DOCUMENT RESUME

ED 242 771 TM 840 188

AUTHOR Koffler, Stephen L.
TITLE Assessing Students' Writing Skills: A Comparison of Direct & Indirect Methods.
PUB DATE Apr 84
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Essay Tests; Evaluation Methods; Graduation Requirements; High Schools; High School Students; *Minimum Competency Testing; *Multiple Choice Tests; Scoring; State Programs; *Test Construction; Testing Programs; Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS *Direct Assessment; New Jersey State Department of Education

ABSTRACT

This research examined the results from direct and indirect writing assessments to determine the most effective method of discrimination. The New Jersey State Department of Education developed a test for ninth-grade students which was designed to measure the ability to apply writing mechanics to written text and to communicate effectively in writing. The instrument combined direct and indirect assessment in a 54-item multiple choice section and a 30-minute essay. This minimum competency test measured minimum writing skills. Essays were holistically scored. Direct writing assessment requires writing samples by examinees to be read and scored by examiners. Indirect assessment requires examinees to respond to items which measure correlates of writing. Both methods are reliable assessments. In states which mandate that students pass a writing test as part of the requirements for receiving a high school diploma, the important criterion is which form of assessment discriminates best between competent and incompetent writers. Results of statistical analysis indicated the indirect assessment provides a better means of discrimination between competent and incompetent writers. However, a combination of both methods, creating a weighted total test score, is considered the most appropriate method. (DWH)

Assessing Students' Writing Skills: A Comparison of Direct & Indirect Methods

Stephen L. Koffler
New Jersey State Department of Education

Paper presented at the annual conference of the American Educational Research Association
New Orleans, La.
April, 1984

Assessing Students' Writing Skills: A Comparison of Direct & Indirect Methods

Stephen L. Koffler
New Jersey State Department of Education

The assessment of students' writing skills has become a focal point for statewide testing programs during the past few years. As a result, administrators of large-scale testing programs are grappling with a variety of new problems which must be resolved and which were not encountered when reading and mathematics programs were developed.

Of primary concern is "how do we measure students' writing skills?" There are two distinct methods --direct assessment requires that writing samples be written by examinees and read and scored by readers. Indirect assessment requires the examinee to respond to items which measure the correlates of writing, usually in a multiple choice format, rather than doing actual writing.

Certainly, there are advantages and disadvantages to each method. Noyes, Sale and Stalnaker (1945) argued for indirect assessment:

> the good candidate who errs on a few of these [multiple choice items] has plenty of opportunity to redeem himself; a mistake on one item does not affect any other item. In writing a theme, however, the candidate who makes a false start almost inevitably involves his whole theme in difficulties even though he may be, generally speaking, a good writer.

Breland(1977) espoused the case for direct assessment:

> Clearly, writing involves much more than constructing sentences. Writing requires a sense for organization of sentences and paragraphs, the proper use of supporting detail and the ability to distinguish fact from opinion--among other things. Most multiple choice tests make no ostensible attempt to measure these other important aspects of writing. Therefore, how can multiple choice tests be of much value at all in assessing

writing ability? Beyond these limitations with respect to writing ability, it is believed that some persons simply do not respond well to the multiple choice mode of testing.

Diederich(1974) argued that requiring students to produce actual samples of writing, especially under test conditions, is the most convincing test of their writing ability. According to Spandel & Stiggins(1982), because indirect methods measure the prerequisites of effective writing -- understanding of the basic elements and conventions of standard English usage -- they represent necessary but not sufficient components of writing skills.

Spandel & Stiggins suggested that if resources and expertise were available to specify the skills to be assessed, develop the exercises, train the readers, and conduct at least two independent readings of the exercises, then direct methods would provide the most appropriate means for generating valid and reliable information about writing skills. However, Quellmalz, Capell & Chou(1982) showed that levels of performance varied on different types of writing tasks. They implied from this that writing for different purposes and audiences draws on different skills and that those skills must be measured separately. Also, numerous studies have reported low reliabilities for direct assessment methods(e.g. Akeju,1972). However, Coffman(1966) and Godshalk, Swineford and Coffman (1966) have shown that high reliability for direct methods can be obtained by requiring multiple samples and multiple readings of each sample.

Clearly, the two modes of assessing writing are satisfactory and unsatisfactory from different points of view. Stiggins(1982) noted that neither method is inherently superior. Rather the use-

fulness of each varies according to the context of the assessment and the decisions to be made. Thus, one of the first tasks which Directors of large-scale testing programs must resolve is which method(s) they should use to assess students' writing skills, especially when passing the test is a requirement for high school graduation. Because of cost and other factors, including the amount of time available, it might be unlikely that more than one essay can be written and scored for each student. The low reliability of one essay might argue for indirect methods. Yet, the face validity of indirect methods poses a problem -- will the public accept a test of writing skills which does not require students to write?

For states which mandate that students must pass a writing test as part of the requirements for being awarded a high school diploma, the most important question is which form of writing assessment is better able to discriminate between competent and incompetent writers. Because the stakes are high -- denying a high school diploma -- it is crucial that the assessment device used be a reliable means of discriminating between those who will be awarded a diploma and those who will not.

Others have examined whether direct and indirect methods measure common or different skills. Yet, little research has examined the effect of a single or combination approach on the discriminant validity of writing assessment. Will a combination approach be more effective than using solely direct or indirect methods to discriminate between masters and nonmasters? And, if so, in what combination? Because of the high school graduation

laws, such analyses become critical.

This research examined the results from direct and indirect writing assessment methods to determine the best procedure for effectively discriminating between competent and incompetent writers. To do this, it was necessary to determine whether an essay, by itself, or a multiple choice test, by itself, was an effective means of discriminating, and whether in combination they formed an even better means of discrimination.

Data Source

In March 1983 a statewide writing test was administered for the first time to approximately 94,000 ninth grade students in New Jersey public schools. That test was designed to measure students' ability to apply writing mechanics to written text and to communicate effectively in writing. In order to assess both aspects of the students' writing ability, the test consisted of a 54-item multiple choice section and a thirty minute essay.

The multiple choice section consisted of items measuring skills in five clusters: sentence structure(11 items), usage(21 items), punctuation(9 items), capitalization(7 items), and spelling(6 items). Since the test was a minimum competency test, the skills measured were minimum skills identified by educators in the state. All items were in a four-choice format. Students received a score for each cluster and for the total multiple choice section (the sum of the correctly answered items).

For the essay section, students were provided with a topic which they had not previously seen and were given 30 minutes to write no more than two pages of text. The time allotment included

a 5 minute pre-writing period and a 5 minute editing period. The topic was as follows:

> Think of something you thought was unfair. It might be something unfair that happened to you, to a friend, or to someone you didn't know. Write an essay that tells what happened and how you felt about it.

The essays were scored during a four day period by 230 N.J. teachers using a holistic scoring process. Each essay was read twice and scored on a 1-6 scale; thus, a student's essay could receive a score from 2 to 12, the sum of the two readers' scores. In cases where two readers awarded scores which differed by three or more points, the essay was read by a third reader to resolve the discrepancy. Only 1.5% of the essays needed a third reading.

Students also received a score representing a combination of the two sections, weighted 60%(essay) and 40%(multiple choice). The total score was derived using a weighted sum of the Z-scores of the two sections. Further, that total score was reported on a scale which ranged from 40-100 with a mean of 80 and a standard deviation of 10.

In addition to the students' test scores, teacher judgment data were collected about each student. Following the administration of the test, teachers were asked to provide their independent judgment about their students' writing skill competence. The following instructions were provided to the teachers (N.J. Dept. of Educ., 1983):

> To accomplish the task, you will enter two ratings for each student on page 3 of his/her answer booklet in the area marked "For Teachers Use Only." On the line labeled "S.S.1," you will enter a rating for the student relative to the skills assessed by the multiple-choice section of the test. On the line labeled "S.S.2," you will enter a rating of the student's skill level in written expression.

For each rating, describe the student in one of three categories--Master, Borderline, Non-Master--on the skills in question. Mark "M" for "Master" if, in your judgment, the student has mastered most of the skills assessed. Mark "N" for "Non-Master" if, in your judgment, the student has not mastered most of the skills assessed. Mark "B" for "Borderline" if the student has some of the skills but cannot confidently be assigned to either the Master or Non-Master group.

In performing this task, please remember the following:

(a) Judge each student relative to the kinds and levels of writing skills assessed or expected by the test. For the multiple-choice section (S.S.1), you may wish to refresh your memory as to the nature of the test questions before providing your rating. For the essay rating (S.S.2), rate the student's mastery of written expression relative to what you believe to be appropriate or adequate for the ninth grade.

(b) Do not guess about a student's skill level. If you are unfamiliar with the student, try to locate a staff member who is sufficiently knowledgeable to provide a valid rating. If you are unable to make a confident judgment of a student's mastery and are unable to identify someone who can do so, please leave the circles blank.

For purposes of this research, a 10% spaced sample of the statewide results was generated. Students who were classified in a special education or limited English speaking program were excluded from the sample. Further, the results for students who became ill or were disruptive during a section(s) were excluded from the section(s). Thus, the sample size for each of the test sections was not identical. Based on a chi-square Goodness of Fit test, it was determined that the sample was representative of the state in terms of the distribution of students in different socioeconomic types of school districts.

The sample consisted of information for 7326 students. For each student, the following data were available: multiple choice score, cluster scores, essay score, total score and the teachers'

mastery/borderline/nonmastery judgments. The data were analyzed with the 1982 version of the Statistical Analysis System.

Results

Table 1 presents summary statistics for each test section for the sample as a whole and for the master, borderline and non-master groups separately. Many more students were judged to be masters than either borderline or nonmasters. Further, as anticipated, there was a relationship between the mastery groupings and the mean test scores. Students who were judged to have a mastery of the tested skills scored higher, on the average, than students considered to be borderline. Similarly, borderline students scored higher than nonmasters. Finally, all sections of the multiple choice test were relatively easy, especially for the master group. Given the very high mean scores for all groups on the capitalization cluster, very little information can be discerned by examining that area.

Reliability data for the test sections are presented in Table 2. For the multiple choice section, the reliability coefficients are based on the Kuder-Richardson 20 formula. Since reliability is a function of test length, the KR20 estimates for some of the clusters are fairly low. The reliability estimate for the essay score is based on the interrater reliability, after the third reading for the essays which required such a reading. A pooled within-cell correlation was computed as the estimate of the inter-rater reliability (the inter-rater reliability was computed by National Evaluations Systems, Inc. as part of their contract with the N.J. Department of Education.)

Table 1

Summary Statistics For The Entire Sample
And For The Master/Borderline/Nonmaster Groups

Test	Group	N	Mean	Median	S.D.
Essay (1 item)	Master	3178	8.0	8	1.8
	Borderline	2612	6.6	7	1.9
	Nonmaster	1372	5.2	5	1.8
	Total	7162	6.9	7	2.1
Multiple Choice (54 items)	Master	3381	47.9	49	5.3
	Borderline	2521	42.9	44	6.7
	Nonmaster	1286	36.6	37	8.7
	Total	7198	44.1	46	7.7
Sentence Structure (11 items)	Master	3381	9.6	10	1.7
	Borderline	2520	8.1	8	2.3
	Nonmaster	1275	6.4	6	2.6
	Total	7176	8.4	9	2.6
Usage (21 items)	Master	3382	18.2	19	2.7
	Borderline	2524	15.9	16	3.3
	Nonmaster	1287	13.2	14	4.1
	Total	7193	16.3	17	3.9
Punct- uation (9 items)	Master	3382	8.3	9	1.0
	Borderline	2523	7.7	8	1.3
	Nonmaster	1286	6.8	7	1.7
	Total	7191	7.7	8	1.5
Capital- ization (7 items)	Master	3382	6.8	7	0.6
	Borderline	2522	6.6	7	0.8
	Nonmaster	1286	5.1	6	1.1
	Total	7190	6.5	7	1.0
Spelling (6 items)	Master	3378	5.2	5	0.9
	Borderline	2519	4.8	5	1.1
	Nonmaster	1280	4.2	4	1.3
	Total	7177	4.8	5	1.2

Table 2

Reliability Coefficients for
The Direct and Indirect Methods

	Essay	M.C.	S.S.	Us.	Pu.	Ca.	Sp.
Reliability	0.69	0.90	0.76	0.581	0.57	0.57	0.48

Table 3 provices the uncorrected and corrected correlation coefficients for the different test sections. The corrected coefficients are provided to correct for attenuation due to unreliability. They were calculated by dividing the uncorrected correlation coefficient by the product of the square root of the two relevant reliability coefficients.

Table 3

Uncorrected & Corrected Correlation Coefficients[a]
Between the Direct & Indirect Methods

	Essay	M.C.	S.S.	Us.	Pu.	Ca.	So.
Essay	1.0	.60	.60	.63	.49	.37	.41
M.C.	.86	1.0					
S.S.	.80		1.0	.71	.55	.45	.41
Us.	.84		.90	1.0	.57	.49	.45
Pu.	.78		.84	.84	1.0	.51	.41
Ca.	.59		.68	.72	.89	1.0	.35
So.	.71		.68	.72	.78	.67	1.0

[a] Uncorrected correlations appear above the diagonal; corrected correlations appear below the diagonal. Coefficients are not included for the relationships between the total multiple choice test and the five clusters because of the dependence between the clusters and the total multiple choice scores.

The correlations, especially those for the relationship between the essay and the total multiple choice section and some of the clusters, are strong and not inconsistent with the findings from other studies (e.g. Breland & Gaynor, 1979; Hogan & Mishler, 1980; Moss, Cole & Khampalikit (grade 10 results),1980);

Examining the Discriminant Ability of the Test Sections

Biserial correlations were computed to compare the students' mastery designation (master v. nonmaster) to their test scores (see Table 4). Biserial correlations were used instead of the more conventional point biserial correlations because it was assumed that the underlying distribution of the mastery judgments was normally distributed (Glass & Stanley, 1970).

From Table 4, it is evident that the total multiple choice section (as well as the sentence structure cluster and the usage clusters) was a better discriminator than the essay. This result might argue for indirect methods instead of direct methods for tests whose primary purpose is for mastery decisions.

Table 4

Biserial Correlations Examining the Relationship Between Mastery/Nonmastery and Test Scores

	Essay	M.C.	S.S.	Us.	Pu.	Ca.	Sp.
Correlation	0.74	0.83	0.76	0.76	0.62	0.52	0.53

To better examine the ability of the essay and multiple choice tests to discriminate between the competent and incompetent writers, passing scores were determined and the relationship of the pass/fail rate to the mastery decisions was examined. There exist many methods for setting passing scores, some based on judgments about items and some based on judgments about examinees. Livingston & Zieky(1981) provide an excellent description of the major procedures used to set passing scores; therefore, a

discussion of the various procedures is excluded from this study.

Passing scores were set based on the Contrasting Groups procedure. The passing score set by that method is the score which best separates the distribution of students judged to be masters from the distribution of those judged to be nonmasters. To determine the passing scores for each test section, the data were analyzed using a two group univariate discriminant analysis based on the ranks of the data (Koffler, 1980). Only students considered by their teachers to be masters or nonmasters were included in the analysis. By including only these two groups, i.e. those for whom their teachers were certain of their mastery/nonmastery status, and by excluding the borderline group, the separation between the clearly competent and clearly incompetent was more discernible.

Table 5 presents the results of the Contrasting Groups analysis. That table also includes the percent of students judged by their teachers to be nonmasters who would have been classified as masters based on their test score(false masters) and the percent of students judged to be masters who would have been classified as nonmasters (false nonmasters). It also includes the total percent of students misclassified.

It should be clear that a goal in setting a passing score is to minimize the percent of students misclassified. Depending upon the costs associated with each misclassification error, one may be more concerned with minimizing either the proportion of false masters or the proportion of false nonmasters. However, for this study, and for most, because the costs were not discernible, it

was assumed that the respective costs of misclassification were equal; thus, the major concern was the total proportion of misclassified students.

Table 5

Passing Scores For the Direct & Indirect Methods

Test	# of Items	Cut Off	% False Masters	% False Nonmasters	Total Percent Misclassified
Essay	1(2-12)	6	44.7%	9.6%	20.4%
M.C.	54	43	30.7%	13.8%	18.1%
S.S.	11	8	39.0%	14.8%	21.7%
Us.	21	16	35.6%	16.5%	21.0%
Pu.	9	8	40.2%	16.8%	23.5%
Ca.	7	7	47.3%	15.1%	24.3%
Sp.	6	4	73.1%	4.3%	23.3%

To obtain an unbiased estimate of the percent of misclassified students, the discriminant analysis was conducted on a randomly drawn subsample of 80% of the original sample. Once the passing scores were determined, the remaining 20% of the sample was used to estimate the percent of false masters and false nonmasters in the population.

The data presented in Table 5 indicate that students were misclassified least often on the basis of the total multiple choice test. This result is in agreement with the biserial correlation coefficient reported in Table 4. Thus, the multiple choice test by itself is a better means of discriminating between the competent and incompetent writers than the essay is by itself.

In all cases, there were a far greater percent of nonmasters

misclassified as masters based on their test scores than masters so misclassified. This may suggest that teachers are more certain of their judgments about students who are competent rather than those who are incompetent. However, for the indirect methods, it may also reflect the relative easiness of the multiple choice test which resulted in very high passing scores (especially for some of the smaller clusters).

Considered separately, the indirect assessment method is a slightly better procedure than the direct method for discriminating between masters and nonmasters. However, an important question to address is whether the multiple choice and essay score can be used in combination to make the pass/fail decision more accurate (i.e. decrease the percent of false masters & false nonmasters). A set of analyses were conducted to examine that issue.

The first set of analyses examined the total test score and were similar to those conducted for the separate test sections. A combination approach was used to determine the a priori classification of a student as a master or a nonmaster because both the essay and multiple choice scores were being used. A student was considered to be a master if he/she were judged to be a master for both the essay and multiple choice sections. Likewise, a student was considered to be a nonmaster if he/she were judged to be a nonmaster on both sections. Students for whom the judgment was not consistent were not included in this analysis. Table 6 provides information indicating the level of agreement between the mastery decisions on both sections. In total the judgments were in agreement for 6152 of the 7202 students (85.4%).

Table 6

Agreement Between Essay & Multiple Choice Mastery Decisions[a]

Essay	Multiple Choice: Master	Borderline	Nonmaster	Total
Master	2936 (86.9%) (92.7%)	226 (8.9%) (7.1%)	5 (0.4%) (0.2%)	3167 (44.0%)
Borderline	417 (12.3%) (15.8%)	2074 (81.8%) (78.8%)	140 (10.9%) (5.3%)	2631 (36.5%)
Nonmaster	25 (0.7%) (1.8%)	237 (9.3%) (16.9%)	1142 (88.7%) (81.3%)	1404 (19.5%)
Total	3378 (46.9)	2537 (35.2%)	1287 (17.9%)	7202

[a] The top figure in parenthesis represents the column percent; the lower figure represents the row percent.

The biserial correlation between the master/nonmaster groups and the total test score was 0.81. This value was greater than the biserial coefficient for the essay but slightly less than that for the multiple choice test. Based on the Contrasting Groups procedure, a passing score of 76.9 was set. Using that passing score, 24.5% of the students in the nonmaster group were misclassified and 12.2% of the masters were misclassified. In total, 15.7% of the students were misclassified. Thus, the classification of students based on a weighted combination of the tests was more accurate than that based on either procedure alone.

The next set of analyses considered the essay and multiple choice test together, but as separate entities. Analyses were

conducted based on multiple decision rules. First, using the essay and multiple choice cut scores (see Table 5), in order for a student to 'pass', he/she had to have an essay score at least equal to 6 and a multiple choice score at least equal to 43. All other students were considered to 'fail'. Second, using the essay and the scores for each cluster, a student was considered to 'pass' if his/her score on each section was at least equal to the passing score.

A final analysis was conducted using only the Sentence Structure and Usage clusters. These two clusters were used because based on the results of a stepwise discriminant analysis, they were the only two which significantly added to the discriminant ability of the essay. That is, when using the essay, the sentence structure and the usage clusters, the ability to discriminate between the masters and nonmasters was better than the discrimination based solely on the essay. However, the addition of the other clusters did not significantly increase the accuracy of the classification.

Table 7 provides information about these analyses. That table shows that there is an increase in the accuracy of the categorization based on the multiple stage procedure, especially for the combination of essay and cluster scores. This is probably due to the greater difficulty in 'passing' (students must score at least equal to the cut score on all sections). Thus, there is a decrease in the percent of false masters. While these results also provide for a more accurate classification of students compared to using either the essay or multiple choice test alone,

these results do not provide for an as accurate classification as when the weighted combination of the two measures is considered.

Table 7

Multiple Decision Rule Approach

Tests	% False Masters	% False Non-Masters	Total Percent Misclassified
Essay & MC	21.6%	15.5%	17.1%
Essay & Clust.	8.8%	34.4%	27.3%
Essay, SS & U.	16.1%	20.3%	19.6%

Summary

When a key purpose for developing a competency test is to make a decision about whether students should be awarded a high school diploma, it is critical that the test be as sensitive as possible to correctly classifying students as either masters or nonmasters. The errors made by awarding or denying diplomas erroneously cannot be quantified, but clearly can be large.

This study has shown that if a choice has to be made between using an indirect or a direct method, the indirect method provides for a better means of discriminating between the competent and incompetent writer. However, the use of either method alone begs the question as to whether either can sufficiently measure the area of writing. A better approach, indeed the most appropriate manner in which to determine whether a student is a master or a nonmaster, is by using a combination of both direct and indirect methods, creating a weighted total test score and basing the pass/fail decision on that total test score.

References

Akeju, S.A., The reliability of General Certificate of Education Examination English Composition Papers in West Africa. Journal of Educational Measurement, 1972, 9, 175-180.

Breland, H.M. Can Multiple-Choice tests measure writing skills? The College Board Review, 1977, 103.

Breland, H.M. & Gaynor, J.L. A comparison of direct and indirect assessments of writing skills. Journal of Educational Measurement, 1979, 16, 119-128.

Coffman, W.E. On the validity of essay tests of achievement. Journal of Educational Measurement, 1966, 3, 151-156.

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

Glass, G.V. & Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall Inc. 1970.

Godshalk, F.I., Swineford, F.E. & Coffman, W.E. The measurement of writing ability, Princeton: Educational Testing Service, 1966

Hogan, T.P. & Mishler, C. Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 1980, 17, 219-227

Koffler, S.L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 161-178.

Livingston, S.A. & Zieky, M.J. Passing Scores. Princeton, N.J.: Educational Testing Service, 1982.

Moss, P.A., Cole, N.S., & Khampalikit, C. A comparison of procedures to assess written language skills at grades 4, 7, and 10. Journal of Educational Measurement, 1982, 19, 37-48.

New Jersey State Department of Education, Minimum Basic Skills Testing Program: District Test Coordinators Manual, 1983.

Noyes, E.S., Sale, W.M. & Stalnaker, J.M. Report on the first six tests in English composition. New York: The College Board, 1945

Quellmalz, E.S., Capell, F.J. & Chou, Chih-Ping. Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 1982, 19, 241-258.

Spandel, V. & Stiggins, R.J. Direct Measures of Writing Skill: Issues and Applications. Portland, Or: Northwest Regional Educational Laboratory, 1980.

Stiggins, R.J. A comparison of direct and indirect writing assessment methods. Research in the Teaching of English, 1982 16(2) 101-114.