# Reliability Analysis for the Internationally Administered 2002 Series GED Tests

GED Testing Service® Research Studies, 2009-3

GED®
TESTING SERVICE

A Program of the American Council on Education®

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

# Reliability Analysis for the Internationally Administered 2002 Series GED Tests

J. Carl Setzer

Yi He

GED Testing Service®

A Program of the American Council on Education®

## Acknowledgments

GED Testing Service thanks Yi He for her contribution to this study while she was a psychometric intern with GED Testing Service in 2009. The authors also would like to thank Carol Ezzelle and Yung-chen Hsu of GED Testing Service for reviewing a previous draft of the study and for their helpful comments. For further information on the study, please contact J. Carl Setzer at carl_setzer@ace.nche.edu.

## Reliability Analysis for the Internationally Administered 2002 Series GED Tests

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 report was to estimate and evaluate the reliability of the internationally administered 2002 Series GED Tests, which have been developed and maintained by the GED Testing Service (GEDTS) since 1963. 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, 2009).

## The Tests of General Educational Development

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 2002 Series GED test battery comprises five content area tests:

- 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)

There are several versions of the GED Tests. 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. Although the vast majority of GED candidates take the tests in either the United States or Canada, a small number of candidates take the tests internationally. Nevertheless, the content and cognitive specifications are essentially the same across each of these test versions.

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, 2009) for further details.

## Reliability Analysis

Several procedures are available for evaluating reliability; each account for different sources of measurement error and thus produce different reliability coefficients. In this report, the reliability of the computer-based GED Tests was evaluated using calculated estimates of the internal consistency reliability, the standard error of measurement, the 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 ($\sigma_X^2$) we observe in test scores is equal to the sum of the true score variance ($\sigma_T^2$) and random error variance ($\sigma_e^2$), 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 each other. 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

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

GEDTS estimates the internal consistency reliability of the computer-based GED Tests (with the exception of the Language Arts, Writing Test composite score) using the $KR_{20}$ reliability coefficient (Kuder & Richardson, 1937). $KR_{20}$ is a special case of the more general coefficient alpha (Cronbach, 1951). The $KR_{20}$ coefficient 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 $\sigma_x^2$ 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 $\sigma_t^2$), 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 Tests measure highly interrelated content areas and the heterogeneity of the GED examinee population is high; therefore, content heterogeneity or examinee homogeneity does not attenuate GED test score $KR_{20}$ reliability estimates. However, differences in the number of items on the content area GED Tests might influence the $KR_{20}$ coefficients.

*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, over 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 $\sigma_X$ 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 each of the individual content area GED Test was set to 410 (on a standard score scale ranging from 200 to 800 with mean equal to 500 and standard deviation equal to 100). 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.[1] 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,

[1] The $KR_{21}$ coefficient is another internal consistency reliability estimate that requires only the mean and variance of the observed scores as well as the maximum possible total score (Kuder & Richardson, 1937). Unlike $KR_{20}$, we only use KR-21 with polytomously scored items. In fact, we rarely use KR-21 because it assumes that all items are of equal difficulty. However, this assumption is easy to violate in practice.

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

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 score CSEM at raw score point $X_0$, $CSEM_{SS(X)}$, was the product of $slope_{X_0}$ and $CSEM_{R(X)}$, i.e., 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 (e.g., if there were two or three raw score points corresponding to one standard score point), some modifications of the Feldt and Qualls (1998) procedure were made:

- When there are three raw score points corresponding to one standard score, the middle raw score was selected as the raw score point to calculate the standard score CSEM. For example, three raw scores, 17, 18, and 19, correspond with the same standard score (which is 400). When calculating the standard score CSEM for 400, the raw score point 18 was chosen as the corresponding raw score point.

- When there are two raw score points corresponding to one standard score, the average of them 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, 20 and 21, corresponded with the same standard score of 410. When calculating the standard score CSEM for 410, (20+21)/2=20.5 was used to calculate the raw score CSEM. The slope was calculated by $[SS_{(21+3)}\text{-}SS_{(20-3)}]/[(21+3)-(20-3)]= (SS_{24}-SS_{17})/7$.

The Language Arts, Writing Test score is derived by combining weighted multiple-choice and essay portions. As such, the raw-to-standard score conversions are not as direct as with other content area GED Tests. Therefore, the approximation method described above could not be applied to the Language Arts, Writing Test.

*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 each content area test simultaneously with an average score requirement for the entire battery. 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.

| | 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.

## Data

The data used in the current study were collected during 2008. Four different test forms within each of the five content areas were converted to computer-based versions for 2008: forms IG, IH, II, and IJ.[2] Only data from the Language Arts, Writing Test; Social Studies Test; Science Test; and Language Arts, Reading Test were available for the current analysis.

The samples used in the current analysis included most international examinees from 2008 who took forms IG, IH, II, and IJ. For each test, there were five or fewer records with invalid data; these records were excluded from the analysis. In addition, only those examinees who scored a 2 or better on the essay were included in the analysis.[3] The computer-based versions of the tests collected very little demographic information about the examinees. Because this information was not collected for the computer-based version of the tests, no demographic information has been provided in this report.

## Results

Table 1 presents the standard score means, standard deviations, and SEMs for the various forms of the computer-based GED Tests. Note that the numbers in Table 1 for the Language Arts, Writing Test refer only to the multiple-choice portion of the test. The data in Table 1 facilitate comparisons among the four subject tests by presenting the statistics in standard score units. Raw score means, standard deviations, SEMs, and $KR_{20}$ reliabilities are also in Table 1. The $KR_{20}$ reliabilities were computed for raw scores only.

[2] The 2002 Series GED Tests consist of 11 different test forms for each content area. These forms were labeled IA through IK. Only forms IG, IH, II, and IJ were converted to a computer-based version in 2008.

[3] American Council on Education policy dictates that an examinee must achieve a minimum score of 2 on the essay in order to obtain a valid Language Arts, Writing Test score.

Because the transformation of raw scores to standard scores is nonlinear, computing these statistics directly for standard scores is not possible. 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 (American College Testing, 1988). The SEMs are quite different for the standard and raw scores because they are a function of the standard deviation of scores as well as the reliability coefficient.

The results in Table 1 indicate that all test forms have $KR_{20}$ reliabilities of at least 0.86; more than 30 percent of the test forms have a $KR_{20}$ of 0.90 or higher. Standard score SEMs range from 25.7 to 35.6 across all content areas, while raw score SEMs range from 2.5 to 3.0.

Table 1

*Sample Sizes (N), Score Means, Standard Deviations (SD), Standard Errors of Measurement (SEM), and $KR_{20}$ Estimates for the Internationally Administered 2002 Series GED Tests.*

| | | STANDARD SCORES | | | RAW SCORES | | | |
|---|---|---|---|---|---|---|---|---|
| | N | Mean | SD | SEM | Mean | SD | SEM | $KR_{20}$ |
| Language Arts, Writing | | | | | | | | |
| Form IG | 261 | 466.4 | 91.1 | 30.8 | 33.7 | 8.6 | 2.9 | 0.89 |
| Form IH | 417 | 485.1 | 88.5 | 31.7 | 36.5 | 7.7 | 2.8 | 0.87 |
| Form II | 392 | 500.8 | 87.0 | 31.8 | 36.0 | 7.7 | 2.8 | 0.87 |
| Form IJ | 341 | 474.2 | 87.3 | 29.6 | 34.8 | 8.5 | 2.9 | 0.88 |
| Social Studies | | | | | | | | |
| Form IG | 224 | 465.7 | 87.9 | 27.2 | 30.4 | 9.7 | 3.0 | 0.90 |
| Form IH | 424 | 488.3 | 80.8 | 25.7 | 33.4 | 9.2 | 2.9 | 0.90 |
| Form II | 420 | 532.0 | 103.6 | 33.0 | 33.6 | 9.0 | 2.9 | 0.90 |
| Form IJ | 342 | 457.4 | 93.0 | 31.0 | 30.1 | 9.1 | 3.0 | 0.89 |
| Science | | | | | | | | |
| Form IG | 259 | 473.4 | 84.9 | 28.3 | 34.3 | 8.7 | 2.9 | 0.89 |
| Form IH | 412 | 510.3 | 92.8 | 33.1 | 36.1 | 7.9 | 2.8 | 0.87 |
| Form II | 432 | 507.9 | 87.5 | 28.4 | 36.0 | 8.4 | 2.7 | 0.89 |
| Form IJ | 357 | 503.9 | 95.0 | 27.9 | 35.6 | 9.5 | 2.8 | 0.91 |
| Language Arts, Reading | | | | | | | | |
| Form IG | 278 | 469.2 | 87.3 | 32.5 | 26.1 | 7.2 | 2.7 | 0.86 |
| Form IH | 429 | 475.1 | 106.5 | 31.6 | 28.2 | 8.3 | 2.5 | 0.91 |
| Form II | 431 | 507.2 | 109.5 | 35.6 | 27.0 | 8.0 | 2.6 | 0.89 |
| Form IJ | 390 | 432.7 | 92.1 | 32.3 | 25.6 | 7.7 | 2.7 | 0.88 |

The standard score CSEM for values between 390 and 430 for the computer-based GED Tests are available in Table 2. Some of the variations in CSEM within forms may be because of 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 410 on science form IH, then 68 percent of the time the interval of 410-15 and 410+15 (i.e., the interval between 395 and 425) 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 395 to 425 approximately 68 times.

The percentages of examinees meeting and not meeting the minimum score requirements, the probability of correct classification (classification accuracy), and false positive and negative classifications are available in Table 3. In terms of classification accuracy, values range from zero to one, and values closer to one are preferable.

Table 2

*Conditional Standard Errors of Measurement at Various Standard Scores for the Internationally Administered 2002 Series GED Tests.*

| | STANDARD SCORE | | | | |
|---|---|---|---|---|---|
| | 390 | 400 | 410 | 420 | 430 |
| Social Studies | | | | | |
| Form IG | 24.9 | 23.9 | 25.2 | 25.2 | 28.8 |
| Form IH | 24.8 | 23.9 | 25.2 | 19.3 | 21.1 |
| Form II | 31.9 | 32.2 | 28.5 | 28.7 | 28.8 |
| Form IJ | 25.0 | 29.2 | 29.2 | 28.6 | 29.0 |
| Science | | | | | |
| Form IG | 12.7 | 15.1 | 13.0 | 17.2 | 18.6 |
| Form IH | 17.3 | 13.0 | 15.0 | 12.5 | 17.8 |
| Form II | 17.1 | 21.8 | 19.2 | 23.0 | 26.3 |
| Form IJ | 19.0 | 22.6 | 25.3 | 25.2 | 24.6 |
| Language Arts, Reading | | | | | |
| Form IG | 25.5 | 25.4 | 28.8 | 29.1 | 28.9 |
| Form IH | 19.2 | 25.2 | 23.9 | 29.0 | 28.8 |
| Form II | 18.8 | 20.6 | 23.5 | 28.6 | 28.5 |
| Form IJ | 25.4 | 25.5 | 24.4 | 25.6 | 24.2 |

The classification accuracy rates are above 0.90 for all forms of the Social Studies Test, Science Test, and Language Arts, Reading Test. For the Language Arts, Writing Test, the classification accuracy rates are slightly less than 0.90, except for form IG, which has a classification accuracy rate of 0.76.

The false positive rates provided in Table 3 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 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 3

*Probability of Correct Classification, False Positive Rates, and False Negative Rates for the Internationally Administered 2002 Series GED Tests.*

| Test/Form | N | Percent Not Meeting Minimum Score | Percent Meeting Minimum Score | Probability of Correct Classification | False Positive | False Negative |
|---|---|---|---|---|---|---|
| Language Arts, Writing | | | | | | |
| Form IG | 261 | 31 | 69 | 0.76 | 0.14 | 0.10 |
| Form IH | 417 | 20 | 80 | 0.87 | 0.06 | 0.07 |
| Form II | 392 | 12 | 88 | 0.88 | † | 0.12 |
| Form IJ | 341 | 19 | 81 | 0.88 | 0.07 | 0.05 |
| Social Studies | | | | | | |
| Form IG | 224 | 27 | 73 | 0.91 | 0.05 | 0.04 |
| Form IH | 424 | 13 | 87 | 0.94 | 0.03 | 0.03 |
| Form II | 420 | 8 | 92 | 0.96 | 0.02 | 0.03 |
| Form IJ | 342 | 28 | 72 | 0.91 | 0.04 | 0.05 |
| Science | | | | | | |
| Form IG | 259 | 24 | 76 | 0.92 | 0.03 | 0.04 |
| Form IH | 412 | 13 | 87 | 0.94 | 0.02 | 0.03 |
| Form II | 432 | 14 | 86 | 0.95 | 0.02 | 0.03 |
| Form IJ | 357 | 14 | 86 | 0.95 | 0.02 | 0.03 |
| Language Arts, Reading | | | | | | |
| Form IG | 278 | 17 | 83 | 0.92 | 0.04 | 0.04 |
| Form IH | 429 | 22 | 78 | 0.94 | 0.03 | 0.03 |
| Form II | 431 | 17 | 83 | 0.93 | 0.03 | 0.04 |
| Form IJ | 390 | 44 | 56 | 0.90 | 0.05 | 0.05 |

† Value is less than 0.001.

## Conclusion

In general, the reliability of test scores obtained from the computer-based GED Tests was high. We generally consider favorable reliability estimates to be in the upper 0.80s and closer to 1.0. The lowest $KR_{20}$ estimate was 0.86. Another favorable result is that the CSEMs in the range near the minimum standard score of 410 are generally smaller than the average SEM. Ideally, we want to see the greatest amount of measurement precision in this score range because this is where decisions and critical inferences occur. Finally, the classification accuracy was generally high for all but the Language Arts, Writing Test.

Comparisons of the SEMs (both average and conditional) and classification accuracy between the computer-based and paper-based versions of these test forms are not necessarily appropriate because of potential differences in the populations who take the two test versions. The examinees who take the computer-based versions are in locations outside the United States and thus are very likely to reflect a different set of demographics. In addition, differences in international curricula might affect test score distributions in such a way that subsequently would affect many of the estimates described in this study.

# References

American College Testing. (1988). *ACT Assessment technical manual.* Iowa City, IA: American College Testing Program.

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). Washington, DC: American Council on Education.

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. (2009). *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