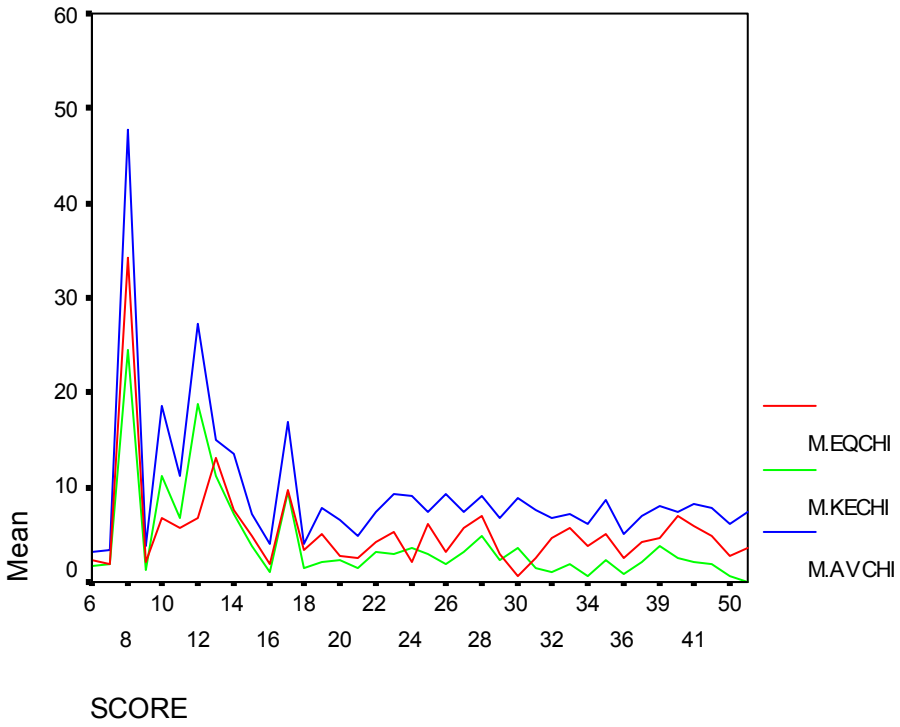

Figure 13. Mean marginal chi squares against total score.

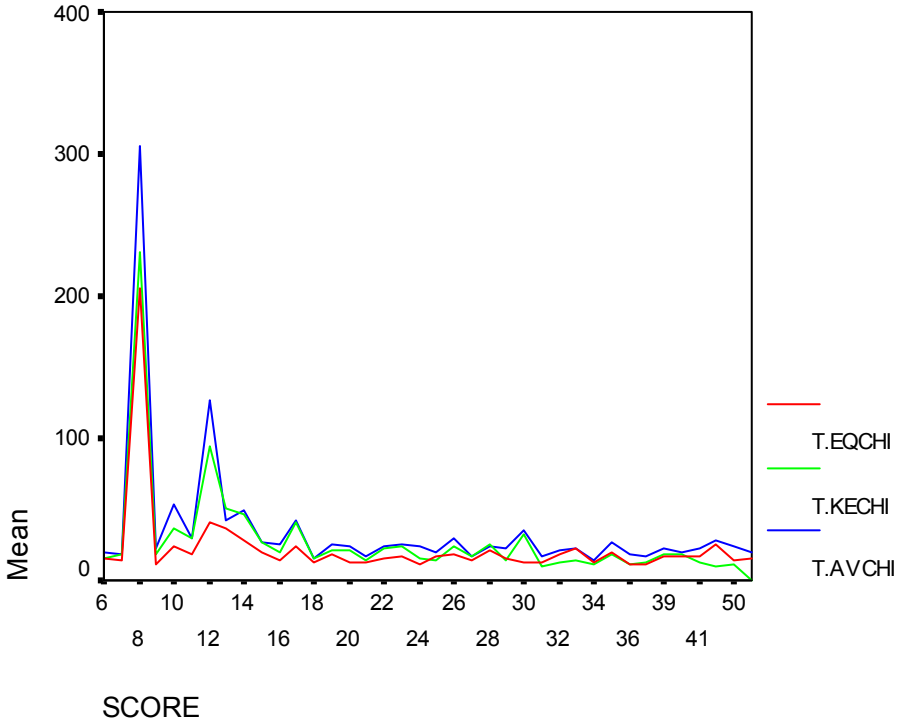

Figure 14. Mean transitional chi squares against total score.

Smoothed lines created in R provided another look at the relationship between final score and chi squares. The significant negative correlation between score and both the marginal and transitional chi squares based on the answer key was evident through the consistent declining slope. As score increased these chi squares decreased. However, not obvious in the correlations was that in both cases the relationship remains relatively flat for low scores. The decline begins at a score of about 20 in the transitional condition and about 25 in the marginal condition. The marginal chi squares based on the average response and on an equal distribution of responses both increased with score before leveling out and declining at the highest score points. In the transitional condition, the chi squares based on the average response and on an equal distribution of response are large and consistent regardless of score. Figures 15 and 16 show the smoothed line graph

between chi squares and test scores.

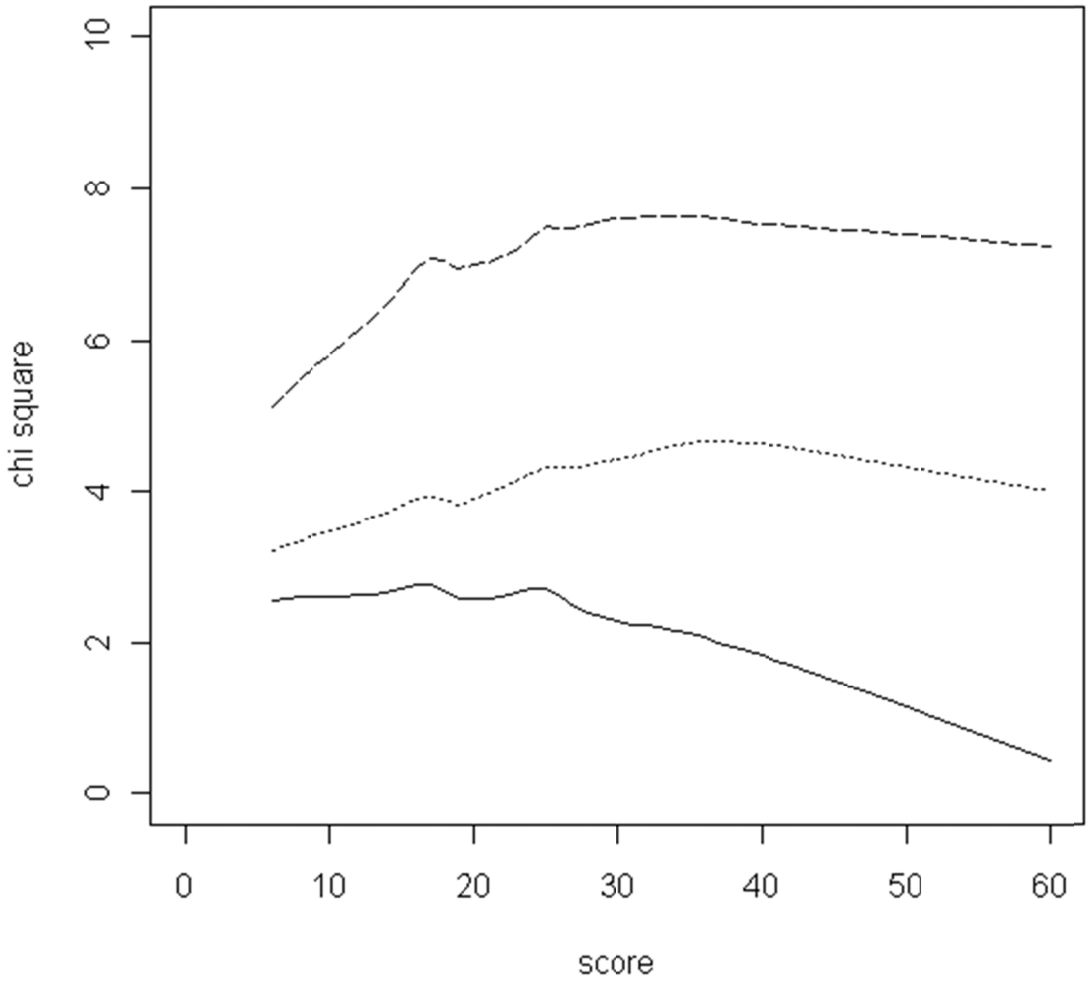

Figure 15. Smoothed line plot of marginal chi square against score.

Figure Key: solid line = answer key, dashed line = average, dotted line = equal

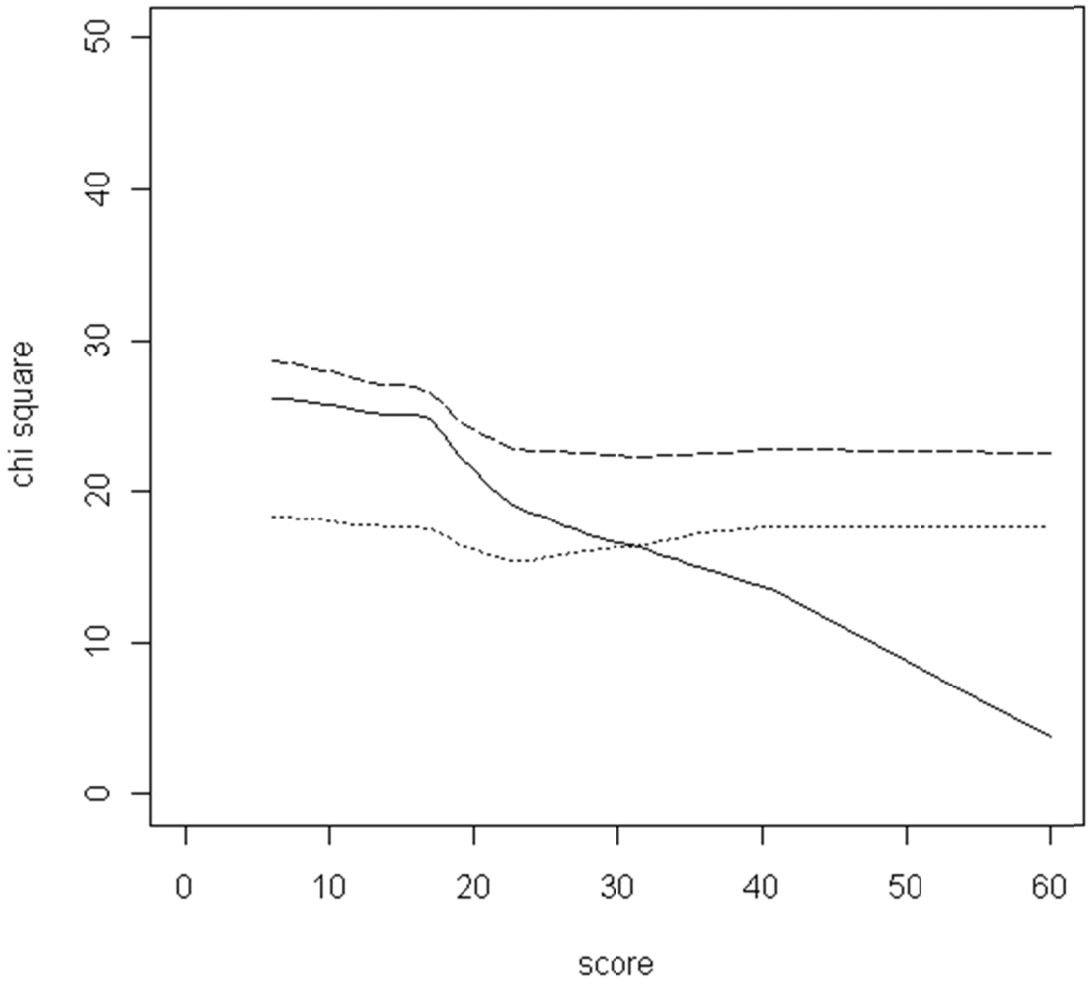

Figure 16. Smoothed line plot of transitional chi square against score.
Figure Key: solid line = answer key, dashed line = average, dotted line = equal

Probability-probability (P-P) plots with various degrees of freedom were used to check the distributions of the marginal and transitional chi squares. Regardless of the degrees of freedom, the condition based on the average response had a poor fit in both marginal and transitional chi squares. Overall, the degrees of freedom required to fit a

chi square to the data were much higher than expected. According to the deviations found in the detrended chi square plots, four or five degrees of freedom in the marginal chi square and 17 to 18 in the transitional chi square provided the best fits for chi squares based on the equal distribution of responses and the answer key. Five degrees of freedom, with a total deviation of 0.77, resulted in the best overall fit for the three marginal chi squares considered together. With a total deviation of 0.60, 18 degrees of freedom described the transitional chi squares best. The maximum absolute values are summarized in Table 25.

Table 25

Absolute Maximum Value of Deviation in Detrended Chi Square P-P Plot for Marginal and Transitional Frequencies

	Expected Value		
df	Equal	Key	Average
	Marginal		
3	.30	.06	.65
4	.15	.20	.50
5	.02	.35	.40
	Transitional		
16	.12	.26	. 45
17	.07	.20	.40
18	.10	.15	.35

For consistent comparisons, five degrees of freedom with $\alpha = .05$ and a critical value of 11.07 and 18 degrees of freedom with $\alpha = .05$ and a critical value of 28.87 were used in significance testing for the marginal and transitional chi squares, respectively. Box plots were used along with standard significance testing to investigate students' responses because of the poor fit and larger than expected degrees of freedom. The results are individually discussed below followed by a summary table of outliers in each condition found in Table 26.

The marginal chi square based on an equal distribution of response choices revealed five students with patterns significantly different than expected. With a cut score of 11.9, the corresponding box plot identified four of these five students as outliers. The response patterns of five students were also found to be significantly different than expected based on the keyed responses. Three of these students had both significant chi squares. However, in the case of the answer key, the box plot identified 11 outliers. The cut value for this box plot was 8.3, lower than that of significance testing. The largest number of students, 19, was captured by the chi square based on the average responses. Only six outliers were found in the box plot and the majority, 12, had no other significant marginal chi squares. The majority of students did not have any significant marginal chi squares, but three students had all three. These students had a mean score of only 12.7, but this was only marginally significantly different than the rest of the population, $F(1,103)=3.9$, $p = .05$.

Compared to the marginal conditions, more response patterns were found to be significantly different than expected in the transitional chi squares regardless of which method was used for calculating the expected value. Eight students had significant chi squares based on the equal distribution of answer options. Four of these students with chi square values greater than or equal to 37.8 were also identified as outliers by the box plot. All eight students with significant chi squares based on an equal distribution of responses also had significant chi squares based on the average response. In addition, seven of these students had significant chi squares based on the answer key. Twenty students had significant transitional chi squares based on the answer key, but only six of these were also identified by the corresponding box plot. The cut point in the box plot was 50, much higher than the chi square critical value. More students were found to be significantly different than expected based on the average response chi square, 29, than in any other condition. With a very high cut score of 53, the corresponding box plot identified only five students with extreme chi squares. The response patterns of most students were not found to be different than expected by any of the three transitional chi squares. Seven students, though, were significant by all three methods. The mean score of these students was significantly lower than the rest of the population, M=14.6, $F(1,103)=6.6$, $p = .01$.

The adjusted box plots on the marginal chi squares did not affect the number of students indentified in the equal distribution of response condition; despite lowering the cut point to 11.9, four students were considered outliers. However, the decrease in cut point to 8.3 in the chi square based on the answer key resulted in 11 outliers and a new cut point of 16 in the average response condition resulted in six outliers.

The box plot adjustments in the transitional chi squares made only a slight difference in the number of outliers identified. A two point drop in cut point led to the classification of six outliers in the chi square based on an equal distribution of responses. Seven students with a chi square based on the answer key greater than or equal to 46.7 were outliers. Six students in the average response condition were found to be outliers based on the new cut point of 50.4.

The response patterns of only three students were found to be significantly different than expected in all six chi squares. These were the only three students who had significant marginal chi squares in all three expected value conditions. The repetitive selection of single response options was obvious in the individual response patterns. In an extreme case, a student selected only A's and C's and left 50% of the answers blank. The item level response patterns for students with all six significant chi squares can be found in Appendix B.

Table 26

Summary of Outliers for all Chi Squares by Method of Identification

	Method for identifying outliers			
Expected Value	Critical Value	Stem & Leaf	Box Plot	Adjusted Box Plot
		Marginal		
Equal	5	4 (≥13.2)	4 (≥13)	4 (≥11.9)
Average	19	6 (≥16.3)	2 (≥17.2)	6 (≥16)
Key	5	12 (≥7.9)	3 (≥(9.1)	11 (≥8.3)
		Transitional		
Equal	8	6 (≥36)	4 (≥37.8)	6 (≥35.2)
Average	29	6 (≥52)	5 (≥53)	6 (≥50.4)
Key	20	8 (≥46)	6 (≥50)	7 (≥46.7)

The students' response patterns were also analyzed with the l_z person-fit statistic. l_z was not significantly correlated with total test score, but was significantly negatively correlated with both the marginal and transitional chi squares based on the answer key and the average response: Table 27. As these chi squares increased, l_z decreased. The greater the chi square, the more likely the student was significantly different than expected. Due to the negatively skewed distribution, small values for l_z would also be considered outliers.

Table 27

Correlations between l_z and Chi Squares

	Marginal		Transitional	
	Average	Key	Average	Key
l_z	-.510**	-.452**	-.533**	-.444**

Note. ** $p < .01$.

Values of l_z had a mean of 0.23, a median of 0.63 and a standard deviation of 1.74 and a range between -5.64 and 2.55. The response patterns of 22 students in the population of 104 had a significant l_z value. The mean test score for these students was slightly, but not significantly, lower than the rest of the population, 21.9 compared to 24.0. Sixteen of these 22 students did not have any significant marginal chi squares and 12 of them were not significantly different than expected in any of the three transitional chi squares. However, six students had significant marginal chi squares based on the average response pattern. Two also had significant marginal chi squares based on the answer key. These two students were significant in all marginal chi squares. Only one student had a significant l_z along with a significant marginal chi square based on the equal distribution of answer options. Eight of the original 22 students with significant l_z values had significant transitional chi squares based on the average response. Eight also had significant transitional chi squares based on the answer key. The response pattern of only two students was significantly different than expected based on the equal distribution of transitions and l_z. These two students had significant transitional chi squares in all three

expected value sources. Table 28 shows the comparison of marginal and transitional chi squares to l_z.

Table 28

Marginal and Transitional Chi Squares Compared to Significance of l_z

		l_z			
		Not Significant	Significant	Not Significant	Significant
Expected Value	Significance	Marginal		Transitional	
Equal	Not Significant	78	21	76	20
	Significant	4	1	6	2
Average	Not Significant	69	16	61	14
	Significant	13	6	21	8
Key	Not Significant	79	20	70	14
	Significant	3	2	12	8
All	Not Significant	80	21	77	20
	Significant	2	2	5	2

Cohen's Kappa was used to quantify the degree of agreement between l_z and the chi squares. With values ranging from -.004 between l_z and the marginal chi square based on an equal distribution of responses to .227 between l_z and the transitional chi square based on the answer key, no Kappas between l_z and any of the chi squares

demonstrated a strong relationship. According to the benchmarks of Landis and Koch (1977), the agreements could be considered slight to fair.

The reliability of the test with all 104 test takers was $\alpha = .920$ with a 95% confidence interval between .896 and .941. Removing students with any indication of patterned responses resulted in either no change or an increase in reliability regardless of the source, but no change exceeded the boundaries of the confidence interval. The largest increase, though still small, resulted from removing the responses of students with either any significant chi squares or any significant transitional chi squares, $\alpha = .930$ in both cases. When students with significant chi squares in any of the three marginal conditions were removed the reliability dropped to $\alpha = .924$. Little variation in reliability resulted from removing students with only one significant chi square. Like the reliability of the entire population, $\alpha = .920$ when students with significant marginal or transitional chi square based on an equal distribution of responses or the marginal chi square based on the answer key were removed. Removing the marginal or transitional chi square based on the answer key resulted in a slight increase in reliability, $\alpha = .924$ and $\alpha = .926$ respectively. The reliability resulting from excluding students with a significant transitional chi square based on the answer key was $\alpha = .925$. Removing students with a significant l_z resulted in a higher reliability ($\alpha = .929$) than that resulting from removing students with single significant chi squares, but was slightly less effective than removing students with multiple significant chi squares.

Guttman's lambda estimate was included as another source of reliability due to the fact that student's in the data set might violate the single true score assumption of

Cronbach's alpha. Guttman's lambda 2 was slightly larger than the corresponding alpha reliability estimate in all conditions, but the same pattern of change in reliability occurred. Lambda 2 was .927 for the entire population. Removing students with any significant chi square increased lambda 2 to .937, slightly larger than the increase to .934 which resulted from removing students with significant l_z coefficients. Removing students with a single significant chi square made little difference in Guttman's reliability estimates.

Political Science

The 1,419 survey takers had an average score of 6.93 on the 10 item questionnaire. With a minimum score of 1 and a maximum score of 10, the median was 7 and the standard deviation was 2.24. With a higher than average mean, the resulting distribution of scores was negatively skewed. All individuals completed the entire survey.

Due to the small number of items in the questionnaire, transitional values were not meaningful, but the marginal frequencies and proportions were calculated. Descriptive statistics, a stem and leaf plot, and box and whisker plots on the largest marginal proportion for each individual revealed a relatively normal distribution with mean of .45 and standard deviation of .10. The values ranged from .30 to 1.0. The 37 students with a marginal proportion of .70 or higher were identified as outliers in the population by the stem and leaf plot in Figure 17. The box plot also identified these 37 students, but with a slightly lower cut point of .65 in Figure 18. The adjusted box plot decreased the cut point to .63, but did not affect the number of students identified.

Frequency Stem & Leaf

165.00 3. 00000000000000
556.00 4. 0000000000000000000000000000000000000000000000
547.00 5. 0000000000000000000000000000000000000000000000
114.00 6. 000000000
37.00 Extremes (>=.70)

Stem width: .10
Each leaf: 12 case(s)

Figure 17. Stem and leaf plot of the largest marginal proportions for each individual.

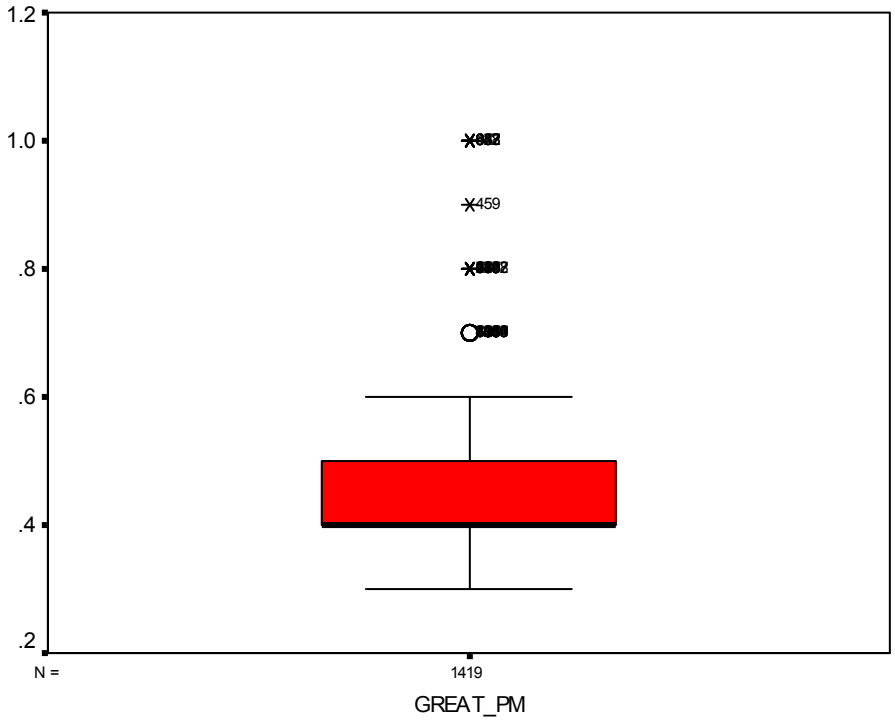

Figure 18. Box and whisker plot of the largest marginal proportions for each individual.

Marginal chi squares with expected values based on an equal distribution of responses and on the average responses were investigated. The marginal chi square based on the answer key could not be run because answer option D was always "Don't Know" leaving only the answer options A, B and C as correct responses. This resulted in

a division by zero for answer choice D, therefore invalidating the chi square. The expected value for the marginal chi square based on the equal distribution of responses was 2.5 for all four answer options. The expected values for the chi square based on the average responses were 2.44, 3.85, 2.31 and 1.40 for response option A, B, C and D respectively. The chi squares were positively correlated with each other, but the chi square based on the average response was negatively correlated with score while the chi square based on an equal distribution of responses was positively correlated with score. The correlations can be found in Table 29.

Table 29

Spearman's Rho Correlations between Marginal Chi Squares and Total Score

Measure	1	2	3
1.Equal	---		
2.Average	.394**	---	
3.Score	.240**	-.428**	---

Note. ** $p < 0.01$.

Plotting the chi squares against score showed a more complicated relationship: Figure 19. The chi squares based on the equal distribution of responses peaked at the lowest total score. A sharp decline followed, but chi squares again increased at a score of four. A slight but steady gain continued through to the highest scores. For the chi squares based on the average response pattern, a steady decline was evident from the

lowest scores up. However, from a score of 7 to 10 there was little variation in chi squares.

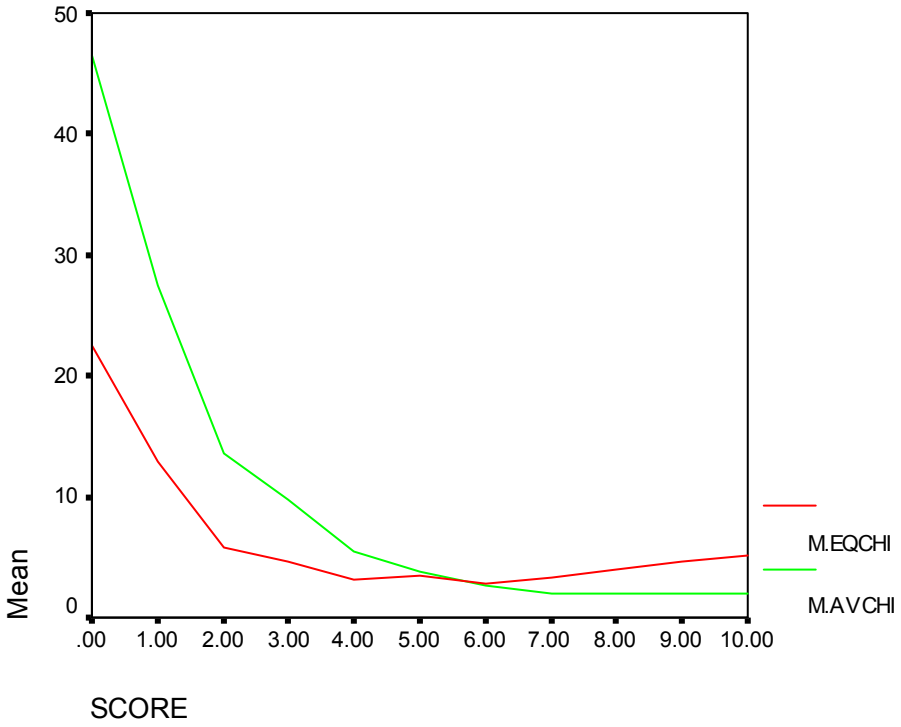

Figure 19. Mean marginal chi squares against total score.

The plot of score by chi square was smoothed using R. The results did not add much to the understanding of the relationship between score and chi square. The rapid and steady decline in the marginal chi square based on the average response as score increased was similar to that in Figure 19, as was the irregular shape between the marginal chi square based on the average response and score evident in Figure 20.

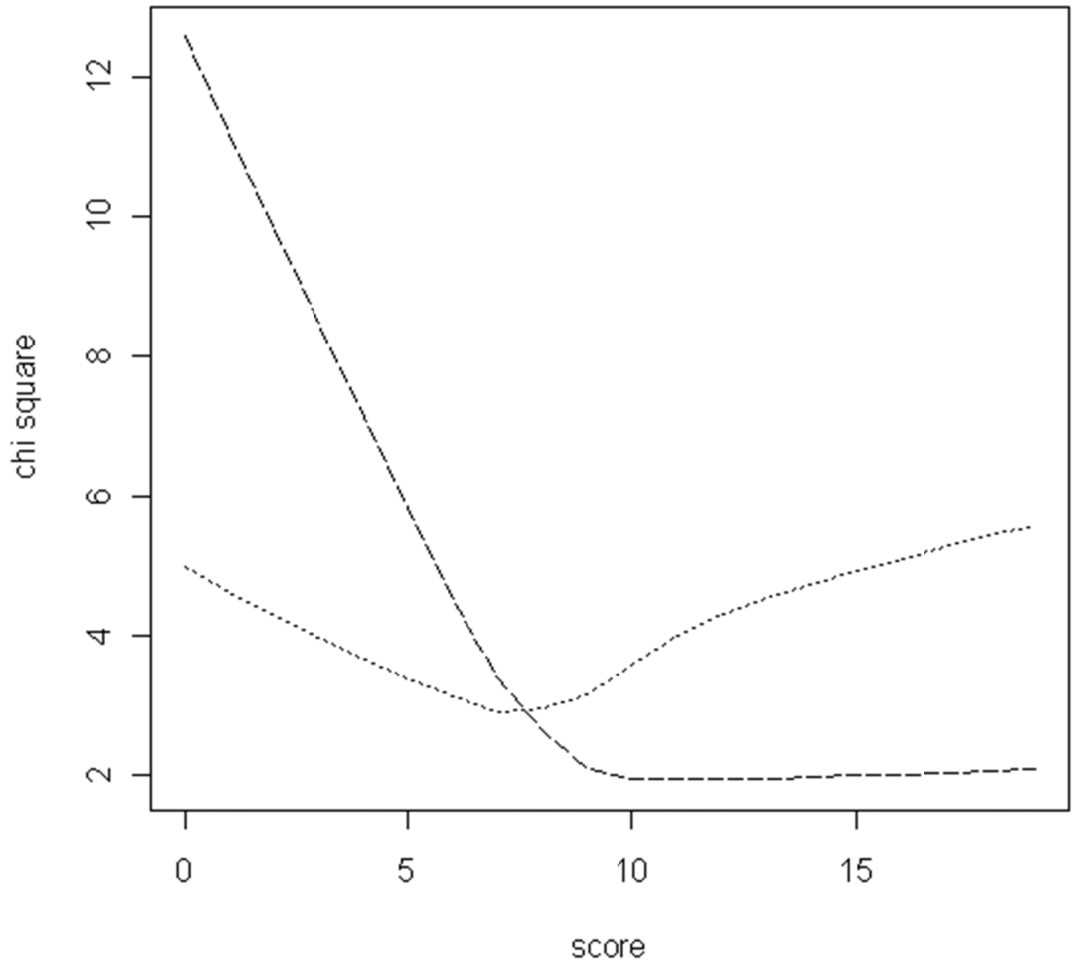

Figure 20. Smoothed line plot of marginal chi square against score.
Figure Key: dashed line = average, dotted line = equal

Probability-probability (P-P) plots were used to check the distribution of the marginal chi squares. According to the maximum deviation in the detrended chi square values found in Table 30, three degrees of freedom had the best overall fit to the data, but was only slightly better than four degrees of freedom. Interestingly, the pattern of fit to the expected value source was reversed with three and four degrees of freedom. With

three degrees of freedom the fit of the chi square based on the average response was 0.13 and 0.28 for the chi square based on the equal response. With four degrees of freedom, the opposite was true – the fit for the chi square based on the average response was 0.29 and 0.13 for the chi square based on the equal response. The chi squares will be tested for significance based on three degrees of freedom, an $\alpha = .05$ and a critical value of 7.82 due to the theoretical correctness. Box plots will also be used to identify outliers.

Table 30

Absolute Values of Maximum Deviations in Detrended Chi Square P-P Plot for Marginal Frequencies

	Degrees of Freedom		
Expected Value	2	3	4
Equal	0.41	0.28	0.13
Average	0.33	0.13	0.29

The chi squares of 95 students were significantly different than expected based on the average response. Eighty-one students had significant chi squares based on the equal distribution of response options. Thirty-five of these students had both significant chi squares. With a cut point of 10, the stem and leaf plot identified 43 outliers. The corresponding box plot of the chi squares based on the equal distribution of responses resulted in 42 outliers based on a cut score of 10.6. The stem and leaf plot of the average response identified 126 outliers. The cut point for an outlier according to the stem and leaf plot, 6.9, was much lower in this case than in all other cases. With a higher cut point of 7.99, the box plot in this condition identified 95 outliers. The revised box plot resulted

in a slight increase of outliers identified in the chi squared based on the equal distribution of responses – 43 at a cut point of 9.7. However, with only a small change in cut point to 7.4, the revised box plot of the chi square based on the average response identified 166 outliers. Table 31 presents a summary of these results.

Table 31

Summary of Outliers for all Chi Squares by Method of Identification

	Method for identifying outliers			
Expected Value	Critical Value	Stem & Leaf	Box Plot	Adjusted Box Plot
Equal	81	43 (≥10)	42 (≥10.6)	43 (≥9.7)
Average	95	126 (≥6.9)	95 (≥7.99)	116 (≥7.4)

The individual response patterns of the 35 students with all significant chi squares reveal the tendency to rely heavily on only one or two response option choices. Many of the students identified repeated selected answer option D. These students would typically be considered low ability due to the lack of knowledge inherent in responding "Don't Know", but the new method argues that these students may be unmotivated and selecting this response so that they do not have to read the questions or think carefully about the answers. The item level response patterns for these students can be found in Appendix C.

The students' responses were also checked for abnormality by the person fit statistic, l_z. l_z was significantly positively correlated with test score and the marginal chi

square based on an equal distribution of expected values. However as seen in Table 32, l_z was not correlated with the marginal chi square based on the average response.

Table 32

Spearman's Rho Correlations between l_z, Marginal Chi Squares and Total Score

	Equal	Average	Score
l_z	.178**	-.019	.354**

Note. ** p < .01.

Values of l_z had a mean of 0.32, a median of 0.49 and a standard deviation of 0.75 and ranged between -2.91 and 1.77. Eighteen students were found to have statistically significant l_z values. Two of these 18 students had significant chi squares based on the equal distribution of responses and four had a significant chi square based on the average response. No other student was found to have all three significant indicators. The majority of students, 1,264, had no indication of abnormality regardless of which method was used. Table 33 compares the significance of chi square and l_z.

Table 33

Marginal Chi Squares Compared to Significant l_z

		l_z	
Expected Value		Not Significant	Significant
Equal	Not Significant	1,322	16
	Significant	79	2
Average	Not Significant	1,310	14
	Significant	91	4

Cohen's Kappa was used to quantify the degree of agreement between l_z and the chi squares. The Kappa value with l_z was 0.02 and 0.05 for the marginal chi square based on an equal distribution and the average response, respectively. Both values indicate only a slight relationship based on Landis and Koch's (1977) guidelines.

The reliability of the test based on the entire population of 1,419 students was low, $\alpha = .680$. The 95% confidence interval had a lower limit of .665 and an upper limit of .713. Removing students with significant chi squares resulted in a further decrease in reliability. Excluding students with significant marginal chi squares based on the average response resulted in the lowest reliability, $\alpha = .552$. The reliability estimate after removing students with significant marginal chi squares based on an equal distribution of responses was .632, but removing students with either significant chi square resulted in $\alpha = .560$. All of these estimates are lower than the confidence interval suggesting that removing these students did lower the estimate of reliability in a meaningful way. Removing students with a significant person-fit statistic slightly increased the test reliability to an α of .683, but this is not beyond the confidence interval boundaries.

Due to the fact that students in the data set can be conceptualized as having two true scores – one while motivated and one while unmotivated – the assumptions of Cronbach's alpha may be violated. Guttman's lambda estimate was included as another source of reliability. Guttman's lambda 2 was slightly larger than the corresponding alpha reliability estimate in all conditions, but the same pattern of decreasing reliability occurred. Lambda 2 was .69 for the entire population. Lambda 2 was slightly increased after removing students with significant l_z values, but at .692 the increase was not

significant. Guttman's reliability estimate was decreased by removing students with significant chi squares.

DISCUSSION

Anecdotally we know a lack of motivation in children and adolescents is prevalent in our society. It crosses all ability levels and socio-economic strata and it seems to be increasing as more and more stories of students that don't apply themselves and youth that detach themselves from responsibility and consequences circulate. We know that we have a heavy reliance on testing, assessment and surveys to provide insights and inform our decisions. We know that educational policy is often based on assessments such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS). We set benchmarks and comparisons are made between schools, states and even internationally. What we do not know are the consequences of such a growing phenomenon. The Unites States typically fall at or below the top 10 in math and science scores, but with our awareness of a cultural lack of motivation can those scores be trusted? It is imperative that we ensure the accuracy of data that serves such an important function. This investigation provided a step in the direction of identifying unmotivated test takers and determining the effect on test results. The power of the new method was low and reliability estimates were unchanged after removing unmotivated test takers, but the problem continues to exist. It continues to demand further research. The following discussion generalizes the findings, successes and issues from all three data sets and provides suggestions for future study.

Little difference was found in the number of outliers identified in the stem and leaf and box plots of the maximum marginal and transitional proportions. In all but the Psych 101 data set, all three methods resulted in the same number of outliers despite the

slight differences in cut point. Surprisingly few outliers were identified in the BTES data set, but due to the large extent of missing data the total proportions for item response options decreased.

As expected, the adjusted box plots resulted in the classification of more outliers than the traditional calculation because of the decrease in cut point. The cut point in the marginal chi squares was reduced by approximately one for all expected value sources. The effect on the cut point in transitional chi squares ranged from two to four with the highest effect found in the chi squares with expected values based on the answer key.

Using the standard critical value in chi square significance testing resulted in more outliers than the stem and leaf plot and both box plots in all cases except for the marginal chi square based on the answer key in the BTES data set and in the marginal chi square based on the average response in the Political Science data set. Which set of cut points to use might be better based on the distributions found in each data set. The distributions of these data sets do not fit nicely into any category. In general, the chi square distribution did not fit the data well. A simulation study might be used to generate the null distribution. There were several specific factors that contributed to the lack of fit including student performance and missing data.

Overall student performance appears to influence the fit of the chi squares. In cases where students perform poorly, the positive skew effects the assumption of an underlying normal distribution of test scores. The chi square with expected values based on the answer key had a particularly poor fit for both the Psych 101 exam and the BTES. In both testing situations, the results identified many more students as outliers than chi

squares based on the average response and on an equal distribution. A possible effect size is also evident. Students taking the BTES had by far the lowest performance of all three data sets. There was a 280% increase in the number of significant marginal chi squares based on the answer key for this group whereas there was an 85% increase for the Psych 101 test takers. The effect size trend also exists for differences between the transitional chi squares based on an equal distribution and the answer key. A 178% increase in the number of significant chi squares from an expected value based on an equal distribution to the answer key on the Psych 101 exam and a 262% increase in the BTES. The change between the transitional chi square based on the average response and the answer key, however, was much smaller in the BTES than in the Psych 101 test. The results of the Political Science survey could not be compared due to the inappropriateness of a chi square based on the answer key. Having the best overall test scores, this result could have provided more insight into the relationship between student performance and chi square outcomes. A future investigation might include tests with high total scores.

Missing data also affected the expected value for chi squares. Those based on the average response were most profoundly influenced. In the case of marginal chi squares, the expected values did not sum to the total items and in the transitional condition, the expected values did not sum to the number of possible transitions – the total number of items minus one. With only a small percentage of missing responses, the affect in the Psych 101 exam was minimal. The expected values based on the average response for the marginal chi square totaled to 99.9. The expected values in the transitional chi square

analysis were 98. However, in the BTES, where missing responses were a serious problem, the total expected value of the average response was only 43 in the marginal chi square and only 40.4 in the transitional chi condition. On a 60 item exam we would expect the marginal expected values to total to 60 and the transitional expected values to total to 59.

The percentage of missing data potentially provides a differential effect on fit. In the BTES test, the chi square based on the average response had the worst fit of all expected value sources regardless of degrees of freedom. The fit gradually improved with an increase in degrees of freedom, but even when applying unexpected large degrees of freedom the maximum value of deviation according to the detrended P-P plot resulted in an unacceptable fit. In the Psych 101 test, however, the chi square based on the average student response had a mediocre fit compared to the other sources of expected value. Depending on the degrees of freedom employed, the fit was sometimes better and sometimes worse than that of an equal distribution of responses and the answer key. This pattern is evident in the Political Science survey where no missing data existed. The chi square based on the average expected value had a better fit than that of the equal distribution for reasonable degrees of freedom. The missing data in the Psych 101 test potentially had no effect on the results. Future research might deliberately vary the extent of missing data to determine at what point the effect is significant.

In the P-P plot analysis of all three data sets, the degrees of freedom providing the best overall chi square fit were counter to theory. However, the directional trend was not consistent. Analysis of the Psych 101 exam found smaller than expected degrees of

freedom fit the data best while the BTES called for much larger than expected degrees of freedom. For the Political Science survey, the theoretically correct degrees of freedom were ultimately used in the analysis, but the fit was only slightly superior to using more than expected degrees of freedom. The missing responses as well as the underlying distribution could be at the root of this discrepancy. Additional research into those questions would shed light here as well. Due to these variations and complications, it is recommended that the selection of the expected value source rest on empirical reasoning.

The source of expected values in the chi square analyses also had differential effects on the identification of unmotivated test takers. The marginal chi squares based on the average response and on an equal distribution of responses resulted in an identification of no more than 10% of the population in all data sets. More students had significant chi squares in the marginal condition with an expected value based on the answer key. This could stem from low overall test performance. Students with poor test scores would have many responses that deviate from the answer key. These same students, especially if the entire population did poorly, could be responding similarly to their peers resulting in lower chi squares based on the average response. Lowered chi squares could also result for these students where an equal distribution of responses was the expected value if they are randomly responding to items that they did not know. Chi squares based on an equal distribution of responses would increase in poor performing test takers only if they used a strategy such as always selecting answer option C when guessing.

Compared to the marginal chi squares, the transitional chi squares were more sensitive to discrepancy from expected values and resulted in more significant outliers. This is true for all test conditions and for all sources of expected value, though the chi square with expected value based on the answer key in the transitional condition appears especially sensitive resulting in the most outliers in all cases. The chi square with expected value based on the answer key resulted in the highest proportion of identified outliers in both data sets where this chi square was appropriate. Thirty percent of the total population of Psych 101 test takers and 28% of the BTES population were found to have significant transitional chi squares based on the answer key. This could be a confounding effect of both the increase in sensitivity of the answer key as an expected value and of the transitional frequency condition. However, this cannot solely be attributed to poor fit because fit varied across data sets, condition and chi square source – the answer key did not always result in the worst fit. A data set where the key sensitivity was not evident could be used in the future to further investigate this potential effect.

This increased sensitivity might be influenced by the small cell size for each transition when the test is relatively short. A single deviation in students' response from the expected would result in a change to two transition cells. Though this is also true for the marginal chi square, the expected values are much smaller in the transitional case due to number of cells – 16 as opposed to four – giving small deviations greater impact. This could be investigated through analysis of a very long test where individual transitional cell values are large.

A comparison of the results of the person fit statistic l_z and the chi squares is difficult due to a lack of consistency across the three datasets. In general, l_z was more likely to be positively correlated with test score and then negatively correlated with the chi squares. In both cases where a correlation was present, the Psych 101 exam and the Political Science survey, l_z was positively correlated with test score. For the Psych 101 exam, l_z was also negatively correlated with all six chi squares. However, in the Political Science survey, l_z was negatively correlated with the marginal chi square based on the average response, but positively correlated with the marginal chi square based on an equal distribution of expected values. No chi squares in this dataset were correlated with total score. Though l_z was not correlated with total score in the BTES, it was negatively correlated with the marginal and transitional chi square based on the average response. There is also the possibility that the l_z values were not particularly good estimates due to the problem of including bad data in the item parameter estimates. A cleaned data set might be used for calibration or more robust estimates may be incorporated to prevent this problem.

Though the disadvantages of l_z model based method is obvious, there are also disadvantages to the new model-free method. There is an inherent low power in model-free estimators. A future investigation might look into defining strata based on score groups and then consider the fit of the chi square within each strata.

As expected, there was not much overlap between the students identified by l_z and those identified by the chi squares. Due to the nature of the statistics, l_z is attempting to determine which response patterns differ from the IRT model and the chi squares are

simply looking for patterns in responding. At most only 50% of the aberrant students were also flagged by l_z. The greatest overlap occurred in the BTES where 50% of the students with all six chi squares and 40% of the student with the marginal or transitional key chi square also having a significant l_z result. The marginal chi square based on an equal distribution of responses in the Political Science survey had the largest overlap with l_z; 28% of those students were found to be significant in both cases. In the Psych 101 dataset, the greatest overlap - 20% of students - was found between the chi squares based on the average response and l_z.

The original test reliability was high for both the Psych 101 and BTES data sets. The low reliability in the Political Science data set is largely a product of the short test length. The influence of the new method on reliability estimates seems dependent on the data set. There was no great change in Cronbach's alpha or Guttman's lambda regardless of the method of exclusion, but in the BTES data set removing students based on significant chi squares resulted in increased reliability while removing these students in the Psych 101 and Political Science data sets decreased the reliability estimates. The person-fit statistic, l_z, led to an increase in reliability for the BTES and Political Science data sets. None of these changes fell outside the range of the 95% confidence interval. However, despite the unremarkable change in reliability, the reliabilities resulting from removing students with significantly patterned response patterns is a more accurate and valid reliability based on the fact that questionable data has been removed.

There are several advantages of this new method. It is easy to use and can be applied to basic test data without consideration of IRT. Also, general results are easily

interpreted; significant chi squares indicate deviance from expected responses. However, there are obvious shortcomings. Some of these plague the entire field of person-fit statistics, and render identifying unmotivated students difficult. Test length influences the use and results of transitional frequencies. A short test does not provide enough information for a transitional chi square to be reliable and in cases of particularly short tests, transitional chi squares are not even possible. However, stem and leaf and box plots on the marginal and transitional proportions alone provide a simple method for identifying potential outliers.

Though the marginal and transitional chi square results provide information above and beyond test score, student performance affects the data distribution which then affects the fit of the chi square model. Additionally, the degree to which students did not respond to all test items impacts the resulting expected values. Though the expected values can be adjusted to account for a restricted number of total items, this patch does not appear to adequately address the problem. Further research may consider additional solutions for this potential data form.

Non-traditional grading schemes that use only a selection of the possible response options also introduce irresolvable complications. Failure to use all four answer options precludes an analysis of a chi square based on expected values drawn from the answer key. Additionally, these methods would require modification if applied to non-academic tests such as opinion questionnaires and surveys. If reversed coding of similar items is incorporated into the questionnaire construction so that a variety of answer options is expected the existing methods can be used to capture unexpected response patterns.

Without reverse coding where there are reasons to expect consistency in responding such as in a section of similar items, the current methods could be modified to identify respondents with unexpected results. These potential avenues will be left for future investigation.

Without further investigation into the individuals behind the responses, the meaningfulness of the results is hard to determine. As they have been conceptualized in this study, the new method succeeded in identifying unmotivated test takers. However, whether patterned responses of this nature truly result from a lack of motivation and not another phenomenon remains unclear. This new method can be used as a data checking tool for identifying students warranting further attention. The new method is one step in the direction of ensuring the accuracy of our data imperative in education and psychology.

APPENDIX A

Item Responses of Students with all Significant Chi Squares on the Psych 101 Exam.

Item	Student Responses																
V1	1	1	4	1	1	1	1	4	4	4	1	1	2	2	1	2	1
V2	4	4	4	4	4	4	4	4	1	1	2	4	4	4	4	4	4
V3	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2
V4	4	2	3	2	4	4	2	3	4	1	4	4	4	1	4	4	2
V5	2	2	4	4	2	2	4	4	4	4	3	2	4	4	2	1	4
V6	2	1	2	2	2	2	3	2	4	2	1	2	1	1	2	2	3
V7	1	2	2	4	1	4	2	4	4	1	3	1	3	4	4	2	3
V8	2	3	2	2	2	2	2	2	1	2	2	2	4	1	3	3	1
V9	4	4	1	4	4	4	2	4	1	4	4	4	4	1	4	4	2
V10	3	4	2	1	3	3	3	3	3	4	1	3	1	4	3	2	2
V11	3	3	3	2	3	3	3	3	3	2	3	3	3	3	3	3	3
V12	4	4	2	4	4	4	2	4	4	2	4	4	4	4	4	2	2
V13	1	4	4	4	1	1	4	1	4	4	1	1	4	1	2	2	4
V14	1	2	1	1	1	1	3	3	4	2	4	1	1	1	3	3	3
V15	2	3	2	2	2	2	2	3	2	1	3	2	3	2	3	2	3
V16	3	4	4	3	3	3	3	3	4	3	3	3	3	3	4	3	4
V17	1	2	2	1	2	1	1	1	4	2	2	1	2	1	1	2	2
V18	3	3	4	2	4	1	3	4	1	3	4	3	3	4	4	2	4
V19	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4
V20	2	2	2	2	2	3	2	3	2	1	1	2	3	2	1	3	3
V21	2	3	2	4	3	2	3	3	1	3	2	2	1	1	2	2	2
V22	2	2	2	2	2	2	2	2	2	4	1	2	2	2	2	2	2
V23	1	3	4	1	1	1	4	1	1	1	1	1	4	4	4	1	1
V24	2	2	2	1	2	2	2	2	4	2	2	2	2	2	3	1	4
V25	1	3	1	3	1	1	2	3	4	3	3	1	3	3	3	4	3
V26	4	4	4	4	1	1	1	4	1	4	4	4	1	1	1	2	1
V27	1	4	1	2	1	1	1	1	1	1	1	1	1	3	1	1	1
V28	3	3	3	1	3	1	1	3	3	3	2	1	2	3	3	2	2
V29	1	2	4	1	1	2	3	2	4	2	1	1	4	3	4	3	2
V30	3	4	4	3	4	3	4	3	1	4	1	3	4	4	4	2	2
V31	1	3	2	2	1	1	1	3	1	3	1	1	4	1	1	2	4
V32	1	4	4	1	4	1	4	1	3	1	1	1	2	1	1	2	2
V33	4	2	3	4	2	4	3	3	2	4	1	4	4	4	3	2	3
V34	2	2	2	2	2	2	2	4	3	2	2	2	3	4	2	3	3
V35	2	2	4	1	1	2	4	2	2	3	1	1	4	2	2	2	2
V36	4	3	3	2	2	3	4	3	4	1	4	3	1	4	3	4	1
V37	3	4	3	4	3	3	4	3	3	1	3	3	3	3	3	2	3

V38	1	1	4	1	1	1	4	4	3	4	1	1	4	3	4	1	4
V39	4	3	3	1	3	4	3	3	1	3	3	3	3	4	3	1	2
V40	3	4	1	4	4	1	4	3	4	4	4	1	3	3	3	2	4
V41	1	2	1	2	1	1	1	4	2	2	2	4	3	2	2	1	2
V42	2	2	2	2	2	2	3	2	2	3	2	2	2	2	2	2	2
V43	3	2	3	1	3	3	2	3	3	1	3	3	2	1	3	4	1
V44	2	2	4	2	2	1	4	1	4	2	1	1	3	2	4	1	4
V45	1	4	1	1	1	1	2	1	1	2	1	1	1	1	2	1	4
V46	3	3	3	2	3	3	3	3	3	3	3	1	3	2	3	3	3
V47	2	2	1	4	1	1	4	2	1	1	2	1	4	4	1	2	4
V48	4	3	3	3	4	4	4	4	4	4	1	3	3	4	4	3	4
V49	4	4	2	4	4	4	4	1	4	4	4	1	3	4	4	4	2
V50	3	4	3	1	1	1	4	3	1	1	1	3	3	2	3	3	2
V51	3	3	3	3	3	3	2	3	3	2	2	3	3	3	4	2	2
V52	2	2	4	2	2	2	2	3	3	3	2	2	2	2	1	3	2
V53	1	3	3	1	1	2	4	1	3	4	1	2	2	3	2	3	4
V54	1	4	3	4	4	1	4	4	4	4	4	1	4	4	4	4	4
V55	3	1	3	3	1	3	3	3	3	3	3	3	2	3	3	3	1
V56	3	2	4	3	4	3	3	2	4	4	4	3	3	4	2	2	2
V57	4	4	2	4	4	4	4	3	4	4	4	4	2	1	3	4	4
V58	1	2	3	1	3	1	3	4	4	1	3	1	2	3	3	1	3
V59	2	2	3	2	2	2	3	4	4	3	2	2	1	3	1	2	3
V60	4	4	4	4	4	4	4	4	4	4	1	4	4	4	4	4	4
V61	3	3	4	3	2	3	4	3	1	4	3	3	1	4	4	2	3
V62	2	4	4	2	4	1	2	1	1	4	1	2	3	3	1	3	3
V63	4	4	4	4	4	1	4	4	1	2	4	1	4	4	4	4	4
V64	3	2	3	4	3	1	1	3	1	4	1	1	4	4	4	4	4
V65	2	2	3	1	2	1	2	2	2	4	2	1	3	1	2	2	2
V66	1	2	1	1	1	1	1	2	3	1	1	1	2	4	3	1	1
V67	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
V68	3	3	1	3	4	3	2	3	2	2	1	3	2	3	3	2	2
V69	4	4	2	4	1	3	4	4	3	4	1	3	4	4	3	4	4
V70	4	4	4	4	4	3	4	4	4	3	4	3	1	4	3	4	2
V71	1	2	3	4	4	1	1	4	3	4	1	2	3	1	2	1	3
V72	2	4	3	4	4	2	4	3	1	2	4	2	3	2	1	2	4
V73	2	1	1	1	4	2	3	2	4	3	1	3	3	4	1	4	3
V74	2	2	1	1	2	2	2	2	4	2	2	2	4	3	2	2	3
V75	3	4	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
V76	2	2	3	2	1	1	2	2	4	2	2	1	2	3	3	2	3
V77	3	1	4	1	1	3	1	1	1	4	4	3	2	2	4	1	4
V78	2	1	1	2	1	2	1	2	2	2	2	2	2	2	1	2	2
V79	1	2	3	3	1	3	4	3	1	1	1	1	1	1	2	3	3

V80	1	4	1	4	4	1	2	1	4	4	1	2		4	4	1	4
V81	2	3	3	3	4	3	4	4	3	4	3	2	3	3	3	3	2
V82	3	4	3	4	2	1	4	3	2	2	4	3	2	4	4	2	3
V83	4	4	3	4	4	2	4	4	4	4	4	4	4	4	4	4	4
V84	2	3	3	1	1	3	1	2	1	4	1	2	1	4	3	1	1
V85	4	4	4	4	4	4	3	2	4	4	1	1	4	1	4	4	2
V86	4	4	4	4	4	4	4	3	2	3	4	4	2	1	4	2	4
V87	3	3	3	4	3	1	4	3	4	4	3	3	4	4	3	2	3
V88	1	2	3	1	1	1	3	2	1	4	1	1	2	4	2	3	4
V89	4	4	3	4	1	4	4	4	4	3	4	1	4	4	4	4	4
V90	1	4	3	1	1	1	4	3	4	4	3	1	4	3	3	2	3
V91	4	4	4	2	4	4	4	2	3	4	1	4	3	4	3	3	4
V92	2	2	2	2	2	2	2	2	4	2	2	2	2	1	3	2	2
V93	3	1	4	1	2	1	3	2	1	1	1	1	3	3	1	1	3
V94	1	4	1	1	1	1	1	1	1	4	1	1	2	3	2	1	3
V95	2	4	2	4	4	1	2	4	2	2	4	2	2	2	1	4	4
V96	4	4	2	4	4	4	4	4	4	4	1	4	4	4	4	4	4
V97	1	2	2	2	1	1	2	2	2	2	1	2	2	1	2	2	2
V98	4	3	4	4	1	4	4	4	4	1	4	4	3	4	2	1	4
V99	3	2	3	1	4	3	3	3	4	3	3	2	3	3	4	1	2
V100	2	4	2	4	4	4	4	4	3	4	4	2	2	2	4	2	4
SCORE	100	44	42	56	63	72	49	58	40	40	56	74	40	51	42	48	36

APPENDIX B

Item Responses of Students with all Significant Chi Squares on the BTES.

Item	Student Responses			
V1	3	3	3	3
V2	1	4	3	1
V3	4	4	3	1
V4	3	1	3	1
V5	2	4	3	1
V6	3	2	3	2
V7	1	2	3	1
V8	2	4	3	2
V9	3	2	3	2
V10	1	2	3	1
V11	1	3	3	1
V12	3	3	1	3
V13	3	2	1	9
V14	4	2	1	3
V15	2	2	1	2
V16	1	1	1	1
V17	3	2	1	3
V18	4	2	1	2
V19	4	1	1	1
V20	3	1	1	4
V21	4	4	1	1
V22	3	3	1	2
V23	1	3	1	1
V24	2	1	1	1
V25	3	1	1	9
V26	1	3	1	9
V27	1	3	1	1
V28	4	3	1	1
V29	3	2	1	9
V30	2	3	1	1
V31	3	3	9	2
V32	3	3	9	9
V33	3	3	9	1
V34	1	3	9	9
V35	1	3	9	9
V36	2	3	9	9
V37	2	1	9	9

V38	1	3	9	1
V39	4	2	9	9
V40	1	9	9	9
V41	2	2	9	3
V42	4	2	9	2
V43	1	4	9	4
V44	3	1	9	1
V45	1	3	9	3
V46	3	1	9	2
V47	3	2	9	1
V48	4	3	9	1
V49	2	2	9	2
V50	2	3	9	2
V51	3	3	9	9
V52	3	2	9	3
V53	1	2	9	9
V54	2	3	9	9
V55	4	1	9	1
V56	4	3	9	9
V57	1	2	9	9
V58	2	3	9	9
V59	3	2	9	2
V60	1	2	9	1
SCORE	60	13	8	17

APPENDIX C

Item Responses of Students with all Significant Chi Squares on the Political Science Survey.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	SCORE
1	1	1	1	3	1	1	1	2	3	4
4	4	4	1	1	1	1	2	1	1	1
4	4	4	4	1	4	2	4	3	4	2
2	1	4	1	2	1	1	1	1	1	4
4	4	4	4	1	1	4	4	4	4	0
4	4	4	4	3	2	4	4	3	4	3
4	4	4	4	3	2	4	4	3	4	3
2	4	4	4	4	1	4	4	4	4	1
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	4	1	4	4	4	4	0
4	4	4	4	4	1	4	4	1	4	0
4	4	4	4	3	1	4	4	4	4	1
2	4	4	4	2	2	4	4	1	4	2
4	4	4	4	2	2	2	2	4	4	3
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	3	1	4	4	4	4	1
4	4	4	4	4	1	4	4	3	4	1
4	4	4	4	3	2	4	4	3	3	3
4	4	4	4	3	1	4	4	4	4	1
2	4	4	4	2	2	4	4	1	4	2
4	4	4	4	2	2	2	2	4	4	3
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	4	4	4	4	4	4	0
4	4	4	4	3	1	4	4	4	4	1
4	4	4	4	4	1	4	4	3	4	1
4	4	4	4	3	2	4	4	3	3	3
4	4	4	4	1	4	4	4	2	4	0
4	4	4	4	1	1	4	4	4	4	0
2	4	4	1	4	1	4	4	4	4	2
2	4	4	1	1	4	4	4	4	4	2
1	4	4	4	3	1	4	4	4	4	1
4	4	4	4	3	2	2	4	4	4	3

REFERENCES

Barnette, J. J. (1996). Responses that may indicate non-attending behaviors in three self-administered educational surveys. *Research in the Schools, 3*(2), 49-59.

Barnette, J. J. (1999). Non-attending respondent effects on internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement, 59*(1), 38-46.

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

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, California: Sage Publications.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*, 191-205.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

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

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

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

Meijer, R. R., & Sijtsma, K (1999). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*(3),261-272.

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

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16,* 345-352.

Rizopoulos, D. (2006). ltm: an R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25.

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

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the Individual Consistency Index. *Journal of Educational Measurement, 20,* 221-230.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

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