An Application of Rasch Model Technology to a Large-City Testing Program

Joanne M. Lenke, John Oswald
The Psychological Corporation
and
Gary Kippel
New York City Public Schools

The purpose of this study was to investigate whether or not a customized version of a norm-referenced test could be developed for a large-city school system to measure the objectives of that school system and still provide norm-referenced information. It was determined that the best approach to the task (aside from collecting national normative data on the customized test) was to use Item Response Theory--more specifically, the Rasch Model. Item Response Theory offers a solution to the whole question of school systems' desire for customization and at the same time, their need for norm-referenced information because (1) when a latent trait model fits a particular set of items, it is possible to estimate an examinee's ability on the same ability scale from any subset of items (Wright, 1977); and (2) an ability scale to which a normed test has been referenced can be useful in providing norms for tests that have been constructed from the pool of items in the normed test (Hambleton, 1980). The Rasch Model, in particular, was appealing because of its provision for simple number-right scoring. This paper reports the method and results for the first year of a customization project.

The new York City Public Schools conducts city-wide testing programs in reading and mathematics in grades 2 through 8 in the spring of each year. After studying the City's testing needs in mathematics for the 1980-81 school year and beyond, the Office of Testing determined that the desired test should:

1. measure the New York City mathematics curriculum, as defined by their *Minimum Standards of Pupil Performance;*

2. provide diagnostic information in mathematics for each student; and

3. describe performance in terms of national norms.

After committee review, the City chose the *Stanford Diagnostic Mathematics Test* (SDMT) because it satisfied two of their three expressed needs: diagnostic utility and empirical national norms. The school district's Mathematics Unit and the test's publisher, The Psychological Corporation, matched the test's content, item by item, with the City's Minimum Standards. For objectives not measured by SDMT, additional items were written.

A plan was developed to obtain Rasch item difficulty estimates (calibrations) for the additional "supplementary" items and for the SDMT items using the item response data of random samples from the New York City student population. (The SDMT items had also already been calibrated on the SDMT fall standardization sample.) Theoretically, once the supplementary items are combined with the SDMT items in a Rasch-calibrated item pool, a new form of SDMT, i.e., a form that matches the New York City Minimum Standards better than the original SDMT, can be designed and will yield national norm-referenced information. The administration of this new form will be the next step in the research.

both sets of data indicate that, even though SDMT was built using classical procedures, fit of the items to the Rasch Model is adequate for the purposes outlined above. The number of non-fitting items in common across both samples is also interesting. At the Red, Green, and Brown levels, most, if not all, of the items identified as non-fitting for the New York City sample were also identified as non-fitting for the standardization sample. Curiously, this was not the case at the Blue Level.

Table 1

Analysis of Fit[1] to the Rasch Model for the SDMT Grades Combined Fall Standardization Sample and the NYC Grade Samples

	Number and Percent of Non-fitting Items					Number of Non-fitting Items in Common Across Samples
	Standardization Sample		New York City Sample			
	N	%		N	%	
Red Level 93 items	5	5	Gr 2	3	3	3
Green Level 114 items	9	8	Gr 3	2	2	2
			Gr 4	4	4	4
Brown Level 117 items	7	6	Gr 5	5	4	3
			Gr 6	4	3	3
Blue Level 117 items	10	9	Gr 7	2	2	0
			Gr 8	3	3	0

[1] Criterion for fit: $MS^{*}_{fit} \geq 1.3$ and Slope < 0.50.

Table 2 shows the correlation between corresponding Rasch ability estimates and SDMT scaled scores, for every possible raw score (except the maximum and those corresponding to a national percentile rank of less than 1). Obviously, if the relationship between the Rasch ability estimates and scaled scores were linear and perfect, the correlation would be 1.00. As can be seen in Table 2, the correlations for each of the two samples range from 0.991 for grade 2 (Red Level) to 0.999 for grades 5 and 6 (Brown Level). These correlations indicate that Rasch ability estimates generated on the basis of items calibrated on very different samples (national vs. large city) tested at opposite ends of the school year (fall vs spring) are highly consistent. The correlations also suggest that the relationship between Rasch ability estimates and the test's scaled score system is sufficiently high to conclude that SDMT scaled scores (and hence national norms) can be predicted adequately from Rasch ability estimates when the test has been calibrated as a subset of a larger pool of items.

Method

As part of the 1981 city-wide spring testing program, New York City administered Form A of SDMT intact, plus one of up to four "forms" of 20 supplementary items to each student in grades 2 through 8 tested (approximately 500,000). Only scores on SDMT itself were reported to schools.

Using a random sample of 2,000 students per grade per supplementary "form," each SDMT plus supplementary item "form" was analyzed as one long test by means of Rasch Model and classical item analysis techniques. Output included traditional item statistics (p-values, biserial and point biserial correlation coefficients) and Rasch Model item difficulties, standard errors, and, for analysis of fit, the mean square fit statistic and slope of the item characteristic curve. It is the Rasch item calibrations that are used to produce the table of raw scores to Rasch ability estimates (Rentz & Bashaw, 1975). Therefore, in order to determine the adequacy of the Rasch Model for estimating national norms for a test that is comprised of items from a calibrated pool, two types of investigation were conducted:

1. A comparison of fit of the SDMT items to the Rasch Model for the SDMT standardization group and for the New York City sample. (Since SDMT was built using classical procedures and the quality of the estimated norms depends on the fit of the items in the test to the latent trait model used to construct the ability scale, examination of fit of the Rasch Model to both sets of data was critical.) In order to evaluate fit of the items to the model, a MS^*_{fit} statistic (Rentz & Rentz) was calculated for each item for each of the two samples. The number and percent of non-fitting items were identified using the criterion $MS^*_{fit} \geq 1.3$ and slope < 0.5.

2. Examination of the relationship between the Rasch scale of ability (derived on the basis of the New York City samples) and the scaled score system developed for SDMT. (The norms for SDMT are scaled-score based.) In order to determine whether the Rasch ability estimates derived from the item calibrations estimated from the New York City samples were monotonically and linearly related to the SDMT scaled score system, Pearson product-moment correlation coefficients were calculated between Rasch ability estimates and the SDMT scaled scores.

Recall that all items in a given New York City test "form (SDMT + supplementary items) were calibrated together to yield a Rasch scale of item difficulties. Item calibrations derived from the performance of the New York City samples on the SDMT items only were used to generate an SDMT raw score to Rasch ability score, by grade. For comparative purposes, item calibrations and Rasch ability estimates were also obtained for each level of SDMT, based on the performance of the grades-combined samples (approximately 2,000 students per sample) from the SDMT fall standardization program.

Results

Table 1 presents the analysis of the fit of SDMT items to the Rasch Model for both samples. Overall fit of the Model to the data is good, since greater than 90% of the items on any level at all grades fit the Model, according to the established criteria. It is interesting to note that the New York City data demonstrated better fit to the Model than the national standardization sample data, perhaps due to the fact that the New York City data are grade based, while the standardization data are based on grades-combined samples. Nevertheless,

Table 2

Correlations Between Rasch Ability Estimates and Thurstone Scaled Scores for Stanford Diagnostic Mathematics Test Total Scores [2]

	Standardization Sample	New York City Sample	
Red Level 93 items	.992	Grade 2	.991
Green Level 114 items	.998	Grade 3 Grade 4	.998 .997
Brown Level 117 items	.999	Grade 5 Grade 6	.999 .999
Blue Level 117 items	.996	Grade 7 Grade 8	.996 .996

[2] Scores corresponding to percentile ranks of less than 1 were eliminated from the analysis.

Discussion

The extent to which the SDMT scaled scores can be predicted from the ability estimates derived through the application of the Rasch Model to the New York City data is an indication of the appropriateness of this approach to the estimation of national norms for tests comprised of items from a calibrated item pool. If the items in the New York City item pool demonstrate adequate fit to the Rasch Model, and the Rasch ability estimates are linearly related to the scaled score system developed for SDMT, the necessary (though not sufficient) conditions exist to suggest that performance on tests made up of items from this pool can be interpreted in relation to national norms.

The results of this study indicate that (1) the SDMT items do fit the Rasch Model and (2) the Rasch ability estimates for SDMT are linearly related to its scaled score system for both the New York City sample and for the SDMT standardization sample. The implications of these results for practitioners are important. Many school districts use nationally normed tests for overall program evaluation of a "national core curriculum." That is, they expect standardized tests to tell them how their students are doing on content that is com-

monly taught nation-wide. Some school districts have also used criterion-referenced tests, often constructed locally, to evaluate performance on their local curriculum. It is a commonly accepted notion that standardized tests seldom, if ever, match the local curriculum well enough for school administrators to answer the question, "How well are our students learning what *we* are teaching them." As testing budgets become tighter and the demands for more relevant test data increase, more school districts are turning to methods, such as the use of Item Response Theory, to customize existing tests. The net result of these efforts will be more comprehensive testing programs at less overall cost.

References

Hambleton, R. K., Latent ability scales: Interpretations and uses. New Directions for Testing and Measurement, 1980, 6, 73-97.

Rentz, R. R. & Bashaw, W. L., Equating Reading Tests with the Rasch Model, Volume I Final Report, Volume II Technical Reference Tables. Athens, Ga.: University of Georgia, Educational Research Laboratory, 1975. ED 127 330 through ED 127 331.

Rentz, R. R. & Rentz, C. C., Does the Rasch Model really work? A discussion for practitioners. ERIC Clearinghouse on Tests, Measurement, and Evaluation, Report 67, 1978.

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