Florida Journal of Educational Research
Fall 1986, Vol. 28, No. 1, Pp. 41-62

Essay Topic Difficulty in Relation to Scoring Models

Patricia Dovell
Dianne Buhr
University of Florida

ABSTRACT. This study examined the difficulty level of essay topics used in large-scale writing assessment in relation to three different scoring models and sought to identify the effects each model would have on passing rates. Models of direct assessment and combined scores produced essentially equivalent pass/fail decisions. A regression model was found to be inappropriate for making decisions about individual students because of the essay scores' discrete scale. Results showed that each model has its advantages depending upon the purpose of the assessment and the nature of the data.

The difficulty level of essay topics used in large-scale assessment of writing is a concern of the testing and measurement profession. In order to make consistent and accurate decisions about examinees' performance over different administrations of an essay examination, scores must be comparable though topics may vary in difficulty.

The objectives of this study were to: (a) define difficulty level, (b) describe three commonly used scoring models, (c) investigate the effects of these scoring models on the pass/fail rates of students, (d) compare and discuss the results obtained using each scoring model, and (e) propose areas for future research.

Topic Difficulty

The literature on essay topic difficulty is sparse.

As pointed out by Breland (1983, p. 19), "The history of direct writing skill assessment is dominated by the issue of reliability." The issue of topic difficulty nevertheless is discussed indirectly by Hoetker (1982) who describes the factors that affect the difficulty of the essay topic and explains why the importance of this subject is now emerging. Although national testing services developed essay topics to measure achievement, their goal was to spread out the examinee scores across a continuum. However, writing competency tests seek to distinguish between two groups of writers--those who are judged to be competent writers and those who are judged to lack competent writing skills. When dealing with competency tests, the essay topic becomes much more crucial because of the need to make accurate distinctions. He illustrates this point with an actual situation in California (p. 381):

> "To cite just one example: students taking the California State University and Colleges Equivalency Examination write a 90-minute essay on a set topic. The topics for the 1974 examination were provided by a committee of English professors, with the details 'explicitly left to the discretion of the Committee on English.' When scores on the 1974 examinations were found to differ drastically from those on the 1973 examinations, the topics were reviewed. It was concluded that the 1974 topic, which called for highly abstract reasoning, was manifestly more difficult than the previous year's topic, which called for reflection upon a personal experience. The report of the affair concludes, It is the intention of the directors.. to give more thought, more attention, and more money to the development of essay questions."

Generally the difficulty level of an essay topic is assumed to be represented by the mean score on responses to that topic. This definition has several drawbacks. First, the mean does not reflect just the difficulty of the topic. In addition to measurement error, the mean is a function of the sample of examinees, essay topics, and readers, plus the possible

interaction of readers with topics and with examinees. Second, given the testing situation in which the student chooses to write on one of two topics, the mean is affected by the students who chose to write on that topic. Rosenbaum (1985, p. 14) describes this problem in his discussion of the College Board's 1982 Advanced Placement Examination in Biology:

> The students selecting one essay rather than another may differ systematically, as was the case with the essay pair consisting of Essay #5 and Essay #6. In particular, the mean number of the multiple choice items answered correctly was 70.66 for the 4129 examinees selecting Essay #5 and 65.23 for the 11547 examinees selecting Essay #6, with a two-sample t-statistic of 16.7.

The idea is that students should not benefit if they choose to write on a topic selected by the least able examinees, and they should not be penalized if they select the topic also selected by the most able examinees. Meyer (1939) observed that students could not select the essay topics on which they would score best. He had students write essays in response to five questions and asked them to identify one of their five responses to be omitted from scoring. He found that (p. 164) "...forty percent [of one group] and ...forty-six percent [of a second group] made better scores when the answer which would have been omitted was counted in place of the poorest of their four choices."

A third problem with defining difficulty level as the mean is that its numerical value is not directly interpretable. Given a scale of 2 through 8, is an essay topic with a mean of 5.1 more difficult than a topic with a mean of 5.3? To answer this question, the standard error of measurement of the mean needs to be calculated and confidence intervals identified to determine if the differences between the means are statistically significant.

Another problem with defining difficulty level as the mean is that students' scores generally cluster around the mean. Hoetker (1982, pp. 378-379) explains the statistical limitations of these data:

> "This 'piling up' of holistic ratings is probably unavoidable, short of major changes in the conventions of the process of scoring (such as, for example, directing raters to assign a predetermined percentage of essays to each quality category). This piling up has two important consequences which researchers generally have not recognized. First, there is typically so little variance among ratings that any study using them is almost guaranteed to yield results of 'no difference.' Second, the fact that holistic ratings pile up around the mean suggests that in most cases parametic statistics, such as t-tests and analysis of variance, are inappropriate, since such methods assume an underlying normal distribution of scores."

In summary, essay topic difficulty level is not directly defined in the literature. It is generally assumed to equal the mean score of the item responses, but there are several problems with this assumption.

Scoring Models

The classic model involves direct assessment. This model is politically and academically popular because it is easily understood and has common sense appeal. Diederich (1974, p. 1) explained the popularity of the direct assessment model with the analogy that "Whenever we want to find out whether young people can swim, we have them jump into a pool and swim."

In the classic or direct assessment model, examinees write an essay or multiple essays, their essays are read (holistically or analytically), and assigned a score within a range, such as 1 to 4. The examinee's score is the direct result of the score assigned by the reader or the sum of the scores assigned by multiple readers Meredith (1984, p. 14) explains how this works:

> "Basically, student scores are derived by one of two procedures: (1) summing or averaging the values assigned by two raters and resolving differences if the ratings are on different sides of the cut point or (2) summing or

averaging the ratings of two or three raters and resolving any discrepancies greater than one point."

To increase the reliability of the classical scoring model, the literature strongly favors three conditions: (1) all examinees write on the same topic; (2) multiple writing samples be collected from each examinee; and (3) multiple readers read each writing sample. Other conditions such as adjudicating split ratings, training readers, and conducting readings with readers assembled on-site in one location are often recommended.

In spite of these recommendations, many large-scale assessments do not follow these conditions. Why not? First, it is politically popular to present examinees with a choice of topics. Secondly, for reasons of test security, new topics must be introduced for each administration. Third, collecting and scoring multiple writing samples is both time-consuming and costly. Multiple-readers are generally used, although the number of readings per writing sample varies across programs.

How does the classical model control for the variance in difficulty level? Testing programs control for it by pilot testing prompts. Williams (1984, p. 8) recommends this method when he writes, "One [way] might be to simply field test as many prompts as possible and then only use those operationally that produce similar score distributions." Freedman and Robinson (1982, p. 396) reiterate this position when they outline criteria for essay topics, "...most important, topics must discriminate accurately between good and poor writers. Our only measure of how well the test discriminates comes from our pretesting." Using the results of the pretest, this model pre-equates the difficulty level of essay topics by rejecting topics that do not elicit suitable and comparable score distributions.

A second model, called the composite score model, reports a composite writing score which combines a score from a direct assessment (writing sample) with a score from an indirect assessment (multiple choice format). How this model works is described by the College Board in its explanation of the scoring for

the General Examination in English Composition (1981, p. 6):

> "...Each essay is read and graded by two professors, the sum of the two grades is combined with your multiple-choice score, and the result is reported as a scaled score between 200 and 800.
>
> CLEP does not report separate scores for the multiple-choice section and essay section of the General Examination in English Composition for two reasons. First, neither of the two sections alone is sufficient to assess reliably a candidate's writing skills. Second, although the format of the two sections is very different, both the multiple-choice format and the essay format measure essentially the same type of ability--that is, expository writing skills."

The composite score model is also used in the Advanced Placement Program (APP) Examination in English. In this program, "The candidates' composite score--multiple-choice plus essay--is reported as a grade of 1 (low) to 5 (high), with credit recommended for grades of 3, 4, and 5." (College Entrance Examination Board, 1977, p. 2).

In order for the composite score model to be valid, it must be demonstrated that both the direct assessment and the objective assessment measure the same skills. When this is the case, the composite score is often preferred over the direct assessment score because two types of measures combined have been shown to be more reliable than one (Godshalk, Swineford, & Coffman, 1966).

Tied to the validity concern is the task of equating essay topics. Breland (1983, p. 19-20) states the problem and the usual solution:

> "A validity issue for which no evidence was found is that related to the equating of essay assessments. Since topics and specific tasks vary in difficulty, and since each administration of a test must necessarily change the topic for security purposes, a not inconsequential problem is how best to equate a score received in one administration with a score

> received in another. This problem is usually handled through a combined essay and objective assessment in which the equating is performed on the combined score using an objective measure. However, if an essay assessment were used in isolation, it is not immediately apparent how equating across administrations could be achieved."

This discussion seems to suggest that the variance in essay topic difficulty is diminished through the composite-score scoring model because the essay score is only one component of the score, and the procedures for equating objective measures are well established.

A third scoring model, called the regression model, adjusts the reported score through regression. For the use of regression to be valid, it is necessary to have a highly-correlated covariate. In the case of an essay examination administered in conjunction with an objective writing test, the objective measure serves as the covariate. The regression model can take one of two forms: the essay score or a combined score can be adjusted using the objective measure to compute the regression coefficient.

The procedure underlying the regression model is explained by Hills (1972, p. 139) in the following example:

> "Consider giving Form A of a final examination to the Fall Quarter class, and Form B to the Winter Quarter class. If one has included in each form a set of items, call them Part X, so that a score on those items can be obtained for each student in both classes, one makes the assumption that the regression of final examination score on X score is the same in both groups. He then computes the regression coefficient in the group to which he wishes to equate."

Although Hills does not address the scoring of essays directly, he does give multiple examples of common material that may be used as an "anchor variable to equate the mean and standard deviation" (p. 140). His examples include "a common final examination, common admissions or academic-potential variables, or anchor sections of common material in examinations" (p. 145).

The regression model, unlike the classical and composite score models, recognizes variance in essay topic difficulty and equates scores through the mathematical manipulation of a second measure. In contrast, the other models handle variance in essay topic difficulty "on the front-end" by pilot testing and reducing the pool of acceptable essay topics to those that produce comparable and desirable scoring distributions.

The regression model incorporates equating methods with the goal of increasing the reliability of scores. However, Meredith and Williams (1984, p. 15), point to a problem inherent in the narrow range of scores assigned during holistic scoring:

> "The test score equating analog for direct writing assessment is not quite as clear as it is in indirect assessment. The applicability of equating direct assessments is compromised by the discreteness of the score scale. In programs where two raters evaluate student responses on one prompt, the score scale may only range from one to four at half-point intervals. A scale of this type may not provide a sufficient number of score points to use in the equating process."

The results of equating for essay topic difficulty when dealing with a narrow range of raw score data will be examined later when the regression model is applied to the data source.

Although two other scoring models, the Rasch model and generalizability models, are recommended for detecting and accommodating differences in topic difficulty and rater variance, the testing program used in this study did not have appropriate data for investigating them. To be specific, these models require that each examinee writes on both topics.

Methods

The classical scoring model sums the raw scores assigned by readers to create a scaled score. It is the model used in the testing program studied. Therefore, the particulars of this method are described later in the data source section of this

paper.

The composite score model reports a composite writing score which combines a score from a direct assessment (writing sample) with a score from an indirect assessment (multiple - choice format). Combined scores for the data were calculated using the following procedures.

First, the score from the direct assessment was converted to the same scale as the indirect assessment by using the linear conversion of 30X + 300. The values of 30 and 300 were used because 300 is the approximate mean and 30 is the approximate standard deviation of the indirect assessment. A 30 point difference between scores was selected because this requires an examinee who fails one subtest to score at least one standard deviation above the mean on the other subtest in order to pass by the combined score.

Second, the scale was shifted so that a raw score of 5 on the direct assessment would be equal to a scale score of 300. This was done so that the mean of the direct assessment (approximately 5) would be converted to the mean of the indirect assessment (approximately 300). The linear conversion of the shifted scale resulted in the following:

Essay Score	Converted Essay Score
2	210
3	240
4	270
5	300
6	330
7	360
8	390

Third, the passing score for the composite scoring model was determined by adding the passing score required on the indirect assessment to the converted passing score on the direct assessment. A passing score of 4 is required on the essay, which is converted to 270. A passing score of 265 is required on the indirect assessment. Thus the passing score required for our composite scoring model is 535.

The regression scoring model adjusts scores to

account for variability in essay topic difficulty. This model equates scores with the use of regression on another measure, in our case multiple - choice writing items. Essay scores within and between administrations for three test administrations were equated by means of a procedure adopted by Hills (1972) from a method suggested by Gulliksen (1950). This procedure uses a common examination to place scores from different essay topics on the same scale. This is done through the regression of essay scores for one topic on the score of the common examination. In this case the common examination was the multiple-choice writing subtest. This procedure assumes that the regression of the essay score on the multiple-choice writing score is the same for two groups taking different topics.

Topic 1 of the March 1983 assessment was arbitrarily selected as the base value, and the regression coefficient of Topic 1 scores on multiple-choice writing scores was obtained. This coefficient (.022398) was used to adjust the mean and the variance for Topic 2 scores, according to the following procedure:

$$\bar{X}_{E2A} = \bar{X}_{E1} + b_1 \, (\bar{X}_{W2} - \bar{X}_{W1})$$

where:
- $\bar{X}_{E2A}$ is the adjusted mean for Topic 2 expressed in terms of Topic 1
- $\bar{X}_{E1}$ is the mean for Topic 1
- b_1 is the regression coefficient for Topic 1 scores on the multiple-choice writing scores
- $\bar{X}_{W1}$ is the mean on the multiple-choice writing subtest for examinees choosing Topic 1
- $\bar{X}_{W2}$ is the mean on the multiple-choice writing subtest for examinees choosing Topic 2

$$o^2_{E2A} = o^2_{E1} + b_1^2 \, (o^2_{W2} - o^2_{W1})$$

where: o^2_{E2A} is the adjusted variance for Topic 2 scores expressed in terms to Topic 1

o^2_{E1} is the variance for scores on Topic 1

b_1^2 is the squared regression coefficient of Topic 1 scores on the multiple-choice writing subtest scores

o^2_{W2} is the variance of the multiple-choice writing subtest scores for Topic 2 examinees

o^2_{W1} is the variance of the multiple-choice writing subtest scores for Topic 1 examinees

The same procedure was followed in adjusting means and variances of essay topic scores obtained from two other 1983 test administrations. The adjusted means and variances were then used to obtain rescaled scores for individuals as follows:

$$X_{E2A} = \overline{X}_{E2A} + \frac{o_{E2A}}{o_{E2}} \, (X_{E2i} - \overline{X}_{E2})$$

where: $\underline{X}_{E2A}$ is the rescaled score for examinee i

$\overline{X}_{E2A}$ is the adjusted mean for Topic 2 scores

o_{E2A} is the adjusted standard deviation for Topic 2 scores

o_{E2} is the standard deviation of Topic 2 scores

$\underline{X}_{E2i}$ is the score for examinee i on Topic 2

$\overline{X}_{E2}$ is the mean score for Topic 2

The rescaled scores were then used to determine the percentage of examinees passing at the cutting score of 4 for each topic for the three test administrations.

In addition, the Pearson product-moment correlation between the essay scores and multiple-choice writing scores was calculated. This step was performed because a highly-correlated covariate is a necessary condition of the regression scoring model.

Data Source

Data from a large-scale writing assessment were studied. The examination is administered three times a year (March, June, and September or October) to college students at the end of their sophomore year. Each student has 50 minutes to write an essay on one of two topics, and the scoring criteria are provided to them. Topics are changed with each administration. The scoring method is a modified holistic scoring, in which scores are assigned with reference to range-finders. The readers are assembled on-site in one, two, or three locations. Readers are trained before and during the reading. Scores of one (low) through four (high) are assigned. Each essay is read twice. The reported score is the sum of the two ratings (2-8). However, split ratings (i.e. 1-3, 2-4, and 1-4) are refereed, and one of the two scores is replaced. Beginning with the Fall 1984 administration of the test, a passing score of 4 was established. Starting with the same administration, total scores of 3 were also refereed. This policy was established to correct the situation in which an essay was failed by one rater (score of 1) and passed by another (score of 2).

In addition, a writing subtest, objectively scored, is included in the assessment. Scores range from approximately 200 to 400, with a mean of approximately 300 and standard deviation of approximately 30. A score of 265 is required to pass.

Table 1 presents descriptive data for six test administrations. The mean, standard deviation, alpha, and number of examinees are listed by topic for the essay subtest, along with comparable data for the objective writing subtest.

Results

The results of applying the direct assessment or classical scoring model to the data are reported in Table 1. Within administrations, the largest difference between means for the two topics was .50 for March 1983. Across administrations, differences were larger, with the greatest difference .65, between

TABLE 1

Descriptive Statistics for Essay and Writing Subtests by Topic

		Essay Mean	Essay Std Dev.	Alpha	N	Writing Mean	Writing Std Dev.	N
Mar. 83	Topic 1	4.79	1.39	.760	15,583	307.37	30.76	15,577
	Topic 2	5.29	1.65	.816	3,461	306.86	31.18	3,460
June 83	Topic 1	4.88	1.52	.799	2,952	306.53	30.51	2,980
	Topic 2	4.69	1.45	.787	7,383	302.34	30.59	7,376
Oct. 83	Topic 1	4.64	1.34	.759	9,546	305.77	31.51	9,537
	Topic 2	4.88	1.49	.791	4,763	310.40	33.65	4,762
Mar. 84	Topic 1	5.11	1.42	.805	8,024	315.59	29.64	8,021
	Topic 2	4.97	1.47	.826	10,317	312.86	29.63	10,311
June 84	Topic 1	4.85	1.68	.883	2,159	306.03	32.08	2,138
	Topic 2	5.00	1.46	.828	8,109	310.65	29.45	8,052
Sept 84	Topic 1	4.90	1.45	.814	9,314	319.28	32.89	9,031
	Topic 2	4.77	1.43	.819	4,554	320.19	33.30	4,429

TABLE 2

Descriptive Statistics for Combined Score

	Mean	Std Dev.	Number
March 83	603.68	64.59	19,037
June 83	595.87	64.91	10,356
October 83	599.01	64.66	14,299
March 84	614.98	63.82	18,334
June 84	609.11	65.71	10,190
September 84	616.59	65.83	13,462

March 1983 topic two and October 1983 topic one. This is about two-fifths of a standard deviation.

The results of applying the composite scoring model to our data appear in Tables 2 through 6. Table 2 presents the results achieved (mean, standard deviation, and N) when the composite score is calculated according to the method previously described. Table 3 contrasts the percentage of examinees passing using the direct assessment scoring model with the composite scoring model. The number and percentage of examinees passing the essay subtest, the multiple-choice writing subtest, and both of these are presented. For all six administrations there is a slight increase in the percentage who pass when the composite scoring model is used compared with the percentage who pass both subtests. Increases range from .6 percent to 4.4 percent of the examinee population. This means that examinees who were near the cutting score but failed on one subtest may be pulled up to pass on the combined score by their performance on the other subtest.

Table 4 shows the number of examinees who passed by the combined score but failed either the essay or the writing subtest. For the first three administrations, about three times as many examinees failed the essay examination as those who failed the writing subtest. However, the numbers were about equal for the later administrations.

Table 5 shows the Pearson product-moment correlation between the essay scores and the objective writing scores. The correlations for the six administrations are consistent, ranging from .49 to .51.

Table 6 presents the passing rate by topic for each of the two subtests, and for the composite score. In only one case does the composite score reduce the difference in passing rates between topics (March 1983). Where the difference is greatest between passing rates for essay topics (6 percent in June 1984), the difference for the combined score is even larger (7 percent).

The results of applying the regression scoring model to our data appear in Tables 7 through 9. Table 7 presents by topic the essay subtest's means and standard deviations adjusted by the regression equation

TABLE 3

Examinees Passing Essay and Writing Subtests

Total N		Number (%) Passing Essay	Writing	Both	Combined
19,070	March 83	15,951(83.6%)	17,948(94.1%)	15,564(81.6%)	16,297(85.5%)
10,389	June 83	8,398(80.8%)	9,607(92.5%)	8,104(78.0%)	8,534(82.1%)
14,339	Oct 83	11,809(82.4%)	13,190(92.0%)	11,398(79.5%)	12,032(83.9%)
18,352	March 84	16,665(90.8%)	17,621(96.0%)	16,416(89.5%)	16,583(90.4%)
10,270	June 84	9,153(89.1%)	9,725(94.7%)	8,908(86.7%)	8,974(87.4%)
14,431	Sept 84	12,410(86.0%)	13,126(91.0%)	11,986(83.1%)	12,105(83.9%)

TABLE 4

Examinees Passing by Combined Scores*
but Failing Essay or Writing

	Pass Both	Passed Combined Failed Other	Failed Writing	Failed Essay	Total Tested
March 83	15,564(81.6%)	733(3.8%)	138(.7%)	595(3.0%)	19,070
June 83	8,104(78.0%)	430(4.1%)	119(1.2%)	311(3.0%)	10,389
Oct 83	11,398(79.5%)	634(4.4%)	129(.9%)	505(3.5%)	14,339
March 84	16,416(89.5%)	168(.9%)	84(.5%)	83(.5%)	18,352
June 84	8,908(86.7%)	66(.6%)	50(.5%)	16(.2%)	10,220
Sept 84	11,986(83.1%)	119(.8%)	73(.5%)	46(.3%)	14,431

*No one passed combined and failed both essay and writing.

TABLE 5

Correlations of Essay and Writing Scores

	PPM Corr.	N
March 83	.49	19,037
June 83	.49	10,356
Oct 83	.51	14,299
March 84	.50	18,334
June 84	.50	10,190
Sept 84	.50	13,462

TABLE 6

Percent of Examinees Passing Based on Writing and Essay Subtest, and by a Combined Score, by Topic

		Pass Essay	Pass Writing	Pass Combined
Mar. 83	Topic 1	83.5	94.2	85.4
	Topic 2	85.0	94.0	86.4
June 83	Topic 1	82.3	93.6	84.0
	Topic 2	80.5	92.1	81.6
Oct. 83	Topic 1	82.3	91.8	83.7
	Topic 2	83.0	92.6	84.9
Mar. 84	Topic 1	92.1	96.6	91.7
	Topic 2	89.9	95.6	89.4
June 84	Topic 1	84.3	91.4	81.9
	Topic 2	90.4	95.6	88.9
Sept 84	Topic 1	89.9	93.8	87.7
	Topic 2	88.7	94.0	86.5

TABLE 7

Adjusted Means and Standard Deviations

	ESSAY				WRITING		
	Mean	Std Dev.	Adjusted Mean	Adjusted Std Dev.	Mean	Std Dev.	N
March 1983							
Topic One	4.789	1.390	-	-	307.371	30.755	15,577
Topic Two	5.292	1.645	4.777	1.395	306.859	31.184	3,460
June 1983							
Topic One	4.882	1.521	4.769	1.388	306.531	30.505	2,982
Topic Two	4.686	1.450	4.676	1.388	302.344	30.587	7,383
Oct. 1983							
Topic One	4.644	1.343	4.753	1.399	305.767	31.505	9,546
Topic Two	4.882	1.494	4.856	1.424	310.395	33.653	4,763

TABLE 8

Rescaled Scores for Examinees

Essay Score:	2	3	4	5	6	7	8
March 1983							
Topic One	2	3	4	5	6	7	8
Topic Two	1.985	2.833	3.681	4.529	5.377	6.225	7.073
June 1983							
Topic One	2.141	3.053	3.965	4.877	5.790	6.702	7.614
Topic Two	2.104	3.061	4.019	4.977	5.934	6.892	7.849
Oct. 1983							
Topic One	1.999	3.040	4.082	5.123	6.165	7.206	8.248
Topic Two	2.110	3.063	4.016	4.969	5.922	6.874	7.827

TABLE 9

Percent of Examinees Passing with Rescaled Scores at a Cutting Score of 4

	PERCENT PASSING		
March 1983	Without Equating	With Equating	N
Topic One	.84	.84	15,577
Topic Two	.85	.69	3,460
June 1983			
Topic One	.82	.60	2,982
Topic Two	.80	.80	7,383
Oct. 1983			
Topic One	.82	.82	9,546
Topic Two	.83	.83	4,763

presented in the "Methods" section of this paper. Adjusted means differ by at most .18. Table 8 presents by topic the rescaled essay scores for examinees. Table 9 compares the percent of examinees passing with and without the regression equation rescaling. There is a dramatic decrease in the percent passing for the Topic 2, March 1983 group (17 percent decrease) and for the Topic 1, June 1983 group (22 percent decrease). This is because the regression equation shifted scores of 4 (the passing score) to scores of 3.681 and 3.965 respectively for those two groups. Conversely, no one who failed on the direct assessment scoring model passed with rescaled scores. This is because the increase in rescaled scores is insufficient to move examinees from one level to the next higher level.

Discussion and Conclusions

The direct assessment or classical model is politically and academically popular. Many educators believe that testing writing ability with a writing sample will encourage the practice of writing in the curriculum. In addition, the ability of multiple-choice tests to measure writing ability is still regarded with suspicion. In fact, many multiple-choice writing items do focus on skills at the sentence level, and rarely transcend the paragraph level. This model appeals to our notion of common sense and is therefore trusted. The obvious disadvantages are the cost and time required to score the essays. In addition, the reliability of the scores is a potential problem, although there is literature available that describes measures to increase reliability.

The direct assessment and composite models produce similar results. The majority of examinees pass both subtests using the direct assessment model and pass the combined score using the composite model. A few examinees who failed one of the two subtests passed by the combined score. This occurred when an above-minimum performance on one subtest pulled up their combined score. Conversely, a few examinees who passed one subtest failed by the combined score. This occurred when their performance on one subtest was

insufficiently strong to compensate for a weak performance on the other subtest.

Since these two models produce similar results, it is legitimate to ask, "Which scoring model should be used?" or "Does it matter which scoring model is used?" The answer to these questions should not depend upon the pass/fail results produced by the model. The answers should depend upon the nature of the construct being measured. If it is believed that there is one construct, writing ability, then it is appropriate to measure it with multiple measures and report a composite score. If there is more than one construct being measured, then it is inappropriate to report only a composite score.

The purpose of the assessment is another factor to be considered when choosing between these models. If the purpose of the assessment is to assure that students have acquired the writing skills necessary to compose an essay, then assigning a separate score to the writing sample is the most direct scoring model to use. If, however, mastery of a content area is being measured, then a composite score is easier to interpret and more inclusive.

If the composite scoring model is selected, and one therefore assumes that one construct is being measured, it is reasonable to expect a strong correlation between the two measures. This expectation was checked (see Table 5) by calculating the Pearson product-moment correlation between the essay subtest and the multiple-choice writing subtest. A cursory glance at Table 5 might lead to the conclusion that these correlations are not impressive. However, it is important to remember that when the population is homogeneous the correlations decrease. This population, college students near the end of their sophomore year, is sufficiently homogeneous to explain this effect. Therefore, the correlations had sufficient magnitude to justify the composite scoring model if one assumes that one construct is being measured by both subtests.

The regression scoring model did not work well with these data because the essay ratings are discrete, restricted data points. It is difficult to find a regression line of best fit because there are

restricted values on each axis. In this case, a 4-paper, that passed, was inappropriately shifted to a 3.96 paper, that failed.

However, if the transformed essay scale is considered as a continuous scale, then 3.5 or above could be considered a passing score. In this case, pass/fail decisions would be the same for the converted scores as for the original essay scores.

Two conditions should be present in order to apply the regression model. First, an adequate number of score points are needed to find the line of best fit. We concluded that four discrete score points are inadequate. Secondly, the purpose of the assessment should be to spread out examinee scores across a continuum. This will allow for minor corrections for essay topic difficulty which will change the rank-order of examinees, but not dramatically reverse individual pass/fail decisions.

Further Research

Because important educational decisions are made based upon the scores from large - scale writing assessments, the need for more research in this area is self-evident. Two areas of inquiry are proposed:

1. A generalizability (G) study using analysis of variance to identify the variance due to topic and raters could contribute directly to an improved analysis of the field-test results of prompts and to improvement in rater reliability.

2. A study using item response theory with polychotomous scores could facilitate the development of a bank of essay topics with varying difficulty. More able students could be challenged by more difficult topics, whereas less able students could develop their skills using easier topics. This kind of research could lead to computer-generated essay topics to provide tailored testing.

In order for this kind of research to achieve optimal results, it is important for educational researchers and writing faculty to work together so that their efforts are harmonious.

References

Breland, H. M. (1983). The direct assessment of writing skill: a measurement review (College Board Report No. 83 - 6). Princeton, N. J.: College Entrance Examination Board.

College Entrance Examination Board (1981). College-level examination program: what your scores mean. New York: Author.

College Entrance Examination Board (1977). Guide to CLEP examinations. New York: Author.

College Entrance Examination Board (1977). Guide to examinations in literature. New York: Author.

Diederich, P. B. (1974). Measuring growth in English. Urbana, Ill.: National Council of Teachers of English.

Freedman, S. W., & Robinson, W. S. (1982). Testing proficiency in writing at San Francisco State University. College Composition and Communication, 37, 393-398.

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

Hills, J. R. (1972). Consistent college grading standards through equating. Educational and Psychological Measurement, 32, 137-146.

Hoetker, J. (1982). Essay examination topics and students' writing. College Composition and Communication, 37, 377-392.

Meredith, V. H., & Williams, P. L. (1984). Issues in direct writing assessment: problem identification and control. Educational Measurement: Issues and Practice, 3 (1), 11-15; 35.

Rosenbaum, P. R. (1985). Model-based direct adjustment. Princeton, N. J.: Educational Testing Service.

Williams, P. L. The application of direct writing assessments in five states. Educational Measurement: Issues and Practice, 1984, 3 (1), 19.

AUTHORS
PATRICIA DOVELL, Coordinator for Test Development, Office of Instructional Resources, 1012 GPA Bldg., University of Florida, Gainesville, FL 32611
DIANNE BUHR, Assistant Testing and Evaluation Director, Office of Instructional Resources, 11012 GPA Bldg., University of Florida, Gainesville, FL 32611