MULTIPLE VIEWS OF L1 WRITING SCORE RELIABILITY

JAMES DEAN BROWN
University of Hawai'i at Manoa

ABSTRACT

This paper provides an in-depth investigation of the reliability of scores on the Manoa Writing Placement Examination (MWPE) based on a reanalysis of the scores of 6875 students who took MWPE during a four year period. Classical test theory analyses indicated that the MWPE scores were reasonably consistent across semesters (with a slight overall rise in scores over the four years), and that they produced traditional reliability estimates ranging from .51 to .74. The standard error of measurement was also examined in relationship to placement decision making. Generalizability theory was used to examine the relative importance of the numbers of topics and ratings to the consistency of scores. The results indicate that consistency would be increased more effectively by increasing the number of topics than by increasing the number of ratings. In addition, phi(lambda) indexes and signal-to-noise ratios were calculated to estimate score dependability at various decision points. The discussion focuses on the relative usefulness of these various approaches to reliability in practical test use, development, and interpretations.

INTRODUCTION

Yancey (1999) argued that first language writing assessment reseach has gone through three waves of development: (a) the era of objective assessment with an emphasis on reliability; (b) an era of objective assessment, which still stressed reliability, but also made room for validity; (c) and an era of portfolio assessment, in which validity, contextual, and interpretive notions superceded concerns about reliability. She also points out that these eras overlap, and as such, her representation of the trends in research are probably accurate. Unfortunately, from my point of view, researchers in writing assessment were far to quick to move from the first era to the second one given that they had not yet fully exploited all the available reliability tools that can help us understand of writing assessment reliability. In this paper, I will review the sorts of things that can be learned from using all the well-established classical theory reliability tools, but also explore what can be learned from generalizability theory that will not only help improve

Second Language Studies, 25(2), Spring 2007, pp. 1-31.

the consistency of writing assessments, but also increase understanding of their validity.

Classical Theory Questions

In general terms, the reliability of any test is the degree to which the scores are consistent. As White (1985, p. 177) put it, "The reliability of a measure is an indication of its consistency, or its simple fairness." White (1990, p. 192) and Huot (1990, pp. 202-203) long ago pointed out that the first language writing assessment literature often focused on classical test theory reliability in the form of interrater correlations (e.g., Godshalk, Swineford & Coffman, 1966, pp. 32-37; Diederich, 1974, pp. 32-41; Faigley, Cherry, Jolliffe, & Skinner, 1985, pp. 109-111; Hayes & Hatch, 1999; Penny, Johnson, & Gordon 2000; Johnson, Penny, & Gordon, 2001; Brown, Glasswell, & Harland, 2004; Lee, 2004). Both White and Huot further argued that this classical theory conception of reliability was integrally related to two ideas:

1. "true scores" [i.e., the concept that the ratings received by a student are estimates of that student's true score (the average of all possible scores if he/she were to theoretically take the test an infinite number of times)]
2. "error" [i.e., the notion that some of the variation in scores is due to uncontrolled, random factors that have nothing to do with students' writing abilities]

White (1990, p. 192) found the classical theory notions of "true scores" and "error" to be objectionable because:

1. the phraseology used in "true scores" suggests "a kind of purity and objectivity of measurement",
2. the use of such language fosters fundamental belief in the tests that are used and test scores that result,
3. measurement takes on its own "reality of numbers and charts, equations and computers...",
4. there are times when differences of opinion about a composition simply cannot be resolved and indeed may be a healthy sign,
5. these concepts cause negative exaggeration of the effects of disagreement.

With reference to the last two points, White (1990, p. 192) amplified as follows:

> Is our difference of opinion [in rating essays] to be called error? Not at all. In fact, historically, such differences about value in most areas of experience tend to be more valuable than absolute agreement; they combine to bring us nearer to accurate evaluation than would simple agreement.

Elsewhere, White (1989, p. 94) put it somewhat differently:

> If we are determining the value of a complex phenomenon, such as a writing sample or a work of art, there is probably not a single correct answer or a single right judgment. A difference in judgment is not only *not* error but is positively valuable; the variety of judgments (as experience has shown) helps us see the work in question more clearly and estimate its value more intelligently than a simple unanimity would.

Thus, a dilemma appears to have existed: on the one hand, classical theory reliability and its attendant notions of "true scores" and "error" have fallen into disrepute in the writing assessment community, while, on the other hand, consistent writing assessment procedures continue to be in demand in real educational settings.

Generalizability Theory Answers

One branch of educational testing has gone largely unheralded in the writing assessment literature, a branch that might prove particularly useful for estimating the consistency of essay scoring schemes. White (1990) alluded to this branch of measurement theory as follows:

> A team headed by the distinguished psychometrist L. J. Cronbach developed what he called 'generalizability theory,' based on a 'consensus score' rather than a true score, some decades ago. A consensus score can yield very useful measurement, reflecting the social process of judgment, and offers sound statistical data. (pp. 192-193)

Indeed, generalizability theory (G theory) allows researchers to examine the consistency of a set of essay scores while experimentally investigating the effects of particular variables of interest (alluded to in a general sense in Shale, 2004), variables that affect both reliability and validity. Since generalizability theory has been readily accessible since 1972 (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and since G theory is particularly well-suited to the problems of essay scoring, it is amazing that no

actual applications of generalizability theory appeared in the composition assessment literature until the late 1980s.

Brown (1988b, 1989, 1990a, 1991) was a series of yearly in-house G-theory analyses of the scores on the Manoa Writing Placement Examination (MWPE). This placement test was taken each year by all incoming freshman students who had not already taken a freshman composition course. Each of the four studies cited above used G theory to examine the relative effects of persons, raters, and prompt topic types on the dependability of scores on the MWPE.

Lane and Sabers (1989) also used G theory. They studied 15 essays written by native speakers of English (grades 3 to 8). The prompt asked the students to write for 10 minutes on "best and worst things about a rainy day" (p. 198). Eight raters were used from a variety of backgrounds. The researchers then used G theory to estimate the relative importance of persons, categories, and raters as sources of error in the rating process.

Sudweeks, Reeve, and Bradshaw (2005) did a pilot study designed to improve the rating process for essays testing the writing ability of college sophomores. They used generalizability theory to estimate the relative importance of persons, occasions, tasks, and raters as sources of error in the rating process.

Only fairly recently, then, have several G studies been performed to examine the relative contribution of various factors to the dependability of essay examination scores with different student populations. However, even these studies were limited in the ways they used G theory. That is, they did not use G theory to its full potential (for much more on G theory and its full potential, see Shavelson & Webb, 1991; Brennan, 2001; Chiu, 2001).

Purpose

The purpose of the present study is to examine a number of different aspects of writing test score consistency, including both classical theory and generalizability theory approaches. Some of these have never been touched on in the literature on first language writing assessment. From the outset, it was clear that there is much more to the concept of reliability than has typically been reported in the writing assessment literature.

Reliability is not just a coefficient; it serves as the basis for sound decision-making practices and is a precondition for test validity. Indeed, reliability functions as a fundamental building block in any test construction, use, and interpretation.

In order to examine test consistency in depth, this study drew upon three classical theory notions (interrater reliability, K-R20 reliability, and the standard error of measurement) and three generalizability theory concepts (generalizability coefficient, the phi(lambda) coefficient, and the signal/noise ratio). As a starting point, the following research questions were posed:

1. How consistent are test results on a holistic rating scale in terms of the percentages of students placed into each level of study from one year to the next?
2. Using classical test theory, to what degree is a holistic rating scale reliable (from an interrater reliability perspective and from the K-R20 perspective)?
3. How can the degree of classical theory reliability be appropriately estimated in test score terms that will help in making fair placement decisions?
4. Using generalizability theory, to what degree do the number of topics and ratings contribute to the consistency of the scores?
5. How dependable are the test scores when they are used in the placement decision processes?

Ultimately, the goal of this study was to use the information gained to improve the consistency of one institution's testing and scoring procedures. However, the demonstrations and explanations provided herein will hopefully also stimulate other researchers to use forms of analysis other than interrater and intrarater coefficients. In other words, the research questions posed above, or very similar ones, are questions that should be asked of any composition assessment procedure. The ultimate importance of reliability should never be underestimated because a set of scores can be reliable without being valid, but they cannot logically be any more valid than they are reliable (i.e., the scores cannot be systematically testing what they purport to test unless they are first shown to be systematic).

METHOD

Participants

All incoming Freshman (and transfer students without English composition credit) at the University of Hawai‘i are required to take the Manoa Writing Placement Examination (MWPE). Typically, the MWPE was administered on four or five dates in the Spring semester and three or four before and during the Fall semester. These examination sessions were conducted primarily at the main (Manoa) campus of the University of Hawaii, but were also offered at sites on the neighbor islands for the convenience of potential students from all parts of the state. A total of 6875 students took MWPE during a four year period covered in this study.

Testing Materials and Procedures

Each student was required to write on two topics. These topics were of distinctly different types:

1. an essay in response to a two-page academic prose reading (e.g., one such reading presented two points of view on genetic engineering);
2. an essay based on personal experience (e.g., one question asked for a discussion of the effects of television on contemporary society and another asked for a discussion of prejudice in education).

Twenty-four different sets of prompts were developed over the four year period covered here, but all sets consisted of pairs as described above.

Students were randomly assigned to particular prompt sets, and the sets were administered in a counterbalanced manner so that a randomly selected half of the students addressed the reading-based prompt first, while the other students addressed the personal-experience prompt first. The students were allowed seventy-five minutes to draft their initial responses, followed by a fifteen minute break. The students then had seventy-five more minutes for drafting their response to the second prompt. After a lunch break, the students were given two additional hours to revise their responses to both topics. They were required to return for the revising session, but could leave as soon as they had completed their revisions. Thus the students were allowed a total of 270 minutes of

writing time if they felt that they needed it.

It is important to note that the MWPE provides unique opportunities for research in the areas of reliability and validity because of its basic design, two types of topics with two ratings each for a minimum total of four ratings for a large number of students.

Scoring Materials and Procedures

Before each scoring session, raters were sent a packet of scoring materials that they were to read in order to become familiar with the rating procedures. The scoring materials contained:

1. sample test questions,
2. example responses by students at various levels of ability,
3. holistic scores based on the five point rubric, and
4. justifications for the holistic scores.

The scoring rubric indicated that the scorers should consider the following six facets of student work: (a) thesis/evidence, (b) organization/ development, (c) grammar, (d) diction/style, (e) spelling, and (f) punctuation. Nevertheless, the score was a "holistic" one based on a fast reading and reference to the descriptions provided in the zero-to-five point scale of the rubric.

All of the examinations were holistically scored in three-hour scoring sessions. The first hour was used for training. Approximately two hours were then devoted to actual scoring. Scorers were directed to consider the fact that the essays had been written in a relatively short period of time under testing conditions. Each essay was rated by a minimum of two readers. All of the readers were faculty members with teaching experience at the University of Hawaii. In cases where a two-point or greater discrepancy was found between scores, the essay was read by a third rater, who resolved the difference (as suggested by Diederich, 1974, p. 35). Such discrepancies occurred in about six percent of the cases.

Placement Decisions

On the basis of their scores, students were placed into one of four categories for English composition instruction:

1. accelerated ENG100A (for those who already wrote well)
2. regular ENG100 (for those with average writing skills)
3. supplemental ENG101 (equivalent of ENG100 plus a one-credit writing tutorial)
4. remedial ENG22 (for those unprepared to do college-level writing)

The placement of each student was based on a composite score which could range from 0 to 20 (2 essays evaluated by 2 readers each using a 5 point scale, or 2 x 2 x 5 = 20). Students were assigned to English composition courses on the basis of the MWPE scores as follows:

1. scores of 15-20 resulted in assignment to accelerated instruction
2. scores ranging from 10-14 meant assignment to regular instruction
3. scores of 7-9 caused assignment to regular instruction plus a supplemental writing tutorial
4. scores in the 0-6 range produced assignment to remedial instruction

RESULTS

Descriptive Statistics

The descriptive statistics shown in Table 1 give the means and standard deviations for the first, second, third, and fourth ratings, as well as for the subscores on each of the topics, and the total. TOPIC A is for the essay based on a reading, and TOPIC B is for the essay based on personal experience. Note that the individual Rating scores are based on a 0-5 point scale, the Subscores are based on a 0-10 point scale (the sum of the two ratings for each topic), and the Total scores are based on a 0-20 point scale (the sum of four ratings for each student).

Table 1
Descriptive Statistics

TOPIC Rating	*M*	*SD*
TOPIC A		
Rating 1	2.83	0.86
Rating 2	2.80	0.84
Subscore A	5.63	1.48
TOPIC B		
Rating 3	2.81	0.86
Rating 4	2.80	0.84
Subscore B	5.61	1.48
TOTAL	11.23	2.47

Percentages in Each Course

In the process of using the MWPE scores to make decisions about the placement of students, one central concern was with the degree to which the test results on this holistic scale were consistent from year to year in terms of the percentages of students placed into each level. Table 2 presents the percentages of students who placed into each of the course levels for each of the four school years.

Table 2
MWPE Score Frequencies

YEAR	ENG22 (0-6)	ENG101 (7-9)	ENG100 (10-14)	ENG100A (15-19)	TOTAL
1	1%	12%	77%	10%	100%
2	1%	14%	77%	8%	100%
3	5%	21%	67%	7%	100%
4	7%	24%	61%	7%	100%

Test Reliability

Two key terms, reliability and validity, will arise often in the ensuing explanation of results. Therefore, brief definitions may be in order. First, as mentioned in the **INTRODUCTION**, test reliability is defined here as the degree to which the scores on a test are consistent. Second, test validity is defined here, in its rather traditional sense, as the degree to which a test is measuring that which it was designed to test (for much more on these topics, see Brown, 2005).

The following two subsections will explore the reliability of the MWPE from two perspectives: the classical test theory approach and the generalizability theory approach. These approaches are treated separately because they provide different types of information, both of which are useful in evaluating and improving the consistency and effectiveness of a test.

Classical theory approach. In classical test theory, reliability coefficients can range from zero, when a test provides scores that are totally inconsistent (or random) to 1.00 when the scores are completely reliable (i.e., the test provides perfectly consistent scores, which are assumed to be exactly equivalent to the students' true abilities). Generally, the higher the coefficient, the more reliable the results can be considered, and the more confidently decisions can be made on the basis of them.

Traditionally, holistic writing scales such as the one used in this study have been evaluated for reliability by calculating an interrater reliability coefficient. For instance, in this study, a Pearson product-moment correlation coefficient was calculated to estimate the degree of relationship between the first and second ratings assigned for the composition on Topic A. As shown in the first column of numbers in Table 3, the correlation between these two sets of scores was .51. Interpreted as a reliability estimate, the .51 indicates that about 51 percent of the variation in ratings was consistent, while 49 percent remained inconsistent, or random—a somewhat discouraging state of affairs. However, this statistic, which is commonly reported as interrater reliability in the literature, is providing an incomplete picture of the reliability of the scores as they were used on the test. The problem is that this interrater correlation only estimates the

reliability of one set of ratings (either set, but only one set) when in fact two sets of ratings were used for each topic on the MWPE (see Guilford, 1954, p. 397 for explanation of when and how to use the Spearman-Brown formula on ratings). The second column of numbers in Table 3 shows the interrater reliability after it was adjusted (using the Spearman-Brown formula) to reflect the reliability of the two sets of scores taken together. For Topic A (based on a reading), the reliability of the two readings taken together was estimated to be .68. For Topic B (the topic based on personal experience), the analogous interrater correlation turned out to be .53, while the reliability estimate adjusted for two sets of scores was .69. The adjusted estimates for Topics A and B both indicate that there was a certain amount of agreement, or consistency, (68 and 69 percent, respectively) for the ratings produced for each topic type, but also a good deal of disagreement (32 and 31 percent, respectively) between the first and second ratings.

Table 3
Summary Classical Theory Reliability Statistics

Statistic	Topic A			Topic B			Mean	K-R20
Number of Raters	1	2	4	1	2	4	4	4
Reliability	0.51	0.68	0.81	0.53	0.69	0.82	0.74	0.70
SEM							1.26	1.35

The pairs of ratings were examined together for three reasons: (a) because that is how they were used in actual decision making, (b) because multiple observations are known to be more reliable than single observations, and (c) because, as pointed out in the **INTRODUCTION**, the scores "combine to bring us nearer to accurate evaluation than would simple agreement" (White, 1990, p. 192).

The amount of unreliable variance found here for the scores of two raters (nearly one-third) would be disheartening if it were not for the fact that the actual placement of students was based on a minimum of four ratings, rather than being based on any two ratings. To examine the reliability of four ratings taken together, three different classical strategies were used:

1. The interrater estimates were further adjusted (using the Spearman-Brown

formula) to take into account the fact that four ratings are always used in the MWPE decision-making process rather than just the two ratings assigned to either topic. These adjustments for four ratings resulted in estimates of .81 and .82 as shown in the third and sixth columns of numbers in Table 3.

2. The interrater correlation coefficients (for all possible combinations of the four sets of scores) were averaged using the Fisher *z* transformation and then adjusted for four ratings. The result was .74 as shown in the seventh column of numbers in Table 3.
3. Ebel's (1979, pp. 282-284) variation of the K-R20 reliability formula was applied. Ebel's formula was adapted for use in just such situations, i.e., situations wherein rating scales are used on multiple questions or used by multiple raters. The result of this formula was .70 as shown in the last column of Table 3.

Another classical theory statistic that can be usefully applied to examining the reliability of composition scores is the standard error of measurement (SEM). The SEM is especially useful because it relates test consistency to the decision-making process. In brief, the SEM is a statistic which expresses (in probability terms) the degree to which scores are likely to fluctuate due to unreliable variance. For example, the SEM of 1.26 found for the *average* interrater reliability (see Table 3) indicates that students' total scores (based on the 20 point scale) on the MWPE would fluctuate ± 1.26 points with 68 percent probability if the students were to take the test again. This has important implications for decision making, especially for those students who are close to the cut-points between courses in the placement decisions. Consider a student who has a score of 6 and is therefore just barely placed into the ENG22 course. The SEM indicates that this student might score within a band of scores that goes as high as 7.26 (6 + 1.26) or as low as 4.74 (6 - 1.26) with 68 percent certainty if he/she were to take the test again. Thus the SEM indicates that the student might place into ENG22 or ENG101 on subsequent administrations of the test.

The SEM provides useful information because it helps in identifying those students who might score in a different placement category on subsequent administrations of the

examination. In the interest of fairness, such students should be handled with special care in the decision-making process. In most cases, *special care* means that additional information should be gathered to help make decisions about those students close to a cut-point. From a practical point of view, educators usually prefer to protect the interests of the students and be fair by not placing them too low. As such, it is often advisable to gather additional pertinent information at least for those students who fall within one SEM below a cut-point.

The classical theory reliability estimates and standard error of measurement are both useful: the former for assessing how consistently raters were scoring and the latter for determining how many points of variation can be expected in scores with regard to placement decisions. However, neither statistic helps directly in deciding how to best restructure the testing and scoring procedures in order to improve their reliability. That is a task for which G theory is particularly well suited so let's now turn to G theory.

Generalizability theory approach. G theory was used in this study to investigate the degree to which the number of different topics and ratings were contributing to the consistent variance in the MWPE scores (see Cronbach et al., 1972; Brennan, 1983). Recall that on the MWPE, each student wrote two essays on topics which were purposely made to be different—one based on a two-page reading and the other based on personal experience. Hence one issue of interest here was whether it was necessary to use two topic types. Would the test scores have been equally consistent with only one? Would three topic types have been more effective? In short, what was the degree to which the number of different topic types affected the consistency of the test scores?

To answer such questions, the number of topic types was designated as one facet in a generalizability study (G study). [Note that this facet was treated as a fixed effect because the types of topics have been the same over the past four years and will not be changed in the foreseeable future.] Since the number of ratings is typically a major source of variance in writing test scores, the ratings variable was included as a second facet in the G study. The question was whether four ratings were needed for each student's essays (two for each of the two topic types). Would two have been enough? Would six ratings

have been more effective? To answer such questions, the second issue of interest in this G study was the degree to which the number of different ratings affected the consistency of the scores. [Note that the ratings facet was treated as a random effect because the actual persons doing the ratings varied considerably from rating session to rating session and from year to year, and because they could reasonably be assumed to be a representative sample of the universe of all possible raters (see Shavelson & Webb, 1981, p. 143).]

In the conducting the G study, Analysis of Variance (ANOVA) procedures were used to isolate the variance components (σ^2) for each facet in the design, as well as for interactions of those facets (see Table 4). The expected observed score variances were then estimated for various combinations of numbers of topic types and numbers of ratings. Next, the estimated errors were calculated for each of the combinations of numbers of topic types and numbers of ratings. Since the purpose of the MWPE was clearly norm-referenced placement, only the expected observed score variances and error for relative decisions were calculated. [Note that G theory can also be used to study the absolute decisions involved in criterion-referenced tests.] Generalizability coefficients (G coefficients) for various combinations of numbers of topic types and ratings were estimated by dividing the sum of the expected variance components for persons and person-by-topic interactions by the expected observed score variance (those same variance components plus the appropriate estimated error).

TABLE 4
ANOVA for G Study (p x r:t) with Estimated Variance Components

Source	*df*	*SS*	*MS*	σ^2
Persons	6874	10495.8097	1.5269	0.2139647
Topics	1.00	0.8624	0.8624	0.0000000
Ratings: Topics	2	2.5543	1.2772	0.0001353
Persons x Topics	6874	4612.6376	0.6710	0.1620537
Persons x Ratings: Topics	13748	4769.4457	0.3469	0.3469192
Total	27499	19811.3097		

The G coefficients shown in Table 5 are analogous to classical theory reliability

coefficients. For example, the table shows a G coefficient of .632 for two topics and four ratings (see column two, row four of the numbers within Table 5), i.e., the conditions under which MWPE placement decisions are currently made. This statistic indicates that the scores are approximately 63 percent consistent and 37 percent inconsistent. G coefficients are also presented for other possible numbers of topics and ratings beyond those actually used in this study. For instance, cutting back to one topic with two ratings would clearly make the procedure much less consistent (G coefficient = .389). If three topics and three ratings were used, it would have the opposite effect of increasing the G coefficient (G coefficient = .698). This is important information in that it makes clear the relative value of increasing or decreasing the numbers in each facet (numbers of topics, or numbers of ratings) separately and together.

Table 5
Generalizability Coefficients for Different Numbers of Topic Types and Ratings

Topics	1	2	3	4	5	6	7	8	9	10	20	30	40	50	100
Ratings															
1	0.296	0.457	0.558	0.627	0.678	0.716	0.746	0.771	0.791	0.808	0.894	0.927	0.944	0.955	0.977
2	0.389	0.561	0.657	0.718	0.761	0.793	0.817	0.836	0.852	0.864	0.927	0.950	0.962	0.970	0.985
3	0.435	0.606	0.698	0.755	0.794	0.822	0.844	0.860	0.874	0.885	0.939	0.959	0.969	0.975	0.987
4	0.462	0.632	0.721	0.775	0.811	0.838	0.858	0.873	0.886	0.896	0.945	0.963	0.972	0.977	0.989
5	0.480	0.649	0.735	0.787	0.822	0.847	0.866	0.881	0.893	0.902	0.949	0.965	0.974	0.979	0.989
6	0.493	0.661	0.745	0.796	0.830	0.854	0.872	0.886	0.898	0.907	0.951	0.967	0.975	0.980	0.990
7	0.503	0.669	0.752	0.802	0.835	0.858	0.876	0.890	0.901	0.910	0.953	0.968	0.976	0.981	0.990
8	0.510	0.676	0.758	0.806	0.839	0.862	0.879	0.893	0.904	0.912	0.954	0.969	0.977	0.981	0.990
9	0.516	0.681	0.762	0.810	0.842	0.865	0.882	0.895	0.906	0.914	0.955	0.970	0.977	0.982	0.991
10	0.521	0.685	0.765	0.813	0.845	0.867	0.884	0.897	0.907	0.916	0.956	0.970	0.978	0.982	0.991
20	0.544	0.705	0.782	0.827	0.856	0.877	0.893	0.905	0.915	0.923	0.960	0.973	0.979	0.984	0.992
30	0.552	0.711	0.787	0.831	0.860	0.881	0.896	0.908	0.917	0.925	0.961	0.974	0.980	0.984	0.992
40	0.556	0.715	0.790	0.834	0.862	0.883	0.898	0.909	0.919	0.926	0.962	0.974	0.980	0.984	0.992
50	0.559	0.717	0.792	0.835	0.864	0.884	0.899	0.910	0.919	0.927	0.962	0.974	0.981	0.984	0.992
100	0.564	0.721	0.795	0.838	0.866	0.886	0.900	0.912	0.921	0.928	0.963	0.975	0.981	0.985	0.992

Because the consistency of decisions based on test scores may vary depending on where the cut-points are placed in the continuum of possible scores, it is often useful to also examine the dependability of scores at various cut-points used for decision making. The phi(lambda) dependability index, or Φ(λ), is one statistic that can be used to study this issue. Phi(lambda) is similar to a reliability coefficient in that it gives an estimate of the degree of score consistency. However, the Φ(λ) index is based on the squared-error loss agreement strategy, which focuses on the degree to which classifications in clear-cut categories have been consistent. In addition, it is known to have "...sensitivity to the degrees of mastery and nonmastery along the score continuum..." (Berk, 1984, p. 246). [See Brown (1990b) for much more on this topic.]

Table 6
Phi(lambda) Dependability Indices and Signal/Noise Ratios at Various Cut Points

Cut Point	Decisions	Φ(λ)	Signal/Noise
2.00		0.98	44.55
4.00		0.97	28.00
6.00	ENG22	0.94	15.48
8.00		0.87	6.97
9.00	ENG101	0.81	4.22
10.00		0.71	2.48
*11.23		0.63	1.72
12.00		0.67	2.01
14.00	ENG100	0.85	5.57
16.00		0.93	13.13
18.00		0.96	24.72
20.00		0.98	40.33

*overall mean = 11.23

The single biggest advantage of Φ(λ) is that dependability estimates can be made for different decision points. Table 6 presents the Φ(λ) estimates for different possible cut scores (when four ratings are used to judge two essays as they were on the MWPE). If the cut score for a placement decision were on the mean of 11.23, the table indicates that the

dependability would be .63, or the lowest Φ(λ) reported in the table. It turns out that the dependability of such decisions is always lowest at the point that corresponds to the mean on the test (Brennan 1980, 1984). Fortunately, none of the actual cut-points used on the MWPE (at the time when these data were gathered) was near the mean. A glance at the table will indicate that the decision dependability for ENG22 is a satisfactory .94, while the corresponding statistics for ENG101 and ENG100 are .81 and .85, respectively.

Brennan (1984) suggests that the signal-to-noise ratios may also provide a "useful alternative coefficient for norm-referenced interpretations" (p. 306). Brennan goes on to define what is meant by the terms signal and noise as follows:

> The signal is intended to characterize the magnitude of the desired discriminations. Noise characterizes the effect of extraneous variables in blurring these discriminations. If the signal is large compared to the noise, the intended discriminations are easily made. If the signal is weak compared to the noise, the intended discriminations may be completely lost. (p. 306)

In the column furthest to the right in Table 6, signal/noise ratios are presented for each of the cut-points. Clearly, the decision discriminations that are made on the MWPE are most difficult at the ENG101 decision point with a ratio of about 4 to 1, while the easiest discriminations are those for the ENG22 decision with a ratio of nearly 15.5 to 1.

DISCUSSION

The primary purpose of this section will be to provide direct answers to the research questions posed at the end of the **INTRODUCTION** section. To help organize the discussion, the research questions will serve as subheadings.

1. How consistent are test results on a holistic rating scale in terms of the percentages of students placed into each level of study from one year to the next?

The MWPE appears to have functioned reasonably well during the four year period in this study as indicated by the descriptive statistics given in Table 1. These statistics further demonstrate that the MWPE scores are normally distributed, reasonably well-

centered, and dispersed as widely as the range of possible scores allows. This is important because it means that the MWPE is functioning efficiently as a test for norm-referenced interpretations.

The distributions of MWPE test scores according to levels of placement also appear to be reasonably consistent across years as shown in Table 2. However, careful inspection of Table 2 will reveal that there is an overall rise in scores that is reflected in the yearly increase in the percentages of high scores and the corresponding decrease in the percentages of low scores. This overall rise may be caused by such factors as better or more focused writing teaching in the local high schools, changes in the types of students coming to the University of Hawaii, breaches of test security, changes in rater behavior, etc. Hence such a rise in scores may be an indication that there is a weakness in the reliability of the testing procedures over time—a weakness due to one or more underlying variables that should perhaps be controlled. However, the only recourse in the present situation is to carefully monitor and study this issue in the hope of determining the underlying cause(s) in the future.

Clearly, in thinking about the reliability of composition assessment procedures, it is useful to examine the relationships among yearly score distributions, as well as the consistency of percentages of students placed in each category from year to year (or from semester to semester, or even from test administration to test administration). However, it is also important to come to grips with the causes of any observed shifts over time because of the additional threats that such shifts may pose to the validity of the test. In other words, if there are major and continuing shifts of scores, validity will be threatened by the fact that, after a certain amount of time, the test will no longer be testing what it was originally designed to measure. Thus reliability and validity appear to be interdependent with regard to this issue of shifts in the percentages of students placed into different courses.

2. Using classical test theory, to what degree is a holistic rating scale reliable (from an interrater reliability perspective and from the K-R20 perspective)?

The overall classical theory reliability results indicate that the holistic rating scale used in the MWPE is reasonably reliable. In general, it appears that roughly 70 percent of

the variance in MWPE scores is consistent while 30 percent is random. This might be worrisome if it were not for the fact that there is a clear trend toward increasing reliability when the results for each of the four years are compared. For instance, the K-R20 estimates for each of the years (reported in Brown 1988b, 1989, 1990a, & 1991) were as follows: .69, .67, .85, and .82, repectively for the four years.

This year-to-year variation of the reliability estimates points to a very important fact about reliability that is often ignored in the writing assessment literature: a reliability estimate is for a particular set of scores not for the testing procedures being used. In other words, if a set of procedures turns out to produce highly reliable scores in one situation, but is then used on a different type of students, it may not be reliable at all. Thus the reliability of the scores produced by a set of composition assessment procedures should be examined every time the examination is administered because, as illustrated here, the reliability may change rather dramatically when the test is administered under different conditions, or to different students.

The reader may also have noticed the wide variety of results reported in Table 3 including estimates ranging from .51 to .82. Such variation illustrates several additional points about classical theory reliability which are often ignored in the composition assessment literature. To begin with, there are a variety of methods for calculating reliability, which may result in indexes of different magnitudes and may have different meanings. For example, the first reliability estimate given in Table 3 is .51 for one rating on Topic A. If, however, two ratings are used on Topic A the reliability is estimated to be .68, and if four raters had been used on Topic A the reliability would have been about .81. A similar set of relationships exists for Topic B. However, four raters were not used on either topic. In fact, two raters were used on each of two topics and the reliability for this situation is probably better estimated by the AVERAGE (.74) and K-R20 estimates, which turned out to be .74 and .70, respectively.

Thus including only the interrater estimates (or even the interrater estimates adjusted for two raters) would have provided an incomplete picture because four raters were used in making the decisions. At the same time, presenting only the interrater estimates adjusted for four raters on each topic would have been misleading because in fact two raters are used on each of two topics for a total of four ratings—a different situation

entirely. Thus the reliability of the scores in this study was best estimated by using the AVERAGE and K-R20 strategies.

In other words, in thinking about the reliability of composition assessment procedures, it is often useful to examine classical theory reliability coefficients from a number of perspectives. The interrater correlations are useful as estimates of the single rater reliability, but additional helpful information may be provided by adjusting the same correlations (using the Spearman-Brown formula) for the numbers of ratings involved in various combinations of topics or other subtest divisions. It may also prove useful to average all possible interrater correlations (using the Fisher *z* transformation) and adjust the result for the total number of ratings, or to use Ebel's variation of the K-R20 formula. Estimation of reliability can take many forms, and proper interpretation of the results is crucial.

It is also important to realize that even elaborate study of reliability is just a start. As Carlson and Bridgeman (1986) pointed out:

> High reliability does not provide sufficient evidence that a test is valid. Instead, the test may be measuring a variable consistently that is not the primary criterion of interest. (p. 146).

Hence it is also important to recognize that classical theory reliability is only a first step.

Reliability, though not sufficient for demonstrating the value of a test, is a logical and necessary precondition for establishing the validity of a test. Logically and mathematically, a set of scores can be no more valid than they are reliable. Logically, if a set of scores is unreliable and inconsistent, the test is likely to be assessing unsystematic factors other than what it was designed to test. Mathematically, it turns out that the squared value of a criterion-related validity coefficient (the result of one strategy for analyzing test validity), called the coefficient of determination, cannot be larger than the reliability coefficient of either of the two tests involved.

Put yet another way, the reliability coefficient for a set of composition scoring procedures is an estimate of the upper limit that validity can attain. Thus it appears that a test cannot be any more valid than it is reliable and that the establishment of test reliability is a necessary first step in studying the quality of the test. In short, reliability is the foundation upon which validity arguments can then be built. Thus, once again,

reliability and validity are clearly interdependent concepts.

3. How can the degree of classical theory reliability be appropriately estimated in test score terms that will help in making fair placement decisions?

Regardless of which type of reliability coefficient is used, the estimate produced is only an abstract indication of the degree of consistency in the scores. As mentioned in the **RESULTS** section above, another approach to reliability that is integrally related to actual score values and decision making is the standard error of measurement. The practical importance of this statistic is found in the way it can be used to establish a band of scores around each decision point within which the decisions will be made with greater care. For example, in the present study, an SEM of 1.26 was found (based on the average interrater reliability) and decision points exist at scores of 6, 9, and 14, for placement into ENG101, ENG100, and ENG100A (for clarification, see Table 2 or 6). The SEM indicates (with 68 percent accuracy) that students with a score of 6 could score anywhere within a band of possible scores ranging from as low as 4.74 to as high as 7.26 if they were to take the test again. Taking this as a premise, decision makers should gather additional information about any students falling within ±1.26 points of a cut-point so that the accuracy of those decisions will be enhanced. Since the scores on the MWPE are always whole numbers, the band of scores that is actually used around each decision point has been rounded to ±1.00. In other words, additional information is gathered for students scoring between 5-7 (6 ±1), 8-10 (9 ±1), and 13-15 (14 ±1). Such additional information can take many forms: another writing test, a portfolio of the students work, an SAT verbal subtest score, a transcript of high school grades in English, and/or anything else that the decision makers view as pertinent.

In short, in thinking about the reliability of composition assessment procedures, it is useful to examine the SEM because of its potential importance in making the actual decisions which result from the test scores. However, it is also important to understand that the SEM may be fundamentally related to the validity of the test, particularly a placement test, in the sense that the SEM can be used to increase the accuracy of the placement decisions, which in turn increases the degree to which the test is measuring

what it was designed to measure. Thus reliability and validity again appear to be interdependent notions.

4. Using generalizability theory, to what degree do the number of topics and ratings contribute to the consistency of the scores?

In the best of all possible worlds (with unlimited resources), the generalizability results shown in Table 5 could be used to create set of examination procedures that would produce almost perfectly consistent scores. For instance, given the results of this study, it is clear that, if the students could be made to write on 100 different topics, each of which is given 100 separate ratings, a nearly perfect G coefficient of .992 would result. Short of that rather impractical set of procedures, the table can be used to determine which realistic restructuring steps would be most helpful. It seems clear from the results in Table 5 that, overall, the facet representing number of topics has a greater effect on the consistency of the scores than does the facet for number of ratings. This conclusion is based on the apparent pattern of coefficients increasing in magnitude more rapidly from column to column (left to right) than they do from row to row (top to bottom) in the table. Thus, to increase the consistency of the MWPE, it would seem to be more important to increase the number of topics than to augment the number of ratings. This means, for instance, that having the students write on four topics with one rater for each topic would probably prove to be more reliable (G coefficient = .775) than the present strategy of having students write on two topics, each of which is rated by two raters (G coefficient = .632).

These results seem to address the issues raised by Hoetker (1982) when he stated that:

> ...we know almost nothing about topic variables because the attention of researchers has been devoted almost entirely to issues of rater reliability, while issues of validity have been ignored, as have the other two sources of variation in essay examination results: students and topics. (p. 380)

Indeed, it appears that G theory can be used to incorporate the study of variation due to students and topics directly into investigation of test reliability (and include ratings as a source of variation, as well). In fact, with G theory as a tool, almost any variable of interest can be included in the study of reliability if the research is properly designed.

Thus in thinking about the reliability of composition assessment procedures, it may prove useful to G theory to study the relative effects on score consistency of those variables of interest in particular testing situations. In fact, if the variables selected are directly related to the validity of the testing procedures, they can be studied from both consistency and validity points of view. For instance, the ratings and topics variables examined in this paper have also been studied from a validity perspective. In Brown, Hilgers, and Marsella (1991), it was found that topics do affect the validity of the test, but more importantly that the topics can be manipulated (on the basis of previous test results) in order to improve the validity of the procedures so that they are more fairly measuring that which they were originally designed to measure. Once more, reliability and validity seem to be inextricably interdependent.

5. How dependable are the test scores when they are used in the placement decision processes?

The phi(lambda) estimates in this study indicated that the dependability of the MWPE scores at the three decision points was fairly high, .94, .81, and .85, for the ENG22, ENG101, and ENG100 decisions, respectively. The signal-to-noise ratios suggested roughly the same thing, but presented the information in a form that is easier for some people to interpret. Thus from a decision consistency perspective, the MWPE appears to be fairly sound. In other words, a fair amount of confidence can be placed in our decision-making processes. Such confidence is particularly justified since two factors (which would both tend to increase reliability) are not accounted for in these results: (a) additional raters are regularly used in those cases where ratings diverge by two or more points, and (b) additional information is gathered for those students who fall within one SEM of a cut-point.

Thus in thinking about the reliability of composition assessment procedures, it is useful to consider the dependability of the decisions that will result from the scores. However, it is also important to realize that this decision dependability is integrally related to the validity of the procedures. It stands to reason that, if the purpose of a test is to accurately place students, and if the accuracy of decisions at various cut-points is estimated by decision dependability, then decision dependability is directly related to the

accuracy with which the test achieves its purpose, or directly related to the test's validity. Thus, reliability and validity have once again been found to be interdependent.

CONCLUSION

As mentioned above, Hoetker (1982) argued that there is too much emphasis placed on reliability at the expense of validity. It turns out that he was neither the first nor the last to take this stance. Coffman (1976) felt similarly that "Most of the research on essay testing already completed has dealt in one way or another with the question of reliability. The research of the future must devote more attention to the question of validity." (p. 298). Ruth and Murphy (1988) pointed out that "Early on in the scientific testing movement there developed an obsession with checking the reliability of essay tests" (p. 42). This view of reliability reached the point that Huot (1990) stated:

> The emphasis on reliability fueled some assumptions and confusion about the concept of validity. The emphasis on reliability and the neglect of validity are probably the major reasons why holistic scoring is now in a vulnerable position. (p. 202)

In short, a number of members of the research community in composition assessment clearly believed that too much emphasis was being placed on reliability, while validity has receiving too little consideration.

There are a number of problems with this stance, as follows:

1. The reasons for what Huot called the "inflated position of reliability" and "neglected status of validity" were numerous and may have had little to do with reliability per se, but rather with the knowledge and background of the researchers involved.
2. Reliability may be prominent in the literature because it is easier to demonstrate than is validity. The study of reliability produces nice neat coefficients, usually interrater correlation coefficients, that are directly interpretable, whereas the validity of any test is relatively difficult to demonstrate (usually requiring more elaborate statistical analyses), and should be done from several points of view.
3. The literature on essay assessment has not generally tapped the rich variety of information that can be reaped from thorough analysis of test reliability. In

general, many of the sources of information about reliability that were used in the present study [examination of score distributions over time, adjusted and average interrater estimates, Ebel's K-R20, the SEM, G theory, phi(lambda), and signal/noise ratios] have been ignored. In other words, the literature on writing assessment has only begun to scratch the surface in terms of the types of practical information that thorough study of reliability can provide for decision makers.

4. Reliability can never be taken for granted for a given test. Reliability, after all, is the degree to which a set of scores is consistent, not a condition inherent in the materials, administration procedures, or scoring strategies used. Even within a particular institution, if the students being tested change in level or become more homogeneous in abilities for some reason, the scores produced may become very unreliable. Hence, reliability should be studied every time a test is administered, or at least on a periodic basis.
5. It is necessary, whether researchers like it of not, to study the reliability of any measures used for composition assessment research (even if the focus is validity) because the results of any study can only be as reliable as the scores upon which it is based (for more on this issue, see Brown, 1988a, pp. 35-36,).
6. As argued above, reliability is a necessary precondition for validity. In other words, a set of scores cannot be any more valid than they are reliable because the degree to which a test is systematically measuring what it was designed to measure cannot be higher that the degree to which it is itself systematic.
7. Reliability and validity are inextricably interrelated and do not make much sense when considered separately.

In short, as White (1985) long ago stated "...unfair and inconsistent scores are meaningless, and meaningless scores, however cheaply obtained, are not worth anything at all" (p. 22).

The point is neither that the study of reliability should be abandoned nor that it has been overemphasized, but rather that the study of reliability should be continued, and increased in breadth and depth so that further investigation of validity can be built on a sound foundation. Indeed, in what should perhaps be the third stage of writing

assessment research, a balance should indeed be maintained between reliability and validity (as suggested by Yancy, 1999), but in the process the traditional definitions given for reliability should be broadened to include the classical theory and generalizability notions discussed in the present paper. In addition, definitions of validity should be expanded to include more of the notions listed in the most recent *Standards for the Assessment of Reading and Writing* (IRA/NCTE, 1994) and *Standards for Educational and Psychological Testing 1999* (AERA, 2000), as well as those discussed in Huot (2002) and Wilson (2006). Perhaps more importantly, new views of writing assessment validity should include the ideas of *values implications* and *consequential validity* of *score interpretations* suggested by Messick (1989a, 1989b, 1996, and elsewhere).

In short, the focus and scope of research on both the reliability and the validity of essay score interpretations should be expanded to encompass the variety of additional perspectives available to testers in education, psychology, second language testing, and other fields. Clearly, much more work needs to be done in the area of first language writing assessment.

REFERENCES

AERA. (2000). *Standards for educational and psychological testing 1999*. Washington, DC: American Educational Research Association.

Berk, R. A. (1984). Selecting the index of reliability. In R. L. Brennan (Ed.) *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins.

Brennan, R. L. (1980). Applications of generalizability theory. In R. A. Berk (Ed.) *Criterion-referenced measurement: The state of the art* (pp. 186-232). Baltimore: Johns Hopkins.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City: ACT Publications.

Brennan, R. L. (1984). Estimating the dependability of scores. In R. L. Brennan (Ed.) *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brown, J.D. (1988a). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University.

Brown, J. D. (1988b). *1987 Manoa Writing Placement Examination*. Manoa Writing Board Technical Report #1. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1989). *1988 Manoa Writing Placement Examination*. Manoa Writing Board Technical Report #2. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1990a). *1989 Manoa Writing Placement Examination*. Manoa Writing Board Technical Report #5. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1990b). Short-cut estimators of criterion-referenced test consistency. *Language Testing Journal*, *7*, 1, 77-97.

Brown, J. D. (1991). *1990 Manoa Writing Placement Examination*. Manoa Writing Board Technical Report #14. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: The amelioration of mean differences. Unpublished ms. Honolulu, HI: University of Hawaii at Manoa.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105–121.

Carlson, S., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Weiner, & R. A. Donovan (Eds.) *Writing assessment: Issues and strategies*. London: Longman.

Chiu, C. W.-T. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. Boston: Kluwer Academic.

Coffman, W. E. (1976). Essay examinations. In R.L. Thorndike *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Diederich, P. B. (1974). *Measuring growth in English*. Champaign, IL: NCTE.

Ebel, R. L. (1979). *Essentials of Educational Measurement* (3rd ed.) Englewood Cliffs, NJ: Prentice-Hall.

Faigley, L., Cherry, R. D., Jolliffe, D.A., & Skinner, A.M. (1985). *Assessing writer's knowledge and process of composing*. Norwood, NJ: Ablex.

Godshalk, F., Swineford, E., & Coffman, W. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*, 354-367.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication*, *33*, 77-392.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*, 201-213.

Hout, B. (2002). *(Re) Articulating writing assessment for teaching and learning*. Logan, UT: Utah State University.

IRA/NCTE (1994). *Standards for the assessment of reading and writing*. Urbana, IL: International Reading Association and National Council of Teachers of English.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18*, 229-249.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied measurement in*

education, 2(3), 195-205.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*, 4–26.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.

Messick, S. (1989b) *Validity.* In Linn, R. L. (Ed.), *Educational measurement (3rd ed.)* (pp. 12 – 103). London: Collier Macmillan

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-256.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating argumentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*, 143-164.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Shale, D, (2004). Essay reliability: Form and meaning. In W. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76-96). New York: The Modern Language Association of America.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133-166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239–261

White, E. M. (1985). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance*. San Francisco, CA: Jossey-Bass.

White, E. M. (1989). *Developing successful college writing programs*. San Francisco, CA: Jossey-Bass.

White, E. M. (1990). Language and reality in writing assessment. *College Composition*

and Composition, *41*, 187-200.

Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.

Yancy, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*, 483-503.