DOCUMENT RESUME

ED 324 363 TM 015 629

AUTHOR	Baghi, Heibatollah
TITLE	The Use of Rasch Model Fit Statistics in Selecting Items for the Maryland Functional Testing Program.
PUB DATE	Mar 90
NOTE	22p.
PUB TYPE	Information Analyses (070) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; Computer Software; Equated Scores; *Error of Measurement; *Goodness of Fit; *Item Response Theory; *State Programs; Test Construction; *Testing Programs
IDENTIFIERS	BICAL Computer Program; *Maryland Functional Testing Program; *Rasch Model

ABSTRACT

The Maryland Functional Testing Program (MFTP) uses the Rasch model as the statistical framework for the analysis of test items and scores. This paper is designed to assist the reader in developing an understanding of the fit statistics in the Rasch model. Background materials on application of the Rasch model in statistical analysis of the MFTP are provided, studies of item fit and causes of item misfit are outlined, and recommendations for uses of fit statistics are made. It is recommended that: (1) since the number of items detected as "misfitting" is a matter of choice in selecting the fit statistic from a wide range of fit statistics produced by BICAL, and since there are no sound statistical procedures for selecting items on the basis of fit indices; test of item fit should not be used with some arbitrary critical value to make automatic decision for eliminating them; (2) items that have total or between "t" values larger than 2.00 should be investigated for flaws in the construction of the item, unusual item content, "order effects" resulting from its position in the test, and level of difficulty; (3) if an easy item does not fit the model because it occurs near the end of the test which low ability students may not reach time for the test might be extended; (4) guessing is inherent in the use of multiple-choice items and such items should not be deleted for this reason; and (5) items flagged by various tests of fit should be studied by subject matter experts to generate possible hypotheses about the reason for misfitting items. Fit statistics produced by BICAL-1980 are described, with emphasis on the sample size effect on fit indices, and computation of these statistics are explained. One data table is included. (TJH)

THE USE OF RASCH MODEL FIT STATISTICS IN SELECTING ITEMS FOR THE MARYLAND FUNCTIONAL TESTING PROGRAM

Heibatollah Baghi, Ph.D.

March, 1990

Program Assessment, Evaluation, and Instructional
Support System
Maryland State Epartment of Education
200 West Baltimore Street
Baltimore, MAryland 21201

The purpose of this paper is to assist the reader in developing an understanding of the fit statistics in the Rasch model. First, background materials on application of the Rasch model in statistical analysis of Maryland Functional Testing Programs will be provided. Second, review of studies of item fit and causes of item misfit will be listed and recommendations for uses of fit statistics will be made. The last section of this paper concentrates on describing fit statistics produced by BICAL-1980. Computation of these statistics is also explained.

Background

The Maryland Functional Testing Program uses the Rasch model as the statistical framework for the analysis of test items and scores. This analysis produces data which are used in (1) item calibration, (2) item selection, (3) equating of scores, and (4) generating scaled scores. The test items are field tested, analyzed, and included in item banks with the associated statistical indices. A major question and the reason for this review is: What are the criteria that should cause an item to be rejected for misfitting the Rasch model? A prior question is: What are the effects of including misfitting items in a test? Is there a sound procedure available for selecting items for the MFTs on the basis of fit statistics.

An advantage of Rasch model is that its simplicity makes it easy to apply the model in solving many measurement problems (Gustafsson, 1980). Items can be described with item parameters and persons can be described with person parameters, both on the same scale. This allows the probabilistic prediction of the response of any person to any item, which can be compared with the actual response to give us an idea about how appropriate the Rasch model is for the data set we are using. Divgi (1981) lists three properties of the Rasch model as follows:

1. The number correct is a sufficient statistic for the estimation of ability. The pattern of right-wrong responses does not provide additional information about ability once the total score is known (but it can tell us something about the items).

2. Conditional maximum likelihood estimates of item parameters equal the true values when samples are infinitely large. Conditional maximum likelihood estimates cannot be obtained with other models.

3. It is possible to compare difficulties of two items or abilities of two persons without having to estimate any other parameters. Rasch (1966) called this phenomenon "specific objectivity".

Given these properties, how do we decide if the Rasch model is appropriate for our test? One of the first issues to consider is the purpose for which the test is being built. Wright (1978) deals primarily with a test to discriminate among people, whereas

Sabers and Jones (1982) are trying to develop a test to measure the extent to which the curriculum has been mastered. They quote Buros (1978) as arguing that selecting items for the purpose of discriminating among people instead of to to measure learning results in a test of questionable validity. This argues against a strict policy of rejecting items for not discriminating among people, as it may result in spurious fit only (Gustafsson, 1980). Gustafsson suggests that for a scale to measure individual differences, highly discriminating items would be most useful, while these might not be best for a scale to assess curriculum. Further Sabers and Jones (1982) say that the Rasch model is a good model not because most items will fit to the model, but because it represents the kind of items most desirable on a measuring instrument.

But are we selecting items to fit a model or a model to fit our items (Sabers and Jones)? The basic principles of a good test are the same regardless of what mathematical model is selected for the analysis. Clear test specifications and tight item specifications will most often provide a reasonably good test by any standards (Popham, 1978).

Item selection occurs at least twice in the history of a test built from an item bank, once in the selection of items for the bank, and again in selection of items for a specific test. If a good job is done at the first stage, the second will be easier. When a large number of items have been generated and judged to be acceptable by non-statistical procedures, the question becomes how to decide which of the item fit the Rasch model.

Many tests are designed to learn about people at a particular point on the ability continuum. If the test will be used to determine mastery, it is at the cutting point that the most precise measurement is desired (Cook and Hambleton, 1979; Lord, 1977, 1980; Wright, 1977). But the design of tests to maximize the amount of information yielded at a point of interest is not a subject for this paper.

Review of Studies of Item Fit

A major line of research has dealt with the fit of various types of test items to the Rasch model, as indicated by those item fit indices included in the BICAL programs or by other ad hoc indices of fit. But first, because most of the studies reviewed here have used fit indices produced by the BICAL program, the reader is encouraged to read the last section of this document in which different fit statistics are defined and the computation of the statistics is explained.

Wright and Panchapakesan (1969) developed a test of fit of items to the model. The test is based on comparisons between the observed and the expected frequencies of correct answers to each item at different levels of ability. This comparison is checking to see if different groups fit statistic indicates that too many high ability people, as identified by total score, missed the item, or that too many low ability people answered that item correctly. The expected value (mean) is one and its standard error is suggested as a rule of thumb for "too large."

A chi-square test is often used to test the fit statistics, but the actual distribution of test statistics are unknown. The chi square, z or t distributions have been relied on, but monte-carlo studies have shown that, while the means of the test statistic distributions may conform to the expected ones, the variance may differ substantially (Gustafsson, 1980).

Using the standardized residual in selecting items for tests has been questioned by George (1979). (The standardized residual has also been called the "mean squared error" and the "fit mean square".) George suggests that using standardized residual is not appropriate for use in selecting items and gives several reasons for the suggestion. He points out that the test is very conservative and rarely rejects items because the sample size is always one that causes the standard deviation used in the denominator of the standardized residual formula to be too high. Further he states that the use of the normal approximation to the binomial is a serious problem. Examples are presented to show the nature and magnitude of error introduced when standardized residuals are used under certain conditions. Some of the conclusions from his data are: (1) Easy items usually appear to fit better than they should, (2) Difficult items tend to misfit because low ability students may guess the answers correctly. When the latter happens a very large value is added to the sum of squared residuals resulting in a large mean square residual, (3) Items having steeper item characteristic curves fit the Rasch model best, even better than those having larger discrimination indices. This would account for the high percentage of fitting items found in tests that have been developed using classical criteria of item quality.

Divgi (1981) has also criticized using the mean squared error statistic as an index of fit to the Rasch model. Being concerned about the correlation between the mean square fit and item difficulty he proposed a new fit statistic based on "...an approximate quadratic depends of the standardized residual on estimated ability." The item responses from eight levels of the Survey Reading Test from the 1978 Metropolitan Achievement Tests were tested for fit using Wright's "fit mean square" and his new fit statistic. From 48% to 82% of the items were identified as misfitting by his new fit statistic, while the fit mean square reflected only 16% of the items. Divgi suggests that many studies have found test data to fit the Rasch model only because their tests of fit lacked power to detect deviations from the model.

Wright, Mead, and Draba (1976) suggest a test of fit for the Rasch model that involves using an analysis of variance on the variation remaining in the data after removing the effect of the fitted model. This allows not only the determination of the general fit of the model, but also assists in pin-pointing guessing and item bias.

Gustafsson (1980) claims that testing individual items for goodness of fit to the model is illogical because the basic requirement of the Rasch model is that the items be homogeneous. What is tested is whether the items fit with each other, not whether they fit the model. He uses this reasoning against what he says is standard procedure for obtaining fit to the Rasch model. First, a set of items is administered to a sample of persons and an overall test of fit is computed. If this test is significant, and it usually is, the fit statistic for each item is computed and those items which do not fit are excluded. This process is repeated until no misfitting items remain. He argues that this process should rarely be used, for several reasons. One reason is that the tests most often used represent only a partial evaluation of fit to the model, and they can fail to detect even very serious deviations from the model. He also argues that this may be trading off between different violations of the model assumptions. He feels this process results only in spurious fit.

Gustafsson (1980) proposes an alternative strategy that begins with identifying a likely source of poor fit to the Rasch model. If there are possible causes such as guessing and speededness, steps should be taken to eliminate those causes. Any item heterogeneity that is severe might be resolved by grouping items into homogeneous subsets, or excluding a few items that are poorly constructed. He further suggests a cross-validation of the derived scale. It would appear that Gustafsson has reflected item statistics altogether.

Canner and Lenke (1980) reviewed the misfitting items from a large number of tests covering grades kindergarten through twelfth grade in a variety of content areas. Their criterion of misfit was the mean square fit statistics adjusted for sample size by the factor (1500/N). They found that many of the misfitting items stood out from the main set of items and appeared to be different in some distinct way. There were some consistencies in the subject matter areas. For example, 89% of misfitting spelling items were those in which the stimulus word was presented as an incorrectly, rather than correctly, spelled word. In the reading comprehension tests, the inferential items misfit more frequently than the literal comprehension items. In one of the mathematics tests they found three of four misfitting items to be dictated items requiring computation. Items dealing with the metric system also tended not to fit the Rasch model. They concluded that:

> .. items measuring knowledge of specific content may not fit the Rasch Model if the item content is not always taught (e.g., metric items) or does not follow a regular pattern of instruction at particular grades and times of year (e.g., spelling skills at second grade level, sounds of blends at first grade level)."(p.11)

Douglas (1981) examined the relationships between selected BICAL tests of item fit and item bias (and score invariance). He used data from the Michigan State University Vocabulary Placement Test given to freshmen. The test was speeded and independent measures of speededness and item bias were available. He found the BICAL "between group t" and the "Discrimination" values to be highly correlated with indices of bias and of stability of item difficulty estimates, while the "total t" and standard error of difficulty values were not. He recommended that the between groups t and the discrimination indices be five more weight as indicators of item fit. Implied in this recommendation is acceptance of the criteria, "detection of biased items." However, the recommendation might not be appropriate for tests that are not speeded or do not contain biased items. An important point made by Douglas is that LOGIST might be preferred over BICAL for speeded tests, even if only one item parameter is being estimated, because BICAL scores unreached items as incorrect while LOGIST does not.

Reckase (1981) summarizes the results of his study of fit to the one parameter model:

> ...there seems to be no good procedure for selecting items with the one-parameter logistic model. Not only do the fit statistics not work well, but no reason can be thought of for selecting items with discrimination parameters equal to the mean discrimination in the pool. Typically, use of the best items in a pool would seem desirable, as opposed to using the mediocre items as suggested by selection on the basis of one-parameter model fit (p. 42).

Attempts have been made to identify causes of item misfit to the Rasch model. Mead (1976) studied the causes of item misfit using residual analysis (for the definition and the computational formula of residual refer to the last section of the paper). Theortically when the residuals are plotted against (Ov - bi) they should fall on a horizontal line through the origin. Systematic sources of item misfit appear as departures from the horizontal line. Mead listed several sources of misfit, and the effect of the misfit on the observed item characteristic curve. He described, for example, that random guessing on an item will be reflected in lowered discrimination of the observed item characteristic curve. The observed item characteristic curve for a biased item would appear to be highly discriminating. This kind of analysis would not be practical as a routine procedures for idenfitying misfitting items. However, familiarity with the technique of residual analysis is useful for staff members involved in the assessment of test items.

Generally, anything which causes a person's response to differ from what is expected by the model is a source of misfit. Mead (1976a) demonstrated that Items near the end of a test will appear to be more difficult than they really are in speeded tests. Also, items near the end of a test appear to have too high a discrimination since many low ability examinees will not reach them. Douglas (1981) reported that "between groups t" and the "discrimination" indices (produced by BICAL program) were more sensitive than the "total t" for detecting poor item fit related to test speededness.

Guessing is reported to be a major source of item misfit for multiple-choice items. Usually when an examinee of low ability answers a difficult item correctly (probably by guessing) a large value will be added to the fit index. However, a lower value will be added when an examinee of high ability misses an easy item (probably due to carelessness) (George, 1979). Also it has been demonstrated that guessing causes easier items to have very high discrimination and difficult items to have too low discrimination (Gustafsson, 1979).

If an item is too easy or too difficult for a subgroup of the population, it may affect the item fit indices. The subgroup of the population might be based on the variable such as sex or ethnic group. Specific study is undertaken by the Program Assessment, Evaluation, and Instructional Support Branch to detect the possible forms of item bias such as sex and ethnic bias.

Lack of unidemensionality of a test might affect the item fit indices. If a given test includes a set of items assessing a different ability from what was assessed by the majority of the items is the test, then the slopes of their item characteristic curves will be affected and as as result the items may not fit the model.

Lack of local independence might affect the item fit incices. The assumption of local independence is equivalent to the unidimensionality assumption. For example in a reading comprehension test, if an item is answered correctly because it is based on a reading paragraph, then a factor other than the ability of interest is affecting the observed response. This phenomenon may increase the item's misfit index.

The sample size effect on fit indices produced by BICAL has been studied systematically. Rentz and Ridenour (1978) reported the results of their study that the "mean square fit statistics tend to inflate, with the same degree of fit, as sample sizes increases" (p. 3). They adjusted for the inflation by rescaling mean squares by a factor of 1500/N, where N is the number of subjects in the sample. Hambleton and Murry (1983) used simulated data and the "t-fit" statistic from the BICAL program (Wright and Mead, 1977) to show that "...the number of misfitting items ranged from 5 to 38 of the 50 items when sample size increased from 150 to 2400." (p. 73) These studies make it clear that the size of mean sqaure fit indices are related to the number of persons in the analysis sample. This indicates that with small sample sizes items which have a high degree of misfit may be undetected, and with a large sample sizes many good items may be identified an misfitting items.

Recommendations

On the basis of review of the literature, the following recommendations should be considered for identifying misfitting items and for using fit statistics in the selection of items for inclusion on the Maryland Functional Tests.

1. The number of items detected as "misfitting" is a matter of choice in selecting the fit statistic from a wide range of fit-statistics produced by BICAL. Test of item fit should not be used with some arbitrary critical value to make automatic decision for eliminating items. There are no sound statistical procedures for selecting items for the MFTs on the basis of fit indices.

2. Items which have total or between t values larger than 2.00 should be investigated for flaws in the construction of the item, unusual item content, "order effects" resulting from its position in the test, and difficulty. When a flawed item is found, it should be revised or thrown out.

3. It is possible that an easy item may not fit the model because it occurs near the end of the test where low ability students may not reach it at all. If this causes a problem, the time for the test might be extended. Otherwise, the misfitting phenomenon may be accepted as part of the measurement error, or as a penalty for slow work on the part of the examinee.

4. One of the common causes of misfitting item is that low ability students answer the difficult item correctly by perhaps guessing. This phenomenon is inherent in the use of multiple choice items and there is no reason to delete such items from the test.

5. Items flagged by various tests of fit should be studied carefully by subjects matter experts to generate possible hypotheses about the reason for misfitting items.

Fit Statistics in BICAL Version 1980

The purpose of this section is to describe the fit statistics produced by BICAL-1980. The formulas and steps involved in the BICAL calculation will be presented. Then the theoretical rationale and uses of the fit statistics is discussed. Commonly, the following equation is called the Rasch model:

$$Pvi=P(xvi=1/\theta v,bi) = \frac{\exp(\theta v- bi)}{1+\exp(\theta v - bi)} \qquad (1)$$

Where:

Θv=ability of person vi

bi=difficulty of item ij

This model simply states that the probability of a correct response (xvi=1), given a person's ability Θv, taking an item with difficulty values of bi is a function of the difference between person ability (Θv) and and item difficulty (bi).

The key concept in understanding item fit to the Rasch model is the residual (rvi). Residual is the difference between the probability of a correct response (Pvi) and the observed outcome.

$$\text{Residual} = rvi = (Xvi - Pvi) \qquad (2)$$

The standardized residuals can be calculated by dividing each residual by its standard deviation

$$Zvi = \frac{(Xvi - Pvi)}{(Pvi \,.\, Qvi)} = \frac{(\text{Observed} - \text{Expected})}{\text{Expected S.D. of Observation}} \qquad (3)$$

Where:

$$Qvi = 1 - Pvi$$

The standard residual is expected to be distributed normally with a mean of zero and a variance of one. The standardized squared residual (Zvi) can be determined by squaring Z scores and summing them across persons:

$$Z^{2}vi = \frac{(Xvi - Pvi)^{2}}{Pvi\ Qvi} \qquad (4)$$

The squared standard residual has an approximate chi-square distribution with one degree of freedom. Standardized squared residuals are summed over all persons to evaluate fit of an item. This is referred to as Fit Mean Square Total in BICAL Version 1977 and is calculated by:

$$\text{Fit Mean Square} = \frac{\sum Z^{2}vi}{N - 1} \qquad (5)$$

N = number of students in the score groups.

More precisely,

> This is the squared standard residual Z inflated to one degree of freedom per person and then averaged over persons. It will be large for an item when there are too many relatively high ability persons who fail on that item and too many relatively low ability

> persons who succeed. What is 'too large' a fit mean square depends on the requirements of the particular situation. The expected values and standard errors of these mean squares are 1 and (2/f) ,where f is the number of persons trying the item. More than three or four standard errors greater than one seems to be a reasonable rule of thumb for 'too large.' But items found to 'misfit' must also be carefully studied for the presence of a substantive explanation for their statistical implausibility before wise decisions concerning their use can be made. (Wright and Mead, 1978)

As demonstrated in equation 1, the Rasch model is based on the assumption that only one item parameter (difficulty = b), and one person parameter (ability = θ) are needed to describe what happens when an examinee attempts an item. When this assumption is correct, knowledge of an item's difficulty and a persons ability allows us to formulate the probability statement of the person correctly answering the item. This probability (expected) can be compared with the observed score to check whether the prediction holds. If the observed and expected number of correct responses are statistically equivalent, then the conclusion is that items and persons fit the Rasch model. The fit of items and persons to the RAsch model is provided in BICAL Program. The probability of a person (v) with ability Ov correctly answering an item (i) with the difficulty bi is Pvi and is obtained using equation 1. The difference between the observed score and the probability of a correct response (expected by the model) is called residual. The residual is small if the observed response is in the direction predicted by the model, or large if it is not. For example, if an examinee has a 0.95 probability of answering an item correctly, and indeed answered the item correctly, therefore receiving a score of one, the residual would be very low, (1-0.95) = 0.05.

To make it interpretable, the residual is divided by its theoretical standard deviation. The standardized residual is distributed normally with a mean of zero and a variance of one (Z-statistics)(See equation 3). When Z-statistics are squared and summed across persons, it produces a statistic that has an approximate X distribution that can be used to evaluate fit of an item (See equation 4).

BICAL Version 1980 produces five item fit statistics. The program also has an option to delete misfitting persons to the model. The important point is that the residual is the basic "building block" for all item fit statistics. It may be used to calculate a Z-score, T-score or chi-square statistic, but it always reflects the difference between what the model expected to happen and what was actually observed. Residuals can be summed across items to obtain person fit, or summed across the examinees to obtain item fit statistics. In "Between Fit" statistic examinees are grouped on the basis of their ability. This statistic shows whether the groups have responded as the model expected. In the following section, the item fit indices produced by the BICAL Version 1980 will be described.

Between Group Fit Statistic

The between groups residual (standardized) for group g on item i is computed using the following equation:

$$Zgi = \frac{Sgi - nr\ pri}{nr.\ Pri\ Qri} \qquad (6)$$

Where:

Sgi = Proportion of examinees in group g answering item i correctly.

nr = number of examinees with total score r

Pri = expected probability of success for ability r on item i

Qri = 1 - Pri

= sum across all abilities r within group g

Equation 5 can be transformed into a mean square among the m groups using the following equation:

$$(Zgi)\ \left(\frac{L}{(L-1)(N-1)}\right) \qquad (7)$$

Where:

L = Number of items
N = Number of groups

As the attached example of the BICAL program demonstrates, six groups of students responses are formed on the basis of total score. In other words, groups are formed of those students whose scores fall within the range of scores in each category. This grouping process results in similar abilities within each student group and different abilities among the six groups. In the attached example, 763 students whose score range was 5-55 form the lowest ability group while 584 students who score range was 75-76 form the highest ability group. The Rasch model expects that groups at the high end of the ability have a larger proportion of students correctly answering an item than groups at the low end of ability.

To compute "Fit Between" statistics, the following steps should be followed. Note that in the following process the ability groups are considered one at a time.

1. Calculate the probability of a correct response for each group

using equation 1.

2. Multiply this probability by the number of examinees within that ability group to obtain the expected number of examinees within that ability group answering the item correctly.

3. Do this for each ability within each ability groups.

4. Sum these expected numbers over all of the abilities in each group. This results in the number of people in the particular group (for example, lowest ability group) that would be expected to answer the item correctly.

5. Subtract the expected value obtained in step 4 from the actual number in the group who answered the item correctly. This produces the residual for a given ability group.

6. Standardize the residual (for each group) by dividing it by the standard deviation of these residuals using equation number five. The standard deviation is found by multiplying the probability of a correct response by the probability of incorrect response weighted by the number within ability O, summing across the abilities in that group, and taking the square root of this quantity.

7. The results obtained in step 6 is the standardized residual or

Z score. Then the mean square statistic can be calculated using equation 6. To do this, simply square the Z scores, summ across the six groups, and weight them by a function of the number of items and the number of groups.

8. Finally the mean square statistic is converted into a t-statistic. The t-statistic has a convenient mean of zero and the standard deviation of one. This statistic is labeled "Fit Between" in the BICAL output. The mean square statistic is converted into a t statistic using the following equation:

$$tgi = a\ Vgi - a + 1.0/a \qquad (8)$$

where:

$$a = [4.5(m-1)]$$

The "Fit Between" calculated using the equation 7 is shown in the attached BICAL output in the table headings "Item Characteristic Curve," "Departure From Expected ICC" and "Item Fit Statistics." The first two sections show the expected and actual performance of the six ability groups for each item. The item characteristic curve shows the proportion of examinees in each of six groups that answered the item correctly. The Rasch model expects this proportion to be smallest for the first group and largest for the sixth group. Simply stated, the proportion answering the item correctly within a group increases as the ability increases. The section labeled "Departure From

Expected ICC" shows the residual for each of the groups on each item. If the residual has a positive sign, it means that too many people answered the item correctly in that group. The negative value of residual means that too many people answered the item incorrectly. If a large positive residual for the first group (the lowest ability group) is followed by negative residuals for the remaining five groups, it indicates that examinees in the first group answered the item correctly perhaps by guessing. The third section in the BICAL output is labeled "Item Fit Statistics." In this section "Fit Between" is reported for each item. A high "Fit Between" statistic indicated that groups of different ability are not responding to an item in the way the model expected. "Fit Between" are assumed to have a standard deviation of one and the mean of zero. Items with "Fit Between" greater than three should be examined. In other words, if a low ability group of examinees answered an item correctly more than expected, there is something strange about that item.

Weighted Mean Square (WTDMNSQ)

For WTDMNSQ, each examinee defines an ability group. The difference between the observed response and what would be expected is the residual. The residuals are squared and summed over examinees. This value is divided by the sum of variances of responses of examinees of ability O to an item of difficulty ˋ. The result is WTDMNSQ which can be expresses in terms of tne following equation:

$$\text{WTDMNSQ} = \frac{(Xvi - Pvi)}{Pvi \cdot Qvi} \qquad (9)$$

The standard deviation of the WTDMNSQ is derived using the following equation:

$$\text{MNSQSD} = \frac{\sum[(P.Q)-4 \quad (P.Q)]^2}{\sum(P \cdot Q)} \qquad (10)$$

It is also called mean square standard error.

The WTDMNSQ has an expected mean of 1. As the observed item responses depart from the expected value of 1, WTDMN Q will increase. The weighted mean square fit statistics and their associated standard deviations are listed in the "Item Fit Statistics" produced by BICAL Program.

Total T Fit Statistic "T-Tests Total"

The total t-statistic is listed in the "Item Fit Statistics' table produced by BICAL program T-Tests Total Statistic is calculated by simply weighting the standardized WTDMNSQ. The purpose of this transformation is to come up with the same distribution for each item.

$$\text{T-Tests Total} = (\text{WTDMNSQ}^{1/3} - 1)\ (3/\text{Si}) + (\text{Si}/3) \qquad (11)$$

where:

WTDMNSQ = weighted mean square for item i

Si = mean square standard deviation for item i

T-Tests Total has the expected mean of zero and the expected standard deviation of one. Wright, et.al. (1980) indicates that in practice T-Test Total (which are assumed by definition to have a standard deviation of one) have been found to have standard deviation as low as 0.7 when the data fit the model. Then he suggests that "...values larger that 1.5 ought to be examined for response irregulation. Certainly "total t fit" values greater than 2.0 are noteworthy" (p. 13).

Error Impact "ERRIMPAC"

The impact of item misfit on item calibration is computed by subtracting one from the square root of the weighted mean square. This statistic is listed in the "Item Fit Statistics" produced by BICAL Program.

$$\text{Error Impact} = (\text{WTDMNSQ}^{1/2} - 1) \qquad (12)$$

The error impact provides a measure of the proportional inflation that the misfit of the item may have on the standard error of the item calibration. The error impact is a function of the difference between weighted mean square WTDMNSQ and its expected value of one. For example, if there are items in a given test, the inflation in measurement error over the test thatcan be attributed to one item's misfit would be equal to $(\text{WTDMSQ}/L)^{1/2}$.

Discrimination Index "DISCINDX"

The discrimination index labeled "DISCINDX" in the BICAL output is calculated using the following steps: (1) finding the difference between the residual for person-item combination and average residual for that item, (2) multiplying the results by the differ nce between the person ability and the item difficulty, (3) summiong over alal persons and (4) dividing this sum by the sum over persons of difference between person abilities and the item's difficulty, and (5) add one to the value obtained in step 4. The computational steps are summarized into this equation.

$$a_i = \frac{(Yvi - Y.i)\ (\theta v - bc)}{(\theta v - bi)^2} + 1 \qquad (13)$$

Where:

$$Yvi = \frac{(Xvi - Pvi)}{Pvi\ .\ Qvi} = \text{standardized residual}$$

Where: Pvi = probability of person v answers correctly item i

Xvi = actual response of person v on item i

Y.i = average of Yi for item i overall persons.

BICAL calculates the average slope for all the items. This average slope is given a value of one. The slope of all items' characteristic cures are compared with the value of one. The difference is represented by the discrimination index. It should be noted that this discrimination index is not the discrimination index used in classical testing theory. It tells us whether the item characteristic curves for the entire test. For example, if the item discrimination value for an item is one, it means that the observed ICC and average ICCs are identical. When the observed ICC is less than the average ICC, the discrimination value will be less than one. A high discrimination value for an item indicates that the particular item discriminates among abilities better than the average items on the test. Wright (1980) suggests that a high discrimination value may be the symptom of problems causing by an interaction between the item and the subjects. An example of this situation would be "speededness," where the low ability students might not reach the items at the end of the test. Low ability students would have large residuals on the final items which results in a high discrimination values for the final items.

```
MATH FORM 268                                                          RECAL. WITH  14 MISFITTING PERSONS OMITTED   F

                  SERIAL ORDER                           DIFFICULTY ORDER                               FIT ORDER
------------------------------------------------------------------------------------------------------------------------------
 SEQ ITEM    ITEM    STD    DISC     FIT ! SEQ ITEM   ITEM   DISC     FIT ! SEQ ITEM   ITEM    ERR   FIT T-TESTS   WTD  MNSQ  DISC P
 NUM NAME    DIFF  ERROR    INDX   TTEST ! NUM NAME   DIFF   INDX   TTEST ! NUM NAME   DIFF  IMPAC  BETWN  TOTAL  MNSQ   SD   INDX B
------------------------------------------------------------------------------------------------------------------------------
   1 I001  -1 79     09      73       40 !   3 I003  -2 53    .97    - 80 !  36 I033    .35    .00  11.57  -9.09   .78   .03  1.36
   2 I002  -1 38     08      73       60 !   6 I006  -2 40    .77    -.00 !  40 I037    .38    .00   9.02  -6.97   .83   .03  1.27
   3 I003  -2 53     13      97    -.80 !  46 I042  -2.37    .99    - 60 !  64 I059   1.04    .00   6.94  -6.94   .87   .02  1.21
   4 I004  -1 44     08     .99   -1 33 !   1 I001  -1.79    .73     .40 !  34 I031    .03    .00   9 62  -6.72   .81   .03  1.28
   5 I005  -1 15     07     .89     .10 !  57 I052  -1.71    .90   -1.29 !  65 I060    .86    .00   6.70  -6.11   .88   .02  1.20
   6 I006  -2 40    .12     .77    -.00 !  55 F005  -1.70    .88    -.39 !  68 I062   1.58    .00   5.38  -6.03   .91   .02  1.18
   7 I007   - 08     05     .54    6.56 !  24 I022  -1.62     73     .54 !  37 I034    .25    .00   7.94  -5.86   .85   .03  1.23
   8 I008   - 91    .07    1.02   -1 88 !  58 I053  -1 54    .99   -1.45 !  44 F004    .16    .00   7.06  -5.58   .85   .03  1.18
   9 I009     50     04     .89    1.49 !   4 I004  -1.44     99   -1.33 !  63 I058    .65    .00   6.52  -5.49   .88   .02  1.21
  10 I010     39     04      80    3.70 !  66 F006  -1.40   1.04   -1.16 !  20 I019   1.32    .00   7.57  -5.18   .91   .02  1.21
  11 F001   - 78     06    1 03    -.37 !   2 I002  -1.38    .73     .60 !  35 I032    .01    .00   8.72  -5.05   .85   .03  1.25
  12 I011   - 93     07    1 18   -3.20 !  61 I056  -1.31   1.17   -3.23 !  38 I035    .29    .00   6.64  -4.95   .87   .03  1.18
  13 I012     06     05     .78    3 30 !  27 I025  -1.29   1.00   -1 32 !  53 I049   1.97    .00   4.17  -3.76   .95   .01  1.13
  14 I013     54    .04    1.00    -.87 !   5 I005  -1.15    .89     .10 !  41 I038    .66    .00   3.79  -3.61   .92   .02  1.13
  15 I014     64     04     .97     .74 !  33 F003  -1.11   1.12   -2 78 !  61 I056  -1.31    .00   5 66  -3.23   .82   .06  1.17
  16 I015   1 01     04      78    5 85 !  45 I041  -1.02     93    -.65 !  12 I011   -.93    .00   5.64  -3.20   .85   .05  1.18
  17 I016     54     04    1.03    - 90 !  56 I051  -1.02    .80     .53 !  73 I067   -.55    .00   3.29  -3.03   .88   .04  1.07
  18 I017   - 78    .06    1 07    -.96 !  12 I011   -.93   1.18   -3 20 !  74 I068   1.34    .00   2.73  -2.95   .95   .02  1.10
  19 I018   1 01    .04      97    2.72 !   8 I008   -.91   1.02   -1 88 !  33 F003  -1.11    .00   3.74  -2.78   .86   .05  1.12
  20 I019   1 32     04    1 21   -5 18 !  62 I057   -.90    .95   -1 06 !  75 I069   -.00    .00   1 63  -2.53   .92   .03  1.07
  21 I020   -.04     05    1.03     .82 !  18 I017   -.78   1 07    - 96 !  26 I024    .04    .00   2.81  -2.42   .93   .03  1.05
  22 F002     08    .05    1.04    -.13 !  11 F001   -.78   1 03    - 37 !  72 I066   -.24    .00   2.39  -2.23   .92   .03  1.05
  23 I021   - 57     06     .76    1.39 !  59 I054   - 71    .97    -.36 !  42 I039    .08    .00   2.17  -2 14   .94   .03  1.07
  24 I022  -1 62     09      73      54 !  70 I064   -.67    .91    - 36 !  54 I050   1.12    .00   3.12  -2.14   .96   .02  1.09
  25 I023     06    .05      91   -1.00 !  23 I021   -.57    .76    1 39 !   8 I008   -.91    .00   1.21  -1.88   .91   .05  1.02
  26 I024     04    .05    1 05   -2.42 !  73 I067   - 55   1.07   -3 03 !  39 I036   -.10    .00   2.34  -1.57   .95   .03  1.08
  27 I025  -1 29     08    1 00   -1 32 !  71 I065   - 48    .99   -1.20 !  32 I030   1.35    .00   4.34  -1.51   .97   .02  1.08
  28 I026   - 11    .05      97     .18 !  72 I066   -.24   1.05   -2 23 !  58 I053  -1.54    .00    .64  -1.45   .91   .07   .99
  29 I027     45     04      93    1.29 !  28 I026   -.11     97     .18 !  77 F007    .07    .00    .59  -1.45   .96   .03  1.00
  30 I028   1.32     04    1 03     .15 !  39 I036   - 10   1 08   -1.57 !   4 I004  -1.44    .00   1.54  -1.33   .92   .06   .99
  31 I029   1 39     04    1 03     .33 !   7 I007   -.08     54    6.56 !  27 I025  -1.29    .00   2.21  -1.32   .92   .06  1.00
  32 I030   1.35     04    1 08   -1 51 !  21 I020   -.04   1.03      82 !  57 I052  -1.71    .00   3.13  -1.29   .91   .07   .90
  33 F003  -1 11     07    1 12   -2 78 !  75 I069   -.00   1.07   -2 53 !  71 I065   -.48    .00    .85  -1.20   .95   .04   .99
  34 I031     03     05    1 28   -6.72 !  35 I032    .01   1.25   -5 05 !  66 F006  -1.40    .00    .41  -1.16   .93   .06  1.04
  35 I032     01    .05    1 25   -5.05 !  34 I031 [illegible] 1.28  -6 72 !  62 I057   -.90    .00   1.59  -1.06   .95   .05   .95
  36 I033     35     04    1 36   -9 09 !  26 I024     04   1.05   -2.42 !  25 I023    .06    .00   5.62  -1.00   .97   .03   .91
  37 I034     25     05    1 23   -5 86 !  25 I023    .06    .91   -1.00 !  18 I017   -.78    .00   1.58   -.96   .96   .04  1.07
  38 I035     29    .05    1 18   -4 95 !  13 I012     06    .78    3.30 !  17 I016    .54    .00   3 01   -.90   .98   .02  1 03
  39 I036   - 10     05    1 08   -1.57 !  77 F007     07   1.00   -1 45 !  14 I013    .54    .00   -.40   -.87   .98   .02  1.00
  40 I037     38     04    1 27   -6 97 !  42 I039     08   1 07   -2.14 !   3 I003  -2.53    .00    .79   -.80   .91   .11   .97
  41 I038     66     04    1 13   -3.61 !  22 F002     08   1.04    -.13 !  45 I041  -1 02     00   2.01   -.65   .97   .05   .93
  42 I039     08     05    1 07   -2 14 !  76 I070    .16    .92      42 !  46 I042  -2.37    .00   -.10   -.60   .94   .10   .99
  43 I040   1 46     04      98     .40 !  44 F004    .16   1.18   -5 58 !  55 F005  -1.70    .00   3.20   -.39   .97   .07   .88
  44 F004     16     05    1 18   -5 58 !  37 I034     25   1 23   -5 86 !  11 F001   -.78    .00     02   -.37   .98   .04  1.03
  45 I041  -1 02     07      93    - 65 !  38 I035    .29   1 18   -4.95 !  70 I064   -.67     00   3 90   -.36   .98   .04   .91
  46 I042  -2 37     12      99    - 60 !  36 I033     35   1 36   -9 09 !  59 I054   -.71     00  -1 31   - 36   .98   .04   .97
  47 I043   1 19     04     .72    7 27 !  40 I037    .38   1 27   -6 97 !  22 F002     08    .00   4 95   -.13  1.00   .03  1.04
  48 I044   1 42     04    1 00     .28 !  10 I010     39    .80    3 70 !   6 I006  -2.40    .00   4.25   -.00  1.00   .10   .77

          ****************************************************************************************************
                                              TABLE CONTINUED
```

```
                     ITEM CHARACTERISTIC CURVE                 DEPARTURE FROM EXPECTED ICC                         ITEM FIT STATISTICS
------------------------------------------------------------------------------------------------------------------------------------------------
SEQ  ITEM !  1ST   2ND   3RD   4TH   5TH   6TH  !  1ST   2ND   3RD   4TH   5TH   6TH  ! * ERR   FIT T-TESTS   WTD  MNSQ   DISC  POINT
NUM  NAME ! GROUP GROUP GROUP GROUP GROUP GROUP ! GROUP GROUP GROUP GROUP GROUP GROUP ! IMPAC  BETWN  TOTAL  MNSQ  SD    INDX  BISER
------------------------------------------------------------------------------------------------------------------------------------------------
 49  I045 !   .55   .73   .85   .92   .95   .99 !   .04  -.03  -.00   .01    00   .01 !  .C1   1.33    .79  1.02  .03    .97     42
 50  I046 !   .47   .61   .71   .79   .87   .93 !   .13   .01  -.03  -.04  -.03  -.03 !  .07   7.96   7 41  1.14  .02    .72     35
 51  I047 !   .28   .35   .50   .63   .76   .89 !   .10  -.02  -.02  - 02  -.02  -.01 !  .04   6.48   6 03  1.09  .01    .81     38
 52  I048 !   .24   .37   .50   .59   .67   .86 !   .08   .04   .02  -.02  -.08  -.03 !  .07   7.67   9.37  1.14  .01    .71     35
 53  I049 !   .17   .40   .51   .67   .86   .94 !  -.02   .00  -.04  - 01   .06   .03 !  .00   4.17  -3.76   .95  .01   1 13    .49
 54  I050 !   .34   .60   .72   .85   .93   .99 !  -.01  -.01  - 03    02   .03   .02 !  .00   3.12  -2 14   .96  .02   1 09    .49
 55  F005 !   .89   .98   .97   .98   .98  1.00 !   .01   .01  -.01  -.00  -.01   .00 !  .00   3.20   -.39   .97  .07    .88     24
 56  I051 !   .84   .93   .96   .96   .97   .99 !   .05   .01  -.00  -.02  -.02  -.01 !  .01   5.36    .53  1.03  .05    .80     26
 57  I052 !   .89   .97    98   .98   .99   .99 !   .01   .01   .00  -.01  -.01  -.01 !  .00   3.13  -1.29   .91  .07     90     30
 58  I053 !   .86   .97   .97   .98   .99  1.00 !  -.00   .01  -.00  - 00   .00   .00 !  .00     64  -1.45   .91  .07     99     33
 59  I054 !   .76   .90   .95   .97   .98   .99 !   .01  -.00   .00   .00  -.00  -.00 !  .00  -1.31   -.36   .98  .04     97    .35
 60  I055 !   .46   .61   .71   .82   .88   .95 !   .10  -.00  -.03  - 02  -.03  -.02 !  .05   5.95   5.75  1.11  .02    .79     37
 61  I056 !   .78   .97   .99   .99  1.00   .99 !  -.05   .02   .02    01   .01  -.00 !  .00   5.66  -3.23   .82  .06   1.17     46
 62  I057 !   .78   .94    95   .96   .98   .99 !   .00   .02  -.00  -.01  -.00  -.00 !  .00   1.59  -1.06   .95  .05     95     35
 63  I058 !   .37   .69   .86   .93   .97   .99 !  -.09  -.02   .04   .04   .03   .01 !  .00   6.52  -5.49   .88  .02   1 21     55
 64  I059 !   .26   .62   .80   .89   .94   .97 !  - 11  -.00   .05   .05   .03   .00 !  .00   6.94  -6.94   .87  .02   1.21     57
 65  I060 !   .31   .66   .84    91    95   .98 !  -.10  -.01   .05   .04   .02   .01 !  .00   6.70  -6.11   .88   02   1.20     56
 66  F006 !   .83    96   .97   .99   .99  1 00 !  -.01   .01   .00   .01   .00   .00 !  .00    .41  -1.16   .93  .06   1 04     34
 67  I061 !   .48   .69   .79   .88   .92   .95 !   .05   .00  -.01    00  -.01  -.03 !  .02   3.60   2.01  1.04  .02    .87     41
 68  I062 !   .19   .47   .67   .80   .90   .94 !  -.07  -.03   .03    05   .05  -.00 !  .00   5.38  -6.03   .91  .02   1.18     54
 69  I063 !   .41   .64   .74   .80   .89   .93 !   .06   .03  - 00  - 04  - 02  -.03 !  .04   5.15   4.59  1.09  .02    .81     39
 70  I064 !   .75   .92   .95   .96   .96   .99 !   .01   .01   .01  - 00  - 02  -.00 !  .00   3.90  - 36    .98  .04     91     35
 71  I065 !   .70   .89   .94   .97   .97   .99 !  -.00   .01    01    01  -.01   .00 !  .00    .85  -1.20    95  .04     99     40
 72  I066 !    62    89   .92   .97   .97   .99 !  - 04   .03    01   .01   .00  -.00 !  .00   2.39  -2.23   .92   03   1 05     45
 73  I067 !   .67    92   .96   .97   .98   .99 !  -.05   .03   .02    01   .00  -.00 !  .00   3 29  -3.03    88   04   1 07     46
 74  I068 !   .29   .53   .69   .82   .92   .97 !  -.01  -.03  -.01   .02   .03   .01 !  .00   2.73  -2.95   .95   02   1 10     50
 75  I069 !    57   .83   .91   .96   .97   .99 !  -.03   .01    02   .02   .00   .00 !  .00   1.63  -2.53   .92  .03   1 07     47
 76  I070 !   .59   .80   .89   .92   .95   .98 !   .03   .00   .01  -.01  -.01  -.01 !  .01   1.12    .42  1.01  .03     92     40
 77  F007 !   .57   .84   .90   .94   .97   .98 !  -.02   .02   .01    00   .00  -.01 !  .00    .59  -1.45    96  .03   1 00     45
------------------------------------------------------------------------------------------------------------------------------------------------
SCORE RANGE   5-55 56-63 64-68 69-71 72-74 75-76                                      *ERROR IMPACT = PROPORTION ERROR INCREASE
MEAN ABILITY   .44  1.55  2.18  2.72  3.39  4.40         PLUS=TOO MANY RIGHT                        DUE TO THIS MISFIT
                                                         MINUS=TOO MANY WRONG
MEAN Z-TEST    .2    .5    .4    .1   -.1   - 0
SD(Z-TEST)    4.1   1.8   2.1   2.2   3.1   2.3
GROUP COUNT   763   790   799   631   839   584

     77 ITEMS CALIBRATED ON    4406 PERSONS
   4406 MEASURABLE PERSONS WITH MEAN ABILITY =    2.37  AND STD. DEV. =    1.19
```

References

Buros, O. K. (1978) Fifty years in testing. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press, 1972-1983.

Canner, J. M., & Lenke, J. M. (1980, April). Some types of test items do not fit the Rasch model: Examples and hypotheses. Paper presented at the annual meeting of the National Council of Measurement in Education, Boston, MA.

Cook, L. L. & Hambleton, R. K. (1979). A comparative study of item selection methods utilizing latent trait theoretic models and concepts. Paper presented at the annual meeting of National Council on Measurement in Education.

Divgi, D. R. (1981, April). Does the Rasch model really work: Not if you look closely. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Douglass, J. B. (1981, April). Item bias, test speededness, and Rasch tests of fit. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

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

Gustafsson, J. E. (1979, April). Testing and obtaining fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14 (2), Summer.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Lawrence, Erlbaum Assoc., Hillsdale, New Jersey.

Mead, R. J. (1976a). Assessing the fit of data to the Rasch model. Paper presented at the Annual meeting of the

American Educational Research Association, San Francisco.

Mead, R. J. (1976b). Assessing the fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago.

Popham, W. J. (1978). Criterion-Referenced Measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Rasch, G. (1966). An item analaysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Reckase, M. D. (1981). To use or not to use (the one- or three- parameter model) that is the question. Paper presented at American Educational Research Assoc. annual meeting, Los Angeles.

Sabers, J. L. & Jones, P.B. (1982). Issues in norming a custom curriculum-referenced test. Paper presented at annual meeting of Florida Educational Research Association, Orlando.

Rentz, R. R., & Ridenour, S. E. (1978, March). The fit of the Rasch model to achievement tests. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, VA.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Mead, R. J. (1977). BICAL: Calibrating and scales with the Rasch model. (Research Memorandum No. 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). BICAL: Calibrating items with the Rasch model. (Research Memorandum No. 23c). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., Mead, R., & Draba, R. (1976). Detecting and correcting item bias with a logistic model. (Research Memorandum No. 22) Chicago: University of Chicago, Statistical Laboratory.

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