# DOCUMENT RESUME

ED 310 149 TM 013 738

| | |
|---|---|
| AUTHOR | Bunch, Michael B.; Littlefair, Wendy |
| TITLE | Total Score Reliability in Large-Scale Writinq Assessment. |
| PUB DATE | Jun 88 |
| NOTE | 18p.; Paper presented at the Conference of the Education Commission of the States/Colorado Department of Education Assessment (Boulder, CO, June 1988). |
| PUB TYPE | Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Cutting Scores; *Essay Tests; *Generalizability Theory; Interrater Reliability; Reader Response; Sample Size; *Scores; Secondary Education; Standardized Tests; *Test Reliability; *Writing Evaluation |
| IDENTIFIERS | *Domain Referenced Tests; GENOVA (Computer Program) |

ABSTRACT

A total of 2,000 essays written by 1,000 students was submitted to generalizability analyses for domain-referenced tests. Each student had written one essay on each of two prompts representing two models of discourse. Each essay was read by six readers and judged on a scale of from 1 to 4. No reader read essays from both prompts. Reader agreement rates and interrater reliability coefficients were computed. Extensive analyses were conducted using GENOVA, a generalizability analysis program. Special consideration was given to the universes of generalizability for readers and prompts. One set of artificially unreliable scores was introduced to increase variance due to readers and the reader-times-essay interaction and, thus, lower reliability. Results indicate that the score reliability of essay tests is multifaceted and can be estimated in a variety of ways depending on the purpose of the assessment and the intended use of results. When pass-fail decisions or determinations of absolute skill levels are to be made, indices that take into account the cut point or points are needed. Seven data tables and two graphs are provided. (TJH)

***********************************************************************
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
***********************************************************************

ED310149

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
□ Minor changes have been made to improve reproduction quality
● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MICHAEL B. BUNCH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

# TOTAL SCORE RELIABILITY IN LARGE - SCALE WRITING ASSESSMENT

Michael B. Bunch
Wendy Littlefair
Measurement Incorporated

Paper presented at the Education Commission of the States / Colorado Department of Education Assessment Conference, Boulder, Colorado, June, 1988.

TM013738

BEST COPY AVAILABLE

Direct assessment of student writing has become a significant part of large-scale assessment in North America. At least 17 states and one Canadian province are known to have writing assessment programs. For many years, Educational Testing Service (ETS) has conducted writing assessment programs as part of their advanced placement (AP) tests in English and history. In recent years, the American College Testing Program (ACT), the General Educational Development Testing Service (GEDTS) and commercial test publishers have introduced direct assessments of student writing. For purposes of this paper, large-scale writing assessment is defined as any direct assessment of writing employing standardized stimuli (prompts) and standardized scoring conditions.

Writing assessment programs serve a variety of purposes and treat scores in many different ways. The ETS AP tests, for example, provide norm referenced scores for placement decisions at colleges and universities. The GED essay score, combined with a multiple choice language score, contributes to decisions regarding the award of a high school equivalence certificate. Commercial writing tests (e.g., *Metropolitan Achievement Test*) generally offer norm referenced interpre-tations based on large national norming samples, and scores serve diagnostic purposes.

In state and district testing programs, writing assessment yields scores that sometimes help to determine whether or not a student passes to the next grade or graduates from high school. In other instances, where pass-fail decisions are not based on essay scores, scores are still interpreted in absolute terms; i.e., a given score point is associated with a well-defined standard. In Rhode Island, for example, a total score of 7 or 8 is considered a Superior response, defined as follows:

> presents good ideas that are developed logically and fully; is well organized from beginning to end; expresses ideas very clearly;shows a generally strong command of sentence structure; uses language effectively; and has relatively few serious errors in grammar and usage. (Rhode Island Department of Education, 1987; p.7)

Indeed, most writing assessment programs not only have such detailed score point definitions; they have examples of student essays that typify each score point as well.

It is this process of making absolute decisions about students, whether those decisions involve advancement/retention or a less momentous action, that creates special psychometric challenges for large-scale writing assessment programs. The selections of the appropriate measure of reliability and the collection and analysis of data to calculate the chosen measure or measures of reliability are crucial decisions faced by program directors.

An informal survey of state assessment programs revealed that the most common measure of essay reliability is reader agreement rate or inter-rater reliability. It may be helpful to examine some of these indices in light of the psychometric and practical demands and constraints of a typical writing assessment program.

- Interpretation of scores is typically criterion referenced or domain referenced, as opposed to norm referenced.
- Individuals, rather than groups, are the focus of measurement.
- In pass-fail programs, students usually are given multiple opportunities to pass.
- The scope of the essay test is acknowledged to be narrow; i.e., no attempt is made to generalize from observed scores to a much wider range of writing tasks.
- Most such tests consist of a single essay scored by two readers.

## Specifying Sources of Error

Assuming that the student is the object of measurement, sources of error in essay scores might include mode of discourse (e.g., narrative, explanatory, persuasive, etc.) prompt, and reader. Other sources have also been identified such as day of the week and time of day the essay was scored (Braun, 1986). Under proper conditions, it is possible to set up experiments in which some or all of these potential sources of error can be examined through application of analysis of variance (ANOVA) techniques. Coffman (1971) strongly recommended ANOVA for reliability estimation for essay tests because other estimates (e.g., test-retest) overestimate reliability.

Cronbach, Gleser, Nanda, and Rajaratnam (1972) developed the theory of generalizability to estimate sources of total score variability and to allow investigators to generalize to specified conditions. Estimation of components of score variability is referred to as a generalizability study or G study. Application and manipulation of those components to a specified decision is referred to as a decision study or D study.

Brennan (1983) extended the work of Cronbach et al (1972) to criterion-referenced or domain-referenced tests. This extension is important in two respects: first, it focuses on the actual decisions about students as well as the process which leads to those decisions; 2) it provides a method for incorporating the cut score or cut scores into the reliability or dependability coefficient.

The present study employs generalizability analyses for domain referenced tests. Results should be applicable to most writing assessment programs in which students receive absolute scores, although certain aspects of the study have implications for norm referenced programs as well.

### Study Design

Data were 2,000 essays written by 1,000 students. Each student had written one essay on each of two prompts representing two modes of discourse. Each essay was read six times and judged on a scale of 1-4. Readers read within prompt only; i.e., one group of readers was trained to score Prompt 1 only and one group was trained to read Prompt 2 only. Within each prompt, 12 readers actually read essays but for any given essay, only six readers would be involved.

### Reader Training

The 24 readers in the study were selected from approximately 150 experienced readers who had just completed a major scoring project. They had received approximately three days of training. Each reader had read three sets of 10 papers representing all score points. After discussion of increasingly ambiguous essays (e.g., sets of solid 2's and 3's followed by mixed sets of high 2's and low 3's), readers practiced scoring sets of papers that represented the entire score range (1-4). They then were required to qualify by scoring two sets of qualifying papers which also represented the entire score range. Any reader who did not qualify was released from the project.

At the outset of this study, all readers were required to qualify again after an abbreviated training session. Throughout the study, propject managers reviewed the criteria with the readers on a daily basis. Also, packets of prescored essays (validity packets) were distributed and scored each day to allow project managers to check scoring accuracy.

### Data Analysis

Three types of analyses were conducted. In order to make results comparable to those typical of most large-scale writing assessment programs, we computed reader agreement rates (percent agreement) and inter-rater reliability coefficients. Extensive analyses were conducted using GENOVA, a generalizability analysis program developed by Joe Crick and Robert Brennan (1983). The basic design was an S×(R:P) design, students crossed with readers who are nested within prompts. Since many testing programs have a pass-fail component, we computed a decision dependability coefficient, $\phi(\lambda)$, using a cut score of 5.5 (on a scale of 2-8) as well as all other possible cut scores. In live scoring, each essay may receive a score of 0, 1, 2, 3, or 4 from each reader. If two readers disagree, and their scores are adjacent, the essay receives the mean of the two scores. Thus, half-point scores are possible, with one exception. An essay with scores of 0 and 1 is resolved by project managers. A score of 5.5 may be obtained by students whose two essays received scores of 1.5 and 4, 2 and 3.5, or 2.5 and 3.

Special consideration was given to the universes of generalizability for readers and prompts. While it is reasonable to assume that readers for this study would be considered a random sample of all possible readers, it is not necessary to assume that the two prompts are representative of all prompts. While prompts within the two domains change from year to year, only one prompt per domain is given each year; thus, it was impossible to estimate prompt:domain (prompt nested within domain) variance. Rather, we treated the two prompts as proxies for the two domains. For conditions in which domains should be considered random, we treated prompts as random. Where domains should be considered fixed, we treated prompts as fixed. No attempt is made to generalize beyond the two domains.

Finally, we introduced one set of artificially unreliable scores. Previous studies have alluded to individual readers who systematically score high or low (cf. Braun, 1986). Therefore, we created a new data set by adding one point to each score (except 4's) for one reader of Prompt 1 and subtracted one point from each score (except 1's) for one reader of Prompt 2. This systematic variation was expected to increase variance due to readers and the reader × essay interaction and thus lower reliability. It was within the limits of reader variation described by Braun (1986).

## Results

**Reader agreement.** Since each essay was read six times, there were 15 possible combinations or comparisons of scores. Table 1 summarizes reader agreement rate by prompt. Agreement rate is expressed in absolute as well as adjacent terms. While absolute agreement includes score pairs 0-0, 1-1, 2-2, 3-3, and 4-4, adjacent agreement includes these combinations as well as 1-2, 2-3, and 3-4.

Table 1 shows that readers had more difficulty agreeing on scores for essays on Prompt 2 than on Prompt 1; i.e., readers gave identical scores on Prompt 2 less often than on Prompt 1. Since readers were randomly assigned to prompts, it is safe to conclude that the difference in agreement rates shown in Table 1 should not be due to differences in groups of readers' abilities to score consistently.

**Table 1**
**Mean Reader Agreement Rate**
**(Entries are percentages)**

| Agreement Type | Prompt | |
|---|---|---|
| | 1 | 2 |
| Absolute | 78.8 | 73.3 |
| Adjacent | 99.9 | 99.9 |

Since each student's total score is made up of scores on two essays, a measure of total score agreement would be helpful. For this index, we looked to the stability of the pass-fail decision, based on a cut score of 5.5. Over all possible combinations of scores from Prompt 1 and Prompt 2, mean agreement rate was 88.7%. Stated somewhat differently, in 11.3% of the cases, groups of readers disagreed as to whether or not a student should receive a passing score.

**Inter-rater reliability.** Within prompt, correlations among scores ranged from .893 to .912 for Prompt 1 and from .910 to .926 for Prompt 2. Median correlations were .904 for Prompt 1 and .919 for Prompt 2. This finding supports Coffman's (1971) contention that correlational estimates of reliability are too high. Note that there were fewer absolute agreements for Prompt 2 and that its median inter-rater correlation was higher than that for Prompt 1. Why? If reader B rates all essays exactly one point higher than reader A

ERIC Full Text Provided by ERIC

rates them, the absolute agreement rate would be 0% but the correlation would be 1.0. Correlational techniques do not take into account mean differences in scores.

Using the median correlations noted above, it is possible to calculate reader reliability in accordance with the Spearman-Brown formula:

$$r_{nn} = \frac{n\, r_{tt}}{1+(n-1)\, r_{tt}}$$

Thus reader reliability (assuming two readings) is .950 for Prompt 1 and .958 for Prompt 2. One should note that these figures are for two readers per essay only. Table 2 shows the estimated reader reliability for 1-6 readers per essay.

**Table 2**
**Reader Reliability Estimates**

| Readers/Essay | Prompt | |
|---|---|---|
| | 1 | 2 |
| 1 | .904 | .919 |
| 2 | .950 | .958 |
| 3 | .966 | .971 |
| 4 | .974 | .978 |
| 5 | .979 | .983 |
| 6 | .983 | .986 |

As Table 2 shows, reliability for two or more readers is quite high. However, these reliability estimates are for readers, not for scores. Given Coffman's (1971) caveat and the data in Table 1, it may be wise to regard these estimates as upper bounds. Correlations between scores across prompts ranged from .190 to .235, with a median of .222, and can be taken as evidence that the prompts do not measure the same trait.

**Generalizability/dependability.** Table 3 shows the results of the generalizability analysis of scores for the 2,000 essays. As noted previously, the design was S×(R:P) with S representing students, P representing prompts, and R:P representing readers nested within prompts.

Clearly, the students themselves accounted for the greatest portion of variance, both alone (.3026634) and in interaction with prompts (.9927460). It is apparent that a significant student × prompt interaction exists. In other words, some students write better essays on one prompt while other students write better essays on other prompts. While this finding should come as no surprise, it does point out the need for careful prompt selection, since selection of the wrong prompt could result in low scores for students who might have received higher scores on different prompts.

Table 3
GENOVA Source Table for S × (R:P) Design

| Source | df | SS | MS | G Study Variance Component |
|---|---|---|---|---|
| Student (S) | 999 | 9712.83 | 9.72 | .3026634 |
| Prompt (P) | 1 | 547.41 | 547.41 | .0902096 |
| Reader: P | 10 | 2.00 | 0.20 | .0000659 |
| S×P | 999 | 6084.50 | 6.09 | .9927460 |
| S×(R:P) | 9990 | 1339.83 | 0.13 | .1341174 |
| Total | 11999 | 17686.58 | | |

GENOVA allows the investigator to specify an unlimited number of situations or decisions to which results may be applied (D studies). It allows for the calculation of the generalizability coefficient ($E\rho^2$) and two dependability indices ($\phi$) and $\phi(\lambda)$. The index $\phi(\lambda)$ is particularly important to analysis of pass-fail program data as well as other programs with absolute score interpretations.

Some important distinctions among $E\rho^2$, $\phi$, and $\phi(\lambda)$ are worth noting. An estimate of $E\rho^2$ is computationally equivalent to the traditional KR-20 estimate of reliability (Kuder and Richardson, 1937). It is appropriate for norm referenced interpretations because it describes the degree of consistency with which student scores are ranked by different readers or across different prompts. The coefficient $\phi$ "is an index reflecting the contribution of the measurement procedure to the dependability of domain-referenced decisions." (Brennan, 1983, p. 108). It is a conservative estimate and is appropriate for describing the dependability of decisions about individual students. The coefficient $\phi(\lambda)$ is an index of dependability of a domain referenced interpretation. "Specifically, $\phi(\lambda)$ reflects how closely the scores $X_p - \lambda$ can be expected to agree over randomly parallel instances of a measurement procedure." (Brennan, 1983, p. 108) It varies as the cut score ($\lambda$) varies. The expression $X_p - \lambda$ is the difference between the cut score ($\lambda$) and the mean of all observations for person p. The interested reader is directed to Brennan (1983) for complete development of these indices.

By manipulating the universes of generalizability, it is possible to derive varying values of $E\rho^2$, $\phi$, and $\phi(\lambda)$. Table 4 shows what these values would be if the indicated numbers of topics and readers had been employed. For this table, prompts are considered a fixed facet; i.e., the universe of generalization for prompts is only the 1-4 prompts hypothetically tested. All $\phi(\lambda)$ coefficients are based on a total cut score of 5.5 for two prompts and comparable cut scores for 1, 3, and 4 prompts.

Table 4
Values of $E_\rho^2$, $\phi$, and $\phi(\lambda)$ for
Varying Numbers of Prompts and Readers

| Prompts | Readers | $E_\rho^2$ | $\phi$ | $\phi(\lambda)$ |
|---|---|---|---|---|
| 1 | 1 | .91 | .91 | .91 |
| 1 | 2 | .95 | .95 | .95 |
| 1 | 3 | .97 | .97 | .97 |
| 1 | 4 | .97 | .97 | .98 |
| 2 | 1 | .92 | .92 | .93 |
| 2 * | 2 * | .96 | .96 | .96 |
| 2 | 3 | .97 | .97 | .97 |
| 2 | 4 | .98 | .98 | .98 |
| 3 | 1 | .93 | .93 | .94 |
| 3 | 2 | .97 | .97 | .97 |
| 3 | 3 | .98 | .98 | .98 |
| 3 | 4 | .98 | .98 | .98 |
| 4 | 1 | .94 | .94 | .95 |
| 4 | 2 | .97 | .97 | .97 |
| 4 | 3 | .98 | .98 | .98 |
| 4 | 4 | .99 | .99 | .99 |

* Typical configuration

Recall that the estimates of reader reliability were .950 for Prompt 1 and .958 for Prompt 2. From Table 4, we see that the estimated score reliability ($E\rho^2$) for one prompt and two readers is .95. Since prompts are fixed in Table 4, the only sources of error are readers and the reader × essay (student) interaction. The obtained coefficient should therefore be close to the previously computed reader reliability coefficient. The fact that it is indicates that the departure of the study from a strictly crossed design was very small. This coefficient was also confirmed with two smaller data sets (20 essays each) in which students and readers were completely crossed. The resultant values of $Ep^2$ for one prompt and two readers were .95 for Prompt 1 and .96 for Prompt 2. Thus, there is ample evidence that the departure from a strictly crossed design in this study did not significantly affect reliability indices.

Table 5 contains the results of the GENOVA analysis of scores with the variations described earlier. Specifically, scores for one reader of Prompt 1 essays were systematically increased, while scores for one reader of Prompt 2 essays were systematically decreased.

Table 5
GENOVA Source Table for Modified Data

| Source | df | SS | MS | G Study Variance Component |
|---|---|---|---|---|
| Student (S) | 999 | 9038.03 | 9.05 | .2617664 |
| Prompt (P) | 1 | 1427.61 | 1427.61 | .2198456 |
| Reader: P | 10 | 1027.74 | 102.77 | .1026307 |
| S×P | 999 | 5899.93 | 5.91 | .9603544 |
| S×(R:P) | 9990 | 1436.09 | 0.13 | .1437526 |
| Total | 11999 | 18829.45 | | |

A comparison of Tables 3 and 5 reveals two facts. First, total variance has increased slightly in Table 5. Second, the variance for readers has increased by a factor of over 1,500. At the same time, the variance component for students (true score) has actually decreased. A greater appreciation of the effect of this change can be obtained by studying Table 6.

Table 6
Comparison of $E\rho^2$, $\phi$, and $\phi(\lambda)$ for Two Data Sets
(Entries are original values / values based on modified data.)

| Prompts | Readers | $E_\rho^2$ | $\phi$ | $\phi(\lambda)$ |
|---|---|---|---|---|
| 1 | 1 | .91/.89 | .91/.83 | .91/.82 |
| 1 | 2 | .95/.94 | .95/.91 | .95/.91 |
| 1 | 3 | .97/.96 | .97/.94 | .97/.94 |
| 1 | 4 | .97/.97 | .97/.95 | .98/.95 |
| 2 | 1 | .92/.91 | .92/.86 | .93/.85 |
| 2 | 2 | .96/.95 | .96/.92 | .96/.92 |
| 2 | 3 | .97/.97 | .97/.95 | .97/.95 |
| 2 | 4 | .98/.98 | .98/.96 | .98/.96 |

While most values in Table 6 may be considered fairly high, one striking point becomes immediately obvious. Consider one prompt and two readers. With the errant readers in the group, $\phi(\lambda)$ is .91 (row 2, last column, second entry). This can be increased to .95 by doubling the number of readers; i.e., 1 prompt, 4 readers yields $\phi(\lambda)$ of .95 with the errant readers in the group. Yet, without the errant readers (or with these readers retrained) a $\phi(\lambda)$ value of .95 is obtained with only two readers per essay. Similarly, at 2 prompts, 2 readers $\phi(\lambda)$ is .92 for the poor group of readers. This coefficient is increased to .96 by doubling the number of readings per essay. Yet a $\phi(\lambda)$ value of .96 is obtained with two readings per essay if the systematically high and low readers are removed or retrained.

For the present study, prompts were considered fixed facets. What if prompts had been considered simply randomly selected representatives of a large unidimensional universe of prompts? Table 7 shows the values of $E\rho^2$, $\phi$, and $\phi(\lambda)$ for the same scores but with prompts considered random. For the types of prompts used in this study, attempts to generalize results to all possible prompts are clearly inappropriate.

Table 7
Values of $E\rho^2$, $\phi$, and $\phi(\lambda)$ for Random Prompts

| Prompts | Readers | $E\rho^2$ | $\phi$ | $\phi(\lambda)$ |
|---|---|---|---|---|
| 1 | 1 | .21 | .20 | .16 |
| 1 | 2 | .22 | .21 | .17 |
| 1 | 3 | .2[illegible] | .21 | .18 |
| 1 | 4 | .23 | .21 | .18 |
| 2 | 1 | .35 | .33 | .32 |
| 2 * | 2 * | .36 | .34 | .33 |
| 2 | 3 | .37 | .35 | .34 |
| 2 | 4 | .37 | .35 | .34 |
| 3 | 1 | .45 | .43 | .43 |
| 3 | 2 | .46 | .44 | .44 |
| 3 | 3 | .47 | .45 | .44 |
| 3 | 4 | .47 | .45 | .45 |
| 4 | 1 | .52 | .50 | .50 |
| 4 | 2 | .53 | .51 | .52 |
| 4 | 3 | .54 | .52 | .52 |
| 4 | 4 | .54 | .52 | .53 |

* Typical configuration

For writing assessment programs with a pass-fail component, a major issue is the dependability of a decision to assign a failing score. The procedures used in this study allow the investigator to estimate the likelihood of incorrect decisions (both false negatives and false positives). Data from Table 6 are presented in a revised format in Figure 1 to reflect the total score error variance and standard errors associated with each D study.

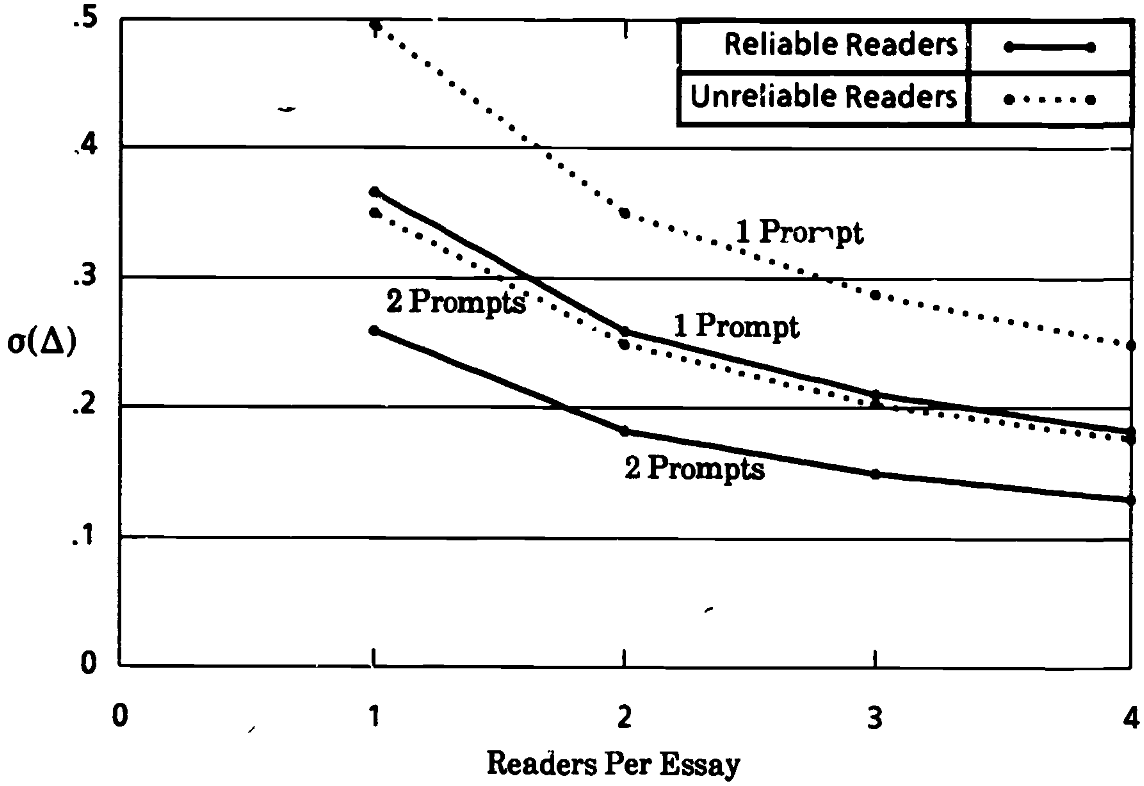

Figure 1. Standard error of measurement ( σ(Δ) ) as a function of numbers of prompts and readers for two sets of readers

The value of σ(Δ) is analogous to the traditional standard error of measurement. Brennan (1981) has provided 68%, 80%, and 90% confidence intervals for estimating an individual's universe score using σ(Δ). Generally speaking, the intervals do not behave exactly like those based on classical measures. Thus, the calculation of probabilities is somewhat cumbersome. A standard error based on $E\rho^2$, while less precise, does allow direct estimation in fairly simple cases (such as one prompt and two or more readers or any number of fixed prompts and any number of readers). The purpose of Figure 1 is to show that the probability of misclassification decreases asymptotically as prompts or readers are added, or as available readers read essays more consistently.

Finaily, it is appropriate to examine the effect of cut score on the decision dependability or φ(λ). Figure 2 shows the effect of cut score o. φ(λ) for two prompts, two readers. The top line (solid) represents fixed prompts, while th bottom line (broken) represents random prompts.

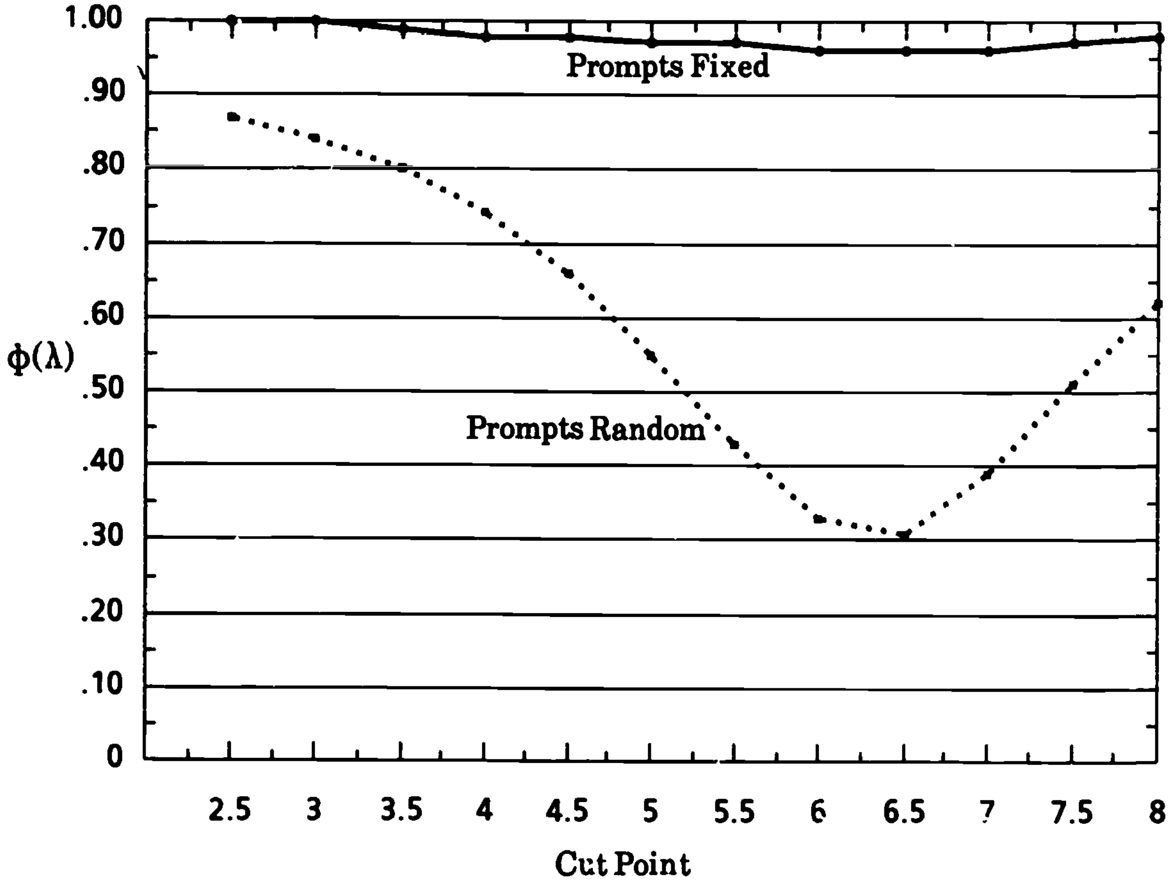

Figure 2. φ(λ) as a function of cut point for two designs

Since the placement of the cut score has very little effect on φ(λ) for fixed prompts, Figure 2 serves for illustrative purposes only. As the cutoff approaches the population mean, φ(λ) decreases. At the point where the mean and cutoff are identical, φ(λ) will take its lowest value. Thus, it may be helpful to compute such a value as a lower bound estimate of the dependability of scores. In programs with fixed descriptors associated with each score point (cf. p. 1 of this paper), a dependability coefficient can be computed for each point.

**Treatment of discrepant scores.** In practice non-identical, non-adjacent pairs of scores are brought to the attention of project managers who assign a third reading. In some projects, non-identical scores are resolved by a third reading. The final score is then based on the third score and only one of the first two scores. For example, if an essay receives scores of 2 and 4, it is sent to a resolution reader. This reader may assign a score of 2, 3, or 4 and drop either the original 2 score or the original 4 score. Whatever the case, the reported score is based on two scores that are more similar than if the unresolved scores were used.

It should be clear that any assessment of reliability would use these final scores rather than the original scores. In some programs (noticeably those with pass-fail components), essays may be read and reread until consensus is reached. For the present study, that would have occurred 21.2% of the time for Prompt 1 and 26.7% of the time for Prompt 2. Under such circumstances, final scores are based on pairs of readings with absolutely no reader variance. If only one prompt is used or if prompts are considered fixed facets, the error variance for such a scenario reduces to zero!

**Multiple attempts.** The discussion of standard errors and confidence intervals was based on a single administration of the test. In most programs, students who fail on the first attempt are afforded two, three, or more opportunities to pass. Thus, the probability of misclassification based on one attempt would need to be raised to the nth power for n attempts. For example, on his first attempt, a student's score is below the cut score and the resulting confidence interval shows a 12% chance that the student's true score was above the cut score. Put another way, given a true score equal to the cut score, the observed score could have occurred 12% of the time by chance. The likelihood of the same or lower observed score occurring twice in a row (given a true score equal to the cut score) would be $.12^2$ or 1.44%. While this would not hold strictly true for single-prompt tests with prompts changing from one administration to the next, it does point out the fact that one needs to consider the likelihood of repeated classification errors. Such errors are much less likely than single errors.

## Discussion

As writing assessment programs continue to proliferate, and as the need to defend the scores assigned in those programs becomes more apparent, questions about reliability will increase. Traditional reliability estimates, while helpful and informative, can not tell the whole story, particularly for multifaceted testing programs. Chapman, Fyans, and Kerins (1984) have employed generalizability analysis in conjunction with Illinois' writing assessment program but looked only at reliability of the process, not dependability of individual (absolute) scores. Their use of signal / noise ratios was an excellent way to sidestep some of the computational problems frequently associated with generalizability analysis while preserving critical information about sources of error.

Score dependability is dependent upon multiple factors: quality of the prompts, consistency of the readers, and placement of the cut score. Measures of reliability which ignore one or more of these factors fail to give a complete picture of the quality of the scoring process. It should also be clear that there is more than one way to increase reliabilty. Breland, Camp, Jones, Morris, and Rock (1987) suggested at least two modes of discourse and two prompts per mode as a way of achieving levels of reliability similar to those of standardized multiple-choice tests (p. 26). Braun (1986) suggested adjusting scores given by systematically high- or low-scoring readers. This procedure would reduce the reader variance component and increase reliability. This will work nicely when reported scores are scale scores or other transformed scores. However, when raw scores are reported in whole- or half-point intervals, scores with a tenth subtracted here and a hundredth added there could cause some credibility problems. If adjustments are made to the readers themselves, error variance can be reduced, reliability will be increased, and scores can be reported without artificial adjustments.

What we have attempted to demonstrate in this paper is that score reliability of essay tests is multifaceted and can be estimated in a variety of ways depending on the purpose of the assessment and the intended use of the results. When pass-fail decisions or determinations of absolute skill levels are to be made, indices that take into account the cut point or points are needed. Obviously, the way one chooses to view the prompts used in a specific assessment (i.e., random or fixed) makes a difference in the interpretation of results. One test can have many applications. Each application will have its own specific reliability. The method of computing an estimate of that reliability is dictated by the intended use of the results.

# References

Braun, H.I. (1986) *Calibration of Essay Readers: Final Report.* Program Statistics Research Technical Report No. 86-68. Princeton, New Jersey: Educational Testing Service.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., and Rock, D.A. (1987) *Assessing Writing Skill.* New York: College Entrance Examination Board.

Brennan, R.L. (1981) *Some Statistical Procedures for Domain REferenced Testing: A Handbook for Practitioners* (ACT Technical Bulletin No. 38) Iowa City, Iowa: The American College Testing Program.

Chapman, C.W., Fyans, L.J., and Kerins, C.T. (1984) Writing assessment in Illinois. *Educational Measurement Issues and Practice, 3*, 24-26.

Coffman, W.E. (1971) Essay examinations. In R.L. Thorndike (Ed.) *Educational Measurement* (2nd Ed.). Washington, D.C.: American Council on Education.

Crick, J.E. and Brennan, R.L. (1983) Manual for GENOVA: A Generalized Analysis of Variance System. ACT Technical Bulletin Number 43. Iowa City, Iowa: The American College Testing Program.

Cronbach, L.J. Glesser, G.C., Nanda, H., and Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

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

Rhode Island Department of Education (1987) *Rhode Island State ASsessment Program 1986-1987 Basic Skills, Health, and Physical Fitness Testing Results.* Providence, Rhode Island : Author.