THE ESTIMATION OF SCALED SCORES AND THEIR FREQUENCY DISTRIBUTIONS FROM ITEM BANK DIFFICULTY VALUES

Jacob G. Beard
&
Lucinda Richards
Florida State University

Thomas Fisher
Florida Department of Education

INTRODUCTION

The purpose of this study was to investigate the application of Rasch "pre-equating" techniques for solving two problems inherent in large scale testing programs. The first problem dealt with the estimation of raw score to scaled score transformations from item difficulty values existing before a test is administered. Such estimates are theoretically possible using Rasch methodology. This use of pre-equating would permit the generation of a raw score to scaled score transformation table as soon as the items to be included in a test have been identified and before the test is administered. The alternative to this procedure is to perform an equating analysis after the test has been administered and scored, necessitating a delay in returning the test scores to the students. The purpose of this part of the study was to compare the score transformations generated through pre-equating with those done after the test was administered, scored, and equated to the reference scale.

The second problem dealt with the estimation of frequency distributions of scores from pre-administration item difficulty values. If such estimations prove to be accurate they would facilitate the setting of new passing scores. Otherwise, when a passing score or standard is to be set for a new or revised test, it would be necessary to administer the test and compute the frequency distributions of scores in order to assess the impact of proposed passing scores. Again, this would substantially delay the reporting of scores to students.

Pre-equating is a procedure for equating a new test to another test, or to a reference scale, prior to the actual administration of the new test. The procedure used in this study was derived from the Rasch model. The procedure requires that the new form be constructed using items from a calibrated item bank, or item pool. In pre-equating, item calibrations, or adjusted difficulty values, are used to compute raw score to scaled score transformation tables.

The procedures used in post-equating with the Rasch model are very similar to those used in pre-equating, the difference being that instead of using pre-calibrated difficulty values as input, the actual current administration data are used to calculate adjusted item difficulty values which are in turn used in calculating ability scores. As the procedures used for the two techniques are the same except for the item difficulty values used, the results obtained should be consistent, depending on the similarity of the two sets of item difficulty values. However, because pre-equating can be performed prior to the actual administration of a new form of a test, it offers a major advantage over the traditional post-equating.

The solutions to these two problems are greatly facilitated by using the Rasch model, and illustrate two of the several advantages in using item response theory in a large-scale testing program.

In summary, the purpose of this study was to compare: (a) raw score to scaled score transformation tables generated by pre-equating and by post-equating procedures, and (b) frequency distributions of scores estimated through pre-equating procedures with those resulting from an actual administration.

PROCEDURES

Pre-Equated Scaled Scores

For the pre-equating, the items contained on the April 1982 Florida Student Assessment Test, Part II, (SSAT-II) were identified. Their pre-existing difficulty values from the item bank, adjusted to the 1978 scale, were located. These adjusted item difficulty values were used as input into an adaptation of the BICAL (Wright and Mead, 1978) computer program. This program used the Rasch model to calculate equated ability logits for each possible raw score. The log ability scores were then transformed to SSAT-II Scaled Scores by the following formula.

$$y = 25(b_i - b_c) + 700$$

Where

y = SSAT-II scaled score

b_i = equated logit corresponding to a students' raw score (adjusted to the 1978 scale)

b_c = logit corresponding to the minimum passing raw score in 1978 (42 out of 60)

The use of this transformation formula provided a raw score to scaled score transformation table, in which a scaled score of 700 was equivalent to the minimum passing score of 42 out of 60 items correct on the 1978 test. The scaled scores have an approximate range of 550 to 800.

For the comparison post-equating procedures, the data from the actual administration of the April 1982 SSAT-II were used to determine raw score to scaled score transformation tables.

Estimation of Frequency Distributions

In order to generate the estimated frequency distributions, the following procedures were used. From the pre-equating results, raw scores and their corresponding scaled scores were located. Using results from a previous, April 1981, administration of the SSAT-II, corresponding scaled scores and their associated raw scores were also located. Using these raw scores from the April, 1981, SSAT-II, the proportion of students achieving each score was identified. This was the estimated proportion for the corresponding raw score in the 1982 test. (Different estimation procedures would be required if the numbers of items in the predictor and predicted tests were different, or if the distributions were expected to differ in shape.)

Because the 1982 and 1981 scaled scores have been equated, and the actual ability/achievement distributions for the two years was assumed to be highly similar, the estimation of the frequency (proportion) distributions of raw scores for the 1982 test was possible as soon as it had been determined which items from the bank would be included in it, and before it was administered to the students.

This process of estimating proportions was performed for the whole range of raw scores. In cases where the scaled scores could not be matched exactly, ability logits were used for interpolation. These estimated proportions were then compared with the actual proportions from the 1982 administration of the test.

RESULTS

Pre-equated Scaled Scores

A comparison of the scaled scores resulting from pre-equating and post-equating can be found in Tables 1 and 2.

These tables show the results for both the mathematics and communications forms of the April, 1982, SSAT-II. The results from pre-equating are similar, but not identical to those produced from post-equating. The scaled scores produced from the pre-equating procedures are systematically larger than those produced from the post-equating procedures for both mathematics and communications sub-tests. In all cases a student having a smaller raw score would pass the test using the pre-equating raw score to scaled score transformation table. These differences are generally small, but their systematic characteristic led to the investigation of the possible cause of the differences.

The statistical procedures involved in pre-equating and post-equating are, in fact, identical except for the item difficulty values used. Accordingly, an examination of the item difficulty values used as input for the pre-equating and post-equating was made. The mean of the post-administration adjusted item difficulty values was smaller than the mean of the pre-administration values for both mathematics and communications. This resulted in a difference between the pre-equated and post-equated score scales so that the post-equated score scale made the tests easier than if the pre-equated score scale had been used. Conversely, using a score scale derived from the pre-equated difficulty values would have resulted in a harder test and more failures.

In order to investigate the instability of item difficulty values, the calibration histories of the items were explored. It was known that items included on the April, 1981, and April, 1982, SSAT-II had been calibrated in various ways and at different times. A large number had been calibrated recently, but some had not been calibrated in a number of years. Some of the items had never been included on full forms of the SSAT-II but had only been included on experimental forms during past administrations of the SSAT-II. Those items that had been calibrated only on data from the administration of experimental forms

Table 1

Pre- and Post-Equated Raw to Scaled Score Transformations: Mathematics

RAW SCORES	SCALED SCORES PRE-EQUATED	POST-EQUATED	RAW SCORES	SCALED SCORES PRE-EQUATED	POST-EQUATED
59	795	791	19	648	647
58	776	772	18	646	644
57	765	760	17	643	642
56	756	752	16	640	639
55	749	745	15	637	636
54	744	739	14	634	633
53	739	734	13	631	630
52	734	730	12	628	627
51	730	726	11	625	624
50	726	722	10	621	620
49	723	719	9	617	616
48	719	715	8	613	612
47	716	712	7	608	608
46	713	709	6	603	602
45	710	707	5	597	597
44	708	704	4	590	590
43	705	701	3	581	581
42	702	699	2	570	570
41	700	696	1	551	551
40	697	694			
39	695	692			
38	693	689			
37	690	687			
36	688	685			
35	686	683			
34	683	680			
33	681	678			
32	679	676			
31	677	674			
30	674	672			
29	672	670			
28	670	667			
27	668	665			
26	665	663			
25	663	661			
24	661	658			
23	658	656			
22	656	654			
21	653	652			
20	651	649			

Table 2

Pre- and Post-Equated Raw to Scaled Score Transformations: Communications

RAW SCORES	SCALED SCORES PRE-EQUATED	POST-EQUATED	RAW SCORES	SCALED SCORES PRE-EQUATED	POST-EQUATED
59	784	779	19	652	650
58	766	761	18	649	648
57	756	750	17	647	646
56	748	743	16	645	643
55	741	736	15	642	641
54	736	731	14	640	639
53	731	727	13	637	636
52	727	723	12	634	633
51	724	719	11	632	630
50	720	716	10	628	627
49	717	712	9	625	624
48	714	709	8	621	620
47	711	707	7	617	617
46	708	704	6	613	612
45	706	702	5	608	607
44	703	699	4	601	601
43	701	697	3	594	593
42	698	695	2	583	582
41	696	693	1	565	565
40	694	690			
39	692	688			
38	690	686			
37	688	684			
36	686	682			
35	684	681			
34	682	679			
33	680	677			
32	678	675			
31	676	673			
30	674	671			
29	672	669			
28	670	667			
27	668	666			
26	666	664			
25	664	662			
24	662	660			
23	660	658			
22	658	656			
21	656	654			
20	654	652			

exhibited a mean decrease in difficulty between the pre- and post administration item difficulty values. These results suggest that improved pre-equating accuracy could be achieved through improvement of the item bank difficulty estimates.

Estimated Frequency Distributions

The results of the frequency distribution estimation for the April, 1982, SSAT-II may be found in Table 3. In this table the actual distributions from the 1982 administration are shown along with the estimated distributions. The results from the pre-equating generally provided good estimates of the actual distributions. Table 3 shows very similar proportions of students achieving the various score levels.

Table 4 shows the accuracy of the use of pre-equating, as compared to post-equating, in estimating the number of failures for different passing scores. The results of the estimation procedure approximate the actual proportions, especially for scores near the current passing raw score of 42.

CONCLUSIONS

The purpose of this paper was to compare scaled scores, frequency distributions, and cumulative proportion distributions generated through pre-equating and post-equating procedures. It was found that scaled scores generated through pre-equating were generally slightly larger for given raw scores than those generated through post-equating. This small but systematic difference appears to be caused by a trend toward the overestimation of item difficulties on experimental test forms, relative to those of actual test administrations.

The frequency distributions and cumulative proportion distributions estimated through pre-equating procedures were quite consistent with those from the actual administrations.

Table 3

Estimated and Actual

Percentage Frequency Distributions

1982 SSAT-II

Communications			Mathematics		
raw score	estimated	actual	raw score	estimated	actual
59	24	23	59	05	07
58	19	19	58	06	07
57	14	14	57	06	07
56	09	10	56	06	07
55	07	06	55	06	05
54	05	05	54	05	05
53	04	03	53	05	04
52	03	03	52	05	05
51	02	02	51	04	05
50	02	03	50	04	04
49	01	02	49	04	03
48	01	02	48	04	03
47	01	01	47	04	03
46	01	01	46	03	03
45	01	01	45	03	03
44	01	<01	44	03	03
43	01	01	43	03	03
42	<01	<01	42	03	02
41	**[a]	**	41	03	03
40	**	**	40	02	02
39	01	**	39	02	01
38	<01	**	38	02	02
37	**	01	37	01	02
36	**	**	36	01	02
35	**	01	35	01	01
34	**	<01	34	01	01
33	**	**	33	01	01
32	**	**	32	01	01
31	**	**	31	01	01
30	**	**	30	01	01
29	**	**	29	01	01
28	**	**	28	<01	<01
27	**	**	27	01	**
26	**	**	26	<01	01
25	**	**	25	**	<01

[a] ** indicates a percentage distribution less than 01.

Table 4

Cumulative Proportion Failing at Various Passing Scores

1982 SSAT-II

Communications			Mathematics		
raw score	estimated	actual	raw score	estimated	actual
54	.192	.190	54	.636	.589
53	.157	.163	53	.587	.553
52	.133	.142	52	.540	.508
51	.115	.124	51	.500	.463
50	.100	.102	50	.457	.423
49	.088	.087	49	.418	.395
48	.078	.074	48	.380	.362
47	.069	.063	47	.345	.330
46	.063	.057	46	.312	.303
45	.057	.051	45	.282	.275
44	.051	.047	44	.253	.247
43*	.046	.040	43*	.225	.217
42	.043	.039	42	.197	.198
41	.039	.035	41	.172	.174
40	.036	.032	40	.148	.155
39	.031	.028	39	.130	.141
38	.029	.026	38	.115	.118
37	.027	.021	37	.102	.103
36	.025	.020	36	.088	.088

* current passing score

It was concluded that pre-equating can be used for establishing raw to scaled score transformations provided that the differences in scale score values of the magnitude shown in Tables 1 and 2 could be tolerated. These differences in estimated and actual values could be further decreased by improving the estimation of item bank difficulty values.

It was also concluded that the frequency distributions which were estimated through pre-equating procedures were sufficiently similar to the actual distributions to permit their use in examining the impact of alternative passing scores. Additional research is needed to determine how accurate the estimations would be when the distributions of raw scores on which the estimations are based vary substantially from those to be estimated.

REFERENCES

Wright, B.D. and Mead, R.J. Research Memorandum No. 23A. Chicago, Ill.: University of Chicago, Department of Education, Statistical Laboratory, 1978.