DOCUMENT RESUME

ED 155 211 TM 007 212

AUTHOR Powers, Stephen; Gallas, Edwin J.
TITLE Implications of Out-of-Level Testing for ESEA Title I MTd##Students.
PUB DATE Mar 78
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); For related documents, see TM 007 215 and 216

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Testing; Compensatory Education Programs; Complexity Level; *Disadvantaged Youth; Elementary Secondary Education; Equated Scores; Grade 4; Grade 7; Grade 9; *Grade Equivalent Scores; Norm Referenced Tests; *Norms; Raw Scores; Reading Achievement; Reading Comprehension; Scores; Standardized Tests; *Testing; *Test Interpretation; Test Reliability
IDENTIFIERS Comprehensive Tests of Basic Skills; Elementary Secondary Education Act Title I; *Out of Level Testing

ABSTRACT

Fourth, seventh, and ninth grade students in Elementary Secondary Education Act (ESEA) Title I programs were tested with the reading comprehension subtests of the Comprehensive Tests of Basic Skills, at each of two levels: on-level for each respective grade, and an easier cut-of-level form. Approximately half of these students were found to be scoring at or below the chance level for the on-level tests. It was judged that in such circumstances it is preferable to use out-of-level tests, particularly at the seventh and ninth grade levels. (CTM)

SESSION NO. 31.12

IMPLICATIONS OF OUT-OF-LEVEL TESTING FOR ESEA TITLE I STUDENTS

by

Stephen Powers, PhD
Title I Research Specialist

Data Analysis by Edwin J. Gallias

Department of Research and Evaluation
Director, Barbara J. Prentice, PhD
Tucson Unified School District
Tucson, Arizona

Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Implications of Out-of-Level Testing for ESEA Title I Students

Introduction

Researchers have advised administrators and evaluators of ESEA Title I programs to test students at their achievement level and not on the basis of the students' grade level (Horst and Fagan 1976; Roberts 1976). Other researchers assert that out-of-level testing is a questionable procedure until test publishers develop out-of-level norms which would aid in obtaining meaningful, derived scores (Long, Schaffran and Kellogg 1977). Most of the major achievement test publishers provide a common metric by which scores which have been obtained through out-of-level testing may be converted to the on-level equivalents. With such a common metric, test publishers indicate that out-of-level testing may be conducted and the results may be converted to the appropriate grade level of the student.

During the 1977-78 school year the Elementary and Secondary Education Act (ESEA) Title I compensatory reading program was provided to approximately 4000 students in grades K-12 of the Tucson Unified School District. In the fall of 1977 a study was conducted with Title I students enrolled in the fourth, seventh and ninth grades. The purpose of the study was: 1) to compare the achievement of students tested with the on-level and out-of-level reading comprehension subtests of the Comprehensive Tests of Basic Skills, Form S, (CTBS/S); 2) to determine if the on-level or out-of-level form of the CTBS/S Reading Comprehension subtests was more suitable for Title I students; 3) to

ascertain if there were significant differences between the out-of-level and on-level test scores when the scores were converted to a common metric, the CTBS/S Expanded Standard Score Scale; 4) to investigate if there were trends in the data which indicated a linear or curvilinear relationship.

By means of the CTBS/S Expanded Standard Score Scale, out-of-level testing scores could be converted to the appropriate grade level scores. Fundamentally, the Expanded Standard Score Scale--an equal interval, normalized scale--was developed following Thurstone's Absolute Scaling Method as described by Gulliksen (1950).

Classic test theory has been formulated either in terms of true score and error or in terms of a definition of parallel test forms. Recently, the concept of domain sampling has appeared in measurement literature. According to Thorndike (1971) page 9: "A somewhat different conception has been offered in recent years, the conception of a domain of admissible tasks from which the test was drawing one sample. Reliability is then conceived as the accuracy with which the sample represents the complete domain from which it was drawn."

When the samples are small, the precision of measurement is poor. Accordingly, the proportion of test material on which students should spend their time should not drop too low. Moreover, increasing the sample size should not only increase the precision of measurement, it should also reduce random sampling errors.

Methodology

In the fall of 1977 the CTBS/S, Reading Comprehension subtests were administered to a selected sample of 89 students enrolled in the ESEA Title I project in the fourth, seventh and ninth grades. At the

first testing session, one-half of each group was tested with the on-level test and the other half was tested with the out-of-level test. Within one week of the first testing, a second testing session was held in which the groups were reversed so as to avoid any bias resulting from the sequence of testing. The number of students and the levels of the CTBS/S involved in the testing are presented in Table 1.

- - - - - - - - - - - - - - - -
Insert Table 1 about here
- - - - - - - - - - - - - - - -

The CTBS/S, Levels 1-4, is a battery of seven tests measuring three basic skills areas: Reading, Language and Mathematics (CTBS/S Test Coordinator's Handbook 1976). The skills areas were classified using Bloom's Taxonomy of Educational Objectives (Appendix A). In the test development, efforts were made to reduce racial and ethnic bias. The K-R 20 reliabilities at each grade level for vocabulary, comprehension and total scores are almost all above .90 with standard errors of measurement from .25 to 1.01 in grade equivalent units. Moreover, it appeared that systematic procedures were followed in test development to ensure content validity. The CTBS/S Reading Comprehension subtests, Levels 1-4 are composed of 45 items and each item in Levels 1-4 contains a multiple choice involving four alternatives. The Reading Comprehension: Passages subtest in Level C included 18 items, each item with a multiple choice of four alternatives. For an outline of the grade levels recommended for administration of the CTBS/S, refer to Table 2.

- - - - - - - - - - - - - - - -
Insert Table 2 about here
- - - - - - - - - - - - - - - -

Before the statistical analysis, raw scores were converted to the CTBS/S Expanded Standard Score Scale. Calculations for the

present study were performed with expanded standard scores unless otherwise noted. As the expanded standard scores are a normalized scale with assumed equal intervals, it was believed this metric was more appropriate for statistical analysis. This choice of a metric is in agreement with the technical advice to ESEA Title I evaluators (Tallmadge and Wood 1976). When it was desired to convert an out-of-level test statistic up to the appropriate on-level statistic, the raw scores were converted to expanded standard scores and the desired statistic was computed with expanded standard scores. Then, the appropriate grade level table was referred to, and by means of the appropriate table, the expanded standard score was used to ascertain the grade level raw score, percentile, stanine, or grade equivalent.

In summary, the purposes of the present study were the following: 1) to compare the achievement of students tested with the on-level and out-of-level reading comprehension subtests of the CTBS/S; 2) to determine if the on-level or out-of-level subtests were suitable by investigating the chance level, floor and ceiling effects, test suitability and a reliability index; 3) to investigate any significant differences between mean expanded standard scores; 4) to study any linear or curvilinear trends in the data.

Results

1. Descriptive statistics were studied to compare the achievement of Title I students who were administered both an on-level and out-of-level CTBS/S Reading Comprehension subtest. Summary statistics are presented in Table 3. Descriptive statistics in expanded standard scores and raw scores are displayed in Appendix B.

Insert Table 3 about here

At the ninth grade level and seventh grade level, the out-of-level testing indicated higher percentiles and grade equivalents than the on-level testing. In the fourth grade level, the percentiles and grade equivalents appeared lower for the out-of-level testing than the on-level testing. Of course, one would not expect to find exactly the same mean ($\bar{X}$), percentile or grade equivalent even if the students were tested with exactly the same test under optimal testing conditions. Variation in testing results would be expected under the best conditions.

2. To determine if the on-level or out-of-level Reading Comprehension subtest was more suitable for Title I students, the following four factors were investigated: a) chance level, b) floor and ceiling effects, c) test's suitability and d) test's reliability.

a. The chance level of a test is a phenomenon which should be investigated when testing Title I students. Since Title I students are selected because of their need for Title I reading programs, the proportion of Title I students scoring at chance level will often reach unacceptable levels. When the number of students scoring at chance level is unreasonably large, this could be an indication that out-of-level testing is a better procedure than on-level testing.

Gulliksen (1950, p. 263) provides guidelines for investigating the chance level of tests. The average chance score ($\bar{X}_c$), the total number of items in a test (K), and the standard deviation of the distribution of chance scores (SD_c) are three quantities which assist the evaluator in investigating the meaningful score range for a test. The following formulas are used to compute the average

chance score and the standard deviation of chance scores:

$$\bar{X}_c = K/A$$

$$SD_c = \sqrt{\frac{K(A-1)}{A}}$$

where K = the total number of items on a test; A = the number of alternative answers for each item. When applied to the data in the present study, a score obtained on Levels 1-4 that is less than 15 or on Level C that is less than 7 would fall within the upper limits of one standard deviation of the distribution of chance scores (Table 4). "Any score within one or two standard deviations (SD_c) of a chance score should not be interpreted as signifying any knowledge of the subject matter of the examination" (Gulliksen 1950, p. 263).

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

Insert Tables 4 and 5 about here

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

When Title I students were tested with the on-level tests, exceedingly large proportions of students at each grade level scored within the limits of one standard deviation of the distribution of chance scores. The proportion of Title I students who scored at the chance level dropped noticeably when they were tested with the out-of-level test. The lowering of the number of students scoring at chance level should provide a larger sampling of students' reading skills, and this in turn should result in a more accurate estimate of the students' reading ability. Table 5 displays the percent of students scoring at chance level.

b. Floor and ceiling effects were investigated using Roberts' (1976) guidelines. When studying floor effects, according to Roberts the following guidelines may be used: if the mean ($\bar{X}$) is higher

than the median (Md) by about one-third of a standard deviation (SD), a floor effect may have been encountered. If the mean is lower than the median by about one-third of a standard deviation, a ceiling effect may have been encountered. At the fourth grade level a floor effect was discovered in the data for CTBS/S Level 1 and a ceiling effect for CTBS/S Level C. It appeared that at the fourth grade level, neither the on-level nor the out-of-level CTBS/S reading comprehension subtest was free of floor or ceiling effects. Refer to Table 6 for the results of the investigation of ceiling and floor effects.

- - - - - - - - - - - - - - - -
Insert Table 6 about here
- - - - - - - - - - - - - - - -

c. To determine the suitability of a test, some criteria are needed. Roberts (1976) suggests that in most instances the level of a test is suitable when the mean raw score of the group is equal to or above a third of the maximum score, and somewhat less than three-quarters of the maximum. Refer to Table 7 for the results of applying Roberts' (1976) guidelines to Title I students.

- - - - - - - - - - - - - - - -
Insert Table 7 about here
- - - - - - - - - - - - - - - -

Three of the tests administered to Title I students were found to be inappropriate when suitability guidelines were applied: Grade 9, Level 4; Grade 7, Level 3; and Grade 4, Level C.

d. One index for estimating a test's reliability is described by Roberts (1976). Tests are constructed by the publishers so that the median score at the appropriate grade level is well above half the number of items in the test. Thus, for the average class, a

ceiling effect is more likely than a floor effect. According to Roberts (1976), "The highest reliability of a test is achieved when the students, on the average, get slightly more than half the items correct."

- - - - - - - - - - - - - - - - -
Insert Table 8 about here
- - - - - - - - - - - - - - - - -

The median number of correct responses was increased in all cases where out-of-level testing was conducted (Table 8). This should increase the test's reliability.

3. To determine if there were significant differences between testing results in expanded standard scores, the out-of-level means were compared with the on-level means using correlated t-tests. At the ninth and seventh grade levels the differences between out-of-level and on-level test scores fell within the range of sampling error. Therefore the null hypothesis that both samples of students' abilities were from the same population could not be rejected. The fourth grade out-of-level and on-level testing appeared significantly different, but further investigation revealed both a floor effect with the on-level testing and a ceiling effect with the out-of-level testing. Accordingly little confidence could be placed in the statistical significance found at the fourth grade level because the score distributions at the fourth grade level were either inflated or depressed. Table 9 presents t-tests comparing out-of-level test means with on-level test means. For t-tests with raw scores, refer to Appendix C.

- - - - - - - - - - - - - - - - -
Insert Table 9 about here
- - - - - - - - - - - - - - - - -

Pearson product-moment correlations were computed between the out-of-level scores and the on-level scores for each grade level

to determine the shared variation between the two testing levels. The proportion of shared variance (r^2) was surprisingly low in the ninth and seventh grade levels although in the fourth grade level the proportion of shared variance more than doubled that of the seventh grade and was almost five times that of the ninth grade level. Correlations of out-of-level test scores with on-level test scores are displayed in Table 10. For correlations using expanded standard scores and raw scores, refer to Appendix D.

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

Insert Table 10 about here

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

The increase in proportion of shared variance for fourth grade students may be related to the decrease in the proportion of students scoring at the chance level in the fourth grade, although there are other possible explanations. If the number of students scoring at the chance level is reduced, this could be one factor which increases the accuracy of the measurement and contributed to a higher correlation between the two levels of the test at the fourth grade level.

4. The possibility of curvilinear trends in the data was investigated following a procedure outlined by Kerlinger (1973). The on-level test scores were regressed on the out-of-level test scores at each of the three grade levels. This was accomplished separately for the expanded standard scores and the raw scores. The out-of-level test scores were squared and entered into the regression equation to determine if the variance accounted for in the on-level scores was significantly increased:

$$y = a + bx + bx^2$$

where y = the predicted on-level scores, b = the b-weight applied to the

predictor scores, x = out-of-level predictor scores and x^2 = the squared out-of-level predictor scores. In the present analysis the trends in raw scores and expanded standard scores were studied. No significant departures from linearity were indicated in the data at any grade level. Therefore, the relationships are best described at each grade level by a linear equation.

Summary and Discussion

The purposes of the present study was: 1) to compare the achievement of Title I students who were administered both an on-level and out-of-level CTBS/S Reading Comprehension subtest, 2) to determine if the on-level or out-of-level Reading Comprehension subtest was more suitable for Title I students, 3) to ascertain if there were significant differences between the out-of-level and on-level mean scores when the scores were converted to the Expanded Standard Score Scale and 4) to investigate if there were linear or curvilinear trends in the data.

1. The achievement of Title I students who were administered out-of-level and on-level tests was compared after the raw scores were converted to expanded standard scores and appropriate statistics computed. The out-of-level means were converted to the appropriate grade levels. Students at the seventh and ninth grade levels attained higher means, percentiles and grade equivalents when the expanded standard score means were converted to the appropriate grade level statistic. Fourth grade students' mean expanded standard score was higher for the on-level testing. It was pointed out, however, that floor effects occurred in the on-level test and ceiling effects occurred in the out-of-level testing at the fourth grade level. Because of the floor and ceiling effects, it was not possible to determine the comparability of out-of-level and on-level means at the fourth grade level

because the score distribution was either artificially depressed or spuriously inflated.

2. To determine if the out-of-level or the on-level tests were more suitable for Title I students, the data were analyzed to determine: a) if a large number of students were at chance level, b) if there were floor or ceiling effects, c) if the levels of the tests were appropriate and d) if the median score indicated a reliable test.

Large percentages of students scored within the chance level on the on-level test. When the same students were tested with out-of-level tests, the chance level dropped noticeably. By lowering the chance level, a larger sample of the students' abilities should be obtained and this should result in a more precise measurement of the students' reading ability.

Using Roberts' (1976) criteria to investigate floor and ceiling effects, it was determined that floor effects occurred in the on-level test and that ceiling effects occurred with the out-of-level test at the fourth grade level. Neither floor nor ceiling effects were indicated in the data at the seventh and ninth grades.

Roberts (1976) provides guidelines for determining when a test is suitable in terms of difficulty. When these guidelines were applied to the tests administered on-level and out-of-level, the tests at the fourth grade level were found to be unsuitable. Moreover, the on-level tests at the seventh and ninth grades appeared to be unsuitable.

The median number of correct responses was compared with the median number of possible responses to obtain a general indication of the test's reliability. The median score at the appropriate grade

level should be above half the number of items in the test (Roberts 1976). When students were administered the out-of-level test, the median number of correct responses was increased over the median correct responses of the on-level test. According to Roberts (1976) this should indicate an increase in reliability.

3. To determine if there were significant differences between the mean expanded standard scores of the out-of-level tests and the mean expanded standard scores of the on-level tests, correlated t-tests were computed. At the seventh and ninth grades no significant differences were found between the out-of-level and on-level mean scores. The differences between the out-of-level means and the on-level means could be the result of sampling error. Thus the information obtained from out-of-level testing did not appear to be significantly different from the information obtained from on-level testing. At the fourth grade level a ceiling effect artificially depressed the out-of-level test scores and a floor effect spuriously inflated the on-level test scores. This is probably the major factor in the statistical significance found between on-level and out-of-level means at the fourth grade level.

Pearson product-moment correlation coefficients were computed between out-of-level scores and on-level scores at each grade level to determine the shared variation between the two test levels. The proportion of shared variance was surprisingly low at the ninth and seventh grade levels. The proportion of variance shared variance was much greater at the fourth grade level. A major factor in the increase in shared variance may have been the reduction of the chance level at the fourth grade level and the resulting increase in precision of measurement.

4. The possibility of curvilinear trends was investigated following a procedure described by Kerlinger (1973), but no such trends were found. The data at each grade level could be described best by a linear equation.

Conclusions

What are the implications of out-of-level testing for ESEA Title I students? Two major questions are often raised by ESEA Title I administrators and evaluators:

1) Can out-of-level test results be converted by means of a common metric to on-level equivalents?

2) Will out-of-level testing increase the precision of measurement of students who are below the grade level for which the test is designed?

The results of this study indicate that when ESEA Title I students are tested with out-of-level and on-level CTBS/S Reading Comprehension subtests, the means obtained when the scores are converted to the CTBS/S Expanded Standard Score Scale are reasonably equivalent within sampling error.

The equating of test means across grade levels is not as important an issue as the increase in measurement precision obtained through testing students at their instructional level. Indeed, the chance level, floor and ceiling effects, the test's suitability (i.e., Robert's criteria) and the medians are indicators of the precision of measurement. In this study the out-of-level tests appeared to be preferable to on-level tests at the ninth and seventh grades because the chance level dropped to lower levels with out-of-level testing, the on-level tests were shown to be unsuitable using Robert's (1976) guidelines and the median number of correct responses was raised.

At the fourth grade level, where ceiling effects occurred with the out-of-level testing and floor effects occurred with the on-level testing, the results could not be equated across grade levels because the scores were either artificially depressed or inflated. Therefore, at the fourth grade level it appears the tests were either unusually difficult (on-level) or exceedingly easy (out-of-level).

In conclusion, if a small bias is introduced into the data by testing out-of-level it is perhaps better to accept the small bias if it is therefore possible to increase substantially the precision of measurement.

REFERENCES

Comprehensive Tests of Basic Skills, Expanded Edition: Test Coordinator's Handbook. CTB/McGraw-Hill, Del Monte Research Park, Monterey, California (1976).

Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, Inc., 1950.

Horst, D. P., & Fagan, B. M. Common Evaluation Hazards: ESEA Title I Evaluation and Reporting System, Technical Paper No. 14, RMC Research Corporation, Mountain View, California, October 1976.

Kerlinger, F. N., & Pedhazur, E. J. Multiple Regression in Behavioral Research. Holt, New York, 1973.

Long, J. V., Schaffran, J. A., & Kellogg, T. M. Effects of out-of-level survey testing on reading scores of Title I, ESEA students. Journal of Educational Measurement (14) (3), fall 1977, pp. 203-213.

Roberts, A. O. H. Out-of-Level Testing: ESEA Title I Evaluation and Reporting System, Technical Paper No. 6. RMC Research Corporation, Mountain View, California, October 1976.

Tallmadge, G. K., & Wood, C. T. User's Guide: ESEA Title I Evaluation and Reporting System. RMC Research Corporation, Mountain View, California, December 1976.

Thorndike, R. L. "Educational Measurement for the Seventies" in Educational Measurement. Robert L. Thorndike (Ed.), American Council on Education, Washington, D.C., 1971.

Table 1. Numbers of Students and Levels of the CTBS/S

Grade	N*	Test Level	Items
Ninth	28	4	45
		3	45
Seventh	33	3	45
		2	45
Fourth	28	1	45
		C**	18

*The same students were tested with the out-of-level and the on-level test.

**At Level C, students were administered the subtest: Reading Comprehension: Passages.

Table 2. CTBS/S Test Levels and Recommended Grades

Test Level	Grades
A	K.0 - 1.3
B	K.6 - 1.9
C	1.6 - 2.9
1	2.5 - 4.9
2	4.5 - 6.9
3	6.5 - 8.9
4	8.5 -12.9

Table 3. Summary Statistics of the CTBS, Levels C-4

Grade	Test Level	N	$\bar{X}$	%ile	Sta.	GE
Ninth	4	28	438	16	3	4.6
	3	28	461	23	4	5.3
Seventh	3	33	387	13	3	3.4
	2	33	402	16	3	3.7
Fourth	1	28	368	27	4	3.1
	C	28	290	16	3	1.9

Table 4. Average Chance Scores and Standard Deviations For Investigating Chance Level of Tests.

Grade	Test Level	Total Items	$\bar{X}_c$	SD_c
Ninth	4	45	11.25	±2.90
	3	45	11.25	±2.90
Seventh	3	45	11.25	±2.90
	2	45	11.25	±2.90
Fourth	1	45	11.25	±2.90
	C	18	4.50	±1.84

Table 5. Percent of Students Scoring at Chance Level on the CTBS.

Grade	Test Level	N	N at Chance Level	Percent
Ninth	4	28	16	57%
	3	28	5	18%
Seventh	3	33	19	58%
	2	33	10	30%
Fourth	1	28	10	36%
	C	28	1	4%

Table 6. Ceiling and Floor Effects on Out-of-Level and On-Level Reading Comprehension Subtests of the CTBS.[a]

Grade	Test Level	$\bar{X}_b$	Md	$\bar{X}$-Md	1/3 SD
Ninth	4	12.8	11.5	1.3	1.9
	3	21.1	23.5	-2.4	2.7
Seventh	3	13.5	12.3	1.2	1.7
	2	18.5	18.9	- .4	2.2
Fourth	1	20.9	17.2	3.7	3.6*
	C	13.9	15.9	-2.0	1.4**

[a]If the mean is lower than the median by about one-third of a standard deviation, a ceiling effect may have been encountered. If the mean is higher than the median by about one-third of a standard deviation, a floor effect may have been encountered (Roberts 1976).
[b]The statistics have been computed from raw scores.
*Floor effects.
**Ceiling effects.

Table 7. The Suitability of the Out-of-Level and On-Level Reading Comprehension Subtests of the CTBS.[a]

Grade	Test Level	$\bar{X}^b$	Maximum Score	Interval	Suitable?
Ninth	4	12.8	45	15.0-33.7	No
	3	21.1	45	15.0-33.7	Yes
Seventh	3	13.5	45	15.0-33.7	No
	2	18.5	45	15.0-33.7	Yes
Fourth	1	20.9	45	15.0-33.7	Yes
	C	13.9	18	6.0-13.5	No

[a]The level of a test is suitable when the raw score mean is equal to or above one-third of the maximum score and somewhat less than three-quarters of the maximum score (Roberts 1976).
[b]The statistics have been computed from raw scores.

Table 8. Medians and Reliability Criteria for Title I Students

Grade	Test Level	Total Items	Criterion	Md
Ninth	4	45	22.5	11.5
	3	45	22.5	23.5
Seventh	3	45	22.5	12.3
	2	45	22.5	18.9
Fourth	1	45	22.5	17.2
	C	18	9.0	15.9

Table 9. T-Tests Comparing Out-of-Level Tests with On-Level Tests Using Expanded Standard Scores.

Grade	df	$\bar{X}_{out}$	$\bar{X}_{on}$	Difference	t	p
Ninth	27	461.0	438.3	22.7	1.18	.247
Seventh	32	402.1	387.0	15.1	1.32	.195
Fourth	27	290.5	367.6	-77.1	-7.29	.000

Table 10. Correlations of Out-of-Level Test Scores with On-Level Test Scores Using Expanded Standard Scores.

Grade	N	r	r_2	p
Ninth	28	.32	.10	.095
Seventh	33	.44	.19	.011
Fourth	28	.67	.48	.011

APPENDIXES

Appendix A

Number of Items in Each Item Classification for the Reading Comprehension Subtest, Levels 1-4 (CTBS/S Test Coordinators Handbook 1976)

Process-Content Category	Level 1	Level 2	Level 3	Level 4
1. Recognition/Application				
Literal Recall	12	8	4	3
2. Translation				
Rewording	8	6	6	6
Context Clues	2	5	7	4
3. Interpretation				
Main Idea	8	5	6	9
Descriptive Words	4	4	4	3
Conclusions	11	8	12	11
4. Analysis				
Structure/Style	--	9	6	9
Total	45	45	45	45

Appendix B

Descriptive Statistics Using Raw Scores and Expanded Standard Scores

A. Raw Scores

Grade	Test Level	N	$\bar{X}$	Md	SD
Ninth	4	28	12.8	11.5	5.7
	3	28	21.1	23.5	8.1
Seventh	3	33	13.5	12.3	5.0
	2	33	18.5	18.9	6.5
Fourth	1	28	20.9	17.2	10.7
	C	28	13.9	15.9	4.3

B. Expanded Standard Scores

Grade	Test Level	N	$\bar{X}$	Md	SD
Ninth	4	28	438.3	424.0	87.0
	3	28	461.0	489.5	86.9
Seventh	3	33	386.9	377.7	62.1
	2	33	402.1	416.2	61.7
Fourth	1	28	367.6	355.8	70.4
	C	28	290.5	292.7	27.2

Appendix C

Correlated T-Tests Comparing Out-of-Level Tests
With On-Level Tests of the CTBS Using Raw Scores and Expanded Standard Scores

A. Raw Scores

Grade	df	$\bar{X}_{out}$	$\bar{X}_{on}$	Difference	t	p
Ninth	27	21.1	12.8	8.3	5.31	.000
Seventh	32	18.5	13.5	5.0	4.59	.000
Fourth	27	13.9	20.9	-7.0	-4.40	.000

B. Expanded Standard Scores

Grade	df	$\bar{X}_{out}$	$\bar{X}_{on}$	Difference	t	p
Ninth	27	461.0	438.3	22.7	1.18	.247
Seventh	32	402.1	386.9	15.2	1.32	.195
Fourth	27	290.5	367.6	-77.1	-7.29	.000

Appendix D

Correlations of Out-of-Level Tests with On-Level Tests
Using Raw Scores and Expanded Standard Scores

A. Raw Scores

Grade	N	r	r^2	p*
Ninth	28	.314	.098	.103
Seventh	33	.443	.196	.010
Fourth	28	.676	.457	.000

B. Expanded Standard Scores

Grade	N	r	r^2	p*
Ninth	28	.321	.103	.095
Seventh	33	.435	.189	.011
Fourth	28	.669	.447	.000

*two-tailed probability