DOCUMENT RESUME

ED 368 788 TM 021 305

AUTHOR	Schumacker, Randall E.; And Others
TITLE	Examining Replication Effects in Rasch Fit Statistics.
PUB DATE	Apr 94
NOTE	24p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Goodness of Fit; Item Response Theory; Research Methodology; *Sample Size; Simulation; Statistical Studies; Test Items; *Test Length
IDENTIFIERS	Power (Statistics); *Rasch Model; Type I Errors; Weighting (Statistical)

ABSTRACT

Rasch between and total weighted and unweighted fit statistics were compared using varying test lengths and sample sizes. Two test lengths (20 and 50 items) and three sample sizes (150, 500, and 1,000 were crossed. Each of the six combinations were replicated 100 times. In addition, power comparisons were made. Results indicated that there were no differences in item and person Rasch fit statistics based on the number of replications. The Type I error rates were close to expected values. It was concluded that the number of replications in Rasch simulation studies are not a major factor influencing fit. Researchers should be more sensitive to the number of persons, number of items, and the correction for degrees of freedom used in the mean square calculation (n vs. n-1). Ten tables are included. (Contains 13 references.) (Author)

EXAMINING REPLICATION EFFECTS IN RASCH FIT STATISTICS

Randall E. Schumacker
University of North Texas

Richard M. Smith
University of South Florida

Joan M. Bush
Dallas ISD, Dallas, Texas

Paper presented at the American Educational Research Association annual meeting in New Orleans, April 7, 1994.

ABSTRACT

Rasch between and total weighted and unweighted fit statistics were compared using varying test lengths and sample sizes. Two test lengths (20 and 50 items) and three sample sizes (150, 500, and 1000) were crossed. Each of the six combinations were replicated 100 times. In addition, power comparisons were made.

Results indicated that there were no differences in item and person Rasch fit statistics based on the number of replications. The Type I error rates were close to expected values. It was concluded that the number of replications in Rasch simulation studies are not a major factor influencing fit. Researchers should be more sensitive to the number of persons, number of items, and the correction for degrees of freedom used in the mean square calculation (n vs. n-1).

EXAMINING REPLICATION EFFECTS IN Rasch FIT STATISTICS

Several studies have been done on the fit of items and persons to the Rasch measurement model. The research literature includes studies related to the effect of (a) test length, (b) sample size, (c) item difficulty distribution, (d) person ability distribution, and (e) the number of steps in each item on the fit statistics. In most cases, computer-simulated data with 10 to 100 replications were used. No research, however, has determined if these results would have been affected by the number of replications.

Rasch FIT STATISTICS

Fit statistics are used to provide a frame of reference for judging the performance of a given item or person on an objectively measured variable. Rasch (1960) suggested several methods for assessing item and person fit to his model. Unfortunately, no computer programs were available at the time that were capable of computing these statistics. As a result, his suggested fit indices have not been widely used. Even so, his influence is reflected in the subsequent work related to fit indices.

The first fit statistics were based upon the overall chi-square (Wright & Panchapakesan, 1969) or the likelihood-ratio chi-square approach (Anderson, 1973). Later, these statistics were converted to weighted or unweighted mean squares. Unweighted mean squares were referred to as "outfit" since more weight is given to items or persons far from the expected logit measure. In contrast, weighted mean squares were referred to as "infit" since more weight is given to unexpected responses nearer the expected item or person logit measure. Most recently, a cube root transformation has been used to convert the mean square to an approximate unit normal with a mean of zero and a standard deviation of one.

In the recent Rasch measurement computer programs, total and between fit statistics are reported for items and persons. These statistics are sensitive to different types of measurement disturbances. The total fit statistic is sensitive to measurement disturbances such as guessing, discrimination, start-up fluctuations, sloppiness, and unexpected correct and incorrect responses. In addition, the total fit statistic is sensitive to changes in the slope of the person or item characteristic curves. Thus, the total fit statistic is sensitive to unsystematic measurement disturbances. In contrast, the between fit statistic is sensitive to systematic measurement such as bias or differential item functioning (Smith, 1991b).

Smith (1982) found that the means and standard deviations of weighted and unweighted between fit statistics were almost identical with a correlation of .99. Both the BICAL and IPARM Rasch programs use the unweighted version of the between fit statistic, however, the IPARM program is less restrictive in the number of ability groups and number of persons per sample. Also, with IPARM, item invariance differences can be tested by group.

SIMULATION STUDIES

Several simulation studies have been conducted related to Rasch fit statistics. For example, Smith has examined (a) the robustness of fit statistics (Smith, 1985), (b) person fit (Smith, 1986), (c) the distributional properties of Rasch standardized residuals (Smith, 1988a), (d) power comparisons of Rasch total and between item fit statistics (Smith, 1988), (e) the distributional properties of Rasch item fit statistics (Smith, 1991a), and (f) separate versus between fit statistics in detecting item bias (Smith, 1993). These findings, however, were typically based on only 10 or 20 replications. It is therefore important to determine if the number of replications effect Rasch fit statistics. As a result, the purpose of this study is to investigate the differences in Rasch item and person between and total weighted and unweighted results based upon varying numbers of replications.

METHODS AND PROCEDURES

In this study, simulated data sets were used which varied in test length and sample size. Two test lengths (20 and 50 items) and three sample sizes (150, 500, and 1000 persons) were completely crossed for a total of six experiments. Each experiment was replicated 100 times.

The data sets were constructed using SIMTEST 2.1 (Luppescu, 1992). For each replication, person abilities were normally distributed and item difficulties were uniformly distributed. The BIGSTEPS Rasch calibration program was used to analyze each of the 600 data sets (Linacre, 1992). Next, the Rasch-calibrated item statistics were used in the IPARM program (Smith, 1986) to generate the (a) item weighted total fit statistic, (b) item unweighted total fit statistic, (c) item unweighted between fit statistic, (d) person weighted total fit statistic, and (e) person unweighted total fit statistic.

Summary statistics were computed for the means, standard deviations, and Type I error rates of each fit statistic for each experiment after 10, 25, 50, and 100 replications.

RESULTS

The data presented in the tables is organized to aid interpretation. The first column of numbers under the mean heading for 20 items indicates the average t value which has an expected value of 0. For example, the -.01, -.06, .01, and -.02 values have expected values of 0. The column of values underneath these are the standard deviations of the mean t values which have expected values of 1.0. For example, the .95, .94, .94, and .84 values have expected values of 1.0. The IPARM program uses n rather than n-1 in calculating the mean square values used in the t calculation, hence these values have not been corrected for degrees of freedom. The column of values next to the mean t values contains the standard deviations of the mean t values: .15, .24, .22, and .21.

The column of values underneath them contain the standard deviations of the mean t standard deviations: .11, .12, .14, and .13. These four columns are repeated for 500 persons and 1000 persons as well as for the columns listed under 50 items.

Tables 1,2, and 3 indicate item fit statistics. Tables 4 and 5 indicate person fit statistics. Tables 7 and 8 indicate the power analysis results for item unweighted and item weighted total fit statistics, respectively. Tables 9 and 10 indicate the power analysis of person unweighted and weighted total fit statistics. Values in these tables indicate the number of items falling in the extreme tails of the distribution compared to the expected number of items or persons. This is based upon a two-tailed hypothesis at the .05 level resulting in .025 percent expectation in each tail.

The item unweighted between fit statistics in Table 1 are affected by systematic differences due to bias or differential item functioning. The results in the table indicate that increasing the number of items, from 20 to 50, reduced the standard deviations of the mean t values and brought the values closer to what is expected (mean t = 0 and standard deviation = 1). The larger sample n resulted in a poorer fit as noted by the mean t values (.19, .22, .18, and .17) and standard deviations (.84, .87, .88, and .88), but this occurred because of the correction for degrees of freedom being n rather than n-1. There was little difference in the expected values as the result of the number of replications.

Table 2 indicates item unweighted total fit statistics (outfit) which are affected by random disturbances such as guessing and sloppiness in responding. These results are not as close as those indicated in Table 1. Dependencies between persons and items as well as the use of n rather than n-1 in calculating the mean square values affects these results. More items reflected a closer fit to the expected values and the number of replications within each column had little effect on the values reported.

Table 3 indicates the item weighted total fit statistics (infit). These values are also affected similar to those in Table 2. When adjustments are made, the values more closely approximate expected values. A correction factor computed as: (n * L)/(n - 1) * (L - 1), where n = # of persons and L = # of items, brings these expected values closer to a mean of 0 and a standard deviation of 1. The negative signs in the tables suggest that the results are conservative and should indicate a lower Type 1 error rate.

Table 1

Means and Standard Deviations of Item Unweighted Between Fit Statistics

	Number of Items			
	20		50	
	Mean	S.D.	Mean	S.D.
<u>150 Persons</u>				
Mean				
10 reps	-.01	.15	-.04	.15
25 reps	-.06	.24	-.01	.13
50 reps	.01	.22	.02	.14
100 reps	-.02	.21	.02	.14
S.D.				
10 reps	.95	.11	.97	.08
25 reps	.94	.12	.95	.07
50 reps	.94	.14	.97	.09
100 reps	.84	.13	.97	.09
<u>500 Persons</u>				
Mean				
10 reps	.06	.27	.07	.11
25 reps	.05	.25	.07	.12
50 reps	.05	.23	.04	.14
100 reps	.05	.22	.03	.14
S.D.				
10 reps	.90	.13	1.00	.13
25 reps	.93	.13	.97	.11
50 reps	.90	.14	.97	.11
100 reps	.92	.14	.97	.10
<u>1000 Persons</u>				
Mean				
10 reps	.19	.14	.02	.14
25 reps	.22	.18	.01	.14
50 reps	.18	.18	.03	.14
100 reps	.17	.18	.03	.13
S.D.				
10 reps	.84	.09	.98	.09
25 reps	.87	.13	1.00	.09
50 reps	.88	.12	.98	.09
100 reps	.88	.12	.97	.09

Table 2

Means and Standard Deviations of Item Unweighted Total Fit Statistics

	Number of Items			
	20		50	
	Mean	S.D.	Mean	S.D.
150 Persons				
Mean				
10 reps	-.17	.06	-.05	.09
25 reps	-.16	.08	-.05	.07
50 reps	-.18	.08	-.05	.07
100 reps	-.18	.09	-.05	.06
S.D.				
10 reps	.75	.14	.85	.06
25 reps	.78	.16	.85	.10
50 reps	.83	.17	.87	.10
100 reps	.84	.16	.87	.10
500 Persons				
Mean				
10 reps	-.29	.12	-.13	.04
25 reps	-.27	.11	-.13	.05
50 reps	-.31	.11	-.14	.05
100 reps	-.32	.10	-.14	.05
S.D.				
10 reps	.83	.15	.89	.08
25 reps	.86	.18	.91	.11
50 reps	.87	.16	.89	.11
100 reps	.86	.17	.88	.10
1000 Persons				
Mean				
10 reps	-.49	.14	-.20	.06
25 reps	-.48	.13	-.19	.05
50 reps	-.48	.11	-.19	.05
100 reps	-.48	.10	-.19	.06
S.D.				
10 reps	.79	.17	.89	.11
25 reps	.84	.17	.88	.10
50 reps	.83	.14	.88	.09
100 reps	.81	.14	.88	.10

Table 3

Means and Standard Deviations of Item Weighted Total Fit Statistics

	Number of Items			
	20		50	
	Mean	S.D.	Mean	S.D.
<u>150 Persons</u>				
Mean				
10 reps	-.23	.04	-.09	.03
25 reps	-.24	.04	-.09	.03
50 reps	-.23	.04	-.09	.03
100 reps	-.23	.04	-.09	.03
S.D.				
10 reps	.79	.10	.77	.07
25 reps	.79	.12	.77	.08
50 reps	.82	.14	.80	.08
100 reps	.83	.12	.80	.09
<u>500 Persons</u>				
Mean				
10 reps	-.45	.04	-.18	.02
25 reps	-.44	.04	-.18	.02
50 reps	-.44	.04	-.18	.02
100 reps	-.44	.05	-.17	.02
S.D.				
10 reps	.82	.18	.77	.08
25 reps	.80	.16	.79	.10
50 reps	.83	.15	.78	.10
100 reps	.82	.15	.78	.09
<u>1000 Persons</u>				
Mean				
10 reps	-.59	.03	-.23	.02
25 reps	-.5[illegible]	.04	-.24	.02
50 reps	-.59	.04	-.24	.02
100 reps	-.60	.04	-.24	.02
S.D.				
10 reps	.76	.14	.84	.13
25 reps	.80	.14	.81	.11
50 reps	.81	.13	.80	.10
100 reps	.81	.13	.79	.09

The person unweighted and weighted total fit statistics are reported in Tables 4 and 5. Table 4 values indicate a good fit with no difference in the number of replications across the experimental conditions specified in the study. These "outfit" expected values were less than the "infit" expected values reported in Table 5. For example, compare -.03, -.04, -.04, and -.04 (Table 4) versus -.07, -.07, -.07, and -.07 (Table 5) for 20 items. Notice that values across the number of replications were almost identical suggesting that the number of replications does not affect the fit statistics.

Table 4

Means and Standard Deviations of Person Unweighted Total Fit Statistics

	Number of Items			
	20		50	
	Mean	S.D.	Mean	S.D.
150 Persons				
Mean				
10 reps	-.03	.02	-.04	.02
25 reps	-.04	.02	-.03	.02
50 reps	-.04	.02	-.04	.02
100 reps	-.04	.02	-.04	.02
S.D.				
10 reps	.89	.04	.94	.07
25 reps	.89	.05	.92	.05
50 reps	.89	.05	.91	.05
100 reps	.89	.05	.92	.05
500 Persons				
Mean				
10 reps	-.04	.01	-.03	.01
25 reps	-.04	.01	-.03	.01
50 reps	-.04	.01	-.04	.01
100 reps	-.03	.01	-.04	.01
S.D.				
10 reps	.89	.02	.93	.03
25 reps	.89	.03	.92	.03
50 reps	.89	.03	.92	.03
100 reps	.89	.03	.92	.03
1000 Persons				
Mean				
10 reps	-.03	.01	-.04	.01
25 reps	-.04	.01	-.03	.01
50 reps	-.04	.01	-.03	.01
100 reps	-.04	.01	-.03	.01
S.D.				
10 reps	.88	.02	.88	.02
25 reps	.89	.02	.89	.02
50 reps	.89	.02	.89	.02
100 reps	.89	.02	.89	.02

Table 5

Means and Standard Deviations of Person Weighted Total Fit Statistics

	Number of Items			
	20		50	
	Mean	S.D.	Mean	S.D.
150 Persons				
Mean				
10 reps	-.07	.02	-.05	.02
25 reps	-.07	.02	-.05	.02
50 reps	-.07	.02	-.05	.02
100 reps	-.07	.02	-.05	.02
S.D.				
10 reps	.89	.05	.91	.05
25 reps	.89	.05	.90	.04
50 reps	.90	.06	.89	.05
100 reps	.90	.05	.90	.05
500 Persons				
Mean				
10 reps	-.07	.01	-.05	.01
25 reps	-.07	.01	-.05	.01
50 reps	-.07	.01	-.05	.01
100 reps	-.07	.01	-.04	.01
S.D.				
10 reps	.89	.03	.92	.04
25 reps	.90	.03	.91	.03
50 reps	.89	.03	.90	.03
100 reps	.89	.03	.90	.03
1000 Persons				
Mean				
10 reps	-.06	.01	-.04	.00
25 reps	-.06	.01	-.04	.01
50 reps	-.06	.01	-.04	.01
100 reps	-.06	.01	-.04	.01
S.D.				
10 reps	.89	.02	.88	.02
25 reps	.90	.02	.89	.02
50 reps	.89	.02	.89	.02
100 reps	.89	.02	.89	.02

POWER ANALYSIS

Tables 6, 7, and 8 indicate the number of items falling in the extreme tails of a two-tailed distribution for item unweighted between fit statistics, item unweighted total fit statistics, and item weighted total fit statistics. Table 6 reflects power results based upon Table 1 results, similarly, Table 7 reflects power results from Table 2 and Table 8 reflects power results from Table 3. The number of items expected in each tail can be easily computed. For example, given a two-tailed test at the .05 level, one would expect .025 percent of the items to fall in each tail (t = +/- 2.00). Therefore, .025 times 20 items taken over 10 replications yields 5 items in each tail. For the other replications listed, the number of items expected would be 12.5; 25; and 50, respectively. These findings can be interpreted by comparing the actual number of items reported versus the expected number of items. For example, given 20 items and 100 replications, one would expect 50 items in each tail; 30 items were indicated yielding an alpha of .03 (60/200).

The difference in these percents across the number of replications was negligible. For example, given 20 items and 150 persons, the percents ranged from .01 (10 replications) to .03 (100 replications). In comparing the number of items indicated in these tables, the Type I error rate is closer to the expected values for the item between fit statistics in Table 6, less item bias is present with more items, and there is no substantial difference over replications.

Table 6

Frequency of Extreme Values for Item Unweighted Between Fit Statistics

	Number of Items			
	20		50	
	t>+2	t<-2	t>+2	t<-2
<u>150 Persons</u>				
10 reps	1	1	7	11
25 reps	4	7	21	18
50 reps	18	12	57	38
100 reps	30	30	115	75
<u>500 Persons</u>				
10 reps	4	2	15	5
25 reps	11	6	39	13
50 reps	20	7	78	37
100 reps	42	16	133	77
<u>1000 Persons</u>				
10 reps	2	1	9	12
25 reps	13	2	29	26
50 reps	16	3	58	44
100 reps	35	11	120	73

Note: The number of items expected in the tail for 20 items across 4 replications is: 5; 12.5; 25; and 50. For 50 items: 12.5; 31; 62.5; 125. For example .025 x 20 items x 10 replications = 5 items.

Table 7

Frequency of Extreme Values for Item Unweighted Total Fit Statistics

	Number of Items			
	20		50	
	t>+2	t<-2	t>+2	t<-2
150 Persons				
10 reps	0	0	10	1
25 reps	5	0	26	3
50 reps	15	1	47	12
100 reps	35	7	92	23
500 Persons				
10 reps	3	1	6	2
25 reps	7	7	19	10
50 reps	11	22	36	25
100 reps	21	35	70	42
1000 Persons				
10 reps	1	2	4	9
25 reps	2	13	11	21
50 reps	4	27	23	37
100 reps	7	40	47	68

Note: The number of items expected in the tail for 20 items across 4 replications is: 5; 12.5; 25; and 50. For 50 items: 12.5; 31; 62.5; 125. For example .025 x 20 items x 10 replications = 5 items.

Table 8

Frequency of Extreme Values for Item Weighted Total Fit Statistics

	Number of Items			
	20		50	
	t>+2	t<-2	t>+2	t<-2
150 Persons				
10 reps	0	2	3	2
25 reps	1	5	6	6
50 reps	6	10	19	21
100 reps	12	25	42	40
500 Persons				
10 reps	0	6	1	2
25 reps	1	14	7	11
50 reps	3	34	15	25
100 reps	6	65	28	45
1000 Persons				
10 reps	0	6	3	9
25 reps	0	18	7	16
50 reps	1	38	13	37
100 reps	3	63	24	65

Note: The number of items expected in the tail for 20 items across 4 replications is: 5; 12.5; 25; and 50. For 50 items: 12.5; 31; 62.5; 125. For example .025 x 20 items x 10 replications = 5 items.

The power results for person unweighted and weighted total fit statisics are in Tables 9 and 10, respectively. The number of persons in the tails were less than expected and in the direction expected.

Table 9

Frequency of Extreme Values for Person Unweighted Total Fit Statistics

	Number of Items			
	20		50	
	t>+2	t<-2	t>+2	t<-2
<u>150 Persons</u>				
10 reps	23	12	30	18
25 reps	62	26	61	37
50 reps	141	45	118	79
100 reps	264	96	242	163
<u>500 Persons</u>				
10 reps	89	38	75	80
25 reps	230	91	200	164
50 reps	430	189	403	335
100 reps	854	381	819	639
<u>1000 Persons</u>				
10 reps	158	80	143	109
25 reps	418	223	372	284
50 reps	800	423	761	579
100 reps	1593	815	1535	1160

Note: The number of items expected in the tail for 150 persons across 4 replications is: 37.5; 94; 187; and 375. For 500 persons: 125; 313; 625; and 1250. For 1000 persons: 250; 625; 1250; and 2500. For example: 150 persons x 10 replications x .025 = 37.5.

Table 10

Frequency of Extreme Values for Person Weighted Total Fit Statistics

	Number of Items			
	20		50	
	t>+2	t<-2	t>+2	t<-2
150 Persons				
10 reps	11	25	18	31
25 reps	33	61	40	69
50 reps	79	123	88	129
100 reps	170	251	180	254
500 Persons				
10 reps	62	99	79	114
25 reps	150	243	164	241
50 reps	255	474	332	466
100 reps	516	942	625	892
1000 Persons				
10 reps	97	188	106	179
25 reps	281	480	291	430
50 reps	535	951	573	856
100 reps	1097	1862	1150	1684

Note: The number of items expected in the tail for 150 persons across 4 replications is: 37.5; 94; 187; and 375. For 500 persons: 125; 313; 625; and 1250. For 1000 persons: 250; 625; 1250; and 2500. For example: 150 persons x 10 replications x .025 = 37.5.

SUMMARY

Rasch between and total weighted and unweighted fit statistics were compared using varying test lengths and sample sizes. Two test lengths (20 and 50 items) and three sample sizes (150, 500, and 1000) were crossed. Each of the six combinations were replicated 100 times. In addition, power comparisons were made.

Results indicated that there were no differences in item and person Rasch fit statistics based on the number of replications. The Type I error rates were close to expected values. It was concluded that the number of replications in Rasch simulation studies are not a major factor influencing fit. Researchers should be more sensitive to the number of persons, number of items, and the correction for degrees of freedom used in the mean square calculation (n vs. n-1).

REFERENCES

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

Linacre, J.M. (1992). *BIGSTEPS. A Rasch test calibration computer program.* MESA Press, Chicago, IL.

Luppescu, S. (1992). *SIMTEST 2.1 A computer program for simulating test data.* MESA Press, Chicago, IL.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. Copenhagen, Denmark.

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

Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433-444.

Smith, R.M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46*, 359-372.

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

Smith, R.M. (1988b, April). *A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances.* A paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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*. MESA Press: Chicago, IL.

Smith, R.M. (1993, April). *A comparison of the Rasch separate calibration and between fit methods of detecting item bias.* A paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

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