JOURNAL OF OUTCOME MEASUREMENT®, 2(1), 66 -78

Using Item Mean Squares to Evaluate Fit to the Rasch Model

Richard M. Smith
Rehabilitation Foundation, Inc.
Marianjoy Rehabilitation Hospital and Clinics

Randall E. Schumacker
University of North Texas

M. Joan Bush
Irving Independent School District

Throughout the mid to late 1970's considerable research was conducted on the properties of Rasch fit mean squares. This work culminated in a variety of transformations to convert the mean squares into approximate t-statistics. This work was primarily motivated by the influence sample size has on the magnitude of the mean squares and the desire to have a single critical value that can generally be applied to most cases. In the late 1980's and the early 1990's the trend seems to have reversed, with numerous researchers using the untransformed fit mean squares as a means of testing fit to the Rasch measurement models. The principal motivation is cited as the influence sample size has on the sensitivity of the t-converted mean squares. The purpose of this paper is to present the historical development of these fit indices and the various transformations and to examine the impact of sample size on both the fit mean squares and the t-transformations of those mean squares. Because the sample size problem has little influence on the person mean square problem, due to the relatively short length (100 items or less), this paper focuses on the item fit mean squares, where it is common to find the statistics used with sample sizes ranging from 30 to 10,000.

An earlier version of this paper was presented at the 1995 Annual Meeting of the American Educational Research Association in San Francisco.

Recent presentations at the Rasch Measurement SIG sessions at AERA have stressed the use of the weighted and unweighted item mean squares as a means to evaluate the fit of the responses to a Rasch model. This evaluation is usually based on a single critical value on the order of 1.2 to 1.3 for both mean squares. The rationale usually given is that the mean square is less affected by sample size than the approximate t-statistic resulting from the cube root transformation of the fit mean square. These arguments are contradictory to the arguments used in the late 1970's and early 1980's when the fit mean square transformations were developed.

HISTORY OF FIT

One of the methods of assessing fit in Rasch measurement models, and the technique that is used in most of the calibration and analysis programs distributed by MESA Press, is based on concatenation of the item/person residual. Other methods, such as those based on the likelihood ratio chi-square, will not be discussed in this paper. There is an approximately parallel history of development for item and person fit statistics based on the item/person residual (Smith, 1989 and Smith, 1991b); however, only the development of the item fit statistics, the object of inquiry in this study will be detailed here.

The item fit statistic, first proposed by Wright and Panchapakesan (1969), was based on person raw score groups which focused on the difference between the observed and expected score for a group of persons with the same raw score on a test. Subsequent developments in fit statistics have been based on the item/person residual. The unweighted item total fit statistic (UT) in the chi-square form, based on the item/person residual (y_{ni}) is

$$x^2(UT)_i = \sum_{n=1}^{N} y^2_{ni}. \qquad (1)$$

The standardized residual y_{ni} is

$$y_{ni} = \frac{(x_{ni} - p_{ni})}{(w_{ni})^{1/2}}, \qquad (2)$$

where x_{ni} is the observed score for each item/person interaction, p_{ni} is the probability of a correct response for each interaction, and $w_{ni} = p_{ni}(1 - p_{ni})$. This chi-square is calculated for each item by summing over all of the persons in the response matrix.

This chi-square can be converted to a mean square by dividing by the number of persons (N),

$$MS(UT)_i = \left(\frac{1}{N}\right)\chi^2(UT)_i = \left(\frac{1}{N}\right)\sum_{n=1}^{N}\frac{(x_{ni} - p_{ni})^2}{w_{ni}}. \quad (3)$$

Note that the degrees of freedom used to convert these and subsequent total fit statistics to mean squares are N rather than the (N-1) used with the Wright Panchapakesan χ^2. This is due to the fact that the (N-1) overcorrects for the loss in degrees of freedom due to using the same x_{ni} to estimate the item and person parameters used in calculating the p_{ni} and to calculate the score residual. Alternative methods for correcting for the loss in degrees of freedom are discussed Smith (1982, 1991b).

The standard deviation of this mean square can be estimated by

$$S[MS(UT)_i] = \frac{\left[\sum_{n=1}^{N}\frac{1}{w_{ni}} - 4N\right]^{1/2}}{N}. \quad (4)$$

These statistics originally were evaluated as fit mean squares (FMS)in BICAL, an early Rasch calibration program. Where MS(UT) has an expected value of one and the standard deviation given in Equation 4. The critical values for detecting misfit with this mean square depend on the number of persons and w_{ni}, so they will vary from item to item and sample to sample. To simplify the critical value problem, the mean square can be standardized to an approximate unit normal by a variety of transformations. This transformation, the unweighted total item fit statistic, is discussed in Wright and Stone (1979).

Later versions of BICAL introduced a log transformation in an attempt to standardize the fit statistics to an approximate unit normal distribution. In this transformation

$$t = \left[ln\left(MS(UT)_i\right) + MS(UT)_i - 1\right]\left[\frac{f}{8}\right]^{1/2}. \quad (5)$$

indicate possible misfit varied from item to item and analysis to analysis depending on the number of persons, the distribution of item difficulties, and the distribution of person abilities.

The last version of BICAL introduced a cube root transformation to convert MS(UT) to approximate unit normals. In this transformation

$$t = \left[\left(MS^{1/3} - 1\right)(3/S)\right] + (S/3), \quad (6)$$

where S is the standard deviation of MS(UT) or MS(UB) given above in equation 4.

Experience with the unweighted fit statistic indicated that when there was a large range of item difficulties and person abilities, unexpected correct responses by low ability persons to difficult items and unexpected incorrect responses by high ability persons to easy items affected the unweighted mean square severely. A relatively small number of anomalous responses can result in unusually large mean squares and t-statistics.

The last version of BICAL also introduced the weighted version of the total item fit statistic, which replaced the unweighted version in that program. The weighted item total fit statistic was developed to diminish the effect of anomalous outliers. In this statistic the squared standardized residual (y_{ni}^2) is weighted by the information function (w_{ni}). The weighted item total fit statistic (WT) in the chi-square form is

$$MS(WT)_i = \frac{\sum_{n=1}^{N} w_{ni}\frac{(x_{ni} - p_{ni})^2}{w_{ni}}}{\sum_{n=1}^{N} W_{ni}} = \frac{\sum_{n=1}^{N}(x_{ni} - p_{ni})^2}{\sum_{n=1}^{N} w_{ni}}. \quad (7)$$

The weighted total mean square is the sum of the weighted squared standardized residuals divided by the sum of the weights. The standard deviation of this statistic is

$$S[MS(WT)_i] = \frac{\left[\sum_{n=1}^{N} w_{ni} - 4\sum_{n=1}^{N} w^2_{ni}\right]^{1/2}}{\sum_{n=1}^{N} w_{ni}}. \quad (8)$$

The weighted version of the total fit statistic is less affected by anomalous responses by persons with ability far from the difficulty of the item. A further description of the weighted total fit statistic can be found in Wright

In recent programs, e.g., BIGSCALE, BIGSTEPS, and FACETS, the unweighted fit statistics (item and person) have become known as OUTFIT statistics and the weighted fit statistics have become known as INFIT statistics.

This study was designed to illustrate the differences between the fit mean squares and the transformed version of the item fit statistics. This comparison focused on the use of a single critical value to determine misfit and effect of sample size and the type of statistic being evaluated (OUTFIT vs. INFIT) on the distribution of the item fit mean square. The Type I error rates of the fit mean square are then compared with those of the transformed t-statistic.

METHODS

In this study 100 replications of simulated data were generated under each of six different conditions which varied the number of persons and the number of items. These conditions were: 150 persons with 20 and 50 item tests, 500 persons with 20 and 50 item tests, and 1000 persons with 20 and 50 item tests. Person abilities were normal with a 0, 1 distribution. Item difficulties were uniformly distributed from -2.0 to + 2.0 logits (See Schumacker, Smith, and Bush (1994) for a complete description of the simulated data.). All simulated data sets were calibrated with the BIGSTEPS program (Wright and Linacre, 1991). For each calibration an item file was generated which contained the weighted and unweighted mean squares and t-statistics for each of the items in that calibration. The mean, standard deviation, minimum value, maximum value, and per cent of cases above given critical values were calculated for each of four statistics, weighted mean squares and t-statistics and unweighted mean squares and t-statistics, in each data set. These summary statistics were then averaged across the 100 replications in each combination of test length and number of persons. The critical values used to calculate the percent of cases with extreme values were fms>1.3, fms>1.2, fms>1.1, fms<.9, fms<.8, and fms<.7 for the mean squares and t>+4, t>+3, t>+2, t<-2, t<-3, and t<-4 for the t-statistics.

RESULTS

t-statistics used in this analysis were obtained from the item file option available in the BIGSTEPS program. The summary information for the weighted mean squares is presented in Table 1 and in Table 2 for the unweighted mean squares.

The means of both mean squares (unweighted and weighted) are very stable about the expected value of 1.00. The average weighted mean square means have a standard deviation of 0.00 across the six conditions, and the unweighted mean square means have a maximum standard deviation of 0.03 across the six conditions. Thus, the number of persons and the length of the test appear to have a small influence on the mean of the unweighted mean squares ($SD \leq .03$), and the influence on the mean of the weighted mean squares cannot be seen in the second decimal point (all $SD = 0.00$).

The standard deviation of the mean squares varies considerably based on the type of mean square (weighted and unweighted) and the number of persons. The mean standard deviation for the unweighted mean squares is approximately double that of the weighted mean square. For example, the mean standard deviation for the unweighted mean square varies from 0.18 with 150 persons to 0.06 for 1000 persons (Table 2). The mean standard deviation of the weighted mean square varies from 0.08 for 150 persons to 0.03 for 1000 persons (Table 1). The standard deviation does not appear to be affected by the number of items, as seen by comparing the top and bottom halves of Tables 1 and 2.

The range of the mean squares is similarly affected. The mean range for the unweighted mean square is 0.72 for 150 persons and 20 items, 0.40 for 500 persons and 20 items, and 0.25 for 1000 persons and 20 items, but the number of items on the test has little effect on the range of the unweighted mean square. Contrast this with the range of the weighted mean square. In the same example given above, the mean range for the weighted mean square is 0.29 for 150 persons and 20 items, 0.16 for 500 persons and 20 items, and 0.10 for 1000 persons and 20 items. These are less than one-half of the range for the unweighted mean squares. As with the unweighted mean square, there appears to be considerable influence resulting from the number of persons and little influence resulting from test length on the range of the mean squares.

To examine the Type I error rates and the influence of mean square type, number of persons and test length, six critical values were chosen

Table 1
Weighted Mean Square Descriptive Statistics

Simulation 1 (150 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.01
S.D.	0.08	1.01	0.05	0.11
Maximum	1.15	0.04	1.08	1.33
Minimum	0.86	0.03	0.81	0.92
Simulation 2 (500 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.04	0.01	0.03	0.06
Maximum	1.08	0.02	1.04	1.15
Minimum	0.92	0.02	0.87	0.96
Simulation 3 (1000 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.03	0.00	0.02	0.04
Maximum	1.05	0.01	1.03	1.09
Minimum	0.95	0.01	0.91	0.97
Simulation 4 (150 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.00
S.D.	0.07	0.01	0.05	0.09
Maximum	1.16	0.03	1.09	1.25
Minimum	0.85	0.03	0.75	0.91
Simulation 5 (500 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	1.00	1.00
S.D.	0.04	0.00	0.03	0.05
Maximum	1.09	0.02	1.05	1.16
Minimum	0.92	0.02	0.86	0.95
Simulation 6 (1000 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	1.00	1.00
S.D.	0.03	0.00	0.02	0.03

Table 2
Unweighted Mean Square Descriptive Statistics

Simulation 1 (150 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.03	0.95	1.11
S.D.	0.18	0.07	0.10	0.53
Maximum	1.45	0.31	1.17	3.17
Minimum	0.73	0.06	0.58	0.86
Simulation 2 (500 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.97	1.04
S.D.	0.10	0.03	0.05	0.19
Maximum	1.25	0.13	1.05	1.80
Minimum	0.85	0.04	0.73	0.91
Simulation 3 (1000 persons, 20 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.02
S.D.	0.06	0.01	0.03	0.10
Maximum	1.14	0.06	1.03	1.34
Minimum	0.89	0.03	0.80	0.95
Simulation 4 (150 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.05
S.D.	0.16	0.03	0.11	0.35
Maximum	1.52	0.28	1.17	3.19
Minimum	0.71	0.05	0.58	0.81
Simulation 5 (500 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.01	0.98	1.02
S.D.	0.08	0.02	0.06	0.16
Maximum	1.25	0.14	1.10	1.99
Minimum	0.82	0.04	0.72	0.88
Simulation 6 (1000 persons, 50 items)				
SAMPLE	MEAN	S.D.	MIN	MAX
Mean	1.00	0.00	0.99	1.01
S.D.	0.06	0.01	0.04	0.09

Table 3
Mean Square Frequency of Extreme Values

Weighted	Simulation Conditions*					
	1	2	3	4	5	6
% > 1.3	0.05	0.00	0.00	0.00	0.00	0.00
% > 1.2	0.60	0.00	0.00	0.32	0.00	0.00
% > 1.1	8.05	0.60	0.00	6.90	0.40	0.04
%< 0.9	8.35	0.65	0.00	6.62	0.20	0.00
%< 0.8	0.00	0.00	0.00	0.08	0.00	0.00
%< 0.7	0.00	0.00	0.00	0.00	0.00	0.00

Unweighted	Simulation Conditions					
	1	2	3	4	5	6
% > 1.3	4.75	1.35	0.05	3.48	0.52	0.12
% > 1.2	10.05	3.90	0.65	8.44	1.76	0.48
% > 1.1	21.40	12.50	4.85	20.70	8.72	4.38
%< 0.9	28.05	11.15	3.10	23.56	8.88	2.70
%< 0.8	8.30	0.50	0.00	6.36	0.48	0.04
%< 0.7	1.30	0.00	0.00	0.84	0.00	0.00
N of Persons	150	500	1000	150	500	1000
N of Items	20	20	20	50	50	50

* Each simulation condition contained 100 replications.

commonly used value for detecting measurement disturbances, occurre less than 1 time per 200 for all sample sizes and test lengths for the weighte mean square and values greater than 1.1 occurred less than 1 time per 20 for sample sizes greater than 500. If weighted mean square critical valu of 1.2 were to be used, then the Type I error rate would approximate .00 With the weighted mean squares the per cent greater than the critical valu is too small in most cases to accurately judge the effect of test length o the statistic.

For the unweighted mean square and sample size of 150, values greate than 1.3 occurred at a rate of approximately 5 per cent. For sample size c 500, values greater than 1.3 occurred at a rate of approximately 1 per cen For sample size of 1000, values greater than 1.3 occurred at a rate of ap proximately .1 per cent. To have a consistent Type I error rate of approxi mately .05, a critical value of 1.3 would be needed with 150 person sample 1.2 with 500 person samples, and 1.1 with 1000 person samples. It is als clear from these data that unweighted mean square is moderately affecte by test length with the per cent above the critical value approximately per cent higher for the 20 item tests than for the 50 item tests.

It is also clear from the values listed in Table 3 that the mean square i not symmetrically distributed about 1.0. Extreme values occur far les frequently below 1.0 then above. This means that symmetrical critical va ues for detecting misfit would operate at different Type I error rates for th upper and lower tails of the distribution.

The results of these simulations suggest that no single critical valu will work with both weighted and unweighted mean squares. It is als clear that no single value will work with different sample sizes. If a crit cal value of 1.2 were chosen, the actual Type I error rate could vary an where from 0.00001 to 0.10 depending on the set of circumstances.

In an effort to contrast the use of the mean square with the transforme t-statistic, the frequency of extreme values for the same simulations we calculated. These are presented in Table 4. In this table the critical valu chosen were +4, +3, +2, -2, -3, and -4. There is no equivalence implie between these values and the values chosen for use in Table 3. They a simple convenient numerical values. The +2.0 value is often used as a indication of misfit with the t-statistic. As is clear from this table, the Ty I error rate for the unweighted t-statistic is approximately twice the valu for the weighted version. However, the differences across the weighte and unweighted version of the t-statistic are less extreme then across t

Table 4
t-statistic Frequency of Extreme Values

Weighted	Simulation Conditions*					
	1	2	3	4	5	6
% > 4.0	0.05	0.00	0.00	0.00	0.00	0.00
% > 3.0	0.10	0.00	0.00	0.02	0.04	0.08
% > 2.0	1.35	0.60	0.40	1.02	0.66	0.64
%< -2.0	0.65	1.70	1.10	0.70	0.70	0.90
%< -3.0	0.00	0.05	0.05	0.06	0.00	0.04
%< -4.0	0.00	0.00	0.00	0.00	0.00	0.00
Unweighted	**Simulation Conditions**					
	1	2	3	4	5	6
% > 4.0	0.10	0.05	0.00	0.08	0.10	0.06
% > 3.0	0.40	0.50	0.10	0.24	0.22	0.26
% > 2.0	2.60	2.45	1.40	2.24	1.80	1.56
%< -2.0	0.35	1.25	0.90	0.40	0.62	0.80
%< -3.0	0.00	0.00	0.00	0.02	0.00	0.00
%< -4.0	0.00	0.00	0.00	0.00	0.00	0.00
N of Persons	150	500	1000	150	500	1000
N of Items	20	20	20	50	50	50

* Each simulation condition contained 100 replications.

items the Type I error rate for the unweighted t-statistic value of +2.0 is 0.026 for the weighted t-statistic the value is 0.0135, approximately a two-fold difference. For the mean square with 150 persons and 20 items the Type I error rate for a value of 1.2 is 0.006 for the weighted version and 0.10 for the unweighted version, approximately a 15-fold difference. This difference is far greater than with the t-statistic. Also, the differences across sample size are less drastic with the t-statistic then with the mean square. The Type I error rates for the unweighted t-statistic critical value of +2.0 with 150 and 1000 persons are 0.026 and 0.014, a multiple of about 2. The Type I error rates for the unweighted mean square critical value of 1.2 with 150 and 1000 persons are 0.10 and 0.0065, a multiple of about 15.

Although it is clear from these simulations that the use of a single critical value for the t-statistic may lead to different Type I error rates for different statistics, sample sizes and test lengths, the effect of these three factors on the statistics is less than those observed for the mean squares. It should be noted that Smith (1982, and 1991b) has proposed several methods for removing the differences found across fit statistics due to the differences in sample size and type of statistic. The values reported in this study were generated by BIGSTEPS which does not employ these corrections. If these corrections were employed, the dissimilarity between the Type I error rates for the t-statistics would be less than those observed here.

DISCUSSION

Clearly these results indicate that the critical value for the mean square used to detect misfit is affected both by the type of the mean square and the number of persons in the calibration. A single critical value, particularly one of 1.2 or 1.3 will not give a .05 Type I error rates for sample sizes of 500 or larger. For the weighted version (INFIT) even a value of 1.1 is too large for sample sizes more than 500. These results have serious implications for BIGSTEPS users since the item fit mean squares have become the preferred method with which the fit of the data to the model is determined. Many authors suggest that the mean square is less sensitive to large sample size that the t-transformation. These results show that this is not the case. The mean squares are more sensitive to sample size and reliance on a single critical value for the mean square can result in an under detection of misfit.

Wright (personal communication, 1996) suggests that the critical value

indicates the direct influence of sample size on the two mean squres, critical value $MS(WT)=1+\frac{2}{\sqrt{x}}$, and

critical value $MS(UT)=1+\frac{6}{\sqrt{x}}$,

where x=the sample size. This formula would yield critical values of 1.16 for sample sizes of approximately 150, 109 for for sample sizes of approximately 500, and 1.06 for sample sizes of approximately 1000 for the weighted mean square. Critical values for the unwieghted mean square would be 1.48 for 1.27 for sample sizes of approximately 150, sample sizes of approximately 500, and 1.19 for sample sizes of approximately 1000. Further research is needed to establish the exact Type I error rate for these approximate critical values and to examine the impact of larger sample sizes (n=1500 and n=2000) on the data.

REFERENCES

Schumacker, R. E., Smith, R. M., Bush, M. J. (1994). Examining replication effects in Rasch fit statistics. A paper presented at the 1994 annual meeting of the American Educational Research Association.

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

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

Smith, R. M. (1989). Item and person fit in the Rasch model. A paper presented at the 1989 annual meeting of the American Educational Research Association.

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

Smith, R. M. (1991b). *IPARM: Item and person analysis with the Rasch model.* Chicago: MESA Press.

Smith, R. M. (1992). *Applications of Rasch measurement.* Chicago: MESA Press.

Wright, B. D. and Linacre, J.M. (1991). *BIGSTEPS: Rasch analysis for all two-facet models.* Chicago: MESA Press.

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

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

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