Reliability and Validity Evidence for the GED® English as a Second Language Test

GED Testing Service® Research Studies, 2009-4

GED® TESTING SERVICE

A Program of the American Council on Education®

ACE®

American Council on Education®

GED®
TESTING SERVICE

GED Testing Service®
One Dupont Circle NW, Suite 250
Washington, DC 20036-1163
(202) 939-9490
Fax: (202) 659-8875
www.GEDtest.org

Reliability and Validity Evidence for the GED® English as a Second Language Test

J. Carl Setzer

GED Testing Service®
A Program of the American Council on Education®

Abstract

The GED® English as a Second Language (GED ESL) Test was designed to serve as an adjunct to the GED test battery when an examinee takes either the Spanish- or French-language version of the tests. The GED ESL Test is a criterion-referenced, multiple-choice instrument that assesses the functional, English reading skills of adults whose first language is not English. The purpose of this report is to provide some background and psychometric information regarding the ESL Test. Sections of the report provide an overview of the test specifications, estimates of reliability (including internal consistency and classification accuracy), and evidence supporting the validity of test score interpretations.

The Tests of General Educational Development (GED® Tests) provide an opportunity for adults who have not completed a formal high school program to certify their attainment of high school–level academic knowledge and skills, and earn their jurisdictions' high school–level equivalency credential, diploma, or certificate. The current GED Tests measure academic skills and knowledge requisite for a high school program of study, with an emphasis on the workplace and higher education. The GED test battery comprises five content area assessments:

- Language Arts, Writing (50 multiple-choice items; single essay)
- Social Studies (50 multiple-choice items)
- Science (50 multiple-choice items)
- Language Arts, Reading (40 multiple-choice items)
- Mathematics (40 multiple-choice items, 10 alternate format items)

Several versions of the GED Tests exist. Specifically, there is currently an English-language U.S. edition, an English-language Canadian edition, Spanish-language GED Tests, French-language GED Tests, and an internationally available computer-based version of the English-language U.S. edition.[1] Although the vast majority of GED candidates take one of the English-language editions, a number of candidates take the tests in either the Spanish or French language. In 2008, more than 29,000 candidates tested using the Spanish-language GED Tests. Less than 1,000 candidates tested using the French-language GED Tests that same year (GED Testing Service, 2009a).

In addition to the five content area tests, GED Testing Service® (GEDTS) offers the English as a Second Language (ESL) Test of Reading Comprehension to jurisdictions. The ESL Test was first developed for Puerto Rico in 1971 alongside the development of the first Spanish-language version of the GED Tests. The ESL Test was not revised again until 1999. It was designed to serve as an adjunct to the GED test battery when an examinee takes either the Spanish- or French-language version of the GED Tests. Each jurisdiction determines the policies regarding the ESL Test.

The GED ESL Test is a criterion-referenced, multiple-choice assessment that assesses the functional, English reading skills of adults whose first language is not English. Four levels of texts are used that represent beginner to intermediate levels of reading comprehension. The texts are authentic examples of general English, drawn from advertisements, forms, newspapers and magazines, instructional handbooks, consumer information, workplace memos, and other daily reading materials. The questions that accompany each passage are either literal or interpretative comprehension items and represent 75 percent and 25 percent of the test, respectively.

The redesign of the ESL Test required the development of new items and subsequent field testing. The ESL Test items were field tested with non-native English-speaking students enrolled in ESL programs at high schools (juniors and seniors), community colleges, and inmates in federal prisons. The population that was tested was diverse and included individuals at all proficiency levels, as determined by their

[1] Details regarding the development of the 2002 Series GED Tests, as well as additional background and technical information, are beyond the scope of this report. However, the reader is referred to the *Technical Manual: 2002 Series GED Tests* (GED Testing Service, 2009c) for further details. Details regarding the internationally available computer-based version can be found in the report *Reliability Analysis of the Internationally Administered 2002 Series GED Tests* (GED Testing Service, 2009b).

instructors. Four tryout forms were developed with at least 60 items each. The texts were presented in order of difficulty, from the easiest to the hardest. The field-testing process ultimately resulted in four operational forms (labeled IA-ID) with 60 items each.

During the development of the 2002 series ESL Test, a committee convened to define the standard of English reading proficiency required when awarding high school equivalency based on taking the GED test battery in a language other than English. The judges on the committee elicited a working definition of the *minimally competent* examinee. That is, "What types of real-world reading should such a person be able to handle? How well should an adult whose first language is not English be able to read in English in order for a state to award a high school equivalency credential?" The result of this process was a passing standard which an ESL examinee must meet in order to pass the ESL Test.

The purpose of this report was to provide some background and psychometric information regarding the ESL Test. The following sections provide an overview of the test specifications, estimates of reliability (including internal consistency and classification accuracy), and evidence supporting the validity of test score interpretations.

Overview of ESL Test Specifications

Each of the four operational ESL Test forms was designed according to the test specifications that were developed prior to the launch of the 2002 Series GED Tests. There are two cognitive and four content levels.

Cognitive Levels

The cognitive levels on the ESL Test consist of Literal Comprehension and Interpretive Comprehension. The definitions of each level and the percentage of questions that measure that level of comprehension on each test are listed below.
Literal Comprehension (75 percent of questions):

- Identifying main subject.
- Recognizing/locating information.
- Identifying supporting details.
- Paraphrasing information.
- Restating opinions.
- Understanding the clear implication of text.

Interpretive Comprehension (25 percent of questions):

- Making inferences.
- Drawing conclusions.
- Identifying main ideas.
- Making generalizations.
- Identifying organization.
- Distinguishing between fact and opinion.
- Identifying persuasive language.
- Using context clues to determine meaning.

The percentage distribution of each type of question was based on the previous ESL Test and an analysis of the initial items written for the new test development.

Content of Passages

The passages on the revised ESL Test were chosen from four increasingly difficult levels to permit a candidate who has taken the GED test battery in French or Spanish to demonstrate his or her functional reading competency in English. The four levels of text are described in **Table 1**. The first level of text, Level 1, relies on graphics, with single words and phrases. Level 2 involves one or two paragraphs with simple sentences and may include graphics. Level 3 includes multiple paragraphs with some additional sentence variety. Finally, Level 4 also includes multiple paragraphs and more complex sentence variety.

Table 1. Levels of Difficulty and Descriptions of Passages

	Graphics	Sentence Structure	Organization	Topic Source	Vocabulary
Level 1: Graphic-based; single words and phrases	All	Single words and phrases; few simple sentences	Graphically organized	Concrete ideas from daily life	High-frequency
Level 2: One or two paragraphs; simple sentences	Some	Simple sentences	One or two paragraphs	Concrete ideas from daily life	Primarily high-frequency
Level 3: Multi-paragraph; some sentence variety	None	Simple sentences with some compound or complex sentences	Multi-paragraph	Concrete ideas from general interest areas	Primarily high-frequency with some context clue; interpretation required
Level 4: Multi-paragraph; sentence variety	None	Sentence variety to include compound and complex sentences	Multi-paragraph	Some abstraction of ideas and concepts	General vocabulary with some idiomatic usage and some abstract words

Context of Passages

While revising the ESL Test, consideration was given to the subject matter of the texts. The test developers gave special consideration to the fact that the population of GED candidates is highly diverse. Therefore, the passages were carefully screened by three different sensitivity reviewers to eliminate passages that may have (a) given an advantage or disadvantage to a particular group of candidates because of references to special circumstances, e.g., subway schedules favor urban dwellers; (b) presented a particular group of people in an unfavorable light; or (c) caused highly emotional reactions. The passages also were written to reflect real-life situations from the workplace as well as daily life, including such topics as child care, consumer information, advertisements, and entertainment. Finally, the passages were written using the same standard for quality writing as that imposed for the Language Arts, Reading Test.

Format of Test

The reading passages on the ESL Test are ordered from Level 1 to Level 4, each constituting two to three passages. Each passage is followed by four to six multiple-choice questions classified by cognitive level. The questions for each passage also range in difficulty, allowing candidates to respond to more difficult texts with some measure of success.

Scoring

All ESL Test items are weighted equally and a summation across items is used to derive a final score. Each ESL Test form raw score is converted to a scaled score for reporting purposes. The standard scale ranges from 20 to 80 with a mean of 50 and standard deviation of 10. Each operational test form was placed on this scale in order to permit comparisons of examinee scores across test forms. GEDTS has set the minimum score requirement at 41.[2]

Test Score Reliability and Standard Errors of Measurement

Reliability refers to the consistency, or stability, of test scores when we administer the measurement procedure repeatedly to groups of examinees (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). If a given test yields widely discrepant scores for the same individual on separate test administrations, and the individual does not change significantly on the measured attribute, then the scores on the test are not reliable. Conversely, if a test produces the same or similar scores for an individual on separate administrations, then the scores from the test are considered reliable. Reliability is inversely related to the amount of measurement error in test scores. That is, the more measurement error present in test scores, the less reliable the test.

Reliability is a crucial index of test quality. Standard practices require test developers to evaluate and report the reliability of their test scores. The purpose of this section is to estimate and evaluate the reliability of the ESL data from test forms IA through ID. The reliability of test scores from other GED Tests versions (i.e., U.S. and Canadian English editions and Spanish- and French-language versions) can be found in the *Technical Manual: 2002 Series GED Tests* (GED Testing Service, 2009a).

Several procedures are available for evaluating reliability; each account for different sources of measurement error and thus produce different reliability coefficients. In this chapter, the reliability of the GED ESL Test was evaluated using calculated estimates of the internal consistency reliability, the standard error of measurement, the

[2] Puerto Rico uses a standard scale ranging from 200 to 800 and requires an average score of 450 across all five content areas and the ESL Test. The Federal Bureau of Prisons has set the minimum score at 400 (also on a standard scale of 200 to 800), although exceptions may apply.

conditional standard error of measurement, and classification accuracy. The following sections briefly introduce each of these areas, along with GEDTS methodologies. More complete descriptions of reliability estimation are available in Anastasi (1988), Feldt and Brennan (1989), and Lord and Novick (1968).

Internal Consistency Reliability

In classical test theory, we model a person's observed test score (*X*) as a function of his or her true score (*T*) and random error (*E*). The function is simply additive such that

$$X = T + E \, .$$

A person's true score is the expected score across parallel replications of the measurement procedure (i.e., a score that is free from measurement error).

The total amount of test score variance (σ_X^2) we observe in test scores is equal to the sum of the true score variance (σ_T^2) and random error variance (σ_e^2), or

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \, .$$

Internal consistency is an estimate of the proportion of total variance in the observed scores that is attributable to the true scores. We also can describe the estimate as the extent to which all the items on a test correlate positively with one another. Given the equation for total variance above, an estimate of internal consistency can be theoretically represented as

$$1 - \frac{\sigma_e^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_X^2}$$

or

$$1 - \frac{\textit{sum of item variances}}{\textit{sum of item variances \& covariances}} = \frac{\textit{sum of item covariances}}{\textit{sum of item variances \& covariances}} \, .$$

GEDTS estimates internal consistency reliability using the KR_{20} reliability coefficient (Kuder & Richardson, 1937). KR_{20} is a special case of the more general coefficient alpha (Cronbach, 1951) and is equivalent to coefficient alpha when test item scores are dichotomous. KR_{20} also is essentially an estimate of the expected correlation of a test with an alternate or parallel test form of the same length (Nunnally, 1978).

The operational formula for the KR_{20} reliability coefficient for dichotomously scored multiple-choice tests is given in Equation 1:

$$KR_{20} = \frac{k}{k-1}\left[1 - \left(\frac{\Sigma p_i q_i}{\sigma_x^2}\right)\right] \qquad \textbf{(1)}$$

where *k* equals the number of items on the test, p_i equals the proportion of examinees answering item *i* correctly (with $q_i = 1 - p_i$), and σ_x^2 equals the variance of the total scores on the test. The variance for the item is $p_i q_i$ when the test item receives a dichotomous score.

The KR_{20} coefficient ranges from zero to one, with estimates closer to one indicating greater reliability. Three factors can affect the magnitude of the KR_{20} coefficient: the homogeneity of the test content (affects $\sum p_i q_i$), the homogeneity of the examinee population tested (affects σ_t^2), and the number of items on the test (k). Tests comprising items that measure similar (i.e., homogenous) content areas have higher KR_{20} estimates than tests comprising items measuring diverse content areas because the covariance among the items is likely lower when the items measure widely different concepts or skills. Conversely, examinee populations that are highly homogenous can reduce the magnitude of the KR_{20} coefficient because the limited amount of total variance in the examinee population limits the amount of covariance among the items. If we assume that all items correlate positively with one another, then adding items to a test increases item covariance, and thus, the KR_{20} reliability coefficient. The GED ESL Test measures highly interrelated content areas and the heterogeneity of the GED examinee population is high; therefore, content heterogeneity or examinee homogeneity does not attenuate ESL Test score KR_{20} reliability estimates.

Standard Error of Measurement

The standard error of measurement (SEM) is an estimate of the average amount of error within test scores. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) define the SEM as “the standard deviation of a hypothetical distribution of measurement errors that arises when a given population is assessed via a particular test or measure” (p. 27). We often use the SEM to describe how far an examinee’s observed test score may be, on average, from his or her “true” score. Therefore, smaller SEMs are preferable to larger ones. We can use the SEM to form a confidence interval around a true score to suggest a proportion of times, during repeated measurements, when the interval contains the true score. Because the SEM is the standard deviation of a hypothetical, normal distribution of measurement errors, we usually expect that an examinee’s observed score will be found within one SEM unit of his or her true score approximately 68 percent of the time.

The SEM is a function of the standard deviation and reliability of the test scores. The equation for the SEM is:

$$SEM = \sigma_X \sqrt{1 - r_{tt}} \qquad \textbf{(2)}$$

where σ_X equals the standard deviation of test scores, and r_{tt} equals the reliability coefficient. (For the SEM reported here, GEDTS uses the reliability coefficient KR_{20}.) We can see in Equation 2 that tests with small standard deviations and larger reliabilities yield smaller SEMs. Because the SEM is a function of the standard deviation of test scores, it is not an absolute measure of error; rather, it is in the metric of raw score units. Therefore, unlike reliability coefficients, we cannot compare SEM across tests without considering the unit of measurement, range, and standard deviation of the tests’ raw scores.

Conditional Standard Errors of Measurement

As described above, the SEM provides an estimate of the *average* amount of error in observed test scores. However, the amount of error in test scores actually may differ at various points along the score scale. For this reason, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) state:

> *Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score* (p. 35).

The minimum standard score requirement for the GED ESL Test was set at 41. Thus, estimating the amount of measurement error in the vicinity of the minimum standard score is important. Because the reported scores are standard scores rather than raw scores, GEDTS reports conditional standard errors of measurement (CSEM, i.e., SEMs at specific points or intervals along the score scale) that are also on the standard score metric.

CSEMs were estimated using an approximation procedure described by Feldt and Qualls (1998). These calculations require estimates of KR_{20} and KR_{21} for the raw scores, the mean and standard deviation of the raw scores, and a constant, *C*, which was determined a priori.[3] This process involves estimating the number of CSEMs within the range of $X_0 \pm C$, where X_0 refers to the raw score of interest. The assumption is that the same range of corresponding standard scores will have the same number of SEMs in scale score units.

To estimate standard score CSEM, three steps were involved. First, the raw score CSEM for a particular raw score point X_0, $CSEM_{R(X)}$, was calculated using Equation 3,

$$CSEM_{R(X)} = \left[\left(\frac{1-KR_{20}}{1-KR_{21}}\right)\left(\frac{X_0(k-X_0)}{k-1}\right)\right]^{1/2}, \qquad \textbf{(3)}$$

where *k* is the number of raw score points and KR_{20} and KR_{21} are reliability estimates. Second, the slope of the function relating standard score to raw score at X_0 was approximated. That is, the slope of the function relating a standard score to raw score at X_0 was calculated using Equation 4,

$$slope_{X_0} = \frac{SS_U - SS_L}{(X_0+C)-(X_0-C)} = \frac{SS_U - SS_L}{2C}, \qquad \textbf{(4)}$$

where *C* is an arbitrary small number of raw score points (here, *C*=4 as recommended by Feldt & Qualls, except where noted), SS_U is the standard score for the raw score point X_0+C, and SS_L is the standard score for the raw score point X_0-*C*. Third, the standard

[3] The KR_{21} coefficient is another internal consistency reliability estimate that requires the mean and variance of the observed scores, as well as the maximum possible total score (Kuder & Richardson, 1937).

score CSEM at raw score point $X_0, CSEM_{SS(X)}$, was the product of $slope_{X_0}$ and $CSEM_{R(X)}$, as shown in Equation 5.

$$CSEM_{SS(X)} = \left(\frac{SS_U - SS_L}{2C}\right) CSEM_{R(X)} \qquad \textbf{(5)}$$

To find the standard score CSEM for a given standard score point rather than a given raw score point, the corresponding raw score point for a given standard score was found from the raw-to-standard conversion table, and then the above three steps were used. When the raw-to-standard conversion was not one to one (i.e., if there were two raw score points corresponding to one standard score point), a modification of the Feldt and Qualls (1998) procedure was made. Specifically, when two raw score points corresponded to one standard score, the average was used to calculate the raw score CSEM, and the interval used to calculate the slope was (low-3, high+3). That is, *C*=3 was used and the interval width was 7. For example, two raw scores, 36 and 37, corresponded with the same standard score of 41. When calculating the standard score CSEM for 41, (36+37)/2=36.5 was used to calculate the raw score CSEM. The slope was calculated by $[SS_{(37+3)}\text{-}SS_{(36\text{-}3)}]/[(37+3)\text{-}(36\text{-}3)]= (SS_{40}\text{-}SS_{33})/7$.

Classification Accuracy

Standard 2.15 in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) states:

> *When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees that would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument* (p. 35).

GEDTS uses a required minimum standard score for the ESL Test. Therefore, it is necessary to adhere to Standard 2.15 and provide appropriate measures of classification accuracy.

GEDTS uses the Livingston and Lewis (LL; 1995) procedure to calculate classification accuracy. The LL procedure essentially compares observed scores with theoretically estimated true scores. To obtain the true scores, the LL procedure estimates a true score distribution using a four-parameter beta distribution. The procedure subsequently compares the true scores with the observed scores in a two-by-two contingency table, as shown below in **Figure 1**.

Figure 1.

	Observed score status	
True score status	Pass	Fail
Pass	A	B
Fail	C	D

Each cell in the table represents a proportion of examinees. For example, cell A represents the proportion of examinees who were classified as *passers* according to both their theoretical true score and their observed score. The sum of the proportions in cells A and D represents the classification accuracy. Cell C represents the proportion of false positives (those who should not have met the passing standard according to their theoretical true score), while cell B represents the proportion of false negatives (those who should have met the required minimum standard score). Ideally, the proportions in cells B and C should be zero, and the sum of cells A and D should be one.

The LL procedure was implemented using the BB-Class software program developed by Brennan (2004). A four-parameter beta distribution was assumed for the true score distribution, and a binomial model was assumed for the observed score distribution conditional on a given true score.

KR_{20} and SEM Results for the GED ESL Test

Table 2 presents the standard score means, standard deviations, and SEM for the ESL test forms in the 2002 GED Test Series. The data presented in Table 2 facilitate comparisons among the four forms by presenting the statistics reported in standard score units. Raw score data and KR_{20} estimates are also presented in Table 2. The KR_{20} estimates were computed for raw scores only. Because the transformation of raw scores to standard scores is nonlinear, it is not possible to compute these statistics directly for standard scores. However, the raw score–to–standard score transformation maintains the rank order of the examinees, and thus, the differences in KR_{20} would be negligible. The SEM, on the other hand, would be quite different because it is a function of the standard deviation of scores, as well as the reliability coefficient.

The information in Table 2 is based on the performance of GED candidates who took the GED ESL Test between 2002 and 2008. The results presented in Table 2 indicate that although there has been some variation in score performance on the forms across years, KR_{20} and SEM estimates have remained consistent. All ESL test forms have KR_{20} estimates of at least 0.93.

Table 2. Sample Sizes (N), Score Means, Standard Deviations (SD), Standard Errors of Measurement (SEM), and KR_{20} Estimates for the 2002 Series English as a Second Language GED Test

		STANDARD SCORES			RAW SCORES			
Form	N	Mean	SD	SEM	Mean	SD	SEM	KR_{20}
IA	7,205	41.8	12.3	3.3	35.8	12.3	3.2	0.93
IB	6,897	43.2	14.1	3.5	37.0	13.7	3.2	0.94
IC	5,970	40.2	13.7	3.1	34.1	14.0	3.2	0.95
ID	6,804	43.6	11.5	2.8	39.9	12.4	3.1	0.94

The standard score CSEM for values between 38 and 42 for the ESL forms are available in **Table 3**. Some of the variations in CSEM within forms may be caused by changes in the constant value, *C*, used in the calculations or whether there was a one-to-one correspondence between the raw and standard scores.

In theory, we can use the test score as an estimate of an examinee's true score, which again is the theoretical average score an examinee would receive if he or she took parallel versions of a test an infinite number of times. Because the test score is not perfectly reliable, there is a certain level of measurement error associated with each test score. We can estimate an interval that contains a person's true score for a given proportion of times over repeated measurements by using the CSEM. For example, if an examinee receives a score of 41 on form IA, then 68 percent of the time the interval of 41-3 and 41+3 (i.e., the interval between 38 and 44) captures his or her true score. In other words, if this person takes the same test (or a parallel version) 100 times, we expect his or her standard scores to fall within the range of 38 to 44 approximately 68 times.

Table 3. Conditional Standard Errors of Measurement at Various Standard Scores for the 2002 Series GED English as a Second Language Test

	STANDARD SCORE						
Form	38	39	40	41	42	43	44
IA	2.7	2.7	3.1	3.0	3.1	3.0	3.0
IB	2.8	2.8	3.3	3.2	3.2	3.2	3.1
IC	2.7	2.7	3.1	3.1	3.1	3.0	3.0
ID	2.7	3.1	3.1	3.5	3.5	3.4	3.4

The percentages of examinees meeting and not meeting the minimum score requirements, the probability of correct classification (classification accuracy), and false-positive and false-negative classifications are available in **Table 4**. In terms of classification accuracy, values range from zero to one, and values closer to one are preferable. The classification accuracy rates are 0.93 or above for all ESL test forms.

The false-positive rates provided in Table 4 reflect the probability of an examinee incorrectly passing the test form, given his or her true score is below the minimum score. Conversely, the false-negative rates indicate the probability that an examinee will not meet the minimum score requirement for the test form, given his or her true score is above the cut score. For most forms, the results indicate that the proportion of examinees who incorrectly met or exceeded the minimum score requirement (false positives) was very close or equal to the proportion of examinees who incorrectly failed to meet the minimum requirement (false negatives). Because the classification accuracy is relatively high, the false-negative and false-positive probabilities are relatively low.

Table 4. Probability of Correct Classification, False-Positive, and False-Negative Rates for the 2002 Series GED English as a Second Language Test

Form	N	Percent Not Meeting Minimum Score	Percent Meeting Minimum Score	Probability of Correct Classification	False Positive	False Negative
IA	7,205	0.50	0.50	0.93	0.04	0.03
IB	6,897	0.44	0.56	0.93	0.03	0.03
IC	5,970	0.52	0.48	0.94	0.03	0.03
ID	6,804	0.59	0.41	0.93	0.04	0.03

Validity of ESL Test Score Interpretations

An investigation into test score validity requires the accumulation of evidence that ideally suggests a specific test score interpretation, or use, is a valid one. Validity is not a property of the test itself, but rather a description of the appropriateness of the interpretations made from test scores. Because validity describes the utility and appropriateness of test score interpretations, it is of paramount importance that test developers provide evidence of validity. As stated in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999):

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations* (p. 9).

According to the Standards, an ideal validation is one that includes several types of evidence that, when combined, best reflect the value of a test for an intended purpose: “Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose” (p. 11). The Standards suggests that test developers report several types of validity evidence, when appropriate. Specifically, evidence may be provided based on test content, response processes, internal structure, relations to other variables, and/or consequences of testing. The sources of validity evidence included in this report are those based solely on relations to other variables.

As clearly noted in the Standards, evidence of validity reported by test developers should reflect the purpose(s) of the test and the types of inferences that are to be made from the test scores. Therefore, in evaluating the validity of the GED ESL test scores, the purpose of the tests must be considered first.

Purpose of the GED ESL Test

As noted earlier in this report, the purpose of the ESL Test is to measure how well an adult whose first language is not English is able to read in English. The validation of ESL test scores must be made with respect to this purpose. Thus, the sources of validity evidence reported in this report help evaluate the ability of ESL test scores to determine whether a GED examinee is able to read in English well enough for the state to award him or her a high school equivalency credential. The validity evidence presented in this report is based upon the relationship of the test scores to other external variables.

ESL Test Score Relationships with Other Variables

Tables 5 through 8 demonstrate how ESL standard scores relate to several examinee background variables. The results reported in these tables were obtained via the data from an equating study. A brief survey was administered to each of the examinees for the purpose of obtaining demographic information as well as additional information on examinees' reading abilities. Only the data from Forms IB and ID were available for this analysis.

In **Table 5**, the percentages of examinees who obtained selected standard scores or higher are listed against the self-reported number of years lived in the United States. Within each row of the table, the percentages decrease as levels of the standard score increase. For example, 96 percent of examinees who had lived in the United States for one year obtained a standard score of 30 or higher. The percentage of those same examinees who scored 41 or higher decreases to 80. As anticipated, the same decreasing pattern holds for each row and test form in Table 5.

We also anticipated that the percentages would have increased within columns as well as across rows. In other words, we expected an overall positive relationship between the number of years lived in the United States and standard scores. This anticipated trend is not supported by the data in Table 5. One speculation is that those examinees who lived in the United States for shorter periods of time were more motivated to complete the GED test battery and thus were involved in greater amounts of preparation. Those who lived in the United States for longer periods of time may not have necessarily undergone formal English-language reading training.

Table 5. Percentage of Examinees in 2003 Achieving Selected Standard Scores or Higher, by Self-Reported Number of Years Living in the United States

		ESL Standard Score ≥						
Years	N	30	34	38	41	44	48	52
					Form B			
0	61	95	92	82	80	75	64	48
1	53	96	92	87	85	70	49	40
2	63	94	89	83	76	59	48	37
3	49	92	84	80	76	65	53	39
4	23	83	78	74	70	65	57	43
5	17	94	94	82	82	71	47	29
>5	69	87	81	71	67	49	41	30

Continued on next page

Table 5 continued

				Form D				
0	54	94	89	80	76	72	57	46
1	51	94	86	86	80	75	63	55
2	64	89	83	78	70	59	48	33
3	50	94	90	86	76	70	58	40
4	24	100	88	79	75	63	54	29
5	17	82	76	71	71	59	41	29
>5	78	92	81	72	60	51	42	26

Data Source: 2003 ESL Equating Study.

In **Table 6**, the percentages of examinees who obtained selected standard scores or higher are listed against the self-reported number of years studying English before coming to the United States. As expected, the percentages decrease as the ESL standard score increases for any given row in the table. In general, the percentages tend to increase within columns, as well. This latter finding suggests that the more years a candidate studied English prior to arriving in the United States and taking the ESL Test, the greater the likelihood of scoring higher on the test.

Table 6. Percentage of Examinees Achieving Selected Standard Scores or Higher, by Self-Reported Number of Years Studying English Before Coming to the United States

		ESL Standard Score ≥						
Years	N	30	34	38	41	44	48	52
					Form B			
0	117	84	79	71	64	51	44	27
1	48	92	81	71	71	60	46	33
2	27	85	81	70	70	56	44	30
3	32	100	97	84	81	75	53	44
4	14	93	79	71	64	57	36	29
5	13	92	92	92	85	54	38	38
>5	91	99	98	95	92	80	66	57
					Form D			
0	140	89	80	72	63	54	44	31
1	39	87	74	67	56	54	36	15
2	30	90	83	80	77	60	57	40
3	28	96	89	86	79	79	71	43
4	16	94	88	75	69	69	50	44
5	13	100	100	92	85	69	62	54
>5	71	100	97	93	91	83	68	55

Data Source: 2003 ESL Equating Study.

In **Table 7A**, the percentages of examinees who obtained selected standard scores or higher are listed against the self-reported number of years studying English after coming to the United States. Again, as expected, the percentages decrease within each row. However, as the number of years spent studying English (after coming to the United States) increase, the percentages of examinees obtaining a given standard score do not necessarily increase. In other words, there does not appear to be a positive relationship between the number of years spent studying English after coming to the United States and standard score.

Table 7B provides similar information with the exception that those who studied English prior to coming to the United States are excluded. Because of the smaller sample sizes, only two groups are compared: those who studied English one year or less after coming to the United States and those who studied two or more years. In this case, the expected relationship holds, in that those who studied English longer after coming to the United States (with no training prior to their arrival) obtained higher standard scores.

Table 7A. Percentage of Examinees Achieving Selected Standard Scores or Higher, by Self-Reported Number of Years Studying English After Coming to the United States

		ESL Standard Score ≥						
Years	N	30	34	38	41	44	48	52
					Form B			
0	65	94	89	78	78	68	58	42
1	106	90	85	79	75	66	51	41
2	60	92	88	82	77	55	45	38
3	48	85	81	73	67	63	42	29
4	26	96	92	85	81	69	62	42
5	13	92	69	62	62	46	38	23
>5	24	92	92	88	79	58	46	33
					Form D			
0	68	93	84	75	72	68	53	41
1	86	93	83	80	73	66	56	43
2	65	89	83	80	71	60	49	35
3	53	96	91	77	62	57	49	32
4	25	96	84	76	72	60	52	32
5	15	87	87	80	73	60	27	13
>5	26	88	85	81	77	65	58	42

Data Source: 2003 ESL Equating Study.

Table 7B. Percentage of Examinees Achieving Selected Standard Scores or Higher, by Self-Reported Number of Years Studying English After Coming to the United States

		ESL Standard Score ≥						
Years	N	30	34	38	41	44	48	52
					Form B			
0 to 1	40	78	73	60	55	50	45	30
2 +	77	87	82	77	69	52	43	26
					Form D			
0 to 1	46	83	67	63	54	46	35	24
2 +	94	93	86	77	67	57	48	34

Note: Data exclude those examinees who studied English prior to coming to the United States.
Data Source: 2003 ESL Equating Study.

Finally, **Table 8** illustrates the relationship between examinees' self-reported reading ability and their standard scores. Generally speaking, the percentage of passers increases as self-reported reading ability increases. For example, none of the examinees who indicated they "can read only a few words and simple sentences" passed Form B, although 23 percent passed Form D. In contrast, the majority of examinees who indicated they could "easily read long novels, college textbooks, and technical information on most subjects" passed Forms B and D.

Table 8. Percentage of Examinees Who Met the Passing Standard, by Self-Reported Reading Ability

	Form B		Form D	
	N	% Passed	N	% Passed
I can read only a few words and simple signs.	9	0	13	23
I can read simple advertisements, forms, menus and schedules.	16	44	19	47
I can read some short news articles, want ads, form letters, and simple instructions.	57	72	45	62
I can read some newspaper and magazine articles, stories, and instructions.	92	73	80	71
I can read most newspaper and magazine articles, editorials, and reviews.	41	88	41	80
I can read some short novels and introductory textbooks.	76	88	71	85
I can read most novels and textbooks, insurance and tax information, and business reports.	35	80	44	82
I can easily read long novels, college textbooks, and technical information on most subjects.	14	71	12	83

Data Source: 2003 ESL Equating Study.

Discussion

The reliability analyses indicated that the test scores are highly replicable. Estimates of internal consistency and classification accuracy were all high (i.e., greater than 0.90). False-positive and false-negative passing rates were correspondingly low. Standard errors of measurement (both overall and conditional) were found to be acceptable.

The analyses that examined the relationship between ESL standard scores and self-reported background variables provided some evidence that the ESL score interpretations are valid. Specifically, some of the analyses examined whether the number of years studying English (either before or after coming to the United States) affected ESL standard scores. Other analyses looked at the number of years the examinee lived in the United States to determine whether there was a positive relationship with ESL standard score. Finally, the relationship between standard scores and self-reported reading ability was examined. The results of these analyses were somewhat mixed. In some instances, there appeared to be a positive relationship between the variable of interest and ESL standard score. In other instances, the same conclusion could not necessarily be made.

Several reasons could contribute to these mixed results. First, the samples used in the analyses were somewhat limited in the sense that there may be some differences with the true examinee population. Recall that the validity analyses used data obtained from the equating studies and not operational data.

Second, the variables used in the analyses relied on self-reported data. There are two potential problems with this situation. First, self-reported data is often suspect because of social desirability biases. Second, if the examinee had a low English-reading ability then he or she may not have understood the survey questions properly.

Finally, the assumptions underlying the analyses and the background variables may have been tenuous. For example, it may not have been appropriate to assume that those who lived in the United States for longer periods of time should necessarily have done better on the ESL Test. Perhaps those who have lived in the United States for shorter periods of time and taken the GED Tests were more motivated to do well on the test and thus studied more intensively. Additionally, aside from a language barrier, the survey questions may not have been entirely clear to the examinees. For example, it may not have been clear to all examinees what was meant by the phrase "number of years spent studying English." Studying may be interpreted formerly or informally, depending on the examinee.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Brennan, R. L. (2004). *Manual for BB-Class: A computer program that uses the beta-binomial model for classification consistency and accuracy (CASMA Rep. No. 9)* [Computer software manual]. Retrieved from www.education.uiowa.edu/casma/computer_programs.htm#classification.

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

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 103–146). Phoenix, AZ: American Council on Education/Oryx Press.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11*(2), 159–177.

GED Testing Service. (2009a). *2008 GED testing program statistical report.* Washington, DC: American Council on Education.

GED Testing Service. (2009b). *Reliability analysis for the internationally administered 2002 Series GED Tests*. Washington, DC: American Council on Education.

GED Testing Service. (2009c). *Technical manual: 2002 Series GED Tests*. Washington, DC: American Council on Education.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

GED®
TESTING SERVICE

A Program of the American Council on Education®

One Dupont Circle NW, Suite 250
Washington, DC 20036-1163
(202) 939-9490
Fax: (202) 659-8875
www.GEDtest.org