MCAS 2001 Grade 10 ELA and Mathematics Model Fit Analyses[1,2]

Ning Han[3]

University of Massachusetts Amherst

December 8, 2003

---

[1] Center for Educational Assessment MCAS Validity Report No. 8. (CEA-540). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

[2] This work was carried out under a contract between the University of Massachusetts Center for Educational Assessment and Measured Progress, Dover, New Hampshire.

[3] The author would like to thank Professor Ronald Hambleton for his direction and suggestions in completing this piece of research for the Massachusetts Department of Education.

1. Overview of Analyses

Background

Item response theory (IRT) has been employed with the MCAS tests to equate scores across different years, analyze item properties, and serve many other psychometric purposes. IRT offers quite a few advantages over classical test theory. However, unless the IRT model adequately fits the data, the benefits of IRT methods may not be realized. Unfortunately, we can never predetermine if a model fits a specific data set. Therefore, the appropriateness of the specified IRT model with the test data set of interest should be established by conducting a suitable goodness-of-fit investigation before any further work is carried out. The purpose of the current work is to apply a new method (to our on-going analyses of MCAS data) based on checking the predicted score distribution and the observed one to assess the fitness of IRT models to MCAS data. In prior research the focus has been on the study of item residuals. This investigation extends earlier work.

Methodology

The model-data fit usually is addressed in two ways. First, the data must conform to the model assumptions such as unidimensionality. Second, the predictive capability of the model should be examined. That is, the predictions from the model should be checked with observed data to see whether the predictions are approximately correct.

Dimensionality can be checked by some widely used general statistical procedures, such as factor analysis, principal component analysis, multidimensional scaling, etc. One straightforward approach is to compute and plot the eigenvalues of the item score response matrix. Usually when the first eigenvalue is bigger than 20% of the sum of the eigenvalues, the data set can be regarded as unidimensional. In the current

work, the eigenvalues of the response matrices are computed by SPSS and plotted by MS EXCEL.

There are many recommended approaches to address the predictions of score distributions from a model. The procedure used in the current research is to compare the observed distributions of the raw scores with the theoretical distributions predicted from the item parameter estimates and ability estimates. This approach was used first by Hambleton and Traub (1973) for dichotomous items. Their general procedures were: (1) The conditional distribution of the test scores for a fixed trait level is obtained by a compound binomial distribution (for polytomous items, the distribution of the conditional probabilities is a compound multinomial distribution.). (2) The expected frequency of examinees having a given score is obtained. (3) The expected frequency and the observed frequency for the group of examinees are compared. See Ferrando and Lorenzo-Seva (2001) for a detail description.

One disadvantage of the approach is its complexity of computation. To compute the conditional probabilities theoretically, a Lord-Wingskey recursive formula (Lord & Wingersky, 1984) is used. For polytomous items, an extension of the formula, which was given by Wang, Kolen, & Harris (2000), is employed instead. Furthermore, even though the conditional probabilities can be obtained theoretically by means of the formula, another serious difficulty will arise in the practice of large-scale education measurements. Table 1.1 is a frequency distribution of the examinees on items 36 to 41 on the 2001 MCAS grade 10 Mathematics test. A considerable amount of data is missing. This is not uncommon for constructed response items. It is obvious that the score distribution

predicted by the Lord-Wingersky formula will differ from the observed one. Therefore, an alternative to Lord-Wingersky formula was necessary.

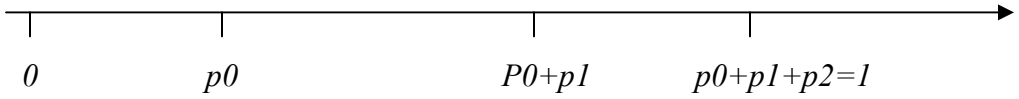
Table 1.1. Frequency Distribution of Examinees on Items 36 to 41, MCAS Grade 10 Mathematics (N=61045)

Item	Score					
	0	1	2	3	4	Missing
36	1643	733	645	4921	3486	22322
37	3026	585	744	2699	3975	38053
38	1651	199	531	4961	6312	40818
39	8282	172	569	2658	8291	34394
40	3687	337	163	5634	802	31200
41	2306	1471	6130	1847	2583	46708

The general idea of the revised method is: Given that the item parameter estimates and ability estimates can be obtained from the calibration of the response matrix, a new response matrix that is consistent with the item parameter estimates and the model can be simulated by the Monte-Carlo method. With the simulated data, test scores for candidates can be calculated by simply summing the item scores and then the distribution of test scores (expected test score distribution under the hypothesis of known IRT model) can be produced. This distribution is then compared to the actual one. Through this method, the missing values can be taken into account and the simulation can be replicated a great number of times so that a reference criterion (confidence intervals) can be set up to interpret the results. As an alternative, a chi-square goodness of fit test and/or Kolmogorov-Smirnov test can be performed as well to assess the difference between the predicted distribution and the observed one.

Since there are both dichotomous items and polytomous items in MCAS tests, mixed IRT models should be employed. Among the widely used models, the 1PL, 2PL, and 3PL logistic models can be used for dichotomous items. Master's Partial Credit Model, Samejima's Graded Response Model, and Muraki's Generalized Partial Credit Model can be used for polytomous items. In the grade 10 Mathematics test, all the polytomous items share a common scale (0-4), therefore, all different combinations of the models (1PL/GRM, 1PL/PCM, 1PL/GPCM, 2PL/GRM, 2PL/PCM, 2PL/GPCM, 3PL/GRM, 3PL/PCM, 3PL/GPCM) will be investigated. In grade 10 ELA, four polytomous items are scored from 0 to 4, one is scored from 2 to 12 and the other one from 2 to 8. PCM can not handle this type of data set. The combinations of the suitable models are 1PL/GRM, 1PL/GPCM, 2PL/GRM, 2PL/GPCM, 3PL/GRM, and 3PL/GPCM.

While the Monte-Carlo simulation of the dichotomous items is routine in IRT research, the simulation of the polytomous item is a little bit tricky. Consider the following example illustrated by the axis: an item is scored 0, 1, and 2 and the probabilities of each score point are *p0*, *p1*, and *p2* (*p0 + p1 + p2 = 1*).



A [0, 1] uniform distribution random number is generated and compared to the probabilities. If the random number generated is less than *p0*, the examinee is scored *0*; if the random number is bigger than *p0* and less than *p0+p1*, the examinee is scored *1*; finally, if the random number is bigger than *p0+p1*, the examinee is scored 2.

Another problem concerns the base examinee group on which the fitness is assessed. Traditionally, the fitness is assessed on an assumed ability distribution. But the abilities or scores obtained from large-scale education assessments, such as MCAS, are seldom distributed normally. An investigation based on the assumed normal distribution is inappropriate. The current work will assess the model data fit using a random sample of the ability distribution observed in the analysis, which will ensure its generalizability. A sample was drawn to avoid carrying out the study with over 60,000 candidates.

2. Item Statistics

A sample of the item statistics that were obtained with two of the models follows:

Table 2.1. Item Statistics (3PL/2PL/GRM, Grade 10 Mathematics)

Item	a	b	c	Threshold Parameters			
Mcc35801	.4466	-.5738	.2280				
Mcc48709	1.1460	-.9192	.2944				
Mcc48790	.7288	-1.0685	.2494				
Mcc43583	.8293	1.0826	.3198				
Mcc35796	.6788	-1.4585	.0000				
Mcc43587	1.3254	.3592	.2765				
Mcc48714	1.4752	.5630	.0883				
Mcc48764	.9553	.5812	.2291				
Mcc48792	1.0934	.7676	.1346				
Mcc48710	.8867	.0745	.2585				
Mcc48739	.9416	.1826	.2512				
Mcc48779	1.0021	1.1982	.2500				
Mcc48708	1.2989	.1385	.2293				
Mcc48706	.9471	-1.2559	.2239				
Mcc48707	.5767	-1.4478	.2143				
Mcc48711	.9986	-.3792	.2107				
Mcc43606	.9185	-.0298	.1875				
Mcc35874	1.5476	.2345	.2515				
Mcc48718	1.7093	1.0065	.3149				
Mcc19107	1.6046	.5223	.1212				

Item	a	b	c	Threshold Parameters			
Mcc48717	1.7201	.1018	.1826				
Mcc48742	1.3239	.2384	.1895				
Mcc48716	1.1748	-.2464	.1531				
Mcc51011	1.2560	-.0626	.3168				
Mcc18968	1.9440	.4412	.1548				
Mcc51016	1.3901	1.6153	.3817				
Mcc48769	1.0168	.0409	.1853				
Mcc35866	1.2059	.2579	.1695				
Mcc48794	.7143	1.6374	.2783				
Mcc48762	1.1303	.1331	.2852				
Mcc48788	.5384	.7427	.2158				
Sac27609	.5320	-.6708	.0000				
Sac43723	1.0260	-.5011	.0000				
Sac35884	1.4489	.4342	.0000				
Sac48719	.7881	.8308	.0000				
Orc26725	1.1535	.7012	.0000	.9385	.2761	-.2294	-.9851
Orc48637	1.5406	-.1050	.0000	1.4088	.3322	-.6391	-1.1019
Orc48639	.9131	-.9666	.0000	1.6147	.6934	-.6363	-1.6718
Orc35916	1.2236	-.2164	.0000	.6322	.3728	-.3300	-.6750
Orc43744	.9503	.2836	.0000	2.0798	1.3513	-.8590	-2.5721
Orc48636	1.2367	-.1093	.0000	1.3053	.8224	-.7902	-1.3375

Table 2.2. Item Statistics (3PL/GRM, Grade 10 ELA)

Item	**a**	**b**	**c**	**Threshold Parameters**									
Mcc34972	1.0074	-1.8761	.2243										
Mcc34973	.2143	-4.3354	.1316										
Mcc34976	.8817	-1.4890	.1766										
Mcc44049	.7466	-1.8302	.1236										
Mcc34977	.8480	.3253	.2551										
Mcc44050	.6749	-.7004	.1726										
Mcc34975	.5754	.8950	.3200										
Mcc34899	.2832	-1.1612	.1595										
Mcc34901	.6086	-.7031	.1163										
Mcc42780	1.1477	-1.0648	.1243										
Mcc24652	.3374	.0623	.1229										
Mcc24656	.7586	-1.2352	.0761										

Item	a	b	c	Threshold Parameters									
Mcc50014	1.6569	-.2561	.1982										
Mcc46679	.9565	-1.7896	.1212										
Mcc46681	.4634	-1.7600	.1097										
Mcc46682	1.1682	-.2255	.2400										
Mcc47104	1.1344	-1.6510	.0886										
Mcc47304	.4717	.1033	.2549										
Mcc47306	.7779	-.5915	.3477										
Mcc47829	.6790	-1.8522	.0945										
Mcc47830	.5023	-.7437	.1590										
Mcc42773	.4858	.3169	.1438										
Mcc35066	1.0659	.4453	.1757										
Mcc35067	.8877	-1.5802	.1215										
Mcc42774	.3290	-.2613	.2000										
Mcc42775	.7994	-.7433	.1659										
Mcc35090	.5066	.7251	.2418										
Mcc42776	1.2933	-.8571	.1985										
Mcc54591	.2698	-1.6134	.1046										
Mcc23225	.8048	-2.1913	.1100										
Mcc43839	.7191	-1.5460	.2722										
Mcc43846	.7513	-1.2229	.1210										
Mcc44114	1.1330	-1.2284	.1495										
Mcc45969	.6673	-.2111	.1203										
Mcc46031	.2540	-1.9556	.1059										
Mcc46159	.8645	-1.0503	.1266										
Orc42782	.7817	-.2632	.0000	2.7534	1.4524	-.9849	-3.2208						
Orc24680	.9968	-.0949	.0000	1.2478	.5736	-.2979	-1.5235						
Orc47832	.8794	-.3071	.0000	2.3396	.7918	-.8561	-2.2753						
Orc45988	.9505	.1542	.0000	2.3001	1.0016	-.8403	-2.4614						
Orc42424	1.0586	-.1809	.0000	3.3244	2.8091	2.1905	1.5878	.5505	-.2675	-1.1934	-2.0432	-2.9009	-4.0574
Orc41414	1.0211	-1.4526	.0000	2.4145	1.6612	.6701	-.2123	-1.6915	-2.8419				

3. Eigenvalues

Eigenvalue plots for the grade 10 Mathematics and ELA tests follow. In both cases, approximately 20% of the variability is associated with the first factor.

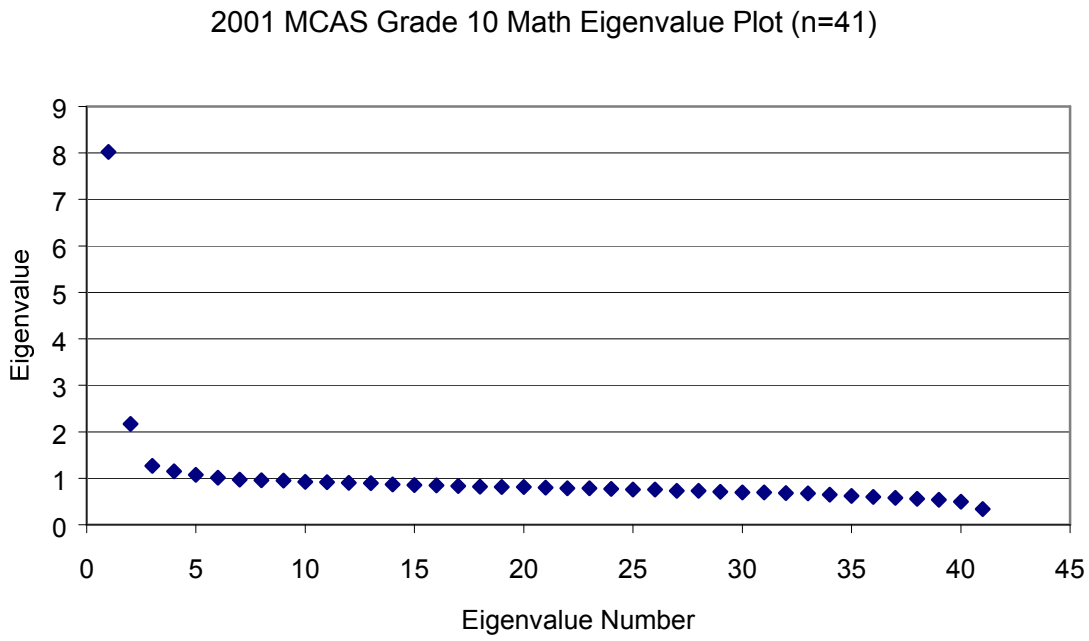


Figure 3.1. Eigenvalues (Grade 10 Mathematics)

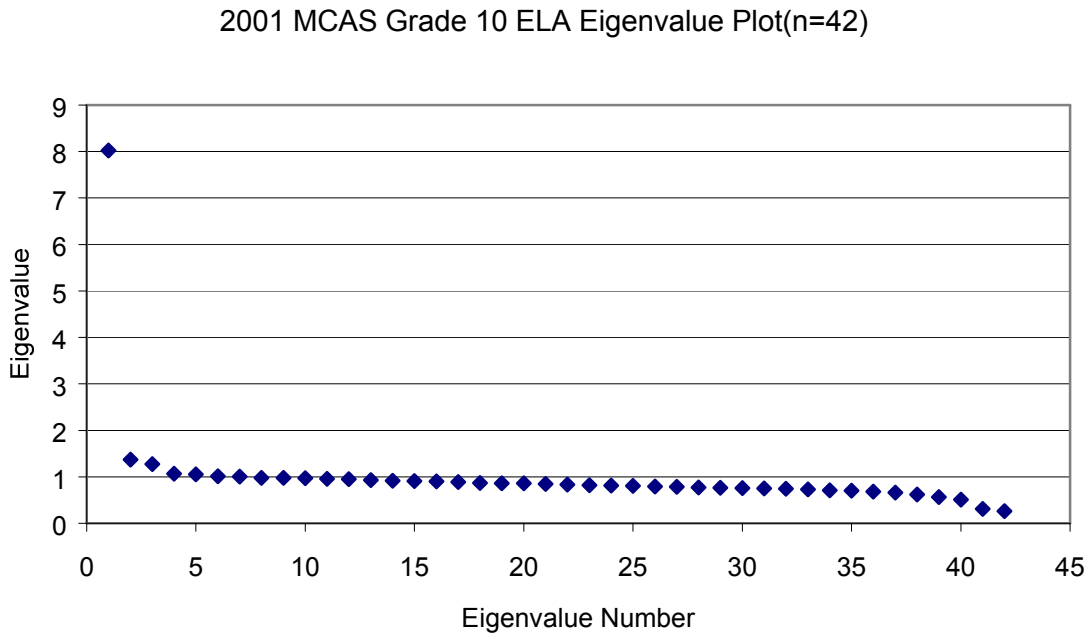


Figure 3.2. Eigenvalues (Grade 10 ELA)

4. Predicted Score Distributions Versus Observed Score Distribution (Mathematics)

In the displays that follow, various IRT models have been used to predict the mathematics observed score distribution. Both the predicted and the observed distribution are displayed.

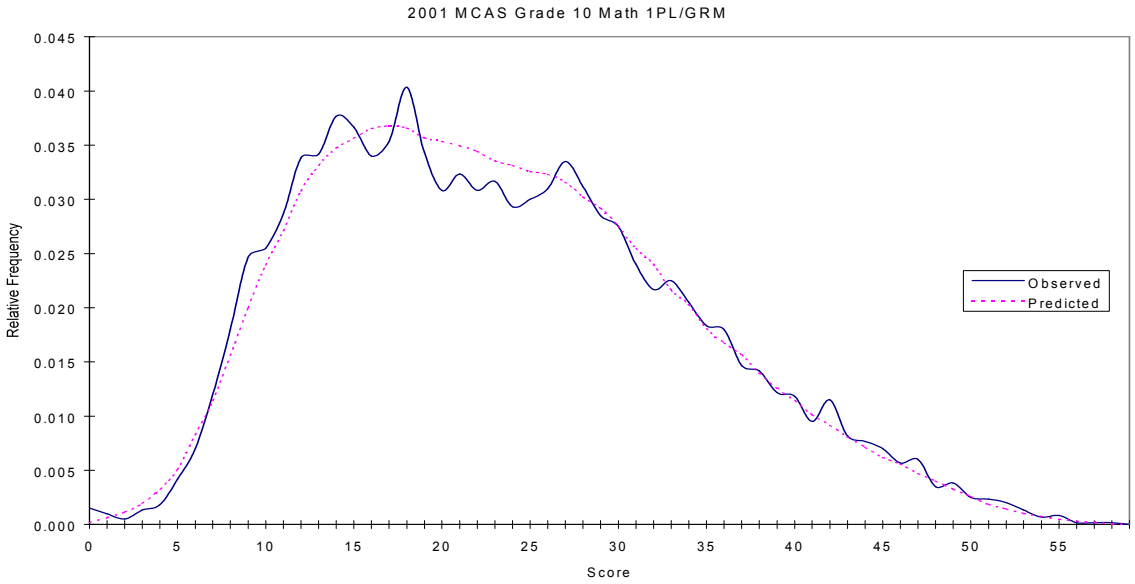


Figure 4.1. 1PL/GRM

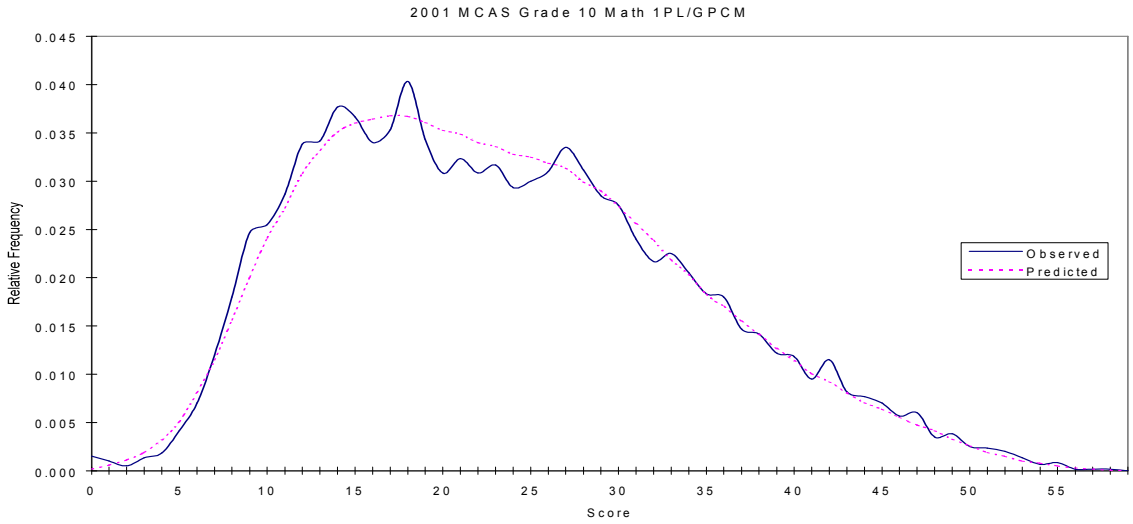


Figure 4.2. 1PL/GPCM

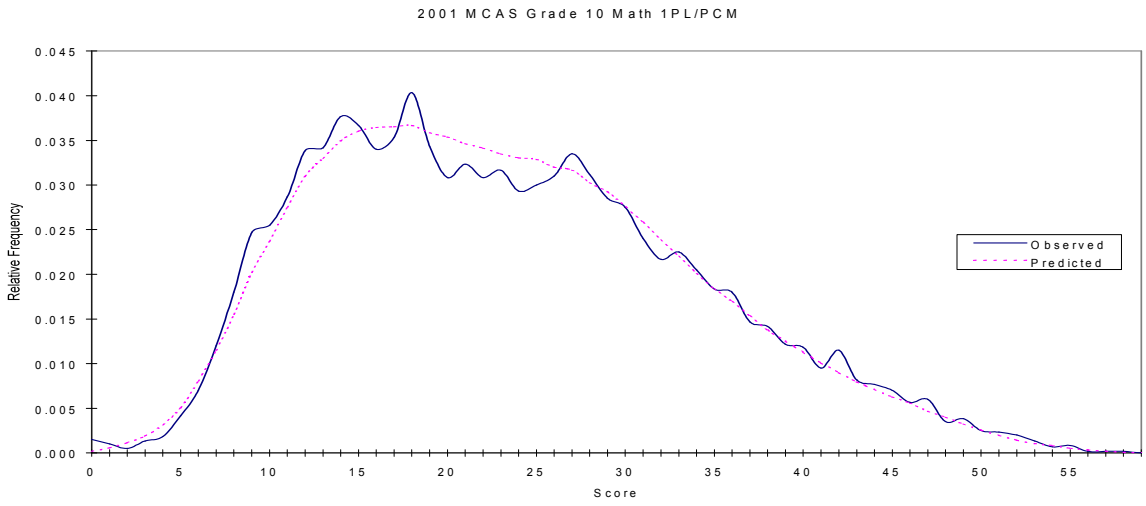


Figure 4.3: 1PL/PCM

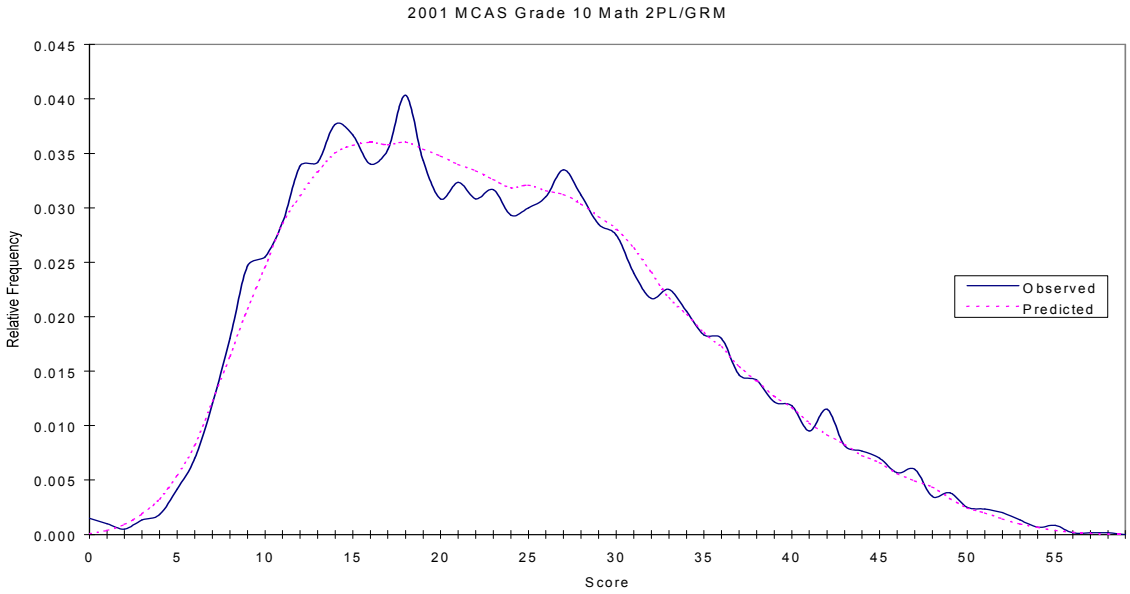


Figure 4.4: 2PL/GRM

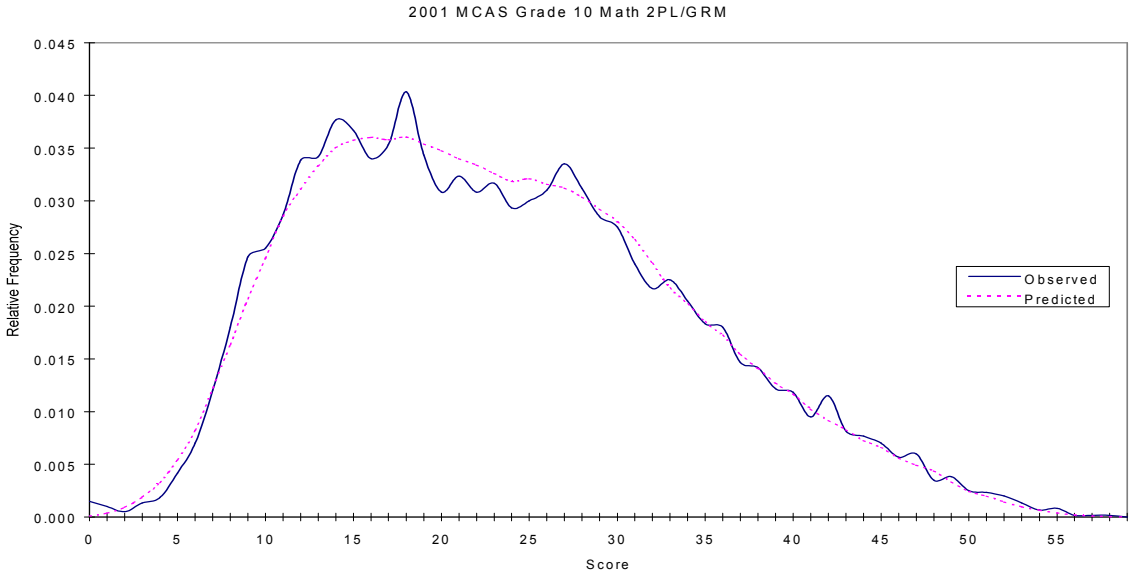


Figure 4.5. 2PL/GPCM

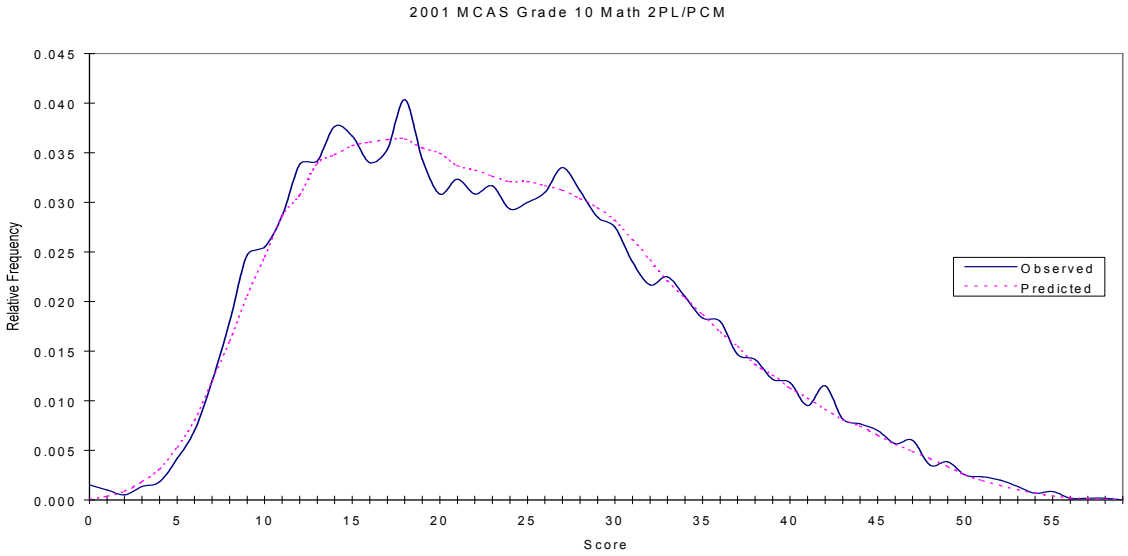


Figure 4.6. 2PL/PCM

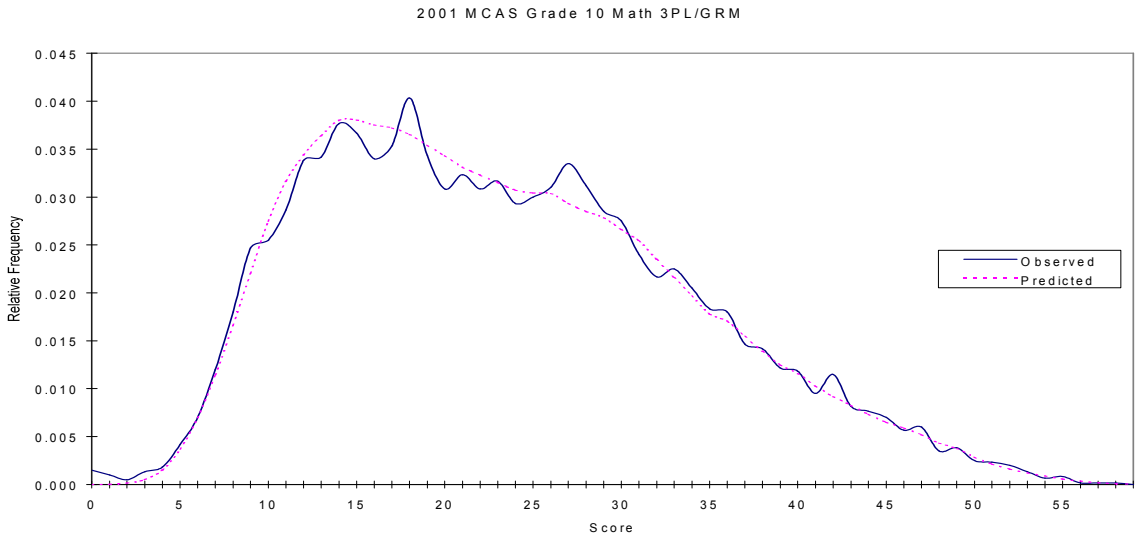


Figure 4.7. 3PL/GRM

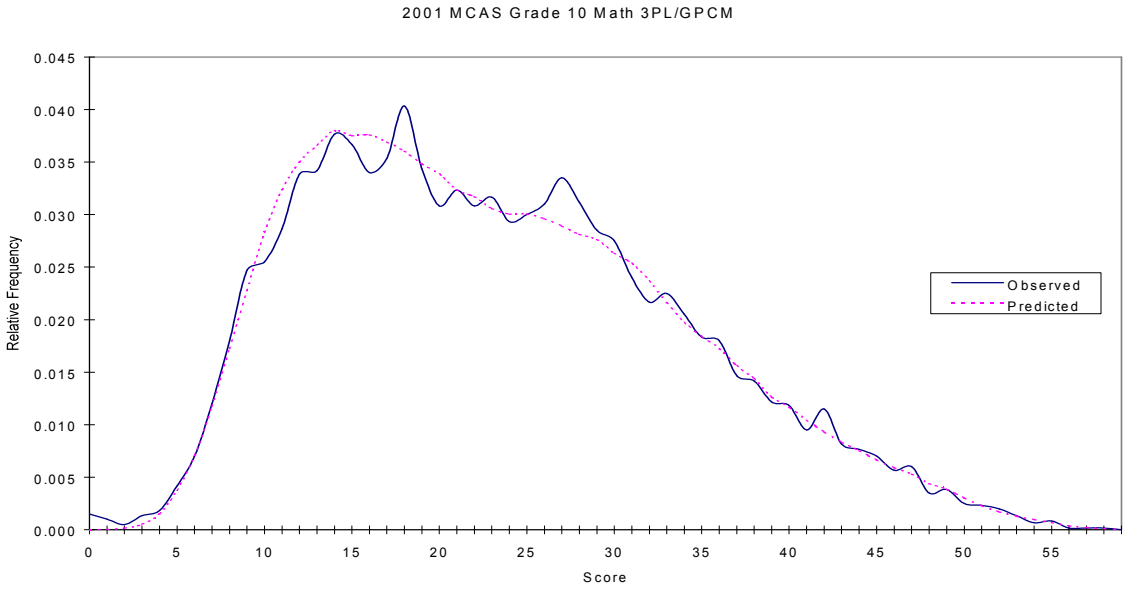


Figure 4.8. 3PL/GPCM

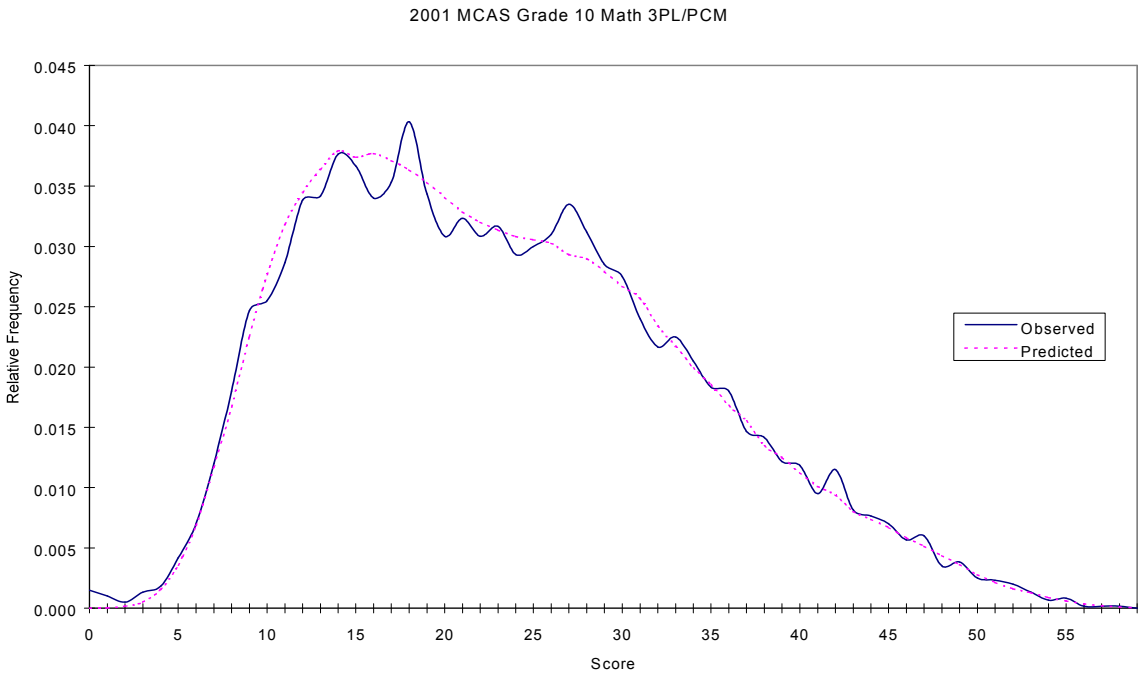


Figure 4.9. 3PL/PCM

5. Predicted Score Distributions Versus Observed Score Distribution (ELA)

In the displays that follow, various IRT models have been used to predict the ELA observed score distribution. Both the predicted and the observed distribution are displayed.

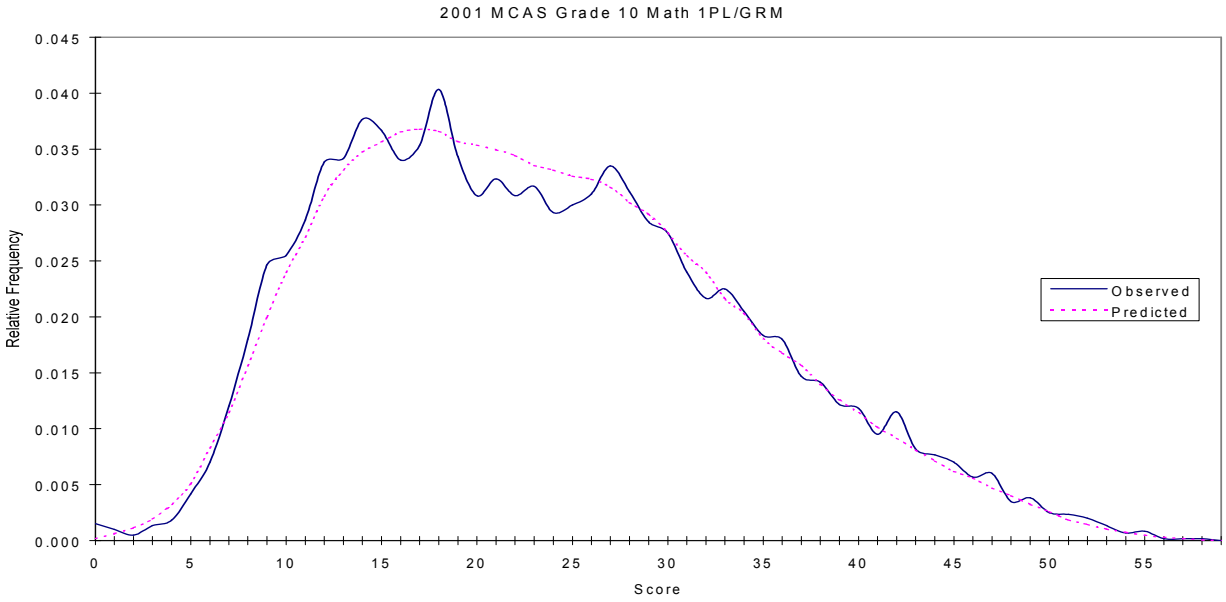


Figure 5.1. 1PL/GRM

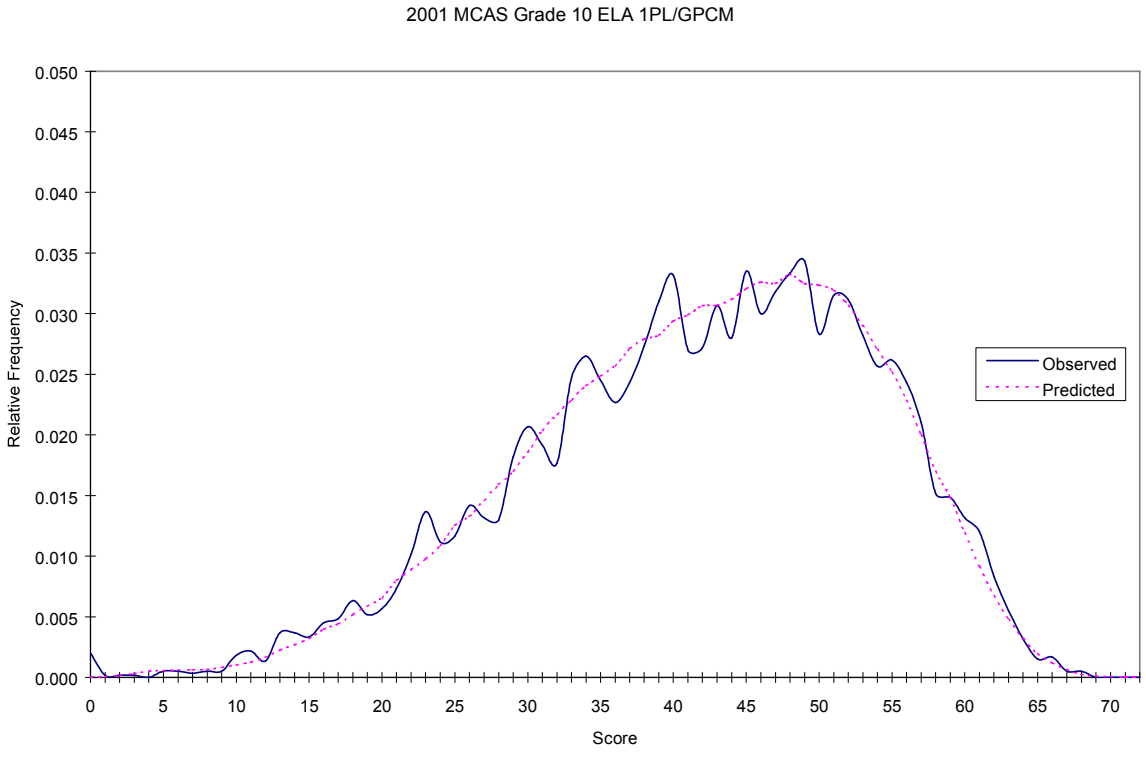


Figure 5.2. 1PL/GPCM

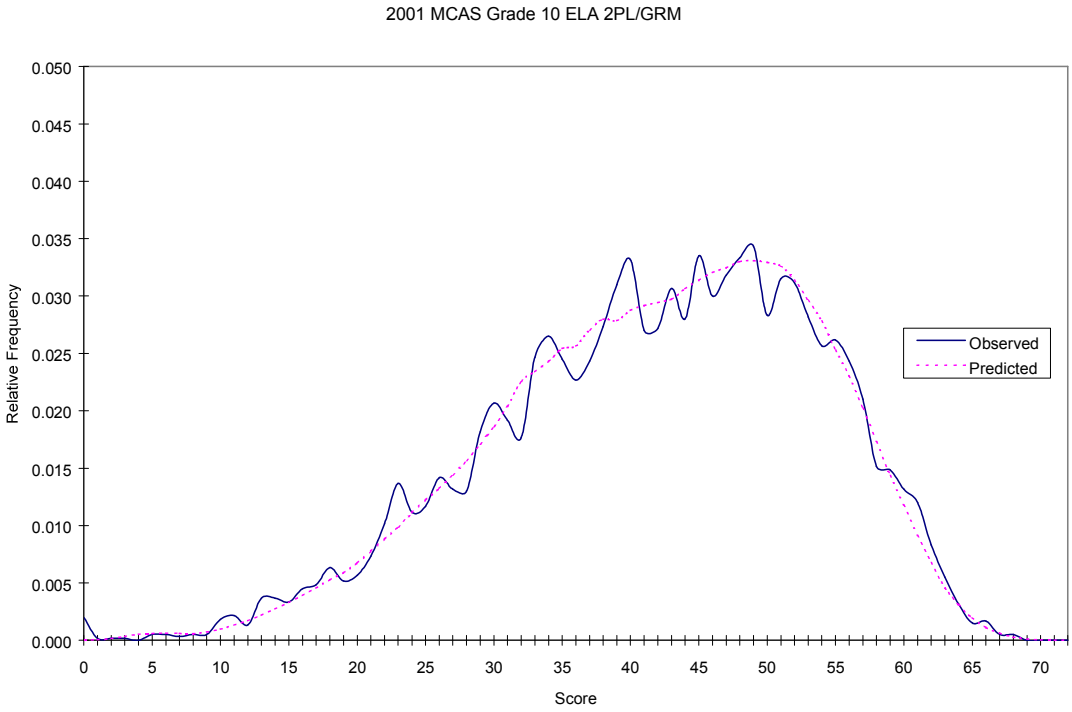


Figure 5.3. 2PL/GRM

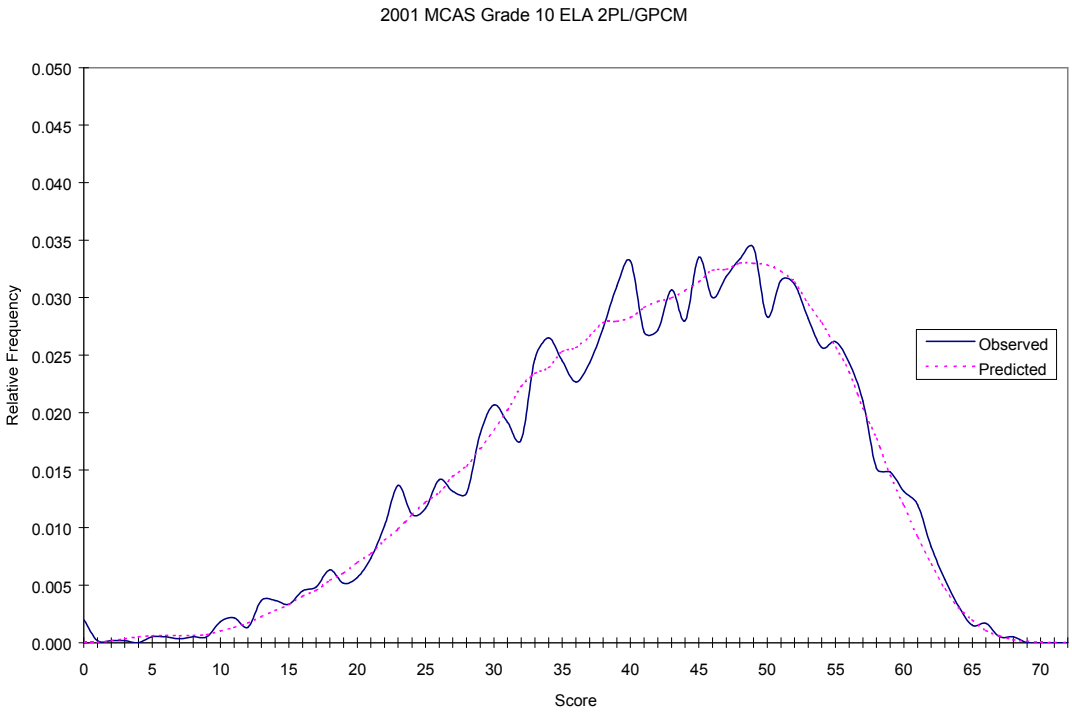


Figure 5.4. 2PL/GPCM

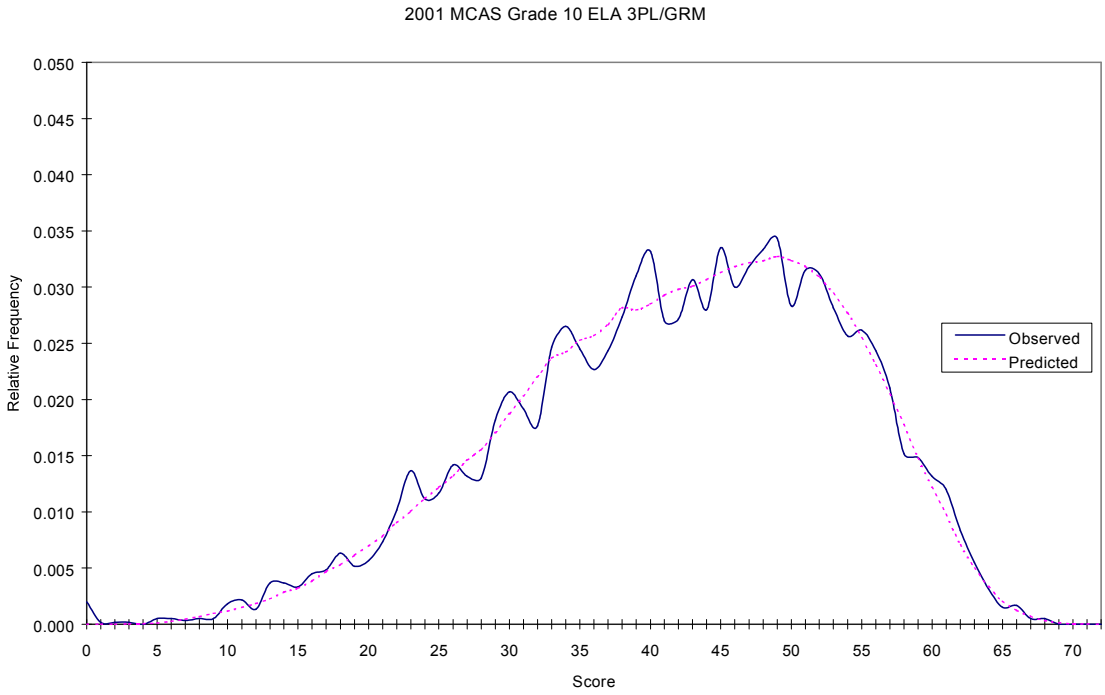


Figure 5.5. 3PL/GRM

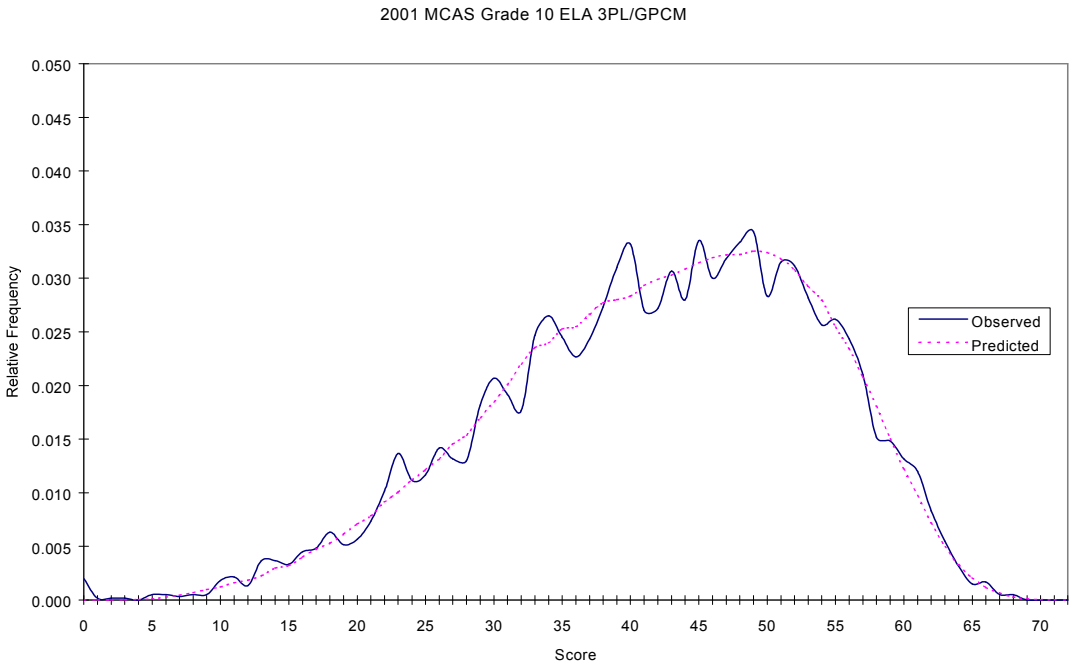


Figure 5.6. 3PL/GPCM

6. Chi-square Tests

Table 6.1. Chi-square Test for Mathematics

Model	Intervals	Chi-square
3PL/GRM	35	32.43
2PL/GRM	35	26.62
1PL/GRM	35	37.62
3PL/GPCM	35	32.90
2PL/GPCM	35	26.63
1PL/GPCM	35	34.72
3PL/PCM	35	32.01
2PL/PCM	35	27.22
1PL/PCM	35	35.22

Table 6.2. Chi-square Test for ELA

Model	Intervals	Chi-square
3PL/GPCM	38	58.83
2PL/GPCM	38	59.34
1PL/GPCM	38	59.60
3PL/GRM	38	60.29
2PL/GRM	38	61.05
1PL/GRM	38	61.47

These results confirm what is clear from sections 4 and 5 and that is that all of the models appear to fit the data fairly well.

7. Conclusions

Three conclusions can be drawn from the results reported in the last section:

1. The eigenvalue plots for the 2001 MCAS grade 10 math and ELA data show that the response data are approximately unidimensional. They meet one of the two main assumptions of IRT because there is a major first factor.
2. Several of the IRT models fit the MCAS data very well. Clearly IRT models can be used with the MCAS data.
3. The differences among the different polytomous IRT models in terms of model fit are practically insignificant.

References

Ferrando, P. J., & Lorenzo, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program EP-Fit. *Educational and Psychological Measurement, 61*(5), 895-902.

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

Lord, F. M., & Winkersky, N. (1984). Comparison of IRT true-score and equipercentile observed-score equating. *Applied Psychological Measurement, 8*, 452-461.

Wang, T., Kolen, M., & Harris, D. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141-163.