*Cuadernos de Investigación*
de la Oficina de Puerto Rico y América Latina

# Reliability Analyses for the English Language Assessment for Hispanics (ELASH), Levels I and II

***Gary L. Marco***

The College Board
Oficina de Puerto Rico y América Latina

Cuadernos de Investigación #8
Marzo 2004

The College Board is a national nonprofit membership association whose mission is to prepare, inspire, and connect students to college and opportunity. Founded in 1900, the association is composed of more than 4,200 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.
For further information, visit www.collegeboard.com.

La Oficina de Puerto Rico y América Latina (OPR/AL) desarrolla programas y servicios similares a los que se ofrecen en los Estados Unidos, pero especialmente diseñados para poblaciones cuyo vernáculo es el español. Estos programas están dirigidos a sistematizar los procesos de evaluación y admisión universitaria, fortalecer la orientación académica y personal y a promover la excelencia educativa. Entre nuestros programas más conocidos se encuentran: el Programa de Evaluación para Admisión Universitaria (PEAU), que incluye la Prueba de Aptitud Académica (PAA) y las Pruebas de Aprovechamiento (PACH), las Pruebas de Ingreso y Evaluación para el Nivel Secundario (PIENSE), el Programa de Nivel Avanzado, el Inventario CEPA (Conoce, Explora, Planifica y Actúa) y el *English Language Assessment System for Hispanics* (*ELASH*).

The College Board está comprometido con el principio de igualdad de oportunidades y sus programas, servicios y política de empleo se rigen por este principio.

The College Board está comprometido con el principio de no discriminación y en combatir el hostigamiento sexual en el reclutamiento de personal, así como en todos los servicios que ofrece y en las actividades que desarrolla.

El College Board basa el empleo en la capacidad personal y la preparación, sin discriminar por razón de raza, color, origen nacional, religión, sexo, edad, condición social, afiliación política, impedimento o cualquier otra característica protegida por la Ley.

Esta publicación está disponible a un costo de $5.00 cada ejemplar, en la Oficina del College Board, Banco Popular Center, Piso 15, Oficina 1501, Hato Rey, Puerto Rico. Puede solicitarla por correo; añada $1.50 para franqueo.

Para comunicarse con nuestra oficina, puede llamar al (787) 759-8625 y a través de nuestro facsímil (787) 764- 4306, o escribirnos a:

*The College Board*
*Oficina de Puerto Rico*
*y América Latina*
*P.O. Box 71101*
*San Juan, Puerto Rico 00936-8001*

# Reliability Analyses for the English Language Assessment System for Hispanics (ELASH), Levels I and II

## Report Prepared for the College Board Puerto Rico and Latin American Office

Gary L. Marco

***Dr. Gary Marco*** was the Executive Director for College Board Statistical Analysis Area for over 20 years at ETS. He also served on the team that designed and produced the new test creation system at ETS. He has conducted research and published numerous articles related to educational measurement, applied statistics, and psychometrics. Currently, he serves as a private consultant to the educational measurement community.


Oficina de Puerto Rico y América Latina

# CONTENTS

## Chapter 1

# Chapter 1

A measure of a test's reliability is commonly used to indicate how consistent test scores are from one occasion to another. There are several types of reliability, depending on how the reliability is assessed. In principle, parallel form reliability, which measures the consistency of scores on two forms of a test—ideally taken within a short time period (a few days or weeks)—is usually considered superior to other types of reliability. This kind of reliability takes into account all sources of measurement error (day-to-day variations in the functioning of test takers as well as form-to-form variations in test items).

Obtaining two scores from a test taker within a short period of time is, however, difficult logistically. Therefore, data from a single test form are commonly used to assess reliability. This type of reliability is called internal consistency reliability because it assesses how consistent scores internal to a test (say, scores on split-halves) are. Day-to-day sources of measurement error are, however, not accounted for by this type of reliability. Even though the reliability estimate may be slightly too high, internal consistency estimates are commonly used to assess score reliability simply because they use data readily available to the statistician.

This study evaluated internal consistency reliability for the English Language Assessment System for Hispanics (ELASH). One purpose of the study was to evaluate whether total test and part test reliability estimates and standard errors of measurement were improved by using the appropriate part or component reliability information to compute reliability coefficients and standard errors. A second purpose was to provide internal consistency reliability information (reliability coefficients and standard errors of measurement) for Level I and Level II total, part, and component tests. (See the section on tests for a description of the part and component tests for the two test levels.) A third purpose was to provide estimates of the scale score standard errors of measurement for the Total and part test scale scores. Part test scale score standard errors could be computed directly from raw score standard errors. The Total scale score standard errors, however, had to be derived indirectly from the part test scale score standard errors. This indirect method was needed because the Total score is not based on the total test raw

score but rather on the average of the part test scale scores. A final purpose was to compare differences in reliability information for Forms A and B of the Listening Comprehension test, each of which were administered to Level I and Level II samples.

## Tests

ELASH consists of two levels (Level I and Level II), each represented by two test forms (Form A and Form B). Each test form consists of three parts and seven components. The breakdown of the parts in terms of components is as follows:

**Level I, Forms A and B:**

**Listening Comprehension-50 items**
Rejoinders-25 items
Short Conversations-15 items
Discourse-10 items

**Language Usage and Indirect Writing-35 items**
Language Usage-21 items
Indirect Writing-14 items

**Vocabulary and Reading-35 items**
Vocabulary-10 items
Reading-25 items

**Level II, Forms A and B:**

**Listening Comprehension-50 items**
Rejoinders-25 items
Short Conversations-15 items
Discourse-10 items

**Language Usage and Indirect Composition-35 items**
Language Usage-17 items
Indirect Composition-18 items

**Idiomatic Expressions and Reading-35 items**
Idiomatic Expressions-5 items
Reading-30 items

Because the Listening Comprehension Test is appropriate for a wide range of abilities, it is the same for Levels I and II. The other two test parts are different; the Level II forms consist of more difficult items.

Reliability analysis were conducted on all three parts and all seven components, as well as on the total test for each test form at each level.

## Samples

The samples for the item and test analyses consisted of the test takers who took Form A or B of ELASH Level I or Level II between June 2000 and April 2002. For these analyses, those test takers who did not reach at least 90% of the items on any of the three test parts were eliminated from the samples. This step was necessary so that for the test takers remaining in the sample, the test was relatively unspeeded. Internal consistency reliability analyses are inappropriate for speeded tests. Table 1 shows the sample scale score statistics for the total and part tests. The scale score means are somewhat higher than those for the full population of test takers because those test takers who did not reach at least 90% of the items were eliminated from the sample. Table 2 provides various raw score statistics for the total, part, and component tests.

## Scale Score Statistics for ELASH Levels I and II, Forms A and B

| Test | Mean | SD | Minimum Scores | Maximum Scores |
|---|---|---|---|---|
| **Level I Form A (Sample Size = 2632)** | | | | |
| Listening Comprehension | 96 | 20 | 54 | 200 |
| Language Usage and Indirect Writing | 104 | 21 | 50 | 150 |
| Vocabulary and Reading | 110 | 26 | 40 | 150 |
| Total Test | 103 | 20 | 59 | 166 |
| **Level I Form B (Sample Size = 5725)** | | | | |
| Listening Comprehension | 105 | 21 | 58 | 200 |
| Language Usage and Indirect Writing | 115 | 24 | 50 | 150 |
| Vocabulary and Reading | 115 | 26 | 41 | 150 |
| Total Test | 112 | 22 | 63 | 167 |
| **Level II Form A (Sample Size = 3293)** | | | | |
| Listening Comprehension | 128 | 27 | 46 | 200 |
| Language Usage and Indirect Writing | 130 | 31 | 43 | 200 |
| Vocabulary and Reading | 131 | 27 | 48 | 200 |
| Total Test | 130 | 26 | 64 | 200 |
| **Level II Form B (Sample Size = 2477)** | | | | |
| Listening Comprehension | 133 | 24 | 58 | 200 |
| Language Usage and Indirect Writing | 135 | 29 | 46 | 200 |
| Vocabulary and Reading | 140 | 27 | 67 | 200 |
| Total Test | 136 | 24 | 79 | 200 |

Table 1

## Raw Score Statistics for ELASH Levels I and II, Forms A and B

| Test | No. of Items | Raw Score Mean | Raw Score SD | Mean % Correct | Mean Pt. Biserial (Item Deleted) | Alpha Reliability |
|---|---|---|---|---|---|---|
| **Level I Form A (Sample Size = 2632)** | | | | | | |
| Total Test | 120 | 73.1 | 20.7 | 61% | 0.36 | 0.95 |
| Listening Comprehension | 50 | 25.1 | 9.3 | 50% | 0.34 | 0.88 |
| Language Usage and Indirect Writing | 35 | 22.9 | 6.9 | 65% | 0.38 | 0.88 |
| Vocabulary and Reading | 35 | 25.1 | 6.8 | 72% | 0.41 | 0.89 |
| Rejoinders | 25 | 12.9 | 4.9 | 51% | 0.32 | 0.79 |
| Short Conversations | 15 | 7.1 | 3.5 | 48% | 0.37 | 0.76 |
| Discourse | 10 | 5.1 | 2.0 | 51% | 0.20 | 0.49 |
| Language Usage | 21 | 14.3 | 4.3 | 68% | 0.38 | 0.82 |
| Indirect Writing | 14 | 8.6 | 3.1 | 62% | 0.36 | 0.74 |
| Vocabulary | 10 | 8.4 | 1.9 | 84% | 0.38 | 0.72 |
| Reading | 25 | 16.7 | 5.4 | 67% | 0.42 | 0.86 |
| **Level I Form B (Sample Size = 5725)** | | | | | | |
| Total Test | 120 | 69.3 | 21.6 | 58% | 0.38 | 0.96 |
| Listening Comprehension | 50 | 22.1 | 9.1 | 44% | 0.34 | 0.89 |
| Language Usage and Indirect Writing | 35 | 22.6 | 7.5 | 65% | 0.42 | 0.90 |
| Vocabulary and Reading | 35 | 24.5 | 7.0 | 70% | 0.43 | 0.90 |
| Rejoinders | 25 | 12.1 | 4.7 | 48% | 0.30 | 0.77 |
| Short Conversations | 15 | 5.6 | 3.1 | 38% | 0.33 | 0.73 |
| Discourse | 10 | 4.4 | 2.3 | 44% | 0.33 | 0.66 |
| Language Usage | 21 | 14.1 | 4.7 | 67% | 0.42 | 0.85 |
| Indirect Writing | 14 | 8.5 | 3.2 | 61% | 0.38 | 0.77 |
| Vocabulary | 10 | 8.4 | 1.9 | 84% | 0.41 | 0.75 |
| Reading | 25 | 16.1 | 5.5 | 65% | 0.43 | 0.87 |
| **Level II Form A (Sample Size = 3293)** | | | | | | |
| Total Test | 120 | 88.8 | 22.1 | 74% | 0.43 | 0.96 |
| Listening Comprehension | 50 | 37.7 | 9.4 | 75% | 0.43 | 0.92 |
| Language Usage and Indirect Composition | 35 | 26.2 | 7.3 | 75% | 0.45 | 0.91 |
| Idiomatic Expressions and Reading | 35 | 24.9 | 6.9 | 71% | 0.43 | 0.89 |
| Rejoinders | 25 | 18.8 | 4.8 | 75% | 0.41 | 0.85 |
| Short Conversations | 15 | 11.7 | 3.5 | 78% | 0.48 | 0.85 |
| Discourse | 10 | 7.2 | 2.0 | 72% | 0.30 | 0.61 |
| Language Usage | 17 | 12.6 | 3.9 | 74% | 0.45 | 0.85 |
| Indirect Composition | 18 | 13.6 | 3.7 | 76% | 0.42 | 0.83 |
| Idiomatic Expressions | 5 | 3.9 | 1.2 | 77% | 0.34 | 0.56 |
| Reading | 30 | 21.0 | 6.1 | 70% | 0.42 | 0.88 |
| **Level II Form B (Sample Size = 2477)** | | | | | | |
| Total Test | 120 | 81.0 | 21.7 | 68% | 0.40 | 0.96 |
| Listening Comprehension | 50 | 34.0 | 9.2 | 68% | 0.40 | 0.91 |
| Language Usage and Indirect Composition | 35 | 25.5 | 6.7 | 73% | 0.40 | 0.88 |
| Idiomatic Expressions and Reading | 35 | 21.6 | 7.5 | 62% | 0.42 | 0.89 |
| Rejoinders | 25 | 17.6 | 4.4 | 70% | 0.35 | 0.81 |
| Short Conversations | 15 | 9.5 | 3.4 | 63% | 0.40 | 0.79 |
| Discourse | 10 | 6.9 | 2.3 | 69% | 0.38 | 0.71 |
| Language Usage | 17 | 11.7 | 3.6 | 69% | 0.39 | 0.80 |
| Indirect Composition | 18 | 13.8 | 3.5 | 76% | 0.38 | 0.79 |
| Idiomatic Expressions | 5 | 3.1 | 1.4 | 63% | 0.36 | 0.59 |
| Reading | 30 | 18.4 | 6.5 | 61% | 0.41 | 0.87 |

Table 2

# Procedures for Computing Reliability Information

Coefficient alpha (Cronbach, 1951; Lord & Novick, 1968) is the primary measure of reliability used in these analyses. It is a widely recognized measure of internal consistency reliability that has stood the test of time. It is equivalent to Kuder-Richardson Formula 20 (KR20) reliability when test items are scored right or wrong, as was the case here. This type of reliability is essentially equivalent to the mean correlation of scores on all possible split-halves adjusted to a full length (by use of the Spearman-Brown formula). The formula for computing coefficient alpha, $a(xx')$ , is:

$$\alpha(xx') = \frac{k}{k-1}\left(\sigma_x^2 - \frac{\sum\sigma_i^2}{\sigma_x^2}\right)$$

where $k$ is the number of test items, $\sigma_x^2$ is the variance of the score on test x, and $\sigma_i^2$ is the variance of the score on item i (scored 1 for a right answer and 0 for a wrong answer). The raw score standard error of measurement, $\sigma(x_e)$ ,is computed from the reliability coefficient as follows:

$$\sigma(x_e) = \sigma(x)\sqrt{1-\alpha(xx')}$$

This standard error is essentially the average standard deviation of observed scores for a given true score. (Since the true score of an individual is unknown, for practical use the standard error is expressed as a band around a given observed scores.) Theoretically, approximately two-thirds of the observed scores will fall within one standard error of measurement of the true score, and about 95% of the observed scores within two standard errors of the true score. Of course, errors of measurement can vary depending on the score. For a test with items scored right or wrong, the standard errors of measurement are smaller toward the extremes of the score scale, especially the top end. Therefore, the standard error of measurement is best applied over the middle part of the score scale for score interpretation purposes.

Raw score reliabilities and standard errors of measurement of the total and part tests were not only computed from the alpha coefficients but also from subtest (part or component) reliability information. The practical question was whether the use of subtest reliability information reduced the total or part standard error or increased the reliability coefficient by a practically significant amount.

The total test or part test raw score standard error of measurement may be calculated from subtest scores as follows:

$$\sigma(x_e) = \sqrt{\sum\sigma^2(g_e)}$$

where, $\sigma^2(g_e)$, is the raw score variance error of measurement for part test g. Substituting the resulting variance error of measurement into the following gives the associated reliability coefficient:

$$1 - \frac{\sigma^2(x_e)}{\sigma^2(x)}$$

The total test raw score reliability information was thus computed in three different ways: Once from the raw score alpha reliability information, once from the reliability information of the three part tests (Listening Comprehension, etc.), and once from the reliability information of the component tests (Rejoinders, etc.). Part test raw score reliability information was in turn computed in two ways: Once from the raw score alpha reliability information and once from the reliability information of the component tests associated with that part. The raw score standard errors of measurement would be expected to decrease (and the reliability coefficients increase) somewhat with the use of reliability information from the more homogeneous subtests.

In addition, scale score standard errors of measurement were computed for the Total scale score and the part test scale scores. Computing the part test scaled score standard errors involved multiplying the raw score standard error of measurement by a "slope" parameter, a parameter that represents the change in a scale score given a one point change in a raw score. Table 3 shows the mean slopes used to convert raw score standard errors of measurement into scale score standard errors. As the table indicates,

these mean slopes apply to a restricted score range in the middle of the score scale, the range over which the slopes differed only slightly from one another and over which the raw scores bore a linear relationship with the scale score. That is, in this part of score range, the raw scores could be converted to a scale score by multiplying the raw score by the mean slope and adding a constant. The raw-score-to-scale-score conversions from which these mean slopes were computed are shown in Table 4. Since only one set of conversions applies to each of the two Listening Comprehension forms, the slopes for Forms A and B are the same across levels.

**Mean Slopes for Converting Raw Scores to Scale Scores and Raw Score Ranges Over Which the Means Were Computed**

| Part | Parameter | Level I Form B | Level I Form B | Level II Form A | Level II Form B |
|---|---|---|---|---|---|
| Listening Comprehension | Mean Slope | 2.2059 | 2.3714 | 2.2059 | 2.3714 |
| | Raw Score Range | 8 to 41 | 10 to 44 | 8 to 41 | 10 to 44 |
| Language Usage and Indirect Writing (Level I) or Composition (Level II) | Mean Slope | 2.7143 | 3.0526 | 2.9500 | 3.1250 |
| | Raw Score Range | 8 to 28 | 8 to 26 | 8 to 27 | 8 to 25 |
| Vocabulary (Level I) or Idiomatic Expression (Level II) and Reading | Mean Slope | 3.0625 | 3.3333 | 3.0556 | 3.1250 |
| | Raw Score Range | 9 to 24 | 11 to 25 | 8 to 25 | 9 to 24 |

Table 3

# Raw-to-Scale Score Conversions for ELASH Levels I and II, Forms A and B

| Listening Comprehension, Levels I and II | | | | Language Usage and Indirect Writing, Level I | | | | Vocabulary and Reading, Level I | |
|---|---|---|---|---|---|---|---|---|---|
| Form A | | Form B | | Form A | | Form B | | Form A | |
| Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score |
| 0 | 40 | 0 | 40 | 0 | 40 | 0 | 40 | 0 | 40 |
| 1 | 40 | 1 | 40 | 1 | 40 | 1 | 40 | 1 | 40 |
| 2 | 40 | 2 | 40 | 2 | 40 | 2 | 40 | 2 | 40 |
| 3 | 40 | 3 | 46 | 3 | 40 | 3 | 43 | 3 | 40 |
| 4 | 41 | 4 | 53 | 4 | 45 | 4 | 50 | 4 | 40 |
| 5 | 46 | 5 | 58 | 5 | 50 | 5 | 56 | 5 | 40 |
| 6 | 51 | 6 | 62 | 6 | 54 | 6 | 61 | 6 | 43 |
| 7 | 54 | 7 | 66 | 7 | 58 | 7 | 66 | 7 | 48 |
| 8 | 58 | 8 | 70 | 8 | 62 | 8 | 70 | 8 | 52 |
| 9 | 61 | 9 | 73 | 9 | 65 | 9 | 73 | 9 | 56 |
| 10 | 64 | 10 | 77 | 10 | 68 | 10 | 77 | 10 | 59 |
| 11 | 66 | 11 | 79 | 11 | 71 | 11 | 80 | 11 | 63 |
| 12 | 69 | 12 | 82 | 12 | 74 | 12 | 83 | 12 | 66 |
| 13 | 71 | 13 | 85 | 13 | 77 | 13 | 86 | 13 | 69 |
| 14 | 73 | 14 | 87 | 14 | 79 | 14 | 89 | 14 | 72 |
| 15 | 76 | 15 | 90 | 15 | 82 | 15 | 92 | 15 | 75 |
| 16 | 78 | 16 | 92 | 16 | 84 | 16 | 95 | 16 | 78 |
| 17 | 80 | 17 | 94 | 17 | 87 | 17 | 97 | 17 | 81 |
| 18 | 82 | 18 | 97 | 18 | 89 | 18 | 100 | 18 | 84 |
| 19 | 84 | 19 | 99 | 19 | 91 | 19 | 103 | 19 | 87 |
| 20 | 86 | 20 | 101 | 20 | 94 | 20 | 106 | 20 | 89 |
| 21 | 88 | 21 | 103 | 21 | 96 | 21 | 109 | 21 | 92 |
| 22 | 89 | 22 | 105 | 22 | 99 | 22 | 111 | 22 | 96 |
| 23 | 91 | 23 | 107 | 23 | 101 | 23 | 114 | 23 | 99 |
| 24 | 93 | 24 | 109 | 24 | 104 | 24 | 118 | 24 | 102 |
| 25 | 95 | 25 | 111 | 25 | 107 | 25 | 121 | 25 | 105 |
| 26 | 97 | 26 | 113 | 26 | 109 | 26 | 124 | 26 | 109 |
| 27 | 99 | 27 | 115 | 27 | 112 | 27 | 128 | 27 | 113 |
| 28 | 101 | 28 | 117 | 28 | 116 | 28 | 132 | 28 | 117 |
| 29 | 103 | 29 | 119 | 29 | 119 | 29 | 136 | 29 | 121 |
| 30 | 105 | 30 | 122 | 30 | 124 | 30 | 141 | 30 | 126 |
| 31 | 106 | 31 | 124 | 31 | 128 | 31 | 147 | 31 | 132 |
| 32 | 108 | 32 | 126 | 32 | 134 | 32 | 148 | 32 | 140 |
| 33 | 111 | 33 | 128 | 33 | 142 | 33 | 149 | 33 | 149 |
| 34 | 113 | 34 | 130 | 34 | 149 | 34 | 149 | 34 | 149+ |
| 35 | 115 | 35 | 133 | 35 | 149+ | 35 | 149 | 35 | 149+ |
| 36 | 117 | 36 | 135 | | | | | | |
| 37 | 119 | 37 | 137 | | | | | | |
| 38 | 122 | 38 | 140 | | | | | | |
| 39 | 124 | 39 | 143 | | | | | | |
| 40 | 127 | 40 | 145 | | | | | | |
| 41 | 130 | 41 | 148 | | | | | | |
| 42 | 133 | 42 | 152 | | | | | | |
| 43 | 137 | 43 | 155 | | | | | | |
| 44 | 141 | 44 | 159 | | | | | | |
| 45 | 145 | 45 | 160 | | | | | | |
| 46 | 151 | 46 | 169 | | | | | | |
| 47 | 157 | 47 | 175 | | | | | | |
| 48 | 166 | 48 | 184 | | | | | | |
| 49 | 181 | 49 | 199 | | | | | | |
| 50 | 200 | 50 | 200 | | | | | | |

Table 4

| Vocabulary and Reading, Level I | | Language Usage and Indirect Composition, Level II | | | | Idiomatic Expressions and Reading, Level II | | | |
|---|---|---|---|---|---|---|---|---|---|
| Form B | | Form A | | Form B | | Form A | | Form B | |
| Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score | Raw Score | Scale Score |
| 0 | 40 | 0 | 40 | 0 | 40 | 0 | 40 | 0 | 40 |
| 1 | 40 | 1 | 40 | 1 | 40 | 1 | 40 | 1 | 42 |
| 2 | 40 | 2 | 40 | 2 | 40 | 2 | 40 | 2 | 57 |
| 3 | 40 | 3 | 43 | 3 | 46 | 3 | 48 | 3 | 67 |
| 4 | 40 | 4 | 50 | 4 | 53 | 4 | 55 | 4 | 74 |
| 5 | 41 | 5 | 56 | 5 | 60 | 5 | 61 | 5 | 80 |
| 6 | 47 | 6 | 61 | 6 | 65 | 6 | 66 | 6 | 85 |
| 7 | 52 | 7 | 65 | 7 | 69 | 7 | 71 | 7 | 90 |
| 8 | 56 | 8 | 69 | 8 | 74 | 8 | 75 | 8 | 94 |
| 9 | 60 | 9 | 72 | 9 | 77 | 9 | 78 | 9 | 98 |
| 10 | 64 | 10 | 75 | 10 | 81 | 10 | 82 | 10 | 101 |
| 11 | 68 | 11 | 79 | 11 | 84 | 11 | 85 | 11 | 105 |
| 12 | 71 | 12 | 82 | 12 | 88 | 12 | 88 | 12 | 108 |
| 13 | 75 | 13 | 84 | 13 | 91 | 13 | 91 | 13 | 111 |
| 14 | 78 | 14 | 87 | 14 | 94 | 14 | 94 | 14 | 114 |
| 15 | 81 | 15 | 90 | 15 | 97 | 15 | 97 | 15 | 117 |
| 16 | 84 | 16 | 93 | 16 | 99 | 16 | 100 | 16 | 120 |
| 17 | 87 | 17 | 95 | 17 | 102 | 17 | 103 | 17 | 123 |
| 18 | 91 | 18 | 98 | 18 | 105 | 18 | 106 | 18 | 126 |
| 19 | 94 | 19 | 101 | 19 | 108 | 19 | 109 | 19 | 129 |
| 20 | 97 | 20 | 103 | 20 | 111 | 20 | 111 | 20 | 132 |
| 21 | 100 | 21 | 106 | 21 | 114 | 21 | 114 | 21 | 135 |
| 22 | 104 | 22 | 109 | 22 | 117 | 22 | 117 | 22 | 138 |
| 23 | 107 | 23 | 112 | 23 | 120 | 23 | 120 | 23 | 142 |
| 24 | 111 | 24 | 115 | 24 | 123 | 24 | 123 | 24 | 145 |
| 25 | 114 | 25 | 118 | 25 | 127 | 25 | 127 | 25 | 148 |
| 26 | 118 | 26 | 121 | 26 | 130 | 26 | 130 | 26 | 152 |
| 27 | 122 | 27 | 125 | 27 | 134 | 27 | 134 | 27 | 156 |
| 28 | 127 | 28 | 128 | 28 | 138 | 28 | 138 | 28 | 160 |
| 29 | 132 | 29 | 133 | 29 | 143 | 29 | 142 | 29 | 164 |
| 30 | 137 | 30 | 138 | 30 | 148 | 30 | 147 | 30 | 169 |
| 31 | 144 | 31 | 143 | 31 | 154 | 31 | 153 | 31 | 175 |
| 32 | 146 | 32 | 150 | 32 | 162 | 32 | 161 | 32 | 182 |
| 33 | 147 | 33 | 159 | 33 | 172 | 33 | 170 | 33 | 192 |
| 34 | 148 | 34 | 175 | 34 | 188 | 34 | 186 | 34 | 200 |
| 35 | 149 | 35 | 200 | 35 | 200 | 35 | 200 | 35 | 200 |

Table 4 (continued)

The Total scale score standard error of measurement had to be computed from the standard errors of the part scale scores, since the Total scale score is the average of the part test scale scores and is not computed directly from total test raw scores. In this case the appropriate standard error of measurement is the average of the standard errors of the part tests and is calculated as follows:

$$\sigma(T_e) = \frac{\sqrt{\sigma^2(L_e) + \sigma^2(W_e) + \sigma^2(R_e)}}{3}$$

where $\sigma(T_e)$ is the Total score standard error of measurement and, $\sigma^2(L_e)$ , $\sigma^2(W_e)$ and $\sigma^2(R_e)$ are the scale score variance errors of measurement for the three part tests: Listening Comprehension, Language Usage and Indirect Writing (Level I) or Indirect Composition (Level II), and Vocabulary (Level I) or Idiomatic Expressions (Level II) and Reading.

## Results

The results of the reliability analysis are shown in Tables 5a through 8b. The tables are grouped by twos. The "a" table in each set gives the total test and part test reliability information; the "b" table gives the component test reliability information. Tables 5a and 5b show the results for Level I, Form A; Tables 6a and 6b, for Level I, Form B; Tables 7a and 7b, for Level II, Form A; and Tables 8a and 8b, for Level II, Form B.

## Level I, Form A: Reliabilities and Standard Errors of Measurement (SEMs) for Total Test and Test Parts

| Source of Reliability Information | Raw Score Reliability | Raw Score SEM | Scale Score SEM |
|---|---|---|---|
| **Total Test (150 Items)** | | | |
| Items | 0.950 | 4.63 | NA |
| Part Tests | 0.951 | 4.58 | 3.93 |
| Component Tests | 0.952 | 4.56 | 3.91 |
| **Listening Comprehension (50 Items)** | | | |
| Items | 0.883 | 3.19 | 7.04 |
| Component Tests | 0.883 | 3.18 | 7.01 |
| **Language Usage and Indirect Writing (35 Items)** | | | |
| Items | 0.878 | 2.41 | 6.55 |
| Component Tests | 0.878 | 2.41 | 6.53 |
| **Vocabulary and Reading (35 Items)** | | | |
| Items | 0.891 | 2.23 | 6.83 |
| Component Tests | 0.893 | 2.21 | 6.78 |

Table 5a

## Level I, Form A: Raw Score Alpha Reliabilities and Raw Score Standard Errors of Measurement (SEMs) for Test Components

| Component | No. of Items | Raw Score Reliability | Raw Score SEM |
|---|---|---|---|
| Rejoinders | 25 | 0.791 | 2.25 |
| Short Conversations | 15 | 0.763 | 1.70 |
| Discourse | 10 | 0.486 | 1.46 |
| Language Usage | 21 | 0.820 | 1.83 |
| Indirect Writing | 14 | 0.743 | 1.56 |
| Vocabulary | 10 | 0.719 | 1.01 |
| Reading | 25 | 0.865 | 1.97 |

Table 5b

## Level I, Form B: Reliabilities and Standard Errors of Measurement (SEMs) for Total Test and Test Parts

| Source of Reliability Information | Raw Score Reliability | Raw Score SEM | Scale Score SEM |
|---|---|---|---|
| **Total Test (120 Items)** | | | |
| Items | 0.956 | 4.52 | NA |
| Part Tests | 0.957 | 4.47 | 4.22 |
| Component Tests | 0.957 | 4.46 | 4.20 |
| **Listening Comprehension (50 Items)** | | | |
| Items | 0.886 | 3.06 | 7.27 |
| Component Tests | 0.886 | 3.06 | 7.26 |
| **Language Usage and Indirect Writing (35 Items)** | | | |
| Items | 0.897 | 2.40 | 7.33 |
| Component Tests | 0.897 | 2.39 | 7.31 |
| **Vocabulary and Reading (35 Items)** | | | |
| Items | 0.900 | 2.20 | 7.34 |
| Component Tests | 0.902 | 2.18 | 7.28 |

Table 6a

## Level I, Form B: Raw Score Alpha Reliabilities and Raw Score Standard Errors of Measurement (SEMs) for Test Components

| Component | No. of Items | Raw Score Reliability | Raw Score SEM |
|---|---|---|---|
| Rejoinders | 25 | 0.769 | 2.25 |
| Short Conversations | 15 | 0.733 | 1.58 |
| Discourse | 10 | 0.663 | 1.34 |
| Language Usage | 21 | 0.848 | 1.82 |
| Indirect Writing | 14 | 0.770 | 1.55 |
| Vocabulary | 10 | 0.749 | 0.97 |
| Reading | 25 | 0.873 | 1.96 |

Table 6b

## Level II, Form A: Reliabilities and Standard Errors of Measurement (SEMs) for Total Test and Test Parts

| Source of Reliability Information | Raw Score Reliability | Raw Score SEM | Scale Score SEM |
|---|---|---|---|
| **Total Test (120 Items)** | | | |
| Items | 0.965 | 4.13 | NA |
| Part Tests | 0.966 | 4.10 | 3.70 |
| Component Tests | 0.966 | 4.08 | 3.68 |
| **Listening Comprehension (50 Items)** | | | |
| Items | 0.922 | 2.62 | 5.79 |
| Component Tests | 0.923 | 2.61 | 5.75 |
| **Language Usage and Indirect Composition (35 Items)** | | | |
| Items | 0.910 | 2.19 | 6.45 |
| Component Tests | 0.910 | 2.18 | 6.44 |
| **Idiomatic Expressions and Reading (35 Items)** | | | |
| Items | 0.893 | 2.26 | 6.92 |
| Component Tests | 0.894 | 2.26 | 6.90 |

Table 7a

## Level II, Form A: Raw Score Alpha Reliabilities and Raw Score Standard Errors of Measurement (SEMs) for Test Components

| Component | No. of Items | Raw Score Reliability | Raw Score SEM |
|---|---|---|---|
| Rejoinders | 25 | 0.850 | 1.85 |
| Short Conversations | 15 | 0.849 | 1.35 |
| Discourse | 10 | 0.611 | 1.25 |
| Language Usage | 17 | 0.847 | 1.53 |
| Indirect Composition | 18 | 0.825 | 1.56 |
| Idiomatic Expressions | 5 | 0.560 | 0.81 |
| Reading | 30 | 0.879 | 2.11 |

Table 7b

## Level II, Form B: Reliabilities and Standard Errors of Measurement (SEMs) for Total Test and Test Parts

| Source of Reliability Information | Raw Score Reliability | Raw Score SEM | Scale Score SEM |
|---|---|---|---|
| **Total Test (120 Items)** | | | |
| Items | 0.959 | 4.40 | NA |
| Part Tests | 0.960 | 4.35 | 4.13 |
| Component Tests | 0.960 | 4.34 | 4.12 |
| **Listening Comprehension (50 Items)** | | | |
| Items | 0.910 | 2.77 | 6.58 |
| Component Tests | 0.910 | 2.77 | 6.56 |
| **Language Usage and Indirect Composition (35 Items)** | | | |
| Items | 0.885 | 2.28 | 7.14 |
| Component Tests | 0.885 | 2.28 | 7.12 |
| **Idiomatic Expressions and Reading (35 Items)** | | | |
| Items | 0.892 | 2.46 | 7.68 |
| Component Tests | 0.892 | 2.45 | 7.66 |

Table 8a

## Level II, Form B: Raw Score Alpha Reliabilities and Raw Score Standard Errors of Measurement (SEMs) for Test Components

| Component | No. of Items | Raw Score Reliability | Raw Score SEM |
|---|---|---|---|
| Rejoinders | 25 | 0.807 | 1.95 |
| Short Conversations | 15 | 0.792 | 1.53 |
| Discourse | 10 | 0.712 | 1.22 |
| Language Usage | 17 | 0.801 | 1.62 |
| Indirect Composition | 18 | 0.795 | 1.60 |
| Idiomatic Expressions | 5 | 0.594 | 0.88 |
| Reading | 30 | 0.875 | 2.29 |

Table 8b

It is important to remember that the Listening Comprehension test is the same for Levels I and II. Thus, reliability information on this test and its three components is available on two samples for Form A and two samples for Form B. For the other parts and components, tests differ by level as well as by form.

## Reliability Information from Different Methods

The reliability information in Tables 5a, 6a, 7a, and 8a permits a comparison of three different reliability and standard error of measurement estimates for the total test. They also permit a comparison of two different estimates for the three-part tests.

The important conclusion one reaches in reviewing the reliability information is that *there is very little difference among the numerical values, for both the total test and the part tests.* The maximum difference in reliabilities among the three reliability estimates is .002 (Level I, Form A) in the case of the total test and also .002 in the case of the part tests (Level I, Form A, Vocabulary and Reading). The standard errors of measurement sometimes changed in the third significant digit, but the differences are of no practical significance. The small impact of different reliability and standard error estimates is also obvious at the scale score level, regardless of whether the Total scale score or the part test scale scores are considered. The largest difference in scaled score standard errors of measurement was .06 (Level I, Form B, Vocabulary and Reading).

It should be noted that the use of subtest reliability information does improve reliability and standard errors of measurement in the expected directions. The total test reliabilities and standard errors of measurement are slightly improved (sometimes in the fourth decimal place, however) when part test reliability information is used. They are improved more, however slightly, when component test reliability information is used. Likewise, the part test reliability information is slightly improved when component reliability information is used to estimate the numerical values. Nevertheless, the improvements are of no practical value.

*Thus, it is unnecessary to compute reliabilities and standard errors of measurement using subtests for the ELASH tests.* The slight improvement in results is simply not worth the extra effort to compute the refined estimates.

## Reliability Coefficients and Standard Errors of Measurement

Because the various methods of estimating reliability yielded very similar results, this section summarizes reliability information provided by the alpha reliability coefficients and associated measures of the standard error of measurement. The two Listening Comprehension test forms were the same for Levels I and II. The other tests forms were built for either Level I or Level II test takers. Ideally, test forms would be constructed to have similar reliabilities for both Level I and Level II test takers unless the testing program wanted to emphasize better measurement for the lower or higher scoring test takers.

At the raw score level, the total test reliability is very high for all four forms, ranging from .950 to .965 (see Tables 5a, 6a, 7a, 8a and Table 9). Table 9 shows the relatively tight clustering of these reliabilities. While these coefficients do not apply to the Total scale score, which is the average of the part test scale scores, they nevertheless indicate that the Total score reliability is undoubtedly high as well.

**Display of Total Test and Part Test Forms in Terms of their Reliability Coefficients**

| Reliability | Total Test | Listening Comprehension | Language Usage and Indirect Writing (Level I) or Composition (Level II) | Vocabulary (Level I) or Idiomatic Expressions (Level II) and Reading |
|---|---|---|---|---|
| .965 - .969 | IIA | | | |
| .960 - .964 | | | | |
| .955 - .959 | IB, IIB | | | |
| .950 - .954 | IA | | | |
| .945 - .949 | | | | |
| .940 - .944 | | | | |
| .935 - .939 | | | | |
| .930 - .934 | | | | |
| .925 - .929 | | | | |
| .920 - .924 | | IIA | | |
| .915 - .919 | | | | |
| .910 - .914 | | IIB | IIA | |
| .905 - .909 | | | | |
| .900 - .904 | | | | IB |
| .895 - .899 | | | IB | |
| .890 - .894 | | | | 1A, IIA, IIB |
| .885 - .889 | | IB | IIB | |
| .880 - .884 | | IA | | |
| .875 - .879 | | | 1A | |

**Table 9**

The reliability coefficients for the three part tests are also high, ranging from .878 to .922. When part test reliabilities are compared within test forms, the reliabilities are highest for Reading and Vocabulary scores in the case of Level I forms and for Listening Comprehension scores in the case of Level II forms (see Table 9). The raw score standard errors of measurement ranged from 2.19 (for Language Usage and Indirect Composition, Level II, Form A) to 3.19 (for Listening Comprehension, Level I, Form A). Because the Listening Comprehension test consists of more items than the other tests, it could be expected to have larger raw score standard errors of measurement.

Ideally, the reliabilities for a given test would cluster together, indicating similar measurement power for the various samples. This clustering is evident for the total test and the Vocabulary and Reading or Idiomatic Expressions and Reading tests. The reliabilities have fairly wide ranges for the other two parts, as may be noted in Table 9.

Among the component tests it would be expected that lower reliabilities would be obtained for components containing only a few items, such as Discourse (10 items), Vocabulary (10 items), and Idiomatic Expressions (5 items). Table 10 shows that this expectation was generally fulfilled. The only strikingly low reliability coefficients are for Discourse scores from Form A of Level I (.486) and Idiomatic Expressions from Forms A and B of Level II (.560 and .594, respectively). The other three Discourse scores yielded reliability coefficients ranging from .611 to .712). As may be noted in Table 2, the mean item-total point biserial (with item deleted from the total) is especially low for Discourse, Level I, Form A (.20 compared with means in the .30s for the other three forms).

**Display of Total Test and Part Test Forms in Terms of their Reliability Coefficients**

| Reliability | Rejoinders | Short Conversations | Discourse | Language Usage | Indirect Writing (Level I) or Indirect Composition (Level II) | Vocabulary (Level I) or Idiomatic Expressions (Level II) | Reading |
|---|---|---|---|---|---|---|---|
| .860 - .879 | | | | | | | IA, IB, IIA, IIB |
| .840 - .859 | IIA | IIA | | IB, IIA | | | |
| .820 - .839 | | | | IA | IIA | | |
| .800 - .819 | IIB | | | IIB | | | |
| .780 - .799 | IA | IIB | | | IIB | | |
| .760 - .779 | IB | IA | | | IB | | |
| .740 - .759 | | | | | IA | | |
| .720 - .739 | | IB | | | | | |
| .700 - .719 | | | IIB | | | IB | |
| .680 - .699 | | | | | | | |
| .660 - .679 | | | IB | | | IA | |
| .640 - .659 | | | | | | | |
| .620 - .639 | | | | | | | |
| .600 - .619 | | | IIA | | | | |
| .580 - .599 | | | | | | | |
| .560 - .579 | | | | | | | |
| .540 - 559 | | | | | | IIB | |
| .520 - .539 | | | | | | IIA | |
| .500 - .519 | | | | | | | |
| .480 - .499 | | | IA | | | | |

## Scale Score Standard Errors of Measurement

As was noted in a previous section of this paper, standard errors of measurement for part test scale scores could be computed directly from the raw score standard errors by multiplying it by an appropriate slope parameter. (See Table 3 for the slope parameters used to compute part test scale score standard errors.) Because the slope parameters do not apply to the full score range, however, scale score reliability coefficients could not be computed. It is likely, though, that because the slopes covered the main part of the score range, the raw score reliability is a reasonable estimate of the scale score reliability for the part tests.

Once the scale score standard errors of measurement were available for the part test scale scores, they could be used to compute a standard error of measurement for the Total scale score. Presumably this standard error would apply to the middle part of the Total scale score range.

Computing reliability for the Total scale scores is even more complicated than the computation for part test scale scores. Conditional standard errors of measurement would be needed for each observed combination of part test scale scores and then averaged. No classical reliability method provides conditional standard errors of measurement, let alone standard errors for combinations of scale scores. Item response theory methods might be used to make such computations, but the methodology has not yet been developed. The raw score total test reliability may be a reasonable estimate of the Total scale score reliability, but we cannot be sure of that. Nevertheless, the high total test raw score reliabilities indicate that the Total scale score reliability would also be very high.

The part test scale score standard errors of measurement range from 5.79 (for Listening Comprehension, Level II, Form A) to 7.68 (for Idiomatic Expressions and Reading, Level II, Form B). The standard errors tend to be highest for Level I, Form B. For the most part the standard errors range between 6.50 and 7.50, roughly 7 score points. The scale score standard errors of measurement for the Total score range from 3.70 to 4.22, roughly 4 score points, for the four forms. Again the standard errors are highest for Level I, Form B.

## Reliability Information for the Same Test from Different Samples

As was mentioned before, Forms A and B of the Listening Comprehension test were administered to two different samples. Thus, reliability information from these samples on the very same test forms may be compared. Tables 5a, 5b, 7a, and 7b provide the relevant information on Form A; and Tables 6a, 6b, 8a, and 8b on Form B. Table 9 displays all of the reliability coefficients from these tables in a single table.

Table 9 shows that the reliabilities are higher for the Level II samples than for the Level I samples regardless of test form. The Form A reliabilities for the Level II sample are considerably higher (by .039) than those for the Level I sample; and the raw score standard errors of measurement, considerably lower (by .57, over half a score point). (See Tables 5a and 7a.) The differences between the Level II and Level I samples for Form B are smaller (.024 for reliability and .29 for the raw score standard error of measurement) but still considerable. (See Tables 6a and 8a.)

As is clear from Table 10, the components of the Listening Comprehension test all yield higher reliabilities and smaller standard errors for Level II samples, except for the Discourse component. For this component Form B of Level I had a higher reliability (.663) than Form A of Level II (.611).

Obviously, Forms A and B of Listening Comprehension are more appropriate for the Level II sample. Tables 1 and 2 identify the factors that contribute to higher reliabilities for

the Level II sample, namely:

- Higher scale score means and percentages correct that are closer to mid-difficulty (0.63 for a test form containing 15 three-choice items and 105 four-choice items), and

- Higher mean item-total point biserial correlations (with item deleted from the total).

It is clear that a test may have considerably higher reliabilities for some samples than others. If the testing program desires similar measurement power for samples at different levels, then the reliabilities should be about equal. If the testing program desires better measurement for one sample over another, then the reliability should be higher for that sample.

## Summary

In this study the way the numerical values were calculated made little difference in the reliability coefficients and standard errors of measurement, whether for the total test or the part tests. Using standard errors of measurement from fairly homogeneous subtests can increase reliabilities and decrease standard errors of measurement. For the ELASH tests, however, the alpha reliability information on the tests was essentially the same as the alternative reliability information based on subtests.

The essential information this study provided consisted of the reliability and standard error of measurement information on the various tests. When raw scores are analyzed, the total test and part test scores turned out to be highly reliable with small standard errors of measurement. The total test reliabilities averaged around .96, and the part test reliabilities around .89. No part test reliabilities below .88 were observed. Among the component tests, only one Discourse form and the two Idiomatic Expressions forms had reliabilities lower than .60. The lower reliability was expected for Idiomatic Expressions, which consists of only five items.

Even though the score conversions were curvilineal, procedures used in this study permitted estimates of scale score standard errors of measurement for the middle part of the score scale. This information had not been available previously to the testing program. The scale score standard errors were on the order of four score points for the Total score and six or seven score points for the part test scale scores.

One further analysis was conducted: Comparing Listening Comprehension reliability information from the two Listening Comprehension forms. This comparison was possible because Level I and Level II samples took each test form. The Listening Comprehension test was more reliable for the Level II samples, indicating that the test has greater measurement power for test takers.

# References

**Cronbach, L. J. (1951).** Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

**Lord, F. M., & Novick, M. R. (1968).** *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.


Oficina de Puerto Rico y América Latina