# Report to the New York City Department of Education

## English Language Arts Testing Program
## **Technical Report**
## Spring 2003

Psychometric and Research Services
Harcourt Educational Measurement
San Antonio, TX

# Table of Contents

Foreword ..... 4

Introduction ..... 5

Test Design and Development ..... 5

Scoring and Data Analysis ..... 6

Key Checks ..... 6
Summary Statistics ..... 6

Completion Rates ..... 8

Scaled Scores ..... 9

Differential Item Functioning ..... 12

Mantel-Haenszel Procedure ..... 12
Subgroups ..... 14

References ..... 16

Appendices ..... 17- 28

A. Raw Score Summary Statistics
Disaggregated By Gender and Ethnicity ..... 17

B. Raw Score to Scaled Score Table ..... 20

C. Raw Score and Scaled Score Values for New York City
Performance Level Cut Scores, Maximum and Minimum
Possible Scores ..... 22

D. Frequency distributions ..... 24

## Table of Tables and Figures

Table 1. ELA Test Design for Spring 2003 ..... 6
Table 2. Raw Score Summary Statistics ..... 7
Table 3. Test Completion Rate ..... 9
Figure 1. Spring 2003 Scaled Score Frequencies ..... 11
Table 4. New York City Scaled Score Summary Statistics ..... 12
Figure 2. General Notation of the *j*th Total Score on the Test ..... 13
Table 5. Summary of DIF Analysis for Field-Test Items ..... 15
Table 6. Summary Statistics: Grade 3 ..... 18
Table 7. Summary Statistics: Grade 5 ..... 18
Table 8. Summary Statistics: Grade 6 ..... 19
Table 9. Summary Statistics: Grade 7 ..... 19
Table 10. Spring 2003 Raw Scores to Scaled Scores ..... 21
Table 11. Raw Score and Scaled Score Cut Scores ..... 23
Table 12. Raw Score to Scaled Score Minimum and Maximum Score Points ..... 23
Table 13. Spring 2003 Frequency Distributions: Grade 3 ..... 25
Table 14. Spring 2003 Frequency Distributions: Grade 5 ..... 26
Table 15. Spring 2003 Frequency Distributions: Grade 6 ..... 27
Table 16. Spring 2003 Frequency Distributions: Grade 7 ..... 28

## Foreword

The Spring 2003 *Technical Report* for the New York City English Language Arts Testing Program provides summaries and documentation of psychometric processes utilized during the administration and data analysis of the customized Reading Comprehension portion of the . The Reading Comprehension test was administered in New York City schools on Tuesday, April 15, 2003. The majority of the processes and data presented refer to the operational test items used for score reporting. Field-test item data were provided separately to New York City, although the statistical procedure and summary results of differential item functioning are presented in this report.

# New York City
# English Language Arts Testing Program
# Technical Report
# Spring 2003

## Introduction

In spring 2003, the New York City English Language Arts Testing Program administered the first customized reading comprehension test developed from the . The aligns well with New York City and State Learning Standards, and uses an established vertical scale and current national norms (2000) that are statistically linked with existing New York City and State scales.

The spring 2003 operational test was administered as part of the two-year (September 16, 2002 to September 15, 2004) contract agreement between the New York City Department of Education (NYCDOE) and Harcourt Educational Measurement (Harcourt) to cooperatively administer customized English Language Arts (ELA) tests to students in grades 3, 5, 6, and 7 in the New York City public schools each spring. The contract agreement establishes Harcourt as the test vendor for the New York City English Language Arts Testing Program. The fall 2002 pilot study (see the *Fall 2002 Pilot Study Technical Report)* showed the consistency of the reading comprehension test with that of the previous vendor. The customization of the spring test and concurrent field testing of test items for further customization were the first steps taken to reach the NYCDOE goal for the ELA tests.

## Test Development and Design

For the spring 2003 ELA tests, the full-length reading comprehension tests for grades 3, 5, 6, and 7 were shortened from 50 multiple-choice test items to 45 multiple-choice test items. This was accomplished by removing one reading passage and its corresponding set of five test items. The choice of passages to remain on the tests was determined to provide the best content alignment to the New York City and State Learning Standards and retain optimal statistical characteristics.

The reading passages and their corresponding set of five test items were replaced with field-test passage-item sets selected from Harcourt item banks to more closely match the New York City and State Learning Standards. Thirteen reading passages and 130 corresponding test items were placed across 26 test forms at each grade level. Therefore, each form comprised 50 items – 45 operational test items and five field-test items. Nine passages (8 operational, 1 field-test) appeared on each test form. Table 1 summarizes the test design for the spring 2003 ELA test administration.

Table 1. ELA Test Design For Spring 2003.

| Grade | Total Items Per Form | Operational Items Per Form | Number Of Test Forms | Field-test Items Per Form | Total Field-test Items Scored | Passage # Removed For Custom-ization |
|---|---|---|---|---|---|---|
| 3 | 50 | 45 | 26 | 5 | 130 | 3 |
| 5 | 50 | 45 | 26 | 5 | 130 | 2 |
| 6 | 50 | 45 | 26 | 5 | 130 | 3 |
| 7 | 50 | 45 | 26 | 5 | 129* | 8 |

* Field-test item #26 on form 8 was deleted due to a miss-key in the scoring program.

## Scoring and Data Analysis

The ELA tests were administered and scored in New York City. Harcourt provided scoring keys for both the operational and field-test test items. Personnel in New York City entered keys, scanned, scored and produced score reports. Harcourt personnel were on-site to conduct analysis to ensure that accurate equating and consistent scaled scores were derived and reported. Scored research files were transmitted to Harcourt for follow-up analysis of operational test items and field-test items.

### Key Checks

To insure that student scores were produced from valid answer keys, classical item analyses were completed on the preliminary research file for both operational and field-test items. Test items with 20 percent or less, or 80 percent or more students answering correctly, or a point biserial correlation statistic of 0.25 or less were flagged. A list of these items were forwarded to and reviewed by content specialists. As a result, the field-test item indicated in Table 1 was not scored.

### Summary Statistics

Analyses of student scores in summary form provide indices that allow for comparisons of averaged student performance and quantitative equivalence of test forms across years. Table 2 presents raw score summary statistics produced from the spring 2003 operational test items at each grade. Additional raw score summary statistics disaggregated by ethnicity and gender are located in Appendix A.

Raw score summary statistics were calculated from valid student scores on the 45 operational test items with each correct response receiving one point. Explanation of raw score descriptive statistics, *standard error of measurement* (SEM), and *reliability coefficients* are presented following Table 2.

Table 2. Raw Score Summary Statistics

| Grade | Number Of Student Scores | Number Of Test Items | Test Mean | Standard Deviation | Standard Error Of Measurement | $KR_{20}$ Reliability Estimate | Mean *p*-Value |
|---|---|---|---|---|---|---|---|
| 3 | 78313 | 45 | 28.3 | 9.31 | 2.83 | 0.91 | 0.63 |
| 5 | 77109 | 45 | 28.6 | 8.74 | 2.85 | 0.89 | 0.64 |
| 6 | 76978 | 45 | 29.6 | 8.54 | 2.79 | 0.89 | 0.66 |
| 7 | 74808 | 45 | 30.7 | 8.45 | 2.71 | 0.90 | 0.68 |

Note: Tabled SEM values were calculated before rounding SD and $KR_{20}$.

- *p*-values represent the percentage of students responding correctly to each test item. The mean *p*-value is derived by dividing the test-mean by the number of test items (45).

- **Point biserial correlation coefficients** provide evidence of how well a test item distinguishes between students who responded correctly to the item and received a high score on the total test, and those students who responded correctly to the item but received a low score on the total test. Point biserial correlation coefficients are best interpreted in conjunction with the *p*-value, item content, and knowledge of alignment of test item content with instruction. Test items with particularly high or low *p*-values tend to have suppressed coefficient values due to the small differentiation between responses of high and low scoring students.

- The **Standard Error of Measurement** is defined in classical test theory as the standard deviation of random inconsistencies (error) when taking a measurement. In this case, the SEM represents inconsistencies occurring in repeated observations of obtained test scores around a student's true test score, which is assumed according to classical test theory to remain constant across repeated measurements of the same trait. It is further assumed that the SEM retains the same value along the entire range of the measurement scale.

$$SEM = s_x\sqrt{1 - r_{xx}},$$

where, $SEM$ = standard error of measurement,
$s_x$ = standard deviation of observed scores,
$r_{xx}$ = test reliability coefficient,

An observed test score comprises true score and error components. If we assume that the error is random and normally distributed, it is possible to use the SEM to determine a *confidence interval* around an observed test score that has an expected probability of containing a score interval with

the student's true score (Feldt & Brennan, 1989). For any test score, potentially there are an infinite number of possible intervals that could be formed around a score. Some of the intervals will contain a student's true score other intervals will not. The confidence interval formed by utilizing the SEM provides a range of scores with a given probability that an interval containing the true score is within the confidence interval. For example, a confidence interval established by subtracting one SEM value from the observed score to form the lower bound of the score interval and adding one SEM value to the observed score to form the upper bound of the score has a 68 percent probability of including an interval with the student's true score. Repeating this procedure with two SEM values produces a confidence interval with a probability of 95 percent.

- **Reliability coefficients** reflect evidence of the accuracy, precision and consistency of a test. It assists in answering the question of whether a group of students who took the same test more than once, assuming the test did not change the students, would be ranked in the same order. If one operational test form is administered during a single test administration, an internal consistency reliability coefficient is typically employed. When test responses are scored as correct and incorrect, the *Kuder-Richardson Formula* 20 ( $KR_{20}$ ) is often employed (Thorndike, Cunningham, Thorndike, & Hagen, 1991).

$$KR_{20} = \left(\frac{n}{n-1}\right)\left(\frac{SD_t^2 - \sum p_i q_i}{SD_t^2}\right)$$

where, $n$ = the number of items in the test,
$SD_t$ = the standard deviation of the test scores,
$p_i$ = the proportion of correct item responses,
$q_i$ = the proportion of incorrect item responses.

$KR_{20}$ reliability coefficients for the 2003 administration of the New York City ELA tests are presented in Table 2.

## Completion Rates

The New York City ELA tests were administered with a 65-minute time limit in spring 2003. Table 3 presents the completion rates for the 50 test items (operational and field-test items) for each test level. Completion rates for the fall 2002 pilot study are also included for comparison. Recall that grades 3, 5, 6, and 7 in spring were administered the same test level as students in grades 4, 6, 7, and 8 in the fall. Completion rate values are rounded to the nearest whole number.

**Table 3. Test Completion Rates**

| Reading Comprehension | | | Completion Rates (%) | |
|---|---|---|---|---|
| Grade Spring | Grade Fall | Number of Items | NYC Spring 2003 Administration | NYC Fall 2002 Pilot Study |
| 3 | 4 | 50 | 98 | 98 |
| 5 | 6 | 50 | 99 | 97 |
| 6 | 7 | 50 | 99 | 98 |
| 7 | 8 | 50 | 99 | 99 |

## Scaled Scores

The derived scales for the spring 2003 administration of the New York City ELA tests were developed to retain the characteristics of the existing New York City scales and provide consistency in score reporting from previous years. The process required linking the existing scale system with the New York City scale system. A brief outline of the process is presented below:

- The 50-item Reading Comprehension test was customized to retain 45 items. The choice of reading passage and associated five items to delete was made to provide a test with similar reliability, content validity, and test difficulty of the original 50-item test and to align with New York City and State Learning Standards.

- The raw score to scaled score conversion for the 45-item test was created first. For each grade level, the Rasch item difficulties for the 45 items from the standardization data were entered into a Harcourt proprietary equating program (CRN) with the appropriate equating constant, to preserve the vertical scale, and multiplicative and additive constants. The resulting scaled score derivations may be used with the published norms.

- A to New York City ELA look-up table was developed to convert all potentially possible scaled scores to all potentially possible New York City ELA scores. This was accomplished through a series of steps.

  - Separately for each grade level, score distributions were compared to New York City ELA score distributions from prior New York City Testing Program administrations

- The comparisons were made by matching scaled scores with New York City ELA scores through mid-interval percentiles. This type of equipercentile equating (Kolen & Brennan, 1995) resulted in a to New York City ELA conversion for each scaled score that appeared in the score distributions compared.

- The ; to New York City ELA conversions were smoothed to obtain the scaled scores not appearing in the original score distributions. The 4th degree polynomial smoothing functions preserved the first 4 moments of the original frequency distributions (mean, variance, skewness, and kurtosis).

- The original New York City ELA scales did not provide differentiated scaled scores for raw scores below chance, resulting in repeated scaled scores for these raw scores. To eliminate the potential for repeated scaled scores at either tail of a scaled score distribution, adjustments were required at the tails of the distributions. The resulting adjustments affected, potentially, one percentage or less of student scores. Since the affected scores were well below the cut score for the lowest performance level, these adjustments have no effect on percentages achieving each performance level.

- Once the scaled-score look-up tables were complete, the raw score to scaled score conversions for the spring 2003 scores were entered into tables to derive the final New York City ELA scaled scores utilized in the spring 2003 ELA administration.

The raw score to scaled score conversion table for spring 2003 is presented in Appendix B. The three performance level cut score points at each grade are shaded for quick reference. The shaded cut-point scores mark the first score in performance levels two, three and four. Table 4 presents summary scaled-score statistics for each grade. A summary of the cut score points and maximum and minimum scaled score values are presented in Appendix C. Figure 1 displays the scaled-score frequencies for each grade graphically. The frequency distribution values are tabled in Appendix D.

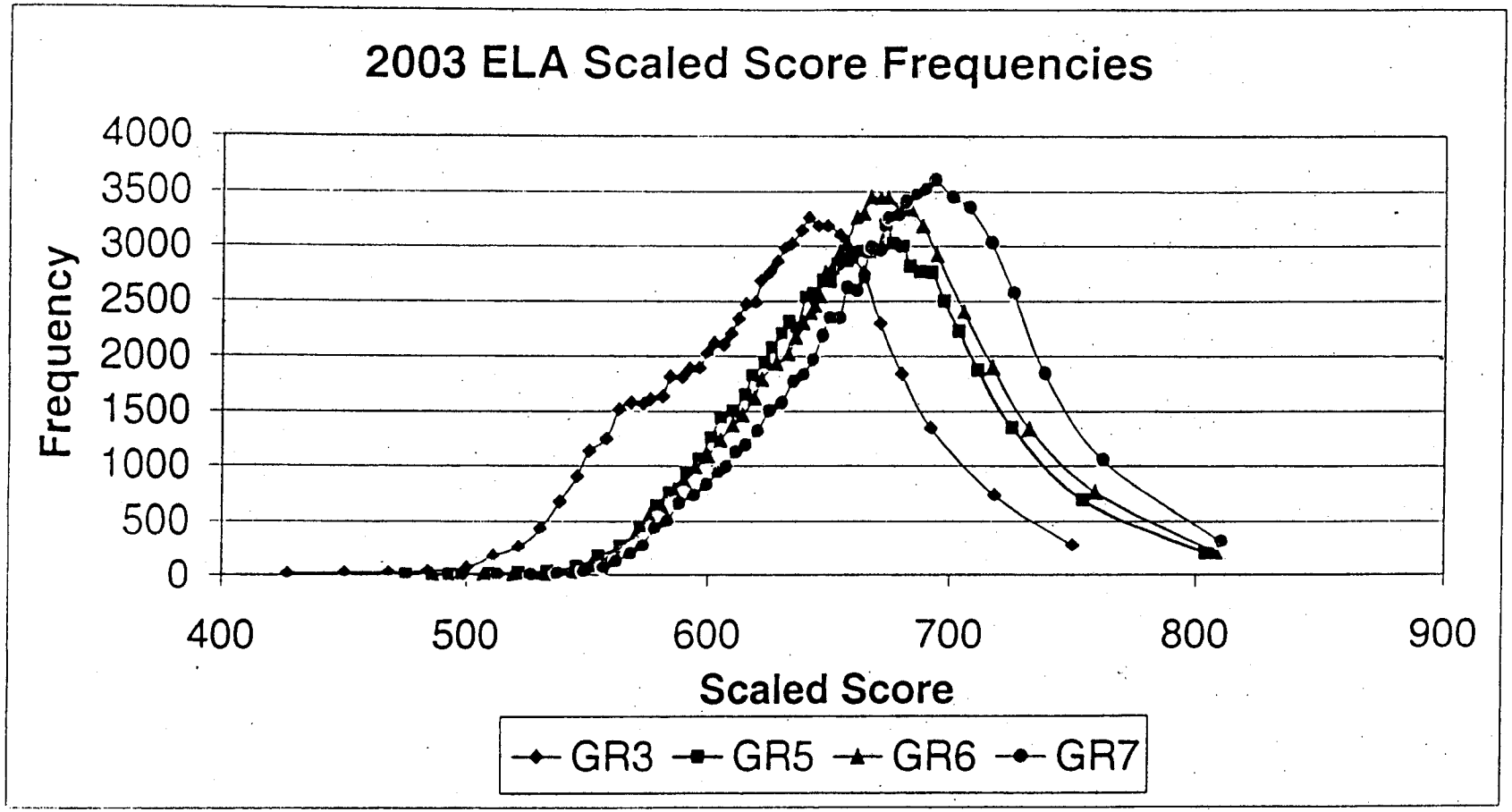


**Figure 1. Spring 2003 Scaled Score Frequencies.**

Table 4. New York City Scaled Score Summary Statistics.

| Grade | Test Mean | Standard Deviation | Minimum Observed Scaled Score | Maximum Observed Scaled Score |
|---|---|---|---|---|
| 3 | 620 | 37.75 | 427 | 750 |
| 5 | 656 | 35.20 | 475 | 805 |
| 6 | 656 | 35.48 | 486 | 808 |
| 7 | 669 | 38.51 | 498 | 810 |

## Differential Item Functioning

Test bias results when systematic errors of measurement occur as a result of sources that are irrelevant to the construct that the test or test items are intended to measure. If test scores demonstrate that a test or test item favors one group over another, it does not necessarily mean that the test or item is biased. Differences in performance among groups on a test item or test may be the result of true subject matter knowledge differences and thus, would be considered an appropriate measure of that construct.

In recent years, there has been a shift in the interpretation of bias. A more neutral term, differential item functioning (DIF) has replaced item bias due to the confusion that the term "bias" created in the literature (Angoff, 1993). DIF refers to a statistical difference in response patterns for subgroups of the population after controlling for ability. DIF occurs when members of a particular group have differing probabilities of success than that of another group after controlling for ability.

### Mantel-Haenszel Procedure

To determine DIF for the 2003 New York City ELA field-test items, the *Mantel Haenszel* (MH) procedure was utilized. The MH procedure compares the odds of a correct response on a studied item for a member of one group with that of a comparable member of another group. The MH procedure provides two statistics:

- The **MH alpha** ($\alpha_{MH}$), which estimates the constant odds ratio, provides an estimate of DIF effect size. A ratio value of 1.00 on the $\alpha_{MH}$ combined-ratio scale (0 to ∞) is consistent with items exhibiting no statistical DIF (Camilli & Shepard, 1994; Dorans & Holland, 1993).

- The **MH chi-square** ($MH\chi^2$) is a statistical test of the null hypothesis ($H_0$: no DIF) and is distributed as an approximate chi-square distribution with one degree of freedom (Holland & Thayer, 1988). The null hypothesis (no

difference in the probability of success) is consistent with items that do not exhibit DIF, and the alternative hypothesis (differing probability of success) is more consistent with items that exhibit statistical DIF.

The purpose of the MH procedure is to compare examinee performance of a reference and a focal group once they have been matched on ability. The reference group typically refers to a traditionally agreed upon majority group, such as, a male or white subgroup. For the 2003 New York City ELA tests the reference groups chosen were the male (gender) and white (ethnic) subgroups. The focal group refers to a group compared to the reference group. The focal groups chosen for the 2003 DIF analysis were female, for gender comparisons, Asian, Hispanic, and African American subgroups, for ethnic comparisons.

A brief description of the MH process follows:

For each item at every score level, j, data from the subgroups are arranged into a 2 x 2 contingency table as depicted in Figure 2 below.

**Figure 2: General Notation for the *j*th Total Score on the Test**

| Group | Score on Studied Item | | |
|---|---|---|---|
| | Correct (1) | Incorrect (0) | Total |
| Reference Group | $A_j$ | $B_j$ | $N_{Rj}$ |
| Focal Group | $C_j$ | $D_j$ | $N_{Fj}$ |
| Total | $M_{1j}$ | $M_{0j}$ | $T_j$ |

where, $A_j$ = the number of examinees in the reference group with score level *j* who answered the item *correctly*

$B_j$ = the number of examinees in the reference group with score level *j* who answered the item *incorrectly*

$C_j$ = the number of examinees in the focal group with score level *j* who answered the item *correctly*

$D_j$ = the number of examinees in the focal group with score level *j who* answered the item *incorrectly*

$N_{Rj} = A_j + B_j$

$N_{Fj} = C_j + D_j$

$M_{1j}$= the number of examinees with score level *j* who answered the item *correctly;* $M_{1j} = A_j + C_j$

$M_{0j}$= the number of examinees with score level *j* who answered the item *incorrectly*; $M_{0j} = B_j + D_j$

$T_j$= the total number of examinees with score level *j* who answered the item; $T_j = A_j + B_j + C_j + D_j$

The $\alpha_{MH}$ odds-ratio, below, compares the reference group with the focal group.

$$\alpha_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j}.$$

In order to test $H_0$, no DIF, the $MH\chi^2$ test of significance, below, is used.

$$MH\chi^2 = \frac{\{|\sum_{j=1}^{S}[A_j - E(A_j)]| - 1/2\}^2}{\sum_{j=1}^{S} VAR(A_j)},$$

where,

$$VAR(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

and

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$$

## Subgroups

Subgroup membership was determined through coding provided on the New York City research file. File preparation for the MH procedure consisted of randomly sampling item response sets within each subgroup to form comparison groups similar in size and of a size compatible with the sensitivity of the $MH\chi^2$. Conducting the MH procedure with extremely large groups tends to overstate the statistical significance of group performance differences.

Reported here is a summary of the $MH\chi^2$ statistics for the embedded field-test items administered across the 26 test forms per grade. $MH\chi^2$ values beyond the 0.001 significance level were flagged as exhibiting statistical DIF. The number of items flagged and indication of the subgroup favored are reported in Table 5.

Reference (R) groups are abbreviated in the table as follows:

- W: White
- M: Male

Focal (F) group abbreviations are

- AS: Asian
- H: Hispanic
- AF: African Americans
- F: Female

Although the research file contained a code for American Indian students, the number of students was not large enough to provide stable statistical values. Therefore, the subgroup of American Indian is not reported. With four sets of comparisons for each test item, there were 520 possible comparisons at each grade for grades 3, 5, and 6 (130 field-test items scored at each grade) and 516 for grade 7 (129 field-test items scored).

**Table 5. Summary of DIF Analysis for Field-Test Items**

| | Reference/Focal Groups | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Favors | | Favors | | Favors | | Favors | |
| Grade | W | AS | W | H | W | AF | M | F |
| 3 | 4 | 1 | 11 | 0 | 9 | 0 | 0 | 0 |
| 5 | 2 | 0 | 3 | 0 | 2 | 0 | 0 | 2 |
| 6 | 1 | 2 | 9 | 1 | 7 | 0 | 2 | 1 |
| 7 | 0 | 0 | 9 | 0 | 7 | 0 | 2 | 5 |

## References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp.3 – 23). Hillsdale, NJ: Erlbaum.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Dorans, M. J. & Holland, P. W. (1993). DIF detection and description: Mantel – Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp.35 – 66). Hillsdale, NJ: Erlbaum.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (3rd ed., pp.105 – 146). New York: Macmillan.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Erlbaum.

Kolen, M.J., & Brennan, R.L. (1995). *Test Equating: Methods and Practices.* New York, NY: Springer-Verlag.

Thorndike, R. M., Cunningham, G. K., Thorndike. R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*, (5th ed.). New York: Macmillan.

# APPENDIX A

## Raw Score Summary Statistics Disaggregated by Gender and Ethnicity

## Table 6. Summary Statistics: Grade 3

| Group | Number Of Students | Number Of Test Items | Test Mean | Standard Deviation | Standard Error Of Measurement | KR-20 Reliability Estimate |
|---|---|---|---|---|---|---|
| Male | 40079 | 45 | 27.5 | 9.72 | 2.84 | 0.91 |
| Female | 38150 | 45 | 29.3 | 9.08 | 2.78 | 0.91 |
| Am. Indian | 472 | 45 | 25.9 | 9.96 | 2.98 | 0.92 |
| Asian | 9023 | 45 | 33.1 | 7.76 | 2.56 | 0.89 |
| Hispanic | 29860 | 45 | 26.9 | 9.16 | 2.89 | 0.90 |
| African Am. | 27954 | 45 | 26.3 | 9.30 | 2.91 | 0.90 |
| Caucasian | 11212 | 45 | 33.5 | 8.61 | 2.51 | 0.92 |

Note: Tabled SEM values were calculated before rounding *SD* and $KR_{20}$.

## Table 7. Summary Statistics: Grade 5

| Group | Number Of Students | Number Of Test Items | Test Mean | Standard Deviation | Standard Error Of Measurement | KR-20 Reliability Estimate |
|---|---|---|---|---|---|---|
| Male | 39153 | 45 | 28.0 | 9.16 | 2.86 | 0.90 |
| Female | 37952 | 45 | 29.2 | 8.56 | 2.81 | 0.89 |
| Am. Indian | 374 | 45 | 25.7 | 8.74 | 2.96 | 0.89 |
| Asian | 8497 | 45 | 33.1 | 7.76 | 2.60 | 0.89 |
| Hispanic | 29308 | 45 | 26.9 | 8.57 | 2.91 | 0.88 |
| African Am. | 27774 | 45 | 26.8 | 8.61 | 2.92 | 0.88 |
| Caucasian | 11124 | 45 | 33.9 | 7.87 | 2.54 | 0.90 |

Note: Tabled SEM values were calculated before rounding *SD* and $KR_{20}$.

## Table 8. Summary Statistics: Grade 6

| Group | Number Of Students | Number Of Test Items | Test Mean | Standard Deviation | Standard Error Of Measurement | KR-20 Reliability Estimate |
|---|---|---|---|---|---|---|
| Male | 39414 | 45 | 28.8 | 8.95 | 2.81 | 0.90 |
| Female | 37568 | 45 | 30.4 | 8.33 | 2.75 | 0.89 |
| Am. Indian | 311 | 45 | 25.7 | 8.61 | 2.95 | 0.88 |
| Asian | 8287 | 45 | 34.1 | 7.57 | 2.52 | 0.89 |
| Hispanic | 29288 | 45 | 28.1 | 8.52 | 2.86 | 0.89 |
| African Am. | 28068 | 45 | 27.9 | 8.46 | 2.86 | 0.89 |
| Caucasian | 11012 | 45 | 34.2 | 7.60 | 2.50 | 0.89 |

Note: Tabled SEM values were calculated before rounding *SD* and $KR_{20}$.

## Table 9. Summary Statistics: Grade 7

| Group | Number Of Students | Number Of Test Items | Test Mean | Standard Deviation | Standard Error Of Measurement | KR-20 Reliability Estimate |
|---|---|---|---|---|---|---|
| Male | 38254 | 45 | 29.7 | 8.87 | 2.73 | 0.90 |
| Female | 36581 | 45 | 31.6 | 8.23 | 2.66 | 0.90 |
| Am. Indian | 281 | 45 | 27.2 | 8.99 | 2.85 | 0.90 |
| Asian | 8146 | 45 | 35.0 | 7.32 | 2.43 | 0.89 |
| Hispanic | 28070 | 45 | 29.2 | 8.50 | 2.78 | 0.89 |
| African Am. | 27301 | 45 | 29.0 | 8.46 | 2.79 | 0.89 |
| Caucasian | 11025 | 45 | 35.2 | 7.32 | 2.41 | 0.89 |

Note: Tabled SEM values were calculated before rounding *SD* and $KR_{20}$.

# APPENDIX B
## Raw Score to Scaled Score Table

## Table 10. Spring 2003 Raw Scores to Scaled Scores

| Raw Scores | Scaled Scores | | | |
|---|---|---|---|---|
| | Grade 3 | Grade 5 | Grade 6 | Grade 7 |
| 1 | 427 | 475 | 486 | 498 |
| 2 | 450 | 493 | 507 | 513 |
| 3 | 468 | 510 | 520 | 527 |
| 4 | 484 | 522 | 533 | 538 |
| 5 | 500 | 534 | 545 | 549 |
| 6 | 511 | 546 | 552 | 557 |
| 7 | 522 | 555 | 560 | 562 |
| 8 | 531 | 564 | 565 | 568 |
| 9 | 539 | 572 | 570 | 573 |
| 10 | 546 | 579 | 576 | 578 |
| 11 | 551 | 584 | 581 | 583 |
| 12 | 558 | 591 | 586 | 588 |
| 13 | 563 | 596 | 590 | 594 |
| 14 | 568 | 601 | 595 | 599 |
| 15 | 573 | 605 | 600 | 604 |
| 16 | 576 | 610 | 605 | 607 |
| 17 | 581 | 615 | 610 | 611 |
| 18 | 584 | [redacted] | 614 | 615 |
| 19 | 589 | 623 | 619 | 620 |
| 20 | [redacted] | 626 | 622 | 625 |
| 21 | 596 | 630 | 628 | 630 |
| 22 | 599 | 633 | 633 | 635 |
| 23 | 602 | 636 | [redacted] | 639 |
| 24 | 606 | 640 | 639 | 643 |
| 25 | 609 | 643 | 642 | [redacted] |
| 26 | 612 | 647 | 644 | 650 |
| 27 | 615 | 650 | 646 | 654 |
| 28 | 619 | 653 | 648 | 657 |
| 29 | 621 | 657 | 651 | 661 |
| 30 | 625 | 659 | 654 | 664 |
| 31 | 628 | 663 | 657 | 667 |
| 32 | 631 | 666 | 661 | 671 |
| 33 | 634 | 669 | 664 | 674 |
| 34 | 638 | 673 | 667 | 678 |
| 35 | 641 | 676 | 671 | 681 |
| 36 | 645 | 680 | 674 | 685 |
| 37 | 649 | 683 | 679 | 689 |
| 38 | 654 | 687 | 683 | 693 |
| 39 | 658 | 692 | 688 | 700 |
| 40 | [redacted] | 697 | 694 | 707 |
| 41 | 671 | [redacted] | 705 | 716 |
| 42 | 680 | 711 | 717 | [redacted] |
| 43 | 692 | 725 | [redacted] | 738 |
| 44 | 718 | 754 | 759 | 762 |
| 45 | 750 | 804 | 808 | 810 |

Note: Cut Score Points are shaded.

# APPENDIX C

## Raw Score and Scaled Score Values For New York City Performance Level Cut Scores, Maximum and Minimum Possible Scores

**Table 11. Raw Score and Scaled Score Cut Scores.**

| Grade | Performance Level | | | | | |
|---|---|---|---|---|---|---|
| | Level 2 | | Level 3 | | Level 4 | |
| | Raw Score | Scaled Score | Raw Score | Scaled Score | Raw Score | Scaled Score |
| 3 | 20 | 592 | 32 | 631 | 40 | 664 |
| 5 | 18 | 618 | 29 | 657 | 41 | 703 |
| 6 | 23 | 636 | 36 | 674 | 43 | 732 |
| 7 | 25 | 647 | 36 | 685 | 42 | 725 |

**Table 12. Raw Score and Scaled Score Minimum And Maximum Score Points.**

| Grade | Score Points | | | |
|---|---|---|---|---|
| | Minimum | | Maximum | |
| | Raw Score | Scaled Score | Raw Score | Scaled Score |
| 3 | 1 | 427 | 45 | 750 |
| 5 | 1 | 475 | 45 | 804 |
| 6 | 1 | 486 | 45 | 808 |
| 7 | 1 | 498 | 45 | 810 |

# APPENDIX D
## Frequency Distributions

## Table 13. Spring 2003 Frequency Distributions: Grade 3

| Raw Score | Scaled Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 0 | 427 | 17 | 0.02 | 17 | 0.02 |
| 1 | 427 | 14 | 0.02 | 31 | 0.04 |
| 2 | 450 | 22 | 0.03 | 53 | 0.07 |
| 3 | 468 | 24 | 0.03 | 77 | 0.10 |
| 4 | 484 | 37 | 0.05 | 114 | 0.15 |
| 5 | 500 | 72 | 0.09 | 186 | 0.24 |
| 6 | 511 | 180 | 0.23 | 366 | 0.47 |
| 7 | 522 | 260 | 0.33 | 626 | 0.80 |
| 8 | 531 | 430 | 0.55 | 1056 | 1.35 |
| 9 | 539 | 671 | 0.86 | 1727 | 2.21 |
| 10 | 546 | 899 | 1.15 | 2626 | 3.35 |
| 11 | 551 | 1132 | 1.45 | 3758 | 4.80 |
| 12 | 558 | 1243 | 1.59 | 5001 | 6.39 |
| 13 | 563 | 1511 | 1.93 | 6512 | 8.32 |
| 14 | 568 | 1574 | 2.01 | 8086 | 10.33 |
| 15 | 573 | 1562 | 2.00 | 9648 | 12.32 |
| 16 | 576 | 1606 | 2.05 | 11254 | 14.37 |
| 17 | 581 | 1632 | 2.08 | 12886 | 16.46 |
| 18 | 584 | 1814 | 2.32 | 14700 | 18.78 |
| 19 | 589 | 1807 | 2.31 | 16507 | 21.08 |
| 20 | 592 | 1884 | 2.41 | 18391 | 23.49 |
| 21 | 596 | 1891 | 2.42 | 20282 | 25.90 |
| 22 | 599 | 2024 | 2.59 | 22306 | 28.49 |
| 23 | 602 | 2117 | 2.70 | 24423 | 31.19 |
| 24 | 606 | 2101 | 2.68 | 26524 | 33.88 |
| 25 | 609 | 2201 | 2.81 | 28725 | 36.69 |
| 26 | 612 | 2335 | 2.98 | 31060 | 39.67 |
| 27 | 615 | 2467 | 3.15 | 33527 | 42.82 |
| 28 | 619 | 2484 | 3.17 | 36011 | 45.99 |
| 29 | 621 | 2680 | 3.42 | 38691 | 49.42 |
| 30 | 625 | 2759 | 3.52 | 41450 | 52.94 |
| 31 | 628 | 2856 | 3.65 | 44306 | 56.59 |
| 32 | 631 | 2974 | 3.80 | 47280 | 60.39 |
| 33 | 634 | 3016 | 3.85 | 50296 | 64.24 |
| 34 | 638 | 3141 | 4.01 | 53437 | 68.25 |
| 35 | 641 | 3261 | 4.17 | 56698 | 72.42 |
| 36 | 645 | 3179 | 4.06 | 59877 | 76.48 |
| 37 | 649 | 3182 | 4.06 | 63059 | 80.54 |
| 38 | 654 | 3095 | 3.95 | 66154 | 84.49 |
| 39 | 658 | 2904 | 3.71 | 69058 | 88.20 |
| 40 | 664 | 2752 | 3.51 | 71810 | 91.72 |
| 41 | 671 | 2292 | 2.93 | 74102 | 94.64 |
| 42 | 680 | 1843 | 2.35 | 75945 | 97.00 |
| 43 | 692 | 1344 | 1.72 | 77289 | 98.72 |
| 44 | 718 | 728 | 0.93 | 78017 | 99.64 |
| 45 | 750 | 278 | 0.36 | 78295 | 100.00 |

## Table 14. Spring 2003 Frequency Distributions: Grade 5

| | | | | Cumulative | |
|---|---|---|---|---|---|
| Raw Score | Scaled Score | Frequency | Percent | Frequency | Percent |
| 0 | 475 | 1 | 0.00 | 1 | 0.00 |
| 1 | 475 | 4 | 0.01 | 5 | 0.01 |
| 2 | 493 | 4 | 0.01 | 9 | 0.01 |
| 3 | 510 | 11 | 0.01 | 20 | 0.03 |
| 4 | 522 | 20 | 0.03 | 40 | 0.05 |
| 5 | 534 | 35 | 0.05 | 75 | 0.10 |
| 6 | 546 | 83 | 0.11 | 158 | 0.20 |
| 7 | 555 | 180 | 0.23 | 338 | 0.44 |
| 8 | 564 | 273 | 0.35 | 611 | 0.79 |
| 9 | 572 | 444 | 0.58 | 1055 | 1.37 |
| 10 | 579 | 630 | 0.82 | 1685 | 2.18 |
| 11 | 584 | 750 | 0.97 | 2435 | 3.16 |
| 12 | 591 | 932 | 1.21 | 3367 | 4.36 |
| 13 | 596 | 1057 | 1.37 | 4424 | 5.73 |
| 14 | 601 | 1254 | 1.63 | 5678 | 7.36 |
| 15 | 605 | 1437 | 1.86 | 7115 | 9.22 |
| 16 | 610 | 1499 | 1.94 | 8614 | 11.17 |
| 17 | 615 | 1646 | 2.13 | 10260 | 13.30 |
| 18 | 618 | 1823 | 2.36 | 12083 | 15.66 |
| 19 | 623 | 1943 | 2.52 | 14026 | 18.18 |
| 20 | 626 | 2075 | 2.69 | 16101 | 20.87 |
| 21 | 630 | 2198 | 2.85 | 18299 | 23.72 |
| 22 | 633 | 2312 | 3.00 | 20611 | 26.72 |
| 23 | 636 | 2253 | 2.92 | 22864 | 29.64 |
| 24 | 640 | 2529 | 3.28 | 25393 | 32.92 |
| 25 | 643 | 2567 | 3.33 | 27960 | 36.24 |
| 26 | 647 | 2687 | 3.48 | 30647 | 39.73 |
| 27 | 650 | 2678 | 3.47 | 33325 | 43.20 |
| 28 | 653 | 2835 | 3.67 | 36160 | 46.87 |
| 29 | 657 | 2863 | 3.71 | 39023 | 50.58 |
| 30 | 659 | 2896 | 3.75 | 41919 | 54.34 |
| 31 | 663 | 2946 | 3.82 | 44865 | 58.16 |
| 32 | 666 | 2947 | 3.82 | 47812 | 61.98 |
| 33 | 669 | 2948 | 3.82 | 50760 | 65.80 |
| 34 | 673 | 3192 | 4.14 | 53952 | 69.94 |
| 35 | 676 | 3020 | 3.91 | 56972 | 73.85 |
| 36 | 680 | 2994 | 3.88 | 59966 | 77.73 |
| 37 | 683 | 2814 | 3.65 | 62780 | 81.38 |
| 38 | 687 | 2766 | 3.59 | 65546 | 84.97 |
| 39 | 692 | 2757 | 3.57 | 68303 | 88.54 |
| 40 | 697 | 2496 | 3.24 | 70799 | 91.78 |
| 41 | 703 | 2227 | 2.89 | 73026 | 94.66 |
| 42 | 711 | 1875 | 2.43 | 74901 | 97.09 |
| 43 | 725 | 1347 | 1.75 | 76248 | 98.84 |
| 44 | 754 | 689 | 0.89 | 76937 | 99.73 |
| 45 | 804 | 207 | 0.27 | 77144 | 100.00 |

## Table 15. Spring 2003 Frequency Distributions: Grade 6

| Raw Score | Scaled Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 0 | 486 | 2 | 0.00 | 2 | 0.00 |
| 1 | 486 | 9 | 0.01 | 11 | 0.01 |
| 2 | 507 | 5 | 0.01 | 16 | 0.02 |
| 3 | 520 | 11 | 0.01 | 27 | 0.04 |
| 4 | 533 | 16 | 0.02 | 43 | 0.06 |
| 5 | 545 | 38 | 0.05 | 81 | 0.11 |
| 6 | 552 | 101 | 0.13 | 182 | 0.24 |
| 7 | 560 | 161 | 0.21 | 343 | 0.45 |
| 8 | 565 | 261 | 0.34 | 604 | 0.78 |
| 9 | 570 | 371 | 0.48 | 975 | 1.27 |
| 10 | 576 | 553 | 0.72 | 1528 | 1.98 |
| 11 | 581 | 637 | 0.83 | 2165 | 2.81 |
| 12 | 586 | 778 | 1.01 | 2943 | 3.82 |
| 13 | 590 | 872 | 1.13 | 3815 | 4.95 |
| 14 | 595 | 982 | 1.27 | 4797 | 6.23 |
| 15 | 600 | 1091 | 1.42 | 5888 | 7.64 |
| 16 | 605 | 1230 | 1.60 | 7118 | 9.24 |
| 17 | 610 | 1372 | 1.78 | 8490 | 11.02 |
| 18 | 614 | 1462 | 1.90 | 9952 | 12.92 |
| 19 | 619 | 1620 | 2.10 | 11572 | 15.02 |
| 20 | 622 | 1792 | 2.33 | 13364 | 17.35 |
| 21 | 628 | 1936 | 2.51 | 15300 | 19.86 |
| 22 | 633 | 2018 | 2.62 | 17318 | 22.48 |
| 23 | 636 | 2174 | 2.82 | 19492 | 25.30 |
| 24 | 639 | 2302 | 2.99 | 21794 | 28.29 |
| 25 | 642 | 2393 | 3.11 | 24187 | 31.39 |
| 26 | 644 | 2462 | 3.20 | 26649 | 34.59 |
| 27 | 646 | 2546 | 3.30 | 29195 | 37.89 |
| 28 | 648 | 2753 | 3.57 | 31948 | 41.47 |
| 29 | 651 | 2805 | 3.64 | 34753 | 45.11 |
| 30 | 654 | 2945 | 3.82 | 37698 | 48.93 |
| 31 | 657 | 3034 | 3.94 | 40732 | 52.87 |
| 32 | 661 | 3261 | 4.23 | 43993 | 57.10 |
| 33 | 664 | 3300 | 4.28 | 47293 | 61.39 |
| 34 | 667 | 3446 | 4.47 | 50739 | 65.86 |
| 35 | 671 | 3434 | 4.46 | 54173 | 70.32 |
| 36 | 674 | 3440 | 4.47 | 57613 | 74.78 |
| 37 | 679 | 3373 | 4.38 | 60986 | 79.16 |
| 38 | 683 | 3341 | 4.34 | 64327 | 83.50 |
| 39 | 688 | 3186 | 4.14 | 67513 | 87.63 |
| 40 | 694 | 2915 | 3.78 | 70428 | 91.42 |
| 41 | 705 | 2400 | 3.12 | 72828 | 94.53 |
| 42 | 717 | 1901 | 2.47 | 74729 | 97.00 |
| 43 | 732 | 1337 | 1.74 | 76066 | 98.73 |
| 44 | 759 | 759 | 0.99 | 76825 | 99.72 |
| 45 | 808 | 217 | 0.28 | 77042 | 100.00 |

## Table 16. Spring 2003 Frequency Distributions: Grade 7

| Raw Score | Scaled Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 0 | 498 | 2 | 0.00 | 2 | 0.00 |
| 1 | 498 | 1 | 0.00 | 3 | 0.00 |
| 2 | 513 | 4 | 0.01 | 7 | 0.01 |
| 3 | 527 | 3 | 0.00 | 10 | 0.01 |
| 4 | 538 | 15 | 0.02 | 25 | 0.03 |
| 5 | 549 | 32 | 0.04 | 57 | 0.08 |
| 6 | 557 | 69 | 0.09 | 126 | 0.17 |
| 7 | 562 | 127 | 0.17 | 253 | 0.34 |
| 8 | 568 | 197 | 0.26 | 450 | 0.60 |
| 9 | 573 | 272 | 0.36 | 722 | 0.96 |
| 10 | 578 | 434 | 0.58 | 1156 | 1.54 |
| 11 | 583 | 494 | 0.66 | 1650 | 2.20 |
| 12 | 588 | 655 | 0.87 | 2305 | 3.08 |
| 13 | 594 | 722 | 0.96 | 3027 | 4.04 |
| 14 | 599 | 821 | 1.10 | 3848 | 5.14 |
| 15 | 604 | 931 | 1.24 | 4779 | 6.38 |
| 16 | 607 | 985 | 1.32 | 5764 | 7.70 |
| 17 | 611 | 1119 | 1.49 | 6883 | 9.19 |
| 18 | 615 | 1184 | 1.58 | 8067 | 10.77 |
| 19 | 620 | 1313 | 1.75 | 9380 | 12.52 |
| 20 | 625 | 1498 | 2.00 | 10878 | 14.52 |
| 21 | 630 | 1575 | 2.10 | 12453 | 16.63 |
| 22 | 635 | 1773 | 2.37 | 14226 | 18.99 |
| 23 | 639 | 1834 | 2.45 | 16060 | 21.44 |
| 24 | 643 | 1964 | 2.62 | 18024 | 24.07 |
| 25 | 647 | 2183 | 2.91 | 20207 | 26.98 |
| 26 | 650 | 2343 | 3.13 | 22550 | 30.11 |
| 27 | 654 | 2342 | 3.13 | 24892 | 33.24 |
| 28 | 657 | 2621 | 3.50 | 27513 | 36.74 |
| 29 | 661 | 2592 | 3.46 | 30105 | 40.20 |
| 30 | 664 | 2721 | 3.63 | 32826 | 43.83 |
| 31 | 667 | 2984 | 3.98 | 35810 | 47.81 |
| 32 | 671 | 2962 | 3.95 | 38772 | 51.77 |
| 33 | 674 | 3259 | 4.35 | 42031 | 56.12 |
| 34 | 678 | 3279 | 4.38 | 45310 | 60.50 |
| 35 | 681 | 3409 | 4.55 | 48719 | 65.05 |
| 36 | 685 | 3462 | 4.62 | 52181 | 69.67 |
| 37 | 689 | 3511 | 4.69 | 55692 | 74.36 |
| 38 | 693 | 3605 | 4.81 | 59297 | 79.17 |
| 39 | 700 | 3442 | 4.60 | 62739 | 83.77 |
| 40 | 707 | 3353 | 4.48 | 66092 | 88.25 |
| 41 | 716 | 3026 | 4.04 | 69118 | 92.29 |
| 42 | 725 | 2570 | 3.43 | 71688 | 95.72 |
| 43 | 738 | 1845 | 2.46 | 73533 | 98.18 |
| 44 | 762 | 1050 | 1.40 | 74583 | 99.58 |
| 45 | 810 | 312 | 0.42 | 74895 | 100.00 |