Chapter V

Summary and Implications

Summary.

The purpose of this study was to examine the calibration efficacy of the one-, two- and three-parameter logistic models to the DRP through descriptive methods and using residual analysis to corroborate results. This involved testing the underlying assumptions of the three models to the DRP.

Both principal components and common factor analysis were used to evaluate the unidimensionality assumption. The number of components extracted from the matrix of phi correlations provided the most stringent evidence of unidimensionality. Approximately 20% of the variance is accounted for by the first component. This result satisfies the traditionally used criterion of unidimensionality (Reckase, 1979).

The degree of speededness was based using the Swineford/ETS measures of speededness. Virtually all examinees reached at least three-quarters of the items and all of the items are reached by more than 90% of the examinees. The DRP may be considered essentially unspeeded. Although no major problem of speededness was uncovered, the ETS criteria are general and do not account for some examinees who might rapidly answer items in the hopes of getting some answers right by chance. As expected, low ability students tended to guess on harder items. However, this behavior was not prevalent enough to consider the DRP speeded.

If items are truly uniform in discrimination, two and three-parameter BILOG calibrations can expect to find leptokurtic distributions of discrimination indices. Histograms of these parameters reveal distributions failing to be leptokurtic to the degree sufficient to demonstrate uniformity in discrimination. The two-parameter BILOG calibration resulted in 34% of the items whose upper and lower discrimination parameters fell outside the Keifer limits of .8 and 1.2, while 70% of the three-parameter calibrated discrimination indices drew confidence intervals inconsistent with the assumption that respective discrimination indices are one. In addition, the use of residual analysis aids in accentuating the failure of the Rasch model to fit low and highly discriminating items. The curvilinear relationship existing between item discrimination and the one-parameter model averaged absolute-value standardized residuals suggests that a model that takes varying discriminations into account better fits the test data. Equal discrimination was also evaluated through the examination of the item-total score biserial correlations. Finding more than 36% of discrimination indices as measured by the biserial correlations falling outside of the mean biserial correlation contradicts the Hambleton index of equal discrimination indices.

If the difference between mean item difficulty and difficulty adjusted for guessing is zero, examinees are obtaining correct answers through appropriately considering each item. This difference which ranged from .019 for the most able students to .142 for the least able students provides the first indication that a lower asymptote may not be zero for all items. If the lower asymptote values obtained from the three-parameter BILOG calibration are close to zero, then there is no need for a lower asymptote. The probability correctly responding to an item on the DRP is .2 as each item consists of five alternatives. The three-parameter BILOG calibration estimated more than 30% of the items to have lower asymptotes greater than .2 indicating that it is likely that an examinee would provide a correct response to some items by guessing. Lower asymptote values ranged from .11 to .35 and have relatively small standard errors. Considering this and the large range of these values, it is likely that lower asymptotes are appropriate. Based on the D'Costa Index, inconsistent response patterns, which can be thought of as guessing, is prevalent among one-third of the examinee population.

Using CTT it is difficult to estimate an examinee’s ability when a test is extremely difficult or very easy. When test data fits an IRT model, estimates of ability are comparable no matter what set of test items are administered. To assess the equivalency of one-, two-, and three-parameter estimates of ability, ability was estimated for each examinee twice, on the easiest 38 and the hardest 38 items. After extensive sample elimination, the correlation of ability estimates obtained between these halves of the DRP ranged from .82 for the one-parameter model to approximately .85 for the two- and three-parameter models. This result means that there is evidence to suggest that the estimate of ability does not depend on the set of items chosen for calibration. The test standard error function can be used to determine the accuracy of the ability estimate across its distribution. The comparison of error functions for the one-, two- and three-parameter models found the three-parameter model to make the most error-free in that it provides the best estimates of ability across the distribution of ability.

A feature of IRT models is the ability to provide consistent estimates of item parameters regardless of the population of examinees tested. If the correlation between the item parameters obtained from groups of examinees expected to perform much differently is high and the correlation of their difference is zero, invariance is established. To be able to compare all three models, b-values of randomly equivalent groups of low and high achieving examinees are correlated. All models had very high correlations between b-values. No significant differences were found between these correlations. However, only the two- and three-parameter models had near zero correlations of b-value differences establishing the notion that test items are being calibrated at similar difficulty levels for these models. The correlation of b-value differences for the one-parameter model was close to one indicating the invariance of item parameter estimates to be implausible for this model.

The frequency of misfit items and the analysis of residuals was used to assess the relative fit of the three models to the DRP. Both these approaches provided evidence of the lack of fit of the Rasch model to the DRP. The difference between theoretically obtained proportions correct to the observed proportion correct were so vast that the great majority of one-parameter residuals are considered outliers. The number of misfitting items is, as expected, overwhelming for this model. The plots of standardized residuals against ability reveal that the one-parameter model provides the least precise estimates of performance, especially for the endpoints of the ability distribution. The two- and three-parameter models provided better estimates across the ability continuum. Plots of one-parameter standardized residuals against classical item difficulty discrimination found hard items to be associated with high residuals. This phenomenon, possibly due to guessing, does not occur when the two- and three-parameter models are fit to the data.

The two- and three-parameter models fit the test data equally as well, suggesting that with the DRP, varying item discriminations are more important than guessing when it comes to model fit. However, since the TIF for these two models are not identical, only one of these models provides an adequate fit. It has been shown that 95% of the three-parameter model predictions are reasonable whereas 84% of predictions are reasonable for the two-parameter model. Target test information functions should be flat if the need is to produce a test that will provide approximately equally precise ability estimates across the range of ability. Based on the overwhelming number of excellent predictions and a TIF that fulfills the test developer's intent to produce a wide-range ability test, the least restrictive three-parameter logistic model provides the most appropriate fit to DRP test data.

Practical Implications of Results.

The results of the present study have raised an important question. The first and foremost conclusion is that direct test on model assumptions, features and predictions have led to doubts about the fit of the Rasch model to DRP test data. Given the acceptance of the results, then what does one make of the assessments being made from DRP test results?

State and local governments of education are responsible for providing assessments of student performance across several grades as well as within a particular grade at selected times throughout the school year. Tests that meet these needs are built using vertical and horizontal equating. Horizontal equating is appropriate when multiple forms of a test are being used. It is generally assumed that the forms are parallel and the ability distribution of the examinees for whom these forms are administered are approximately equal. Vertical equating consists of constructing a single scale that allows one to compare examinee ability across different levels, such as grade level. Different populations are administered different tests of varying difficulty and the ability distribution of the examinees at the various levels will not be the same. While the horizontal equating of tests have shown a great deal of promise for all latent trait models, the current psychometric literature indicates approaching the vertical equating of tests calibrated by the Rasch model with extreme caution. Several factors may account for the lack of suitability of the Rasch model for vertical equating. Model misfit due to the violation of an assumption has frequently been cited for poor vertical equating results. This lack of fit is generally associated with systematic linking errors. Systematic errors are serious because these types of errors compound over successive equatings. The results of this study indicated that the Rasch model does not provide a satisfactory fit to the DRP. The presumption is that as the number of test forms in the equating chain increases, an increasing amount of scale drift (equating error) is likely to result. For the DRP, the attempt to investigate that amount of scale drift when equating across forms is thereby a subject that requires intense review.

Limiting our definition of achievement in a subject to items that fit a unidimensional IRT model is a mistake which inevitably leads to detriments in the measurement of achievement. Model utility is not necessarily defined in terms of whether items fit a particular model because this is not the only nor is it the best indicator of model appropriateness. The attainment of the assumptions and features of the IRT model must be validated as well. Through the verification of the attainment of model features, assumptions and predictions, the results of the present study suggest that the publisher of the DRP should consider the calibration of the three-parameter model to the DRP. The assessment of ability scores with the one-parameter model when the three-parameter model seems to fit the data better is inadvisable in such a high-stakes environment. Better estimates of ability are clearly obtained through the use of the three-parameter model.

The 1995-96 school year was the first year students were penalized for failing the LPT. The Virginia State Department of Education should in turn take appropriate measures to amend ability scores of those individuals who were not eligible for passage to the next grade and/or graduation from high school based on DRP results. In addition, measures should be taken by the publisher of the DRP to ensure the validity of vertical equating so that school officials can accurately compare student performance across grades.

APPENDIX I

Item Analysis of the 77 item DRP

	Item Response (%)					#	Difficulty	Item	index	item*test
Item	1	2	3	4	5	Omit	*p*	reliab.	disc.	Pearson
1	98.0+	.2	1.1	.3	.3	1	.98	.03	.04	.21
2	.8	1.3	96.8+	.5	.4	1	.96	.05	.07	.26
3	1.3	1.3	1.1	95.2+	1.0	0	.95	.09	.12	.40
4	1.7	3.4	.7	.8	93.2+	0	.93	.10	.16.	.40
5	5.3	84.1+	5.4	1.6	3.3	5	.84	.11	.21	.31
6	98.0+	.6	.4	.4	.5	0	.98	.04	.05	.25
7	.4	.4	.3	97.7+	.5	0	.97	.03	.04	.22
8	.5	92.3+	1.6	4.6	.7	3	.92	.10	.18	.39
9	1.3	2.0	94.1+	.8	1.5	3	.94	.08	.13	.36
10	6.0	10.2	3.4	4.4	75.7+	3	.75	.22	.46	.52
11	95.2+	.9	.9	1.2	1.6	1	.95	.08	.13	.39
12	3.1	.4	1.3	94.5+	.6	0	.95	.11	.16	.47
13	1.6	.3	97.2+	.6	.2	0	.97	.05	.08	.32
14	5.0	90.5+	1.7	1.3	1.3	1	.91	.09	.15	.31
15	94.5+	.8	.4	3.7	.5	0	.95	.08	.14	.37
16	1.9	92.8+	1.9	1.4	1.7	4	.93	.12	.19	.45
17	1.2	97.3+	.6	.5	.2	1	.97	.05	.07	.31
18	11.9	.6	2.9	.5	83.8+	2	.84	.15	.32	.42
19	89.7+	2.7	1.9	1.6	4.0	0	.90	.14	.24	.45
20	1.3	6.9	2.5	87.5+	1.6	1	.88	.16	.28	.47
21	89.3+	3.6	1.2	1.4	4.1	6	.89	.14	.25	.45
22	7.3	2.4	9.8	78.6+	1.7	3	.79	.18	.38	.45
23	6.4	6.0	70.0+	9.0	8.3	4	.70	.23	.52	.50
24	3.5	2.1	9.6	1.0	73.7+	0	.74	.23	.51	.52
25	2.8	1.2	14.0	3.0	78.6+	4	.79	.19	.39	.47
26	92.7+	2.2	2.0	1.7	1.2	0	.93	.12	.19	.46
27	7.0	87.8+	1.2	3.0	.8	2	.88	.16	.30	.49
28	2.9	2.3	88.3+	5.0	1.3	2	.88	.16	.27	.48
29	1.4	2.1	4.9	90.4+	1.0	1	.90	.12	.21	.40
30	89.2+	5.3	1.2	2.3	1.9	1	.89	.14	.25	.44
31	3.3	76.1+	2.6	7.4	10.2	4	.76	.17	.36	.41
32	2.0	1.3	82.2+	1.4	12.8	2	.82	.19	.39	.49
33	93.2+	2.3	.8	2.3	1.2	2	.93	.12	.18	.48

	Item Response (%)					#	Difficulty	Item	index	item*test
Item	1	2	3	4	5	Omit	*p*	reliab.	disc.	Pearson
34	1.8	81.0+	4.6	4.6	7.6	6	.81	.19	.39	.48
35	2.8	77.2+	8.8	3.6	7.5	1	.77	.20	.43	.47
36	2.9	6.6	81.9+	6.9	1.5	1	.82	.20	.40	.52
37	11.6	81.9+	1.5	3.0	1.8	2	.82	.17	.35	.43
38	31.8	52.2+	9.1	3.1	3.1	9	.52	.27	.67	.53
39	7.6	2.3	81.0+	6.9	1.9	5	.81	.21	.41	.54
40	5.4	12.0	5.4	69.6+	7.0	9	.70	.25	.59	.55
41	91.5+	1.9	2.8	1.4	2.1	3	.92	.14	.24	.50
42	1.9	4.2	7.1	1.9	84.5+	4	.85	.20	.38	.55
43	8.6	10.7	47.9+	23.0	9.4	8	.48	.25	.66	.51
44	6.5	11.3	76.7+	1.1	3.9	7	.77	.18	.38	.43
45	6.8	3.5	17.7	2.6	69.0+	5	.69	.21	.44	.44
46	14.0	7.6	16.9	56.0+	5.0	8	.56	.25	.61	.51
47	17.9	2.4	2.9	74.8+	1.6	6	.75	.20	.44	.46
48	13.1	2.4	5.6	2.9	75.7+	5	.75	.23	.48	.53
49	17.4	11.8	5.3	62.0+	3.1	6	.62	.26	.62	.53
50	2.1	7.5	86.2 +	1.7	2.1	7	.86	.17	.33	.50
51	72.5+	3.5	8.7	7.1	7.5	12	.73	.22	.49	.49
52	84.9+	7.0	3.3	3.0	1.3	7	.85	.17	.35	.47
53	20.7	2.9	53.5+	9.0	13.3	9	.54	.19	.46	.38
54	3.4	7.9	10.5	57.2+	20.4	9	.57	.27	.69	.55
55	14.9	15.7	8.9	5.5	54.1+	15	.54	.23	.56	.47
56	13.4	5.8	10.5	9.1	60.5+	12	.61	.23	.52	.47
57	19.7	10.8	7.8	48.3+	12.5	14	.48	.26	.64	.52
58	20.4	8.2	15.3	39.6 +	15.2	23	.40	.18	.45	.37
59	26.2	31.3+	18.2	17.2	6.1	17	.31	.15	.34	.31
60	30.2+	20.0	8.3	13.0	27.1	25	.30	.12	.30	.26
61	19.1	47.7+	9.8	10.5	11.4	29	.48	.22	.55	.49
62	4.4	9.9	42.5+	6.0	35.9	22	.43	.15	.35	.30
63	82.3+	3.9	4.6	4.1	4.0	21	.82	.18	.36	.46
64	7.3	61.7+	13.1	8.4	8.3	22	.62	.26	.63	.53
65	67.7+	7.4	9.4	7.7	4.5	22	.70	.25	.59	.55
66	26.6+	10.6	35.4	7.8	8.1	22	.27	.09	.24	.21
67	26.6	10.0	14.1	11.1	37.0+	22	.37	.16	.36	.34
68	6.4	44.5+	17.1	8.2	22.5	23	.45	.17	.41	.34
69	6.6	9.5	27.1	42.9+	12.6	24	.43	.19	.46	.39
70	13.7	53.6+	5.1	3.2	23.1	23	.54	.20	.47	.41
71	24.4	16.7	44.1+	5.9	7.7	22	.44	.15	.36	.31
72	4.6	35.1+	16.1	10.7	32.3	23	.35	.16	.39	.34
73	53.1+	5.0	13.9	7.9	18.4	32	.53	.14	.35	.28
74	14.2	29.9	14.0	33.6+	6.8	28	.34	.14	.33	.29
<u>75</u>	13.5+	34.8	15.3	25.5	9.3	29	.14	-.05	-.13	-.15
76	28.1	11.1	16.7	11.7	30.7+	30	.31	.16	.42	.36
77	15.9	10.2	30.8+	16.2	25.2	31	.31	.07	.13	.16

<u>Note.</u> +: the keyed response, _: a poorly discriminating item

Appendix II

D'Costa B Indicee for Low Performing Examinees

Id Number	Total Correct	B Index*
555	34	.98
751	32	.93
565	34	.92
562	28	.90
1751	25	.89
1754	19	.88
1957	34	.87
172	39	.86
1604	19	.86
578	34	.85
418	29	.84
568	28	.83
575	32	.81
412	28	.81
841	23	.81
1753	29	.81
759	22	.80
1557	37	.80
1567	37	.80
1560	32	.80
1823	38	.80
619	38	.79
1776	29	.79
1971	34	.79
364	39	.78
165	37	.77
159	31	.77
1806	31	.76
158	34	.75
1413	36	.75
1807	35	.75
580	39	.74
1841	31	.74
15	38	.73
954	17	.73
9	32	.72
1954	36	.72
1565	34	.71
1770	39	.71
1821	35	.71
769	37	.70
1415	35	.70
1620	39	.70
8	35	.69
1575	17	.69
1777	33	.69
1804	37	.69

* B indices calculated using D'costa (1994) BSWINDEX program

D'Costa B Indices for Low Performing Examinees

Id Number	Total Correct	B Index*
179	37	.68
609	23	.68
991	22	.68
1410	32	.68
1578	36	.67
1571	31	.67
1778	17	.67
1956	30	.67
1603	30	.66
1040	37	.65
1596	23	.65
621	38	.64
1394	28	.63
1588	25	.62
1599	25	.62
1439	38	.60
1589	37	.56
1918	13	.54
1628	18	.52
1945	27	.51
771	38	.50
1779	38	.47

* B indices calculated using D'costa (1994) BSWINDEX program

References

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Belmont, California: Wadsworth, Inc.

Andersen, E.B. (1973). A goodness-of-fit test for the Rasch model. Psychometrika, 38, 123-139.

Balassiano, M., & Ackerman, T. (1995a). An indepth analysis of the NOHARM estimation algorithm and implication for modeling the multidimensional latent ability space. Unpublished manuscript, University of Illinois, Faculty of Education, Urbana-Champaign.

Balassiano, M., & Ackerman, T. (1995b). An evaluation of the NOHARM estimation accuracy with a two-dimensional latent space. Unpublished manuscript, University of Illinois, Faculty of Education, Urbana-Champaign.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Berger, M.P., & Knol, D.L. (1990). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Birnbaum, A. (1968) Some Latent Trait Models and their Use in Inferring an Examinees' Ability. In F.M. Lord & M.R. Novick, (Eds.), Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Blalock, H.M. (1979). Social sciences. New York:McGraw-Hill.

Bormuth, J.R. (1969). Development of readability analyses. Final Report, Project No. 7-0052, Contract No. OEG-3-7-070052-0326, Office of Education, Bureau of Research, U.S. Department of Health, Education and Welfare, March.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago, 1970.

Bormuth, J.R. (1985). A response to "Is the degrees of reading power test valid or invalid? Journal of Reading,42-47.

Carroll, J.B. (1945). The effect of difficulty and chance success on correlation between items or between tests. Psychometrika, 10, 1-19.

Carver, R.P. (1985). Measuring readability using DRP units. Journal of Reading Behavior, 17, 303-316.

Cross, L. H. (1995). Review of the degrees of reading power. Mental Measurement Yearbook, 258-261.

D'Costa, A. G. (1993). Extending the Sato caution index to define within and beyond ability caution indexes. Paper presented at the annual meeting of the National Council for Measurement in Education, Atlants, GA.

De Champlain, A. (1995). An overview of nonlinear factor analysis and its relationship to item response theory. Paper presented at the meeting of the Americal Educational Research Association, San Francisco, CA.

Divgi, D.R. (1981). Does the Rasch model really work? Not if you look Closely. Paper presented at the annual meeting of NCME, Los Angeles, CA.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

Drasgow, F., & Parsons, C.K. (1983). Applications of unidimensional item response models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

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

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.). Educational Measurement. New York, Macmillan.

Hambleton R.K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), Applications of Item Response Models. Vancouver, BC: Education Research Institute of British Columbia.

Hambleton, R.K., & Rogers, H.L. (1986). Evaluation of the plot method for identifying potentially biased test items. In S.H. Irvine, . Newstead, and P. Dann (Eds.), Computer-based Human Assessment. Hingham, MA: Kluwer-Nijhoff.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement. 10, 287-302.

Hambleton R.K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston: Kluwer-Nijhoff.

Hambleton, R.K., & Traub, R.E. (1973). Analysis of empirical data using two logistic test models. British Journal of Mathematical and Statistical Psychology, 26, 195-211.

Hambleton R.K., Murray, L.., & Williams, P. (1983). Fitting item response models to the Maryland functional reading test results. Paper presented at the annual meeting of the American Educational Research Association, Montreal Quecec.

Hambleton, R.K. , Rogers, H.L., & Arrasmith (1986). A Comparison of the Mantel-Haenzel statistic and item response theory methods of identifying differential item performance. Paper presented at the annual meetings of AERA and NCME, San Francisco, CA.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory, volume 2. Sage Publications, Newbury Park, CA.

Hashway, R.M. (1978). Objective mental measurement. Praeger Publishers, New York, NY.

Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Hertzman, M. (1936). The Effects of the relative difficulty of mental tests on patterns of mental organization. Archives of Psychology, No. 197.

Hulin, C.L. Drasgow, F., & Parsons, L.K. (1983). Item response theory. Homewood, IL: Doe-Jones Irwin.

Humphreys, L. J. (1985). General intelligence: An integration of factor, test and simplex theory. In B.B. Wolman (Ed.), Handbook of Intelligence. (pp. 201-224). New York: Wiley.

Humphreys, L. J. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 77, 327-333.

Kifer, E.W., Mattson, I. & Carlid, M. Item analysis using the Rasch model. Sweden: Institute for the Study of International Problems in Education, Stockholm University, (1975).

Kingston, N.M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 9, 281-288.

Koslin, B., Zeno, S., Koslin, S., Wainer, H., Ivens, S. (1987). The DRP: An effectiveness measure in reading. New York: The College Board.

Lord, F.M. A theory of Test Scores. Psychometrika Monograph, No. 7. Iowa City, IA, (1952).

Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F.M. (1980). Application of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Ludlow, L.H. (1986). A graphical analysis of item response theory residuals. Applied Psychological Measurement 10, 217-222.

Miller, R. H. (1988). Teachers as researchers: Does the DRP predict student achievement? Paper presented at the anunal meeting of the Florida Reading Association, Orlando, Fl.

Mislevy, R. J., & Bock, R. D. (1990). BILOG [Computer software]. Mooresville, IN: Scientific Software.

Murray, L. N., & Hambleton, R.K. (1983) Using residual analysis to assess item response model-test data fit. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. Journal of Educational Measurement, 31, 200-219.

Ozcelik D.A., & Derberoglu, G. (1991) Contributions of the Rasch model to objectivity in measurement. Studies in Educational Evaluation, 17, 167-198.

Powers, D.E. & Leung, S.W. (1995). Answering the new SAT reading comprehension questions without the passages. Journal of Educational Measurement, 32, 105-129.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedagogiske Institut.

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-55.

Reckase, M.D. (1979). Unifactor latent trait models applied to multidimensional tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building unidimensional tests using multidimensional items. Journal of Educational Measurement, 25, 193-203.

Rudner, L.M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-220.

Ryan, J.P. (1980). Testing the appropriateness of the one-parameter model for the analysis of basic skills assessment data. Paper presented at the annual meeting of the American Educational Research Association, New York.

Sato, T. (1980). The S-P Chart and the Caution Index., Tokyo, Japan: NEC Educational Information Bulletin, 80-1, C&C Systems Research Laboratories, Nippon Electric Co.,

Snyder, J. (1993) Assessment of children's reading: A comparison of sources of evidence. National Center for Restructuring Education, Schools and Teaching. Teachers College, Columbia University, New York, New York.

Spearman, C. (1927). The abilities of man: Their nature and measurement. London: Macmillan.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 589-618.

Swineford, F. (1956). Technical Manual for users of test analysis. Statistical Report 56-42. Princeton, NJ: Educational Testing Service.

Tatsouka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnoses. In Fredericksen, Glaser, Lesgold and Shafto (Eds), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, NJ: Erlbaum.

Taylor, W.L. (1953). "Cloze procedure": A new tool for measuring readability. Journalism Quarterly, 30, 415-433.

Tuinman, J.J. (1974). Determining the passage dependency of comprehension questions on five major tests. Reading Research Quarterly, 9, 206-23.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.). Applications of Item Response Theory. (pp. 57- 70). Canada: Educational Research Institute of British Columbia.

Traub, R. E., & Wolf, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D.C. Berliner (Ed.), Review of Research in Education - Volume 9. Washington, D.C.: American Educational Research Association.

Wang, M. (1988). Measurement bias in the application of a unidimensional model to multidimensional item-response data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wherry R.J., & Gaylord R.H. (1944) Factor patterns of test items and tests as a function of the correlation coefficient: content, difficulty, and constant error factors. Psychometrika, 9, 237-244.

Wood, R. (1978). Fitting the Rasch model - A heady tale. British Journal of Mathematical and Statistical Psychology, 31, 27-32.

Wright, B. D. (1968). Sample free test calibration and person measurement. In Proceedings of the 1967 ETS Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service. 85-101.

Wright, B. D., & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. Psyckometrika, 50, 399-410.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.

Vita

Monique V. Granville obtained a B.S in psychology from Howard University and M.S. in statistics from the City of New York-Baruch College. She has ten years of statistical consulting experience. Her training is diverse encompassing medical, educational and marketing arenas.