WILEY NCME national council on measurement in education

Scaling Performance Assessments: A Comparison of One-Parameter and Two-Parameter Partial Credit Models
Author(s): Anne R. Fitzpatrick, Valerie B. Link, Wendy M. Yen, George R. Burket, Kyoko Ito and Robert C. Sykes
Source: *Journal of Educational Measurement*, Vol. 33, No. 3 (Autumn, 1996), pp. 291-314
Published by: National Council on Measurement in Education
Stable URL: http://www.jstor.org/stable/1435374
Accessed: 01-11-2015 04:43 UTC

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.org/stable/1435374?seq=1&cid=pdf-reference#references_tab_contents

You may need to log in to JSTOR to access the linked references.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Wiley and National Council on Measurement in Education are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Educational Measurement.*

http://www.jstor.org

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Journal of Educational Measurement
Fall 1996, Vol. 33, No. 3, pp. 291–314

Scaling Performance Assessments: A Comparison of One-Parameter and Two-Parameter Partial Credit Models

Anne R. Fitzpatrick, Valerie B. Link, Wendy M. Yen, George R. Burket, Kyoko Ito, and **Robert C. Sykes**
CTB/McGraw-Hill

In one study, parameters were estimated for constructed-response (CR) items in 8 tests from 4 operational testing programs using the 1-parameter and 2-parameter partial credit (1PPC and 2PPC) models. Where multiple-choice (MC) items were present, these models were combined with the 1-parameter and 3-parameter logistic (1PL and 3PL) models, respectively. We found that item fit was better when the 2PPC model was used alone or with the 3PL model. Also, the slopes of the CR and MC items were found to differ substantially. In a second study, item parameter estimates produced using the 1PL-1PPC and 3PL-2PPC model combinations were evaluated for fit to simulated data generated using true parameters known to fit one model combination or the other. The results suggested that the more flexible 3PL-2PPC model combination would produce better item fit than the 1PL-1PPC combination.

The National Assessment of Educational Progress (NAEP) and many states are implementing large-scale, high-stakes assessments that include constructed-response (CR) items. In these assessments a CR item requires an examinee to construct a written response that is scored by raters using a rating scale having 2 or more score points. In some testing programs, such as the Maryland School Performance Assessment Program, CR is the only type of item that is used. In others, such as Alabama's end-of-course tests in algebra and geometry, California's Learning Assessment System, Indiana's Performance Assessment for School Success, and Michigan's High School Proficiency Test, multiple-choice (MC) and CR items are used together to assess the skills of interest.

The appropriate means of scaling CR items is an unresolved issue (Loyd, Engelhard, & Crocker, 1993; Mehrens, 1992). NAEP and others have used a mixture of models, applying the three-parameter logistic (3PL) model (Birnbaum, 1968) to scale the MC items and the two-parameter partial credit (2PPC) model

A version of this article was presented at the Annual Meeting of the National Council on Measurement in Education, April, 1995.

The authors thank the Alabama Department of Education, the Michigan Department of Education, and the Maryland State Department of Education for their permission to use data from their testing programs in this research. The authors also thank Maurice Piccari for carrying out some of the statistical analyses used in this article.

Finally, the authors are grateful to Rebecca Zwick and two anonymous reviewers for very constructive comments on an earlier version of this article.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

(Muraki, 1992; Yen, 1993) to scale the CR items (Isham & Allen, 1993; Mazzeo, Yamamoto, & Kulick, 1993).

An attractive alternative to the 3PL-2PPC combination of models is the one-parameter logistic (1PL) model in combination with the one-parameter partial credit (1PPC) model. The 1PL model was developed by Georg Rasch (1960, 1966) to scale dichotomously scored items, and the 1PPC model was developed by Masters (1982), who extended the 1PL model for use in scaling polytomous items, that is, items with more than two ordered score categories. This alternative approach has been regarded as attractive because of its simplicity, its comprehensibility to practitioners and lay audiences, and its modest sampling requirements (Hambleton, 1989; Lord, 1983).

The merits of using the 1PL and 3PL models were hotly debated in the late 1970s and early 1980s (cf. Divgi, 1986; Traub, 1983; Wright, 1977), when applications of item response theory (IRT) to MC items were being tried for the first time. The recent interest in using CR items in high-stakes achievement tests has inspired new studies of the properties of IRT models that can be used to scale the CR items in these tests (e.g., Donoghue, 1994; Maydeu-Olivares, Drasgow, & Mead, 1994; Muraki, 1992; Muraki, 1993; Reise & Yu, 1990; Wainer & Thissen, 1993; Yen, 1993).

The purpose of this article is to investigate and compare the behavior of the 1PPC and 2PPC models in both real data and simulated data contexts where CR items are present. Two studies were conducted. In the first study, real data from four different testing programs were calibrated using each of the two models. Where MC items were present, a mixed model (Thissen & Steinberg, 1986) was used; more specifically, the 1PL or 3PL model was used in combination with the 1PPC or 2PPC model. For this study the 3PL model was preferred over the two-parameter logistic (2PL) model for the analyses of the MC items because it would enable us to detect the presence of any nonzero asymptotes that occurred.

Because findings from analyses of real data cannot indicate which combination of models is more accurate, the second study was a simulation study designed to address this question. In this study, true item parameters were defined, and each model combination was evaluated in terms of its fit to data generated using these parameters.

Study 1

Method

Description of the test data. Table 1 shows the configuration of the eight tests analyzed in the first study. The Grade 3 reading and mathematics tests and the Grade 5 science test were part of the 1992 Maryland School Performance Assessment Program (MSPAP), a large-scale, statewide assessment program administered in Grades 3, 5, and 8 for the purpose of school improvement. Approximately 50,000 students are tested annually in each grade, and each student is administered one of three nonparallel test forms used in each grade. In the current study, random samples of students who took one of the three test forms in 1992 were selected for our analyses.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 1

Tests Analyzed in Study I

Grade/ Test	No. of Items	Maximum Possible Score	Item Type								
			MC[1]	2-L	3-L	4-L	5-L	6-L	7-L	9-L	10-L
3 MSPAP Reading	15	35	0	1	8	6	0	0	0	0	0
3 MSPAP Math Content	41	57	0	25	16	0	0	0	0	0	0
5 MSPAP Science	22	34	0	11	10	1	0	0	0	0	0
9 MWT Writing	29	39	27	0	0	0	0	0	2	0	0
10 Alabama Geometry	57	74	51	0	2	0	2	1	1	0	0
11 MHSPT Reading	36	41	35	0	0	0	0	0	1	0	0
11 MHSPT Mathematics	46	89	39	0	0	0	0	0	4	1	2
11 MHSPT Science	49	75	41	0	0	0	7	0	1	0	0

[1] MC refers to multiple choice; 2-L, 3-L, etc. refers to CR items scored with rubrics having 2 score points, 3 score points, etc.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Each MSPAP form assesses reading, writing, language usage, mathematics, social studies, and science skills using CR items. Students' responses to these items are scored once by raters who are Maryland teachers trained in the scoring process. Some responses are scored using 2-point (dichotomous) scoring rules, whereas others are scored using 3-point or 4-point scoring rules.

The Grade 9 Maryland Writing Test (MWT) was administered in 1988 to students in Maryland to assess their writing skills. As Table 1 shows, the test consisted of 27 MC items and 2 CR items. The CR items were narrative and explanatory essay prompts. These prompts were scored by two raters using a modified holistic score scale that ranged from 1.0 to 4.0 in increments of 1.0. An examinee's final score on each essay was the average of the two ratings, unless there was a discrepancy of more than 1 point between the two ratings. In such an instance, the discrepancy was resolved by a third reader. Given these scoring procedures, the final score scale for each essay had seven possible score points ranging from 1.0 to 4.0 in increments of 0.5. These seven points were treated as seven score levels for each of the two items analyzed. For the current study, a random sample of approximately 3,000 students was drawn from the population of ninth grade students who took the 1988 MWT.

The Grade 10 geometry test was one of 12 pilot forms developed by CTB/ McGraw-Hill for the Alabama State Department of Education to try out items for a first-year geometry end-of-course assessment. The piloted items were written to assess three broad categories: (a) shapes and measures, (b) logical reasoning, and (c) connections between algebra and geometry. The 12 pilot forms were administered in May, 1993, to approximately 24,000 students in geometry classes across the state. Students' responses to the CR items were scored once by trained raters using scoring rules developed by the Alabama State Department of Education and CTB. The current study used a sample of students who were administered one of the pilot forms.

The Grade 11 tests in reading, mathematics, and science were three of 29 test forms administered in the fall of 1994 to try out items for the Michigan High School Proficiency Test (MHSPT). The fall study involved a stratified sample of approximately 29,000 eleventh graders in Michigan. Sets of forms were spiraled within classrooms. Each form was included in two sets, and samples of approximately 500 students were randomly assigned to the forms within a set. These samples contributed the three data sets used in the current study.

Students' responses to the CR items in each MHSPT pilot form were graded by two readers. Items with 3 or more score points were read a third time if the two readers disagreed by more than 1 point. In such a circumstance, a student's item score was obtained by summing all three ratings, multiplying this sum by two thirds, and then rounding to the nearest whole number. Otherwise, the student's item score was obtained by summing the two readers' ratings.

Computer program. The tests were scaled using PARDUX, a microcomputer program designed and written by George Burket (1991) using algorithms specified by Wendy Yen. In its present form, PARDUX estimates parameters for dichotomous and multilevel items using marginal maximum likelihood procedures implemented with an EM algorithm. It permits users to apply either the

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

1PL, 2PL, or 3PL model, the 1PPC or 2PPC model, or a combination of these models. PARDUX also permits the evaluation of model-data fit using a variety of statistics including Yen's (1984) Q_1, a fit statistic that will be described below.

For the parameter estimation process, PARDUX defines a prior true θ distribution to have a mean of 0.0 and standard deviation of 1.0. For the current analyses, a maximum of 25 estimation cycles was used, and a convergence criterion was set at 0.01. Following the parameter estimation, maximum likelihood trait estimates (θs) for examinees were obtained.

The calibration models. All of the tests analyzed were calibrated twice. The five tests that contained both MC and CR items were calibrated once using a 1PL-1PPC model combination and a second time using a 3PL-2PPC model combination. The three tests that contained only CR items were calibrated once using the 1PPC model and a second time using the 2PPC model.

The 1PL model states that the probability of an examinee correctly answering item j is

$$P_j(\theta) = P(x_j = 1 \mid \theta) = \frac{\exp[A(\theta - B_j)]}{1 + \exp[A(\theta - B_j)]}, \tag{1}$$

where A is a common discrimination value, and B_j is a location parameter for the item.

The CR items had m_j score levels assigned integer scores that ranged from 0 to $m_j - 1$. The 1PPC model states that the probability of an examinee with ability θ having a score at the kth score level on item j is

$$P_{jk}(\theta) = P(X_j = k - 1 \mid \theta) = \frac{\exp(z_{jk})}{\sum_{i=1}^{m_j} \exp(z_{ji})}, \qquad k = 1, \ldots, m_j, \tag{2}$$

where

$$z_{jk} = \alpha(k - 1)\theta - \sum_{i=0}^{k-1} \gamma_{ji}, \tag{3}$$

and $\gamma_{j0} = 0$. The α is the common item discrimination, and γ_k/α can serve as a kind of location parameter, since it refers to the point at which the trace lines for adjacent score levels intersect.

The 3PL model (Birnbaum, 1968) is traditionally expressed as

$$P_j(\theta) = P(X_j = 1 \mid \theta) = C_j + \frac{(1 - C_j)}{1 + \exp(-1.7\, A_j(\theta - B_j))}. \tag{4}$$

As Chang and Mazzeo (1994), among others, have noted, the 2PPC model is a general form of Equation 2 above, where α_j replaces the α in Equation 3; the 2PPC model specifies that the discriminations can vary over items. When $C_j = 0$ and $m_j = 2$, the parameters of Equations 2 and 4 are related in that $\alpha_j = 1.7\, A_j$ and $\gamma_{jk} = 1.7 B_j A_j$.

Some plots displaying observed and predicted item functioning will be presented. For multilevel items having many score levels, it is difficult to interpret

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

plots for these many levels. Therefore, our plots will compare the observed item performance with the predicted item performance as expressed by the item response function. This function can be defined as

$$E(X_j \mid \theta) = \sum_{k=1}^{m_j} (k - 1)P_{jk}(X_j = k - 1 \mid \theta).$$

Analyses. The alternative approaches to calibrating the eight tests under study were evaluated by considering the calibrations that resulted and item fit. Item fit was assessed using the Q_1 statistic described by Yen (1984) for the dichotomously scored items and a generalization of this statistic for the multilevel items. As described by Yen, Q_1 is distributed approximately as a chi-square with seven degrees of freedom. It has the form

$$Q_{1j} = \sum_{i=1}^{I} \frac{N_{ji}(O_{ji} - E_{ji})^2}{E_{ji}} + \sum_{i=1}^{I} \frac{N_{ji}[(1 - O_{ji}) - (1 - E_{ji})]^2}{1 - E_{ji}},$$

where N_{ji} is the number of examinees in cell i for item j. O_{ji} and E_{ji} are the observed and predicted proportions of examinees in cell i that attain the maximum possible score on item j, where

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a \in i}^{N_{ji}} P_j(\hat{\theta}_a).$$

There are 10 independent cells created by dividing examinees into deciles using their trait estimates. As Yen (1984) has noted, the assumed distribution and degrees of freedom of Q_1 may be violated when examinees' observed trait values are used to define the cells.

The generalization of Q_1 for multilevel items can be stated as

$$Q_{1j} = \sum_{i=1}^{I} \sum_{k=1}^{m_j} \frac{N_{ji}(O_{jki} - E_{jki})^2}{E_{jki}},$$

where

$$E_{jki} = \frac{1}{N_{ji}} \sum_{a \in i}^{N_{ji}} P_{jk}(\hat{\theta}_a).$$

O_{jki} is the observed proportion of examinees in cell i who perform at the kth score level.

It is well known that chi-square statistics are affected by sample size and extreme expectations (Stone, Ankenmann, Lane, & Liu, 1993), and that their degrees of freedom are a function of the number of independent observations entering into the calculation minus the number of parameters estimated. Items with more score levels have more degrees of freedom, which makes it awkward to compare the statistics for items that differ in the number of score levels. To facilitate this comparison in the current study, the following standardization of the Q_1 statistic was used:

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 2

Descriptive Raw Score Statistics for Tests Analyzed in Study I

Grade/ Test	Raw Score N	Raw Score Mean	Raw Score SD	Mean OPM[1]	Coeff. α	Item Type	Item-Test Correlations Mean	SD	Min	Max
3 MSPAP Reading	7,502	7.69	5.40	0.22	.84	CR:	0.55	0.06	0.44	0.68
3 MSPAP Math Content	1,185	17.35	9.76	0.30	.90	CR:	0.45	0.10	0.18	0.61
5 MSPAP Science	3,024	8.05	5.23	0.24	.80	CR:	0.45	0.10	0.27	0.57
9 MWT Writing	3,044	33.40	5.42	0.86	.84	MC:	0.48	0.06	0.32	0.58
						CR:	0.68	0.02	0.66	0.70
10 Alabama Geometry	2,113	23.42	9.27	0.32	.84	MC:	0.28	0.09	0.13	0.47
						CR:	0.50	0.09	0.41	0.66
11 MHSPT Reading	547	22.48	8.06	0.55	.84	MC:	0.38	0.10	0.20	0.55
						CR:	0.70	----[2]	----[2]	----[2]
11 MHSPT Mathematics	571	34.88	19.89	0.39	.88	MC:	0.36	0.08	0.12	0.50
						CR:	0.68	0.04	0.62	0.73
11 MHSPT Science	549	35.18	15.29	0.47	.91	MC:	0.37	0.10	0.12	0.50
						CR:	0.61	0.10	0.42	0.67

[1] OPM refers to the Observed Proportion of Maximum, which is the mean raw score divided by the maximum possible score.

[2] Statistic not available or not informative because test contains only one CR item.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 3

Summary of Standardized Q_1 Fit Results in Study I

Grade/ Test/ Model Comb.	N	No of Items	Mean Z	SD Z	Percentage Distribution of Z's				
					$Z<2$	$2\leq Z<3$	$3\leq Z<5$	$5\leq Z<10$	$10\leq Z$
3 MSPAP									
Reading	7,502	15							
1PPC			16.79	18.61	7	19	0	20	53
2PPC			9.55	9.21	0	20	20	33	27
3 MSPAP									
Math Content	1,185	41							
1PPC			3.18	4.75	56	12	14	12	5
2PPC			0.44	1.25	88	10	2	0	0
5 MSPAP									
Science	3,024	22							
1PPC			5.07	4.59	41	0	14	32	14
2PPC			1.01	1.70	82	14	0	5	0
9 MWT									
Writing	3,058	29							
1PL/1PPC			7.14	10.30	31	3	38	3	24
3PL/2PPC			4.12	7.91	66	0	6	17	10

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

10 **Alabama**									
Geometry	2,113	57							
1PL/1PPC			6.40	1.55	30	7	19	19	25
3PL/2PPC			1.19	3.23	74	14	7	3	2
11 **MHSPT**									
Reading	547	36							
1PL/1PPC			2.34	2.50	55	14	8	22	0
3PL/2PPC			0.76	1.10	86	11	2	0	0
11 **MHSPT**									
Mathematics	571	46							
1PL/1PPC			3.24	4.31	46	15	22	11	6
3PL/2PPC			0.76	1.56	85	4	11	0	0
11 **MHSPT**									
Science	549	49							
1PL/1PPC			3.30	4.53	48	16	16	14	6
3PL/2PPC			0.77	1.47	88	6	4	2	0

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

$$Z_{Q_{1j}} = \frac{Q_{1j} - df}{\sqrt{(2df)}}.$$

This statistic has a mean of zero and a standard deviation equal to 1.0 when the chi-square assumptions hold.

To use this standardized statistic to flag items for potential misfit, it has been our practice to set a critical value for Z in light of sample size. Items with Zs exceeding this critical value are then inspected for their fit by comparing their observed and predicted trace lines. For our tryouts of MC items for new, norm-referenced tests, our experience has indicated that with sample sizes of about 1,000 students, items with $Z \geq 4.6$ often have severe misfit and are therefore candidates for deletion from the pool of items. With samples of half that size, which occur in the current study, a criterion value of $Z \geq 2.3$ would have roughly the same statistical power to detect potentially misfitting items. Therefore we used this value of Z, rounded to 2.00, as one basis for evaluating fit in our study.

Following the calibration and examination of item fit, the correlation between the estimated θs produced by the two calibrations of each test was computed.

Results

Table 2 provides descriptive statistics for the eight tests analyzed. The table shows that in general the tests were difficult, with all but two of the tests having an observed percent of maximum (OPM)[1] at or below 0.47. Internal consistency was moderately high; the lowest value of 0.80 was associated with the relatively short Grade 5 science test.

With respect to the item-test (Pearson product) correlations, Table 2 shows that they were considerably higher for the CR items than for the MC items. This is a predictable result. CR items with multiple score points typically produce more observed item score variance than will MC items that are dichotomously scored. Since the correlations were calculated with the item of interest included in the total test score, more covariance with the total test score also is to be expected.

In their sizable variability, the item-test correlations given in Table 2 for the MC items resemble those reported by Birnbaum (1968, p. 402). Birnbaum as well as Divgi (1986) concluded that this variability in the biserials of MC items was indicative of real differences in the discriminating power of these items, a quality that fixed-slope models like the 1PL and 1PPC models are not designed to accommodate. Further discussion of this point is presented below.

The item-test correlations for the CR items in the Grade 3 reading test and the MC items in the Grade 9 writing test were slightly less variable than were those for other tests. Most likely this was due to the extreme difficulty and extreme easiness, respectively, of the items in these tests. In the Grade 3 reading test, only 2 of the 15 items had item OPMs[2] greater than 0.27. In the Grade 9 writing test, 22 of the 27 MC items had p-values greater than 0.85.

Results of the Q_1 fit analyses that were carried out following the item calibrations are given in Table 3. Examination of the mean Zs shows that without exception the fit was substantially better (i.e., the mean Zs were substantially

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

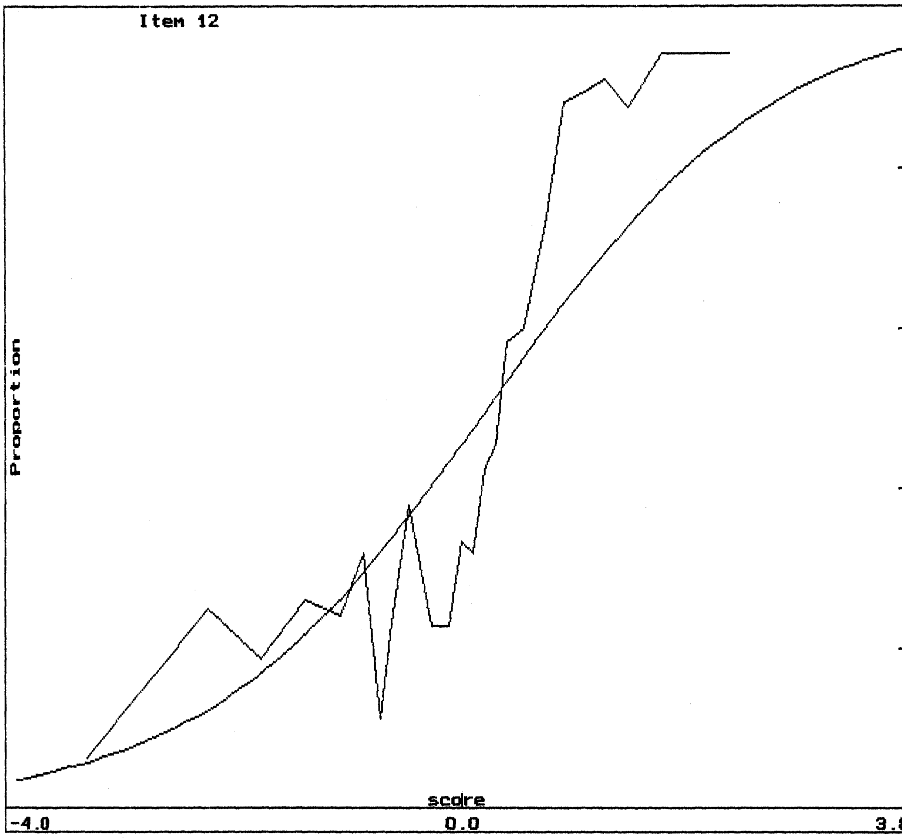

FIGURE 1. *Observed item trace line and 1PL prediction for Grade 11 mathematics test, MC Item 12* (A = *0.47;* Z = *6.41)*

lower) when the items were scaled using the 2PPC model or the 3PL-2PPC model combination than they were when the items were scaled using the 1PPC model or the 1PL-1PPC model combination.

It is useful also to look at the distributions of *Z*s given in Table 3. Were the criterion for flagging set at $Z \geq 2.00$, with the exception of the Grade 3 reading test, vastly greater percentages of items would be flagged under the 1PPC model or the 1PL-1PPC model combination. Due to the exceptional difficulty of the Grade 3 reading test, there were extreme expectations in some cells of the fit analyses, which produced the very high *Z*s seen in the table for both calibrations of this test.

Figure 1 shows the observed trace line and a 1PL trace line for an MC item from the Grade 11 mathematics test; this item would have been flagged for misfit. With an item OPM equal to 0.47, it was only moderately difficult. However, it was exceptionally discriminating ($A_j = 1.62$) when its parameters were estimated using the 3PL model, and it had a nonzero asymptote. Figure 2 shows the observed trace line for the same item and the trace line defined by 3PL parameter

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

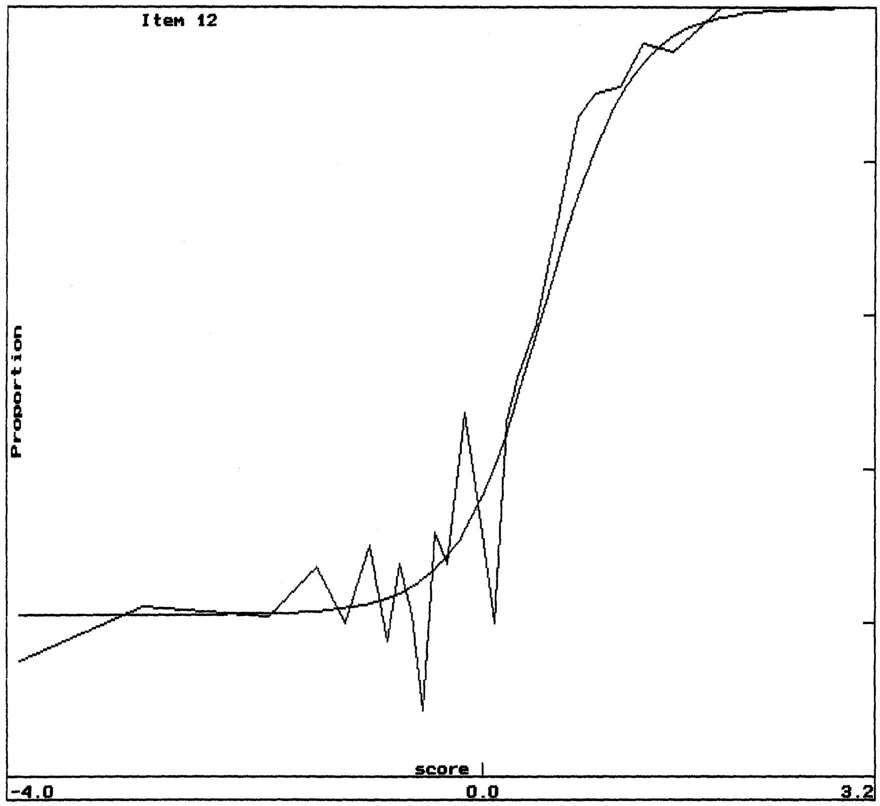

FIGURE 2. *Observed item trace line and 3PL prediction for Grade 11 mathematics test, MC Item 12 ($A_j = 1.62$; $Z = -0.51$)*

estimates. As the lower *Z* reflects, the latter trace line fit the observed data better. Use of the 1PL trace line for scoring would result in underpredictions of performance at the extremes of the ability range and an overprediction of performance in the middle of this range.

Figures 3 and 4 show plots for an MC item in the Grade 9 MWT, and Figures 5 and 6 show plots for a CR item in the same test. With OPMs equal to 0.92 and 0.85, respectively, both the MC and CR items were relatively easy. The 1PL-1PPC calibration produced a common discrimination equal to 0.76. Study of Figures 3 and 4 suggest that this value was somewhat too low for the highly discriminating MC Item 28, which was estimated by the 3PL model to have a discrimination value equal to 1.80. In contrast, Figures 5 and 6 suggest that the 1PL-1PPC common discrimination of 0.76 was too high for the apparently less discriminating CR Item 1, which was estimated by the 2PPC model to have a discrimination equal to 0.54.

To explore the effect of item type on item discrimination a bit further, Table 4 presents summary statistics in the form of 3PL and 2PPC estimates for the five tests that contained both MC and CR items. The table shows that the item

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

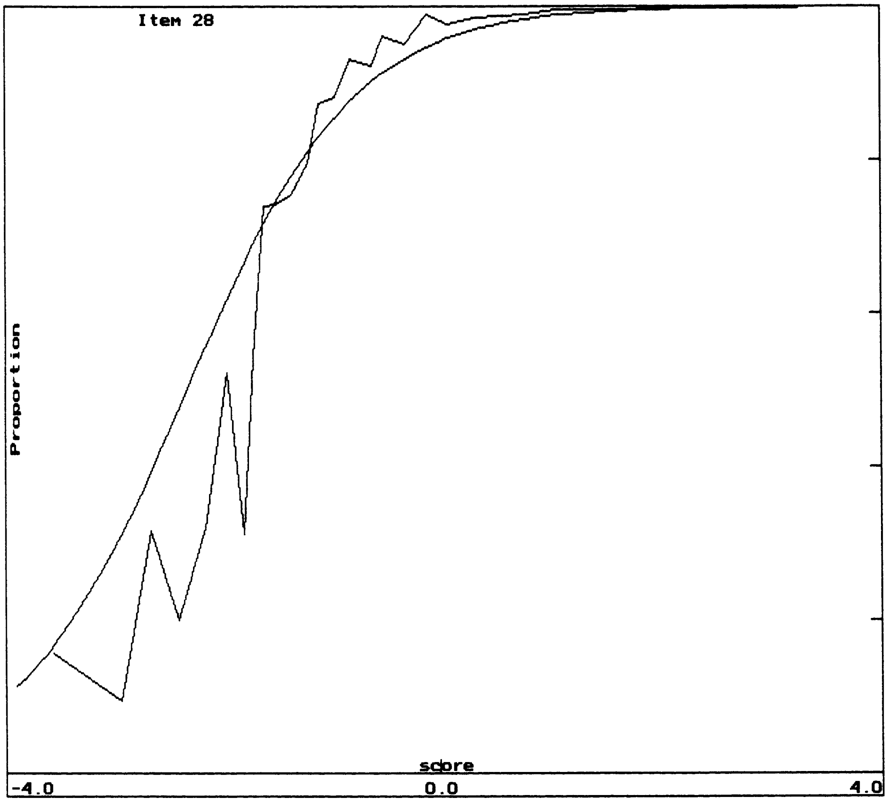

FIGURE 3. *Observed item trace line and 1PL prediction for Grade 9 writing test, Item 28* (A = *0.76;* Z = *39.23)*

discriminations of the MC items were substantially higher than those of the CR items. The MC items also showed considerable variability in their discriminations. These findings help to explain why the fixed-slope models did not fit very well the data from tests that included mixed item types.

The final analysis examined the correlations between the θ estimates produced by the two model combinations. It showed that these correlations generally were extremely high (rs $\geq$ .95), except in the case of the Grade 10 geometry test, where r was equal to .88. This finding of high correlations between θ estimates produced by fixed-slope and variable-slope IRT models is common (see Hambleton & Cook, 1983; Traub & Lam, 1985; van de Vijver, 1986).

A plot of the estimates for the geometry test is given in Figure 7. The plot shows that there was a curvilinear relationship between the two sets of θ estimates. More specifically, it shows that when the geometry items were scaled using the 1PL-1PPC model combination, there was scale shrinkage at the lower end of the θ scale and scale expansion at the upper end. This phenomenon is known to occur when items have nonzero lower asymptotes, that is, nonzero guessing parameters (van de Vijver, 1986; Yen, 1985, 1986).

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

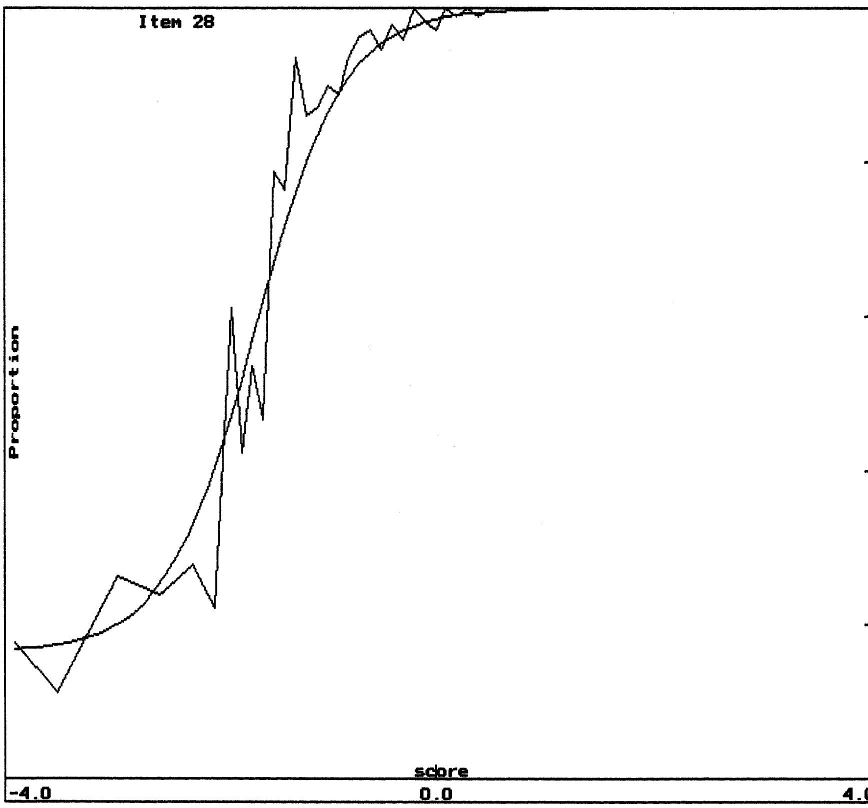

FIGURE 4. *Observed item trace line and 3PL prediction for Grade 9 writing test, Item 28* (A_j = *1.06; Z = 6.25)*

Study 2

Method

Description of simulated test data. Four sets of simulated item responses were generated using sets of true item parameters designed to emulate the item parameters that might characterize state performance assessments. Two of the data sets reflected responses to short, 12-item tests, and two reflected responses to longer, 36-item tests. As Table 5 shows, in both the short and long tests, the MC items and the two-level, three-level, and four-level CR items each represented 25% of the test questions.

Two sets of true item parameters were developed for each test length. The first set, called the ''R'' set, consisted of true parameters that reflected the assumptions of the 1PL-1PPC model combination, that is, equal discriminations and lower asymptotes equal to zero for the MC items. The second set, called the ''P'' set, reflected the assumptions of the 3PL-2PPC model combination, that is, unequal discriminations and nonzero asymptotes for the MC items.

This approach was used because it cannot be known a priori whether the

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

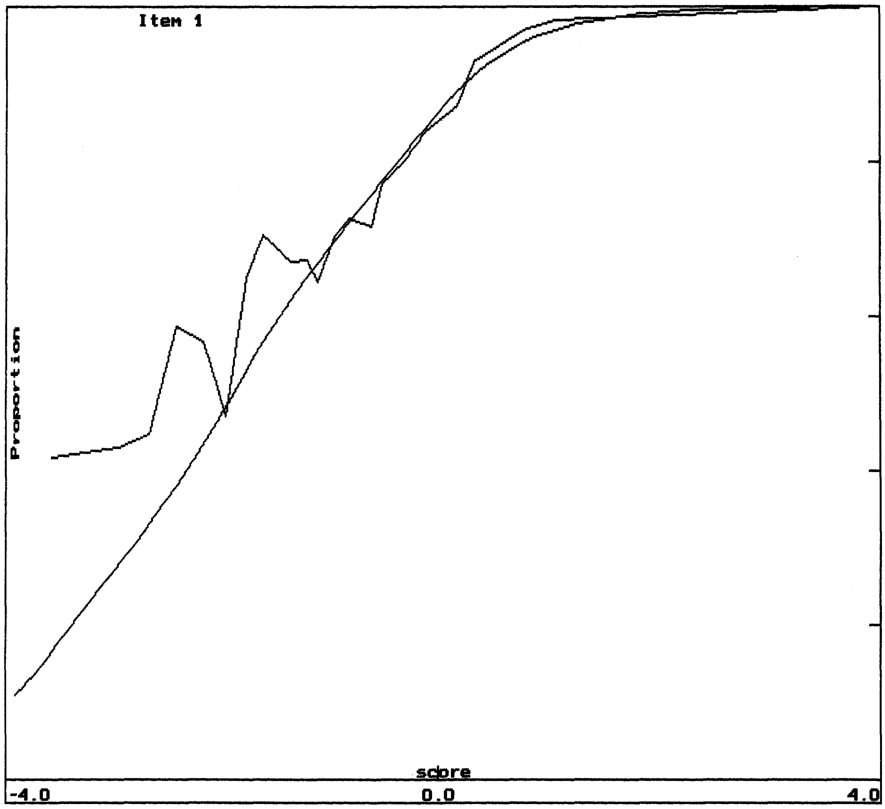

FIGURE 5. *Observed item trace line and 1PPC prediction for Grade 9 writing test, Item 1* (A = *0.76;* Z = *7.33)*

"true" parameters of items meet the assumptions of the 1PL-1PPC model combination or the 3PL-2PPC model combination. Data generated using one or another invariably will bias the results in favor of that model. One way to avoid this bias is to generate data using both model combinations. (For another approach see Wainer and Thissen [1987].)

For all data sets, 2,000 trait values (θs) were generated from a normal (0, 1) distribution. Summary statistics describing the true item parameters for the four tests are given in Table 6.

Computer programs. Simulated responses were generated using the true item and ability parameters, the appropriate model combination, and the microcomputer program BSSTSIM1, written by Burket (1992). Item parameters were then estimated using the simulated responses in combination with PARDUX.

Analyses. All four tests were scaled twice, once using the 1PL-1PPC model combination and a second time using the 3PL-2PPC model combination. As before, in both calibrations a maximum of 25 estimation cycles and a convergence criterion equal to 0.01 were used. Fit between the data and the item parameters that were estimated was evaluated as in Study 1.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

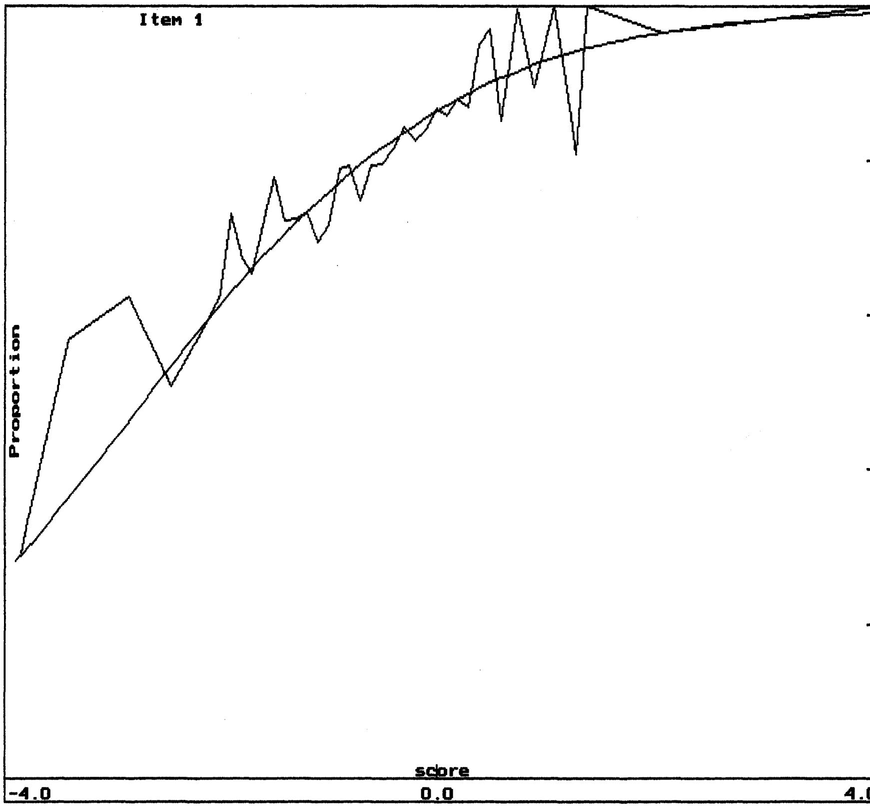

FIGURE 6. *Observed item trace line and 2PPC prediction for Grade 9 writing test, Item 1* (A_j = *0.54; Z = 0.88)*

Results

Descriptive raw score statistics for the tests analyzed are given in Table 7. The OPMs for these tests indicate that they were somewhat difficult, although not quite as difficult as the Grade 3 reading test analyzed in Study 1. The coefficient alphas were similar to those described in Study 1, although the item-test correlations were slightly more variable than those described in the first study.

Table 8 summarizes the fit results calculated after each of the four tests was calibrated using each of the two model combinations. The results for the two short tests (R12 and P12) were similar. The 3PL-2PPC model produced more good-fitting items having *Z*s less than 2.00. However in the case of R12, this model also produced two items with *Z*s $\geq$ 5.00, leading to a slightly higher mean *Z* than that associated with the 1PL-1PPC model combination.

For the two longer tests (R36 and P36), Table 8 shows that estimates from the 1PL-1PPC model combination produced slightly better fit for R36. However, the 3PL-2PPC model combination clearly produced superior fit results in the case of P36.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 4

Comparison of Item Discrimination by Item Type[1]

Grade/ Test	MC Items			CR Items		
	N	Mean	SD	N	Mean	SD
9 MWT Writing	27	1.91	0.47	2	0.54	0.01
10 Alabama Geometry	51	1.46	0.63	6	0.87	0.22
11 MHSPT Reading	35	1.55	0.68	1	0.37	0.00
11 MHSPT Mathematics	39	1.70	0.62	7	0.55	0.21
11 MHSPT Science	41	1.45	0.53	8	0.79	0.24

[1] To facilitate comparisons between the MC and CR items, the discriminations for both types of items are expressed in the 2PPC metric.

Several other results given in Table 8 suggest that the 3PL-2PPC model combination generally may fit data that do not meet its assumptions better than does the 1PL-1PPC model combination. As an example, compare the findings for the R36 and P36 tests. For R36 the difference in the mean *Z*s for the 1PL-1PPC and 3PL-2PPC model combinations was about 0.60. In contrast, for P36 this difference was about 3.80.

The distributions of fit values are also informative. When the 3PL-2PPC model combination was applied to the R36 test data, the misfit for 92% of the items was less than 2.00. In contrast, when the 1PL-1PPC model combination was applied to the P36 test data, only 47% of the items had *Z*s less than 2.00.

Conclusions

The findings from Study 1 suggest that there are differences in the effectiveness of the 1PPC and 2PPC models either alone or in combination with the 1PL and 3PL models, respectively, as means of calibrating items from current testing programs that include CR items. Specifically, the analyses show that the 2PPC model alone or in combination with the 3PL model provided uniformly better fit than did the 1PPC model used alone or in combination with the 1PL model.

The raw score and IRT statistics for the real test data indicated that the

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

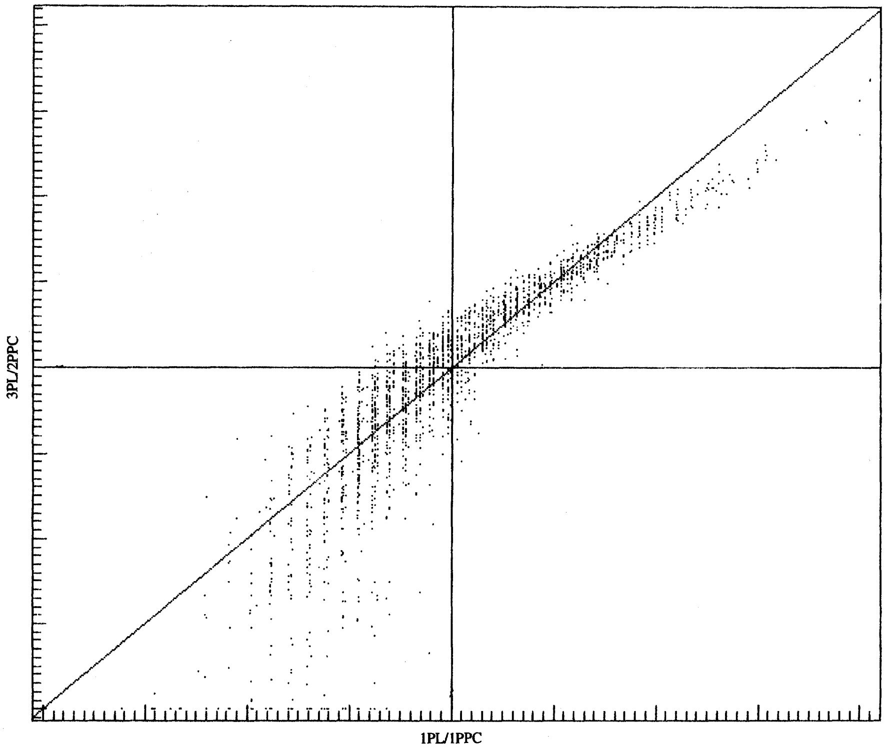

FIGURE 7. *Comparison of θ estimates produced by 1PL-1PPC and 3PL-2PPC model combinations for Grade 10 geometry test*

Table 5

Tests Analyzed in Study II

Test	No. Of Items	Maximum Possible Score	Item Type			
			MC[1]	2-L	3-L	4-L
R12	12	21	3	3	3	3
R36	36	63	9	9	9	9
P12	12	21	3	3	3	3
P36	36	63	9	9	9	9

[1] MC refers to multiple choice; 2-L, 3-L, etc. refers to items scored with rubrics having 2 score points, 3 score points, etc.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 6

Means and Standard Deviations of True Item Parameters Used in Study II

Test	Parameter	No. Of Items	Mean	SD
R12	A_j	3	0.63	0.00
R36	A_j	9	0.55	0.00
P12	A_j	3	0.86	0.16
P36	A_j	9	0.77	0.26
R12	B_j	3	-0.22	0.61
R36	B_j	9	0.18	0.49
P12	B_j	3	0.38	0.68
P36	B_j	9	0.43	0.97
R12	C_j	3	0.00	0.00
R36	C_j	9	0.00	0.00
P12	C_j	3	0.20	0.02
P36	C_j	9	0.22	0.02
R12	α_j	9	1.08	0.00
R36	α_j	27	0.93	0.00
P12	α_j	9	0.85	0.24
P36	α_j	27	0.93	0.25
R12	γ_{j1}	9	2.68	0.78
R36	γ_{j1}	27	1.55	1.35
P12	γ_{j1}	9	1.64	0.93
P36	γ_{j1}	27	1.81	1.22
R12	γ_{j2}	6	-0.27	2.14
R36	γ_{j2}	18	-0.06	2.81
P12	γ_{j2}	6	-1.04	2.79
P36	γ_{j2}	18	-0.31	2.82
R12	γ_{j3}	3	-0.84	0.54
R36	γ_{j3}	9	-0.47	0.79
P12	γ_{j3}	3	-0.41	0.19
P36	γ_{j3}	9	-0.65	0.62

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 7

Descriptive Raw Score Statistics for Tests Analyzed in Study II

Test	N	Raw Score Mean	Raw Score SD	Mean OPM[1]	Coeff. α	Item-Test Correlations Mean	Item-Test Correlations SD	Item-Test Correlations Min	Item-Test Correlations Max
R12	2000	6.94	5.05	0.33	.78	0.52	0.18	0.25	0.78
R36	2000	23.04	12.43	0.37	.90	0.42	0.12	0.26	0.68
P12	2000	8.79	4.40	0.42	.72	0.49	0.12	0.31	0.75
P36	2000	24.58	12.05	0.39	.89	0.44	0.18	0.10	0.76

[1] OPM refers to the Observed Proportion of Maximum, which is the mean raw score divided by maximum possible score.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Table 8

Summary of Standardized Q_1 Fit Results in Study II

Test/ Model Comb.	N	No. Of Items	Mean Z	SD Z	Percentage Distribution of Z's				
					$Z<2$	$2\le Z<3$	$3\le Z<5$	$5\le Z<10$	$10\le Z$
R12	2,000	12							
1PL/1PPC			2.57	1.60	33	8	58	0	0
3PL/2PPC			3.01	2.62	42	17	25	17	0
R36	2,000	36							
1PL/1PPC			-0.58	0.93	100	0	0	0	0
3PL/2PPC			0.03	1.47	92	0	8	0	0
P12	2,000	12							
1PL/1PPC			3.70	2.39	33	0	17	50	0
3PL/2PPC			3.20	2.96	42	17	8	33	0
P36	2,000	36							
1PL/1PPC			3.51	5.09	47	8	14	22	8
3PL/2PPC			-0.32	0.80	100	0	0	0	0

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

discriminations of MC and CR items differed substantially from one another, and that within item type they differed also. It is likely that the considerable variability in item discrimination, as well as the guessing on the MC items, produced the poorer fit performance by the 1PPC model alone or in combination with the 1PL model.

This implies that, were the 1PL-1PPC model combination used for scaling MC and CR items, the different discriminations of these two item types might make one or the other appear to fit poorly. Closer study of some results for the Grade 11 mathematics test that we analyzed illustrates this point: Of the 7 CR items in this 46-item test, 4 had Z-statistics above 9.00 when scaled using the 1PPC model.

Faced with an item pool containing poor-fitting items, the testing director has two alternatives; both are distasteful. First, the fit of items could be ignored. However, misfit is evidence that an item does not fit the model. When misfit occurs, the parameters for the item do not correctly characterize the item. If these parameters are used, say, to assess bias in the item or to anchor the item in an equating, errors will be introduced to the degree that the parameters fail to capture how the item actually behaves. Alternatively, the poor-fitting items could be removed from the item pool. However, this might seriously deplete the entire pool or eliminate large numbers of particular types of items.

In the simulation study, the percentages of items with good fit tended to be larger when the 3PL-2PPC model combination was used. Also, this more flexible model tended to produce better item fit across data sets with dissimilar properties. Thus, in the absence of knowledge about items' true parameters, it seems that choosing this more flexible model combination will provide greater opportunities for good model-data fit.

This conclusion is based on the analyses done here. Certainly further research is warranted. Practical constraints limited the availability of real data that we could examine and the variety of simulations that we could conduct. The topic of appropriate models for test data is an important one, and it deserves further attention by measurement experts.

Notes

[1]OPM is equal to the mean raw score on the test divided by the maximum score obtainable on the test.

[2]An item OPM is equal to the mean raw score on an item divided by the maximum score obtainable on the item.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Burket, G. R. (1991). PARDUX [Computer software]. Monterey, CA: CTB Macmillan/McGraw-Hill.

Burket, G. R. (1992). BSSTSIM1 [Computer software]. Monterey, CA: CTB/McGraw-Hill.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Chang, H., & Mazzeo, J. (1994). Empirical Bayes estimates of domain scores under binomial and hypergeometric distributions for test scores. *Psychometrika, 59,* 391–404.

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.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31,* 295–311.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147–200). New York: American Council on Education and Macmillan.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31–49). New York: Academic Press.

Isham, S. P., & Allen, N. L. (1993, April). *Scale anchoring of extended constructed-response items in NAEP.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Lord, F. M. (1983). Small *N* justifies the Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51–61). New York: Academic Press.

Loyd, B., Engelhard, G., & Crocker, L. (1993, April). *Equity, equivalence, and equating: Fundamental issues and proposed strategies for the National Board for Professional Teaching Standards.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18,* 245–256.

Mazzeo, J., Yamamoto, K., & Kulick, E. (1993, April). *Extended constructed-response items in the 1992 NAEP: Psychometrically speaking, were they worth the price?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11,* 39, 20.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363.

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

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Stone, C. A., Ankenmann, R. D., Lane, S., & Liu, M. (1993, April). *Scaling QUASAR's performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions

Thissen, D. L., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

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). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36,* 19–48.

van de Vijver, F. J. R. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement, 10,* 45–57.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339–368.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103–118.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14,* 219–226.

Yen, W. M. (1984). Effects of local independence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

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

Yen, W. M. (1986). Choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–325.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Authors

ANNE R. FITZPATRICK is Senior Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940-5703; afitzpat@mcgraw-hill.com. *Degrees:* BA, Mount Holyoke College; MEd, EdD, University of Massachusetts at Amherst. *Specialization:* psychometric methods.

VALERIE B. LINK is Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; vlink@mcgraw-hill.com. *Degrees:* BS, MS, PhD, Florida State University. *Specialization:* IRT scaling and equating.

WENDY M. YEN is Vice President, Research and Measurement Services, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wyen@mcgraw-hill.com. *Degrees:* AB, MA, PhD, University of California, Berkeley. *Specializations:* item response theory, test theory.

GEORGE R. BURKET is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; gburket@mcgraw-hill.com. *Degree:* PhD, University of Washington. *Specializations:* scaling, norming, item response theory.

KYOKO ITO is Research Scientist I, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; kito@mcgraw-hill.com. *Degree:* PhD, University of Illinois at Urbana-Champaign. *Specialization:* educational measurement.

ROBERT C. SYKES is Senior Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940. *Degrees:* BS, Michigan State University; MSPH, University of North Carolina; PhD, University of Chicago. *Specializations:* methodology of the behavioral sciences, psychometrics.

This content downloaded from 128.184.220.23 on Sun, 01 Nov 2015 04:43:05 UTC
All use subject to JSTOR Terms and Conditions