Mid-Western Educational Researcher

Volume 12 | Issue 2 | Article 4

1999

# Multimethod Analysis of Mathematics Achievement Tests

Dimiter M. Dimitrov
*Kent State University*

# *Multimethod Analysis of Mathematics Achievement Tests*

Dimiter M. Dimitrov
Kent State University

## *Abstract*

*Multimethod analysis of mathematics achievement tests is illustrated by combining psychometric and statistical methods in the analysis of results from the California Achievement Test-Mathematics administered to seventh-graders from North-East Ohio.*

Taken into account were the category objectives and thinking skill levels defined for the two parts of the test, Computations and Concepts and Applications. The goal is to provide educational analysts results they can use in making informed decisions about teaching mathematics within local educational settings.

Data related to validity, reliability, scaling, norming, and equating are commonly provided with nationally standardized mathematics achievement tests (see, e.g., CTB/ McGraw-Hill, 1986). However, the results reported for local student populations are usually limited to classical item parameters and descriptive statistics of students' scores on such tests. Additional test data at state and district levels may provide research analysts information they can use to further support their decisions about teaching mathematics in local educational environments.

The purpose of this paper is to provide information that may help in making informed decisions based on CAT-M results, by combining Item Response Theory (IRT) and statistical methods in the analysis of results from the California Achievement Test-Mathematics (CAT-M) administered to seventh-graders from North-East Ohio. This study addresses a number of questions:

1. Which IRT model fits the CAT-M data for the target population?
2. How does the CAT-M work at different ability levels?
3. Does the average item difficulty change across different category objectives and thinking skill levels of the CAT-M?
4. Is the relative standing of students the same across different CAT-M items?
5. How many items are needed per CAT-M category objective and thinking skill level in order to obtain given reliability?
6. How can students' abilities be predicted from CAT-M scores?

## Method

Results from the CAT-M (CTB/McGraw-Hill, 1985) of 4135 seventh-graders from a large urban area in North-East Ohio were used. The two parts of the CAT-M, Computation Test and Mathematics Concepts and Applications Test, were analyzed separately. The Computation Test included 50 items grouped by one factor, **Category Objective** (CO), with 10 levels: (1) Subtract fractions, (2) Multiply whole numbers, (3) Multiply decimals, (4) Multiply fractions, (5) Divide whole numbers, (6) Divide decimals, (7) Divide fractions, (8) Integers and percents, (9) Subtraction of whole numbers and decimals, and (10) Addition of whole numbers, decimals, and fractions (CTB/McGraw-Hill, 1986).

The Concepts and Applications Test included 55 items grouped by two factors. The first factor, Category Objective (CO), has six levels, (1) Numeration, (2) Number Sentences, (3) Number Theory, (4) Problem Solving, (5) Measurement, and (6) Geometry. The second factor, Thinking Skill (TS), has three levels, (1) Recall and recognition, (2) Inference, and (3) Evaluation.

The IRT analysis included the calculation of (a) data fit statistics, (b) item and test characteristics,(c) students' ability scores, and (d) descriptive statistics for test scores of students with different abilities. The computer programs RASCAL (Assessment Systems Corporation, 1995a) and XCALIBRE (Assessment Systems Corporation, 1995b)were used for the IRT analysis, while SPSS (SPSS Inc., 1997) and MicroFACT (Waller, 1995) were used for the statistical analysis.

A two-way unbalanced ANOVA was conducted for the Concepts and Applications Test with two fixed factors, CO and TS, with the dependent variable being the IRT difficulty of the items. It was performed through the SPSS procedure MANOVA/METHOD = SEQUENTIAL. Of special interest was the interaction between the two factors in order to see if the difference between the average item difficulties of different category objectives varied across the three thinking skill levels.

To answer the research question related to the prediction of students' abilities on CAT-M scores, a regression analysis was conducted with the independent variable being the test score and the dependent variable being the ability score. The ability scores of all 4135 students were calculated XCALIBRE.

Generalizability theory study (G-study) and related decision study (D-study) were conducted for the CAT-M tests by the use of the GENOVA program (Crick and Brennan, 1983). For the Computation Test, students (S) were the object of measurement and items (I) represented a random facet nested within the fixed facet Category Objective (CO). Thus, the appropriate G-study design in this case was the partially nested design S x (I:CO) (see, e.g., Shavelson and Webb, 1991, p. 75). With the Concepts and Applications Test, a G-study was conducted for the partially nested design S x (I:TS), with items nested within the fixed facet Thinking Skill (TS).

Related D-studies were conducted with both the S x (I:CO) and S x (I:TS) designs for the estimation of the GT coefficients $E\rho^2$ and $\Phi$. The **generalizability coefficient**, $E\rho^2$, is analogous to the reliability coefficients in classical test theory. It is suitable for decisions about the relative standing of students on the test scale. The **index of dependability**, $\Phi$, introduced by Brennan and Kane (1977) as a generalizability index for absolute decisions, is suitable for criterion-referenced analysis and decisions (see, e.g., Shavelson and Webb, 1991, pp. 83-97).

## Results

The IRT assumption about unidimensionality of the data was tested using MicroFACT (Waller, 1995), which performs the iterated principal factor analysis on tetrachoric correlations for binary response data. The results indicated the presence of a dominant factor underlying the students' performance on each test. For the Computation Test, 36.72% of the total variance was explained by the first factor versus 1.54 % explained by the second factor. For the Concepts and Applications Test, this ratio was 42.46 % versus 0.48% in favor of the first (dominant) factor.

The results of the IRT analysis showed that the one-parameter IRT (Rasch) model did not fit the CAT-M data. The RASCAL $\chi^2$ fit statistic indicated misfit of 44 items from the Computation Test and 45 items from the Concepts and Applications Test, with $\chi^2$ values of those items exceeding the critical value, $\chi^2(19)= 30.14$, at the level of significance $\alpha = .05$.

For data fit of the 2- or 3-parameter IRT models, XCALIBRE reported a standardized residual statistic for each item. This statistic is normally distributed and values in excess of 2.0 indicate misfit with a type I error rate of 0.05. The results showed that the data did not fit the 2-parameter IRT model. Standardized residuals in excess of 2.0 for 8 items from the Concepts and Applications Test and 20 items from the Computation Test were found. For each test, the data fit the 3-parameter IRT model because none of the standardized residuals exceeded 2.0.

The internal consistency reliability of each test was found to be 0.90. The information curves of the two tests are given in Figure 1. The average amount of information provided by the Computation Test was found to be 9.31 versus 7.39 provided by the Concepts and Applications Test. Thus, for the local population of seventh-graders, the Computation Test provided more accurate estimates of students' abilities as compared to the Concepts and Applications Test (see, e.g., Allen and Yen, 1979, pp. 262-267). This is especially true for students with ability scores between 0.0 and

Table 1
*Item Parameter Estimates for the Computation Test*

| | Parameter | | | | | Parameter | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | *a* | *b* | *c* | PC | Item | *a* | *b* | *c* | PC |
| 1 | .43 | -2.55 | .14 | 87 | 26 | .43 | -1.37 | .14 | 75 |
| 2 | .65 | -2.11 | .14 | 89 | 27 | 1.45 | .80 | .20 | 42 |
| 3 | .46 | -2.00 | .14 | 94 | 28 | .37 | .44 | .17 | 54 |
| 4 | .77 | .07 | .15 | 56 | 29 | 1.77 | 1.37 | .12 | 23 |
| 5 | 1.64 | .56 | .17 | 45 | 30 | 1.96 | 1.14 | .09 | 23 |
| 6 | .67 | -.90 | .13 | 73 | 31 | .75 | -1.36 | .13 | 81 |
| 7 | .54 | -.37 | .13 | 63 | 32 | 1.00 | -1.13 | .12 | 81 |
| 8 | .74 | -.53 | .14 | 68 | 33 | .76 | -.66 | .11 | 69 |
| 9 | .91 | .28 | .13 | 50 | 34 | .95 | -.58 | .12 | 70 |
| 10 | .94 | .35 | .14 | 50 | 35 | .89 | -.34 | .13 | 65 |
| 11 | .53 | -2.41 | .14 | 88 | 36 | .86 | -1.26 | .13 | 82 |
| 12 | 1.65 | .45 | .14 | 46 | 37 | .87 | -1.57 | .14 | 86 |
| 13 | 1.76 | .80 | .10 | 33 | 38 | .97 | -.79 | .13 | 75 |
| 14 | 1.86 | .42 | .13 | 45 | 39 | .93 | -.29 | .13 | 64 |
| 15 | 1.83 | .67 | .11 | 37 | 40 | 1.07 | 1.62 | .13 | 23 |
| 16 | .67 | -1.52 | .13 | 82 | 41 | 2.11 | 1.11 | .13 | 28 |
| 17 | .62 | -1.80 | .13 | 85 | 42 | 2.31 | 1.20 | .17 | 30 |
| 18 | .78 | -1.34 | .13 | 82 | 43 | 1.81 | 1.28 | .17 | 28 |
| 19 | .79 | -1.13 | .13 | 79 | 44 | 2.16 | 1.30 | .12 | 23 |
| 20 | .71 | -.42 | .12 | 65 | 45 | 1.48 | 1.12 | .19 | 34 |
| 21 | .47 | -2.61 | .14 | 88 | 46 | .83 | 2.44 | .16 | 22 |
| 22 | .66 | -1.17 | .14 | 78 | 47 | .84 | 1.49 | .12 | 27 |
| 23 | .71 | -.61 | .14 | 69 | 48 | .34 | -.09 | .15 | 59 |
| 24 | .91 | .49 | .16 | 48 | 49 | .93 | 1.80 | .16 | 26 |
| 25 | .78 | .39 | .15 | 50 | 50 | 1.09 | 2.51 | .09 | 11 |

*Note*: Used was the 3-parameter IRT model, with *a* = discrimination parameter, *b* = difficulty parameter, and *c* = "guessing"

| | Parameter | | | | | Parameter | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | *a* | *b* | *c* | PC | Item | *a* | *b* | *c* | PC |
| 1 | .68 | -3.00 | .18 | 97 | 29 | .91 | -.94 | .15 | 77 |
| 2 | .58 | -2.32 | .17 | 89 | 30 | .72 | -.18 | .18 | 63 |
| 3 | .83 | -2.50 | .17 | 95 | 31 | .70 | -.18 | .19 | 63 |
| 4 | .93 | -.41 | .18 | 68 | 32 | .87 | .10 | .16 | 56 |
| 5 | .76 | -1.70 | .18 | 87 | 33 | .89 | .23 | .17 | 54 |
| 6 | .83 | -1.81 | .17 | 89 | 34 | .46 | .78 | .19 | 49 |
| 7 | .70 | -2.01 | .17 | 89 | 35 | .97 | -.19 | .16 | 62 |
| 8 | .76 | -1.51 | .18 | 85 | 36 | 1.26 | 1.09 | .18 | 35 |
| 9 | .51 | -1.88 | .18 | 84 | 37 | .93 | -.54 | .17 | 70 |
| 10 | .61 | -.97 | .18 | 75 | 38 | .99 | .64 | .18 | 45 |
| 11 | .54 | -.79 | .18 | 71 | 39 | 1.02 | -.51 | .16 | 70 |
| 12 | .65 | -.60 | .18 | 70 | 40 | .58 | .10 | .17 | 57 |
| 13 | .76 | -1.39 | .18 | 83 | 41 | .73 | .67 | .18 | 47 |
| 14 | .67 | -1.38 | .17 | 82 | 42 | .77 | 1.03 | .17 | 78 |
| 15 | .74 | -.35 | .18 | 66 | 43 | .51 | 1.44 | .17 | 79 |
| 16 | .70 | -1.10 | .18 | 78 | 44 | .81 | -.16 | .16 | 61 |
| 17 | .56 | -.72 | .17 | 70 | 45 | .86 | .69 | .21 | 47 |
| 18 | .80 | -.95 | .16 | 76 | 46 | 1.49 | .73 | .14 | 38 |
| 19 | .33 | .99 | .21 | 51 | 47 | .90 | .62 | .16 | 45 |
| 20 | 1.13 | -.58 | .18 | 73 | 48 | 1.20 | .91 | .20 | 40 |
| 21 | 1.00 | -.79 | .18 | 76 | 49 | .86 | .56 | .14 | 45 |
| 22 | .64 | -.24 | .17 | 63 | 50 | .88 | 1.61 | .15 | 27 |
| 23 | .55 | .33 | .19 | 54 | 51 | .99 | .28 | .17 | 52 |
| 24 | .56 | .13 | .20 | 58 | 52 | 1.04 | 2.17 | .13 | 19 |
| 25 | .62 | .42 | .18 | 52 | 53 | .88 | 1.15 | .16 | 35 |
| 26 | .83 | -.40 | .18 | 67 | 54 | 1.05 | 1.64 | .17 | 28 |
| 27 | .79 | -.74 | .18 | 73 | 55 | 1.20 | 2.05 | .12 | 17 |
| 28 | .94 | -1.52 | .17 | 87 | | | | | |

2.0 on the logit scale, i.e. students above the average and below the top on the ability range of the target population. Beyond this interval, both tests do not work particularly well.

Table 1 provides estimates of *a* (discrimination parameter), *b* (difficulty parameter), and *c* ("guessing parameter") for the Computation Test. The table also shows the percent of correct answers (PC) for each item, based on 4135 students. The item difficulties were spread without any big gaps within the logit interval (-2.61 to 2.51). The item discrimination power varied within the relatively large interval (0.37 to 2.31). The "guessing" parameter, *c*, was quite small in magnitude and variability. This indicates that, for each item, there is small probability for students with low ability to answer the item correctly. The same pattern of findings was observed for the item parameter estimates of the Concepts and Applications Test (see Table 2).

Table 3
*Test Score Means and Standard Deviations by Eight Ability Levels of the Students*

| Computation Test | | | Concepts and Applications Test | | |
|---|---|---|---|---|---|
| Ability Interval | | | Ability Interval | | |
| From - To | *M* | *SD* | From - To | *M* | *SD* |
| Below $P_5$ (-2.10) | 11.39 | 3.25 | Below $P_5$ (-1.87) | 14.81 | 3.95 |
| $P_5$ - $P_{10}$ (-1.49) | 15.75 | 2.35 | $P_5$ - $P_{10}$ (-1.34) | 20.09 | 2.17 |
| $P_{10}$ - $P_{25}$ (-.63) | 19.65 | 2.42 | $P_{10}$ - $P_{25}$ (-.67) | 25.22 | 2.19 |
| $P_{25}$ - $P_{50}$ (.08) | 25.27 | 2.24 | $P_{25}$ - $P_{50}$ (.02) | 31.59 | 2.30 |
| $P_{50}$ - $P_{75}$ (.77) | 31.13 | 2.35 | $P_{50}$ - $P_{75}$ (.70) | 38.39 | 2.08 |
| $P_{75}$ - $P_{90}$ (1.33) | 37.50 | 2.29 | $P_{75}$ - $P_{90}$ (1.41) | 44.39 | 1.81 |
| $P_{90}$ - $P_{95}$ (1.64) | 42.15 | 1.55 | $P_{90}$ - $P_{95}$ (1.83) | 48.43 | .92 |
| Above $P_{95}$ | 45.54 | 1.75 | Above $P_{95}$ | 51.51 | 1.27 |

*Note:* Given in parentheses are the values of the percentiles $P_5$, $P_{10}$, ..., $P_{90}$, $P_{95}$ on the ability scale (in logits).

Table 4
*Generalizability Study of the S x (I:CO) Design for the Computation Test and the S x (I:TS) Design for the Concepts and Applications Test*

| | | Computation Test | | Concepts and Applications Test | |
|---|---|---|---|---|---|
| Source of Variation | Variance Component | Estimated Variance Component | Percentage of Total Variance | Estimated Variance Component | Percentage Of Total Variance |
| Students (S) | $\sigma_S^2$ | .0290 | 14 | .0219 | 9 |
| Items (I) | $\sigma_I^2$ | .0288 | 14 | .0624 | 26 |
| S x (I:CO), E[a] | $\sigma^2_{S\ x\ (I:CO),\ E}$ | .1491 | 72 | | |
| S x (I:TS), E[b] | $\sigma^2_{S\ x\ (I:TS),\ E}$ | | | .1530 | 65 |

[a] For the Computation Test , with Category Objective (CO) fixed facet.
[b] For the Concepts and Applications Test, with Thinking Skill (TS) fixed facet.

Table 3 shows means and standard deviations of CAT-M scores for students at eight ability levels. Boundaries of the ability intervals are the percentiles $P_5$, $P_{10}$, $P_{25}$, $P_{50}$, $P_{75}$, $P_{90}$, and $P_{95}$ on the ability scale (in logits).

Table 4 shows results from the G-studies conducted for the Computation Test, with the S x (I:CO) design, and for the Concepts and Applications Test, with the S x (I:TS) design. With each of the two designs including a fixed facet, the variance due to interaction between subjects and items is inseparable from the variance due to random error in each of the variance components $\sigma^2_{S\ x\ (I:CO),E}$ and $\sigma^2_{S\ x\ (I:TS),E}$. It should be noted, however, that the "guessing" part of the random error variance was relatively small (see the *c*-values in Tables 1 and 2). For the Computation Test, the variance component $\sigma^2_{S\ x\ (I:CO),E}$ accounted for the largest part of the total variance, 72% . Hence, the relative standing of students on the computation scale changes a great deal across items. This was also true for the Concepts and Applications Test where the variance component $\sigma^2_{S\ x\ (I:TS),E}$ also explained the largest part, 65%, of the total variance. Table 5 shows D-study results about relations between number of items and reliability coefficients $E\rho^2$ and $\phi$. For relative decisions with the Computation Test, for example, a reliability of .90 or above ($E\rho^2 \geq .90$) requires at least six items within each category objective of the test. Similarly, for absolute decisions with the Concepts and Applications Test, a reliability of .90 or above ($\phi \geq .90$) requires at least 30 items per thinking skill level of the test.

Table 6 shows results from the 6 x 3 two-way ANOVA, using the item difficulty as the dependent variable and the fixed factors CO and TS of the Concepts and Applications Test as independent variables. The non-significance of the main effects, CO(F(5,39) = 2.06, p = .092) and TS(F(2, 39) = 1.49, p = .237), indicates that the average item difficulty is the same across all category objectives and, separately, across all thinking skill levels. The significance of the interaction between the two factors, CO x TS(F(6,39) = 2.62, p = .031), shows that the difference between the average item difficulties of the category objectives varies across the thinking skill levels of the test.

Table 5
*Decision Study of the S x (I:CO) Design for the Computation Test and the S x (I:TS) Design for the Concepts and Applications Test*

| Computation Test | | | Concepts and Applications Test | | |
|---|---|---|---|---|---|
| Number of Items | $E\rho^2$ | $\Phi$ | Number of Items | $E\rho^2$ | $\Phi$ |
| 1 | .661 | .620 | 6 | .720 | .646 |
| 2 | .796 | .766 | 10 | .811 | .753 |
| 3 | .854 | .830 | 15 | .866 | .821 |
| 4 | .886 | .867 | 20 | .896 | .859 |
| 5 | .907 | .891 | 25 | .915 | .884 |
| 6 | .921 | .907 | 30 | .928 | .901 |
| 7 | .932 | .919 | 35 | .938 | .914 |
| 8 | .940 | .929 | 40 | .945 | .924 |
| 9 | .946 | .936 | 45 | .951 | .932 |
| 10 | .951 | .942 | 50 | .955 | .938 |

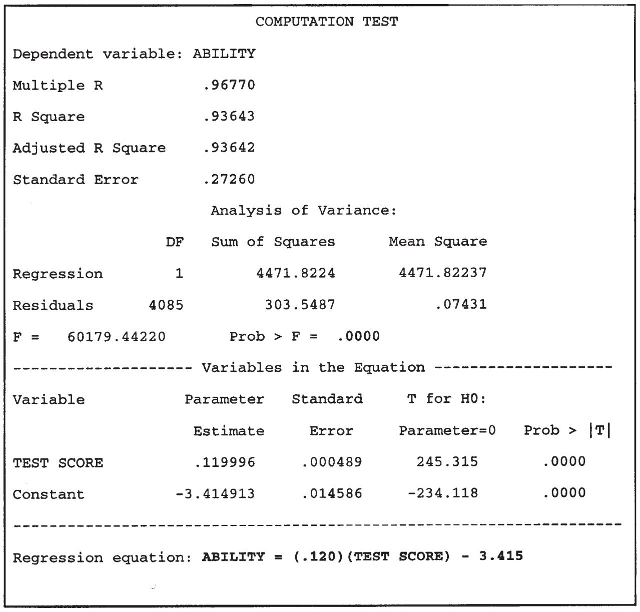
Table 6

*Unbalanced 6 x 3 (CO x TS) ANOVA design with Dependant Variable the Item Difficulty for the Concepts and Applications Test*

| Source | SS[a] | df | MS | F | p-value |
|---|---|---|---|---|---|
| Model | 29.34 | 13 | 2.26 | 2.23 | .027 |
| Category Objective (CO) | 10.41 | 5 | 2.08 | 2.06 | .092 |
| Thinking Skill (TS) | 3.02 | 2 | 1.51 | 1.49 | .237 |
| CO x TS | 15.91 | 6 | 2.65 | 2.62 | .031 |
| Within + Residual | 39.47 | 39 | 1.01 | | |
| Total | 68.82 | 52 | 1.32 | | |

[a] SEQUENTIAL Sums of Squares Source via SPSS (Windows, v. 6.1).

Regression analysis was conducted in an attempt to find a simple model for predicting students' abilities on CAT-M scores. Students with ability scores beyond the interval bounded by ±3.0 on the logit scale, representing about 1% of the 4135 students for each CAT-M test, were excluded from the regression analysis in order to avoid the "outliers" effect. Figure 2 represents an edited SPSS output from the simple linear regression analysis conducted for the Computation Test. The Multiple R of 0.97 indicates an extremely high positive correlation between observed and predicted ability scores of the students. Also, $R^2 = 0.94$ shows that 94% of the differences in the ability scores of the students are explained by differences in their test scores. The regression equation in Figure 2 provides simple and significant prediction of the abilities on test scores. Its graphical representation is given in Figure 3. Almost identical regression results were found for the Concepts and Applications Test (see Figure 4). With this test, 97% of the students' ability variance was explained by the test score variance and, again, the simple linear regression provided highly significant prediction of the abilities on test scores (see, also, Figure 5).



```
                          COMPUTATION TEST
Dependent variable: ABILITY
Multiple R           .96770
R Square             .93643
Adjusted R Square    .93642
Standard Error       .27260
                         Analysis of Variance:
                DF   Sum of Squares      Mean Square
Regression       1        4471.8224       4471.82237
Residuals     4085         303.5487           .07431
F =   60179.44220       Prob > F =  .0000
------------------- Variables in the Equation -------------------
Variable          Parameter   Standard      T for H0:
                   Estimate      Error     Parameter=0   Prob > |T|
TEST SCORE          .119996    .000489       245.315       .0000
Constant          -3.414913    .014586      -234.118       .0000
-----------------------------------------------------------------
Regression equation: ABILITY = (.120)(TEST SCORE) - 3.415
```

*Figure 2.* Edited SPSS output from the simple linear regression of ability scores on test scores for the Computation Test.

## Discussion

Along with the standard information about CAT-M results, provided to local educational analysts, there are additional findings that should be taken into account for the target population of seventh-graders. In the context of the research questions in this study, several findings are important.

First, the Rasch and 2-parameter IRT models did not fit the data for the CAT-M with the target population. This finding suggests that the items differed in discriminating seventh-graders with different ability scores and that there were "guessing" effects, although they were found to be relatively small. The CAT-M data did fit the 3-parameter IRT model for the target population.

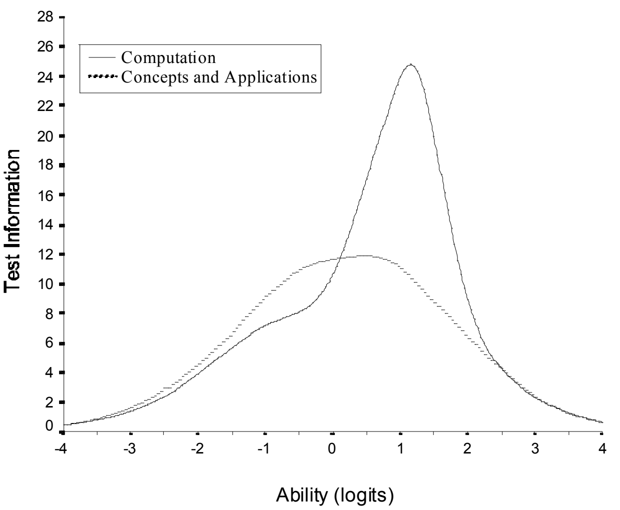


*Figure 1.* Test information curves for the Computation and Concepts and Applications Tests.

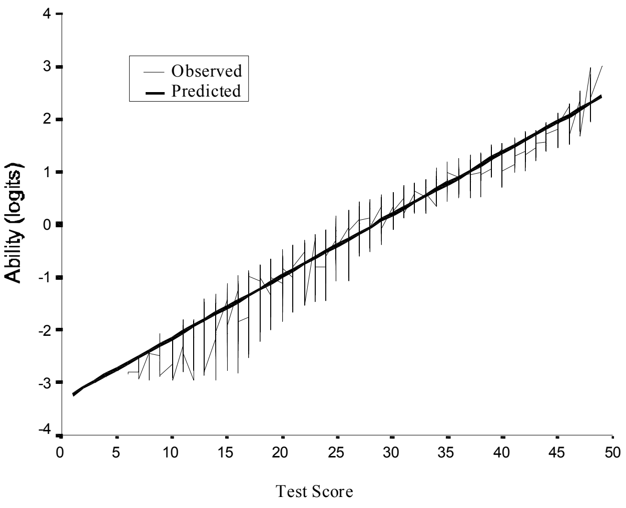
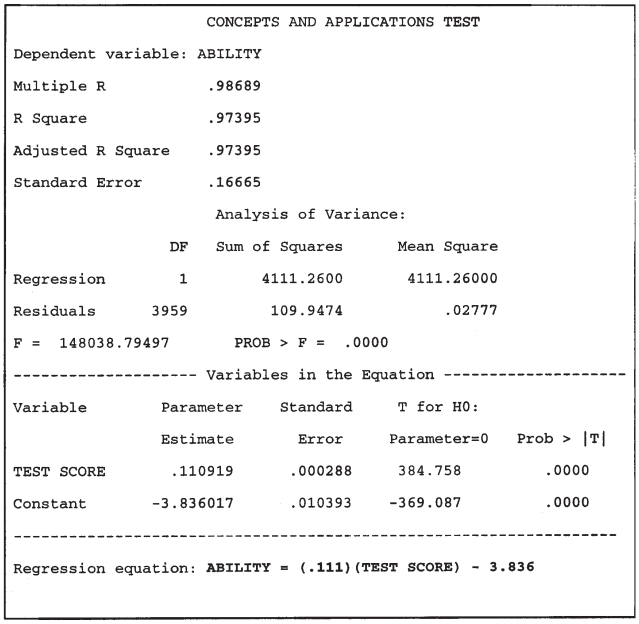


*Figure 3.* Simple linear regression of ability scores on test scores for the Computation Test.



```
                      CONCEPTS AND APPLICATIONS TEST

Dependent variable: ABILITY

Multiple R           .98689

R Square             .97395

Adjusted R Square    .97395

Standard Error       .16665

                      Analysis of Variance:

                 DF   Sum of Squares      Mean Square

Regression        1        4111.2600       4111.26000

Residuals      3959         109.9474           .02777

F =  148038.79497       PROB > F =  .0000

-------------------- Variables in the Equation --------------------

Variable        Parameter    Standard     T for H0:

                Estimate       Error     Parameter=0   Prob > |T|

TEST SCORE        .110919     .000288     384.758         .0000

Constant       -3.836017     .010393    -369.087         .0000

-------------------------------------------------------------------

Regression equation: ABILITY = (.111)(TEST SCORE) - 3.836
```

*Figure 4.* Edited SPSS output from the simple linear regression of ability scores on test scores for the Concepts ans Applications Test.

Second, the Computation Test provided more information and, hence, more accurate estimates of students' abilities than the Concepts and Applications Test, within the range from 0.0 to 2.0 on the logit ability scale. Beyond this interval (i.e., for students with ability below the average and for high ability students) neither test worked particularly well. The results in Table 3 show how students at eight different ability levels performed on the CAT-M.

Third, for the Concepts and Applications Test, the difference between the average difficulty of items from different category objectives varied a great deal across the thinking skill levels. Fourth, the G-study results show that the relative standing of seventh-graders on the CAT-M scale changed a great deal across different items of the test. Fifth, the D-study results provided information about the number of items required to obtain desired reliabilities for both relative and absolute (criterion-related) decisions. Sixth, the regression analysis provided a simple and highly significant model for the prediction of students' abilities on CAT-M scores.

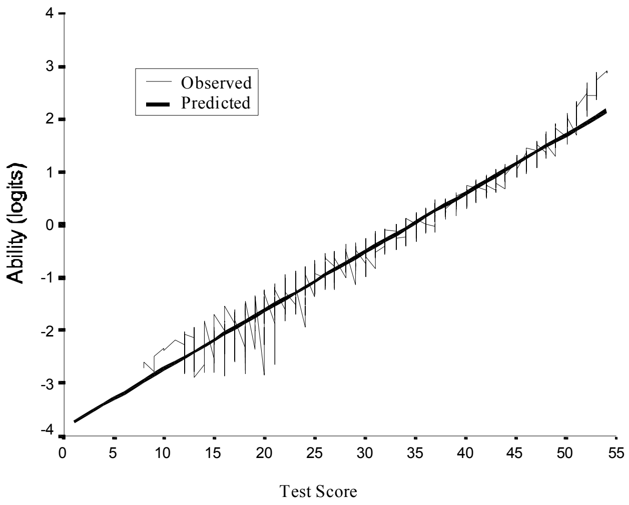


*Figure 5.* Simple linear regression of ability scores on test scores for the Concepts and Applications Test.

In conclusion, reports and interpretations of results of local student populations on nationally standardized mathematics are commonly based on descriptive statistics of test items and student total scores. The analysis illustrated in this article may help local educators and test analysts in interpreting test results by taking into account the ability levels of the students and the interaction between test factors such as item difficulty, category objectives, and thinking levels. In general, it provides valuable feedback for making informed decisions about teaching mathematics within local educational settings. Future research in this area will focus on relationships between psychometric and cognitive characteristics of the items. Also, one can apply the multimethod approach in the analysis of results from science, language, and other standardized tests administered to students representing large local populations.

## References

Assessment Systems Corporation. (1995a). *User's Manual for RASCAL Rasch analysis program* (Windows version 3.5). St. Paul, MN: Author.

Assessment Systems Corporation. (1995b). *User's Manual for XCALIBRE marginal maximum-likelihood estimation program* (Windows version 1.0). St. Paul, MN: Author.

CTB/McGraw-Hill. (1985). *California Achievement Tests*. Forms E Level 17. Del Monte Research Park, Monterey, CA: Author.

CTB/McGraw-Hill. (1986). *California Achievement Tests*. Forms E and F. (Technical Bulletin No. 2). Del Monte Research Park, Monterey, CA: Author.

Crick, J., E., and Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. ACT Technical Bulletin, No. 43. The American College Testing Program. Iowa City, IA.

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability theory: A primer.* SAGE. Newbury Park, CA.

SPSS Inc. (1997). *SPSS (Windows version 7.5): User's guide*. Chicago, IL: Author.

Waller, N. G. (1995). MicroFACT 1.0. *A Microcomputer factor analysis program for dichotomous and ordered polytomous data*. Assessment Systems Corporation. St. Paul, MN.