ED 242 756 TM 840 167

AUTHOR	Breland, Hunter M.
TITLE	The Direct Assessment of Writing Skill: A Measurement Review. College Board Report No. 83-6.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.
REPORT NO	CB-R-83-6; ETS-RR-83-32
PUB DATE	83
NOTE	30p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101 ($5.00).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Essay Tests; Evaluation Needs; Higher Education; Interrater Reliability; *Measurement Techniques; Scoring; Secondary Education; Technological Advancement; *Test Reliability; *Test Validity; *Writing Evaluation; *Writing Skills
IDENTIFIERS	*Direct Assessment

ABSTRACT

Direct assessment of writing skill, usually considered to be synonymous with assessment by means of writing samples, is reviewed in terms of its history and with respect to evidence of its reliability and validity. Reliability is examined as it is influenced by reader inconsistency, domain sampling, and other sources of error. Validity evidence is presented, which shows reported relationships between direct assessment scores and criteria such as class rank, English course grades, and instructors' ratings of writing ability. Evidence on the incremental validity of direct assessment over and above other available measures is also given. It is concluded that direct assessment makes a contribution but that methods need to be developed to improve its reliability and reduce its costs. New automated methods of textual analysis and new kinds of direct assessment in which more than a single score is produced are suggested as two approaches to better direct assessment. (Author)

College Board Report

No. 83-6

The Direct Assessment of Writing Skill: A Measurement Review

Hunter M. Breland

The Direct Assessment of Writing Skill: A Measurement Review

Hunter M. Breland
Educational Testing Service

College Board Report No. 83-6

ETS RR No. 83-32

College Entrance Examination Board, New York, 1983

Acknowledgment

The author expresses thanks for helpful comments from Brent Bridgeman and Garlie Forehand on an earlier draft of this report. Debra Smolinski prepared the manuscript.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101. The price is $5.

CONTENTS

Tables

ABSTRACT

Direct assessment of writing skill, usually considered to be synonymous with assessment by means of writing samples, is reviewed in terms of its history and with respect to evidence of its reliability and validity. Reliability is examined as it is influenced by reader inconsistency, domain sampling, and other sources of error. Validity evidence is presented, which shows reported relationships between direct assessment scores and criteria such as class rank, English course grades, and instructors' ratings of writing ability. Evidence on the incremental validity of direct assessment over and above other available measures is also given. It is concluded that direct assessment makes a contribution but that methods need to be developed to improve its reliability and reduce its costs. New automated methods of textual analysis and new kinds of direct assessment in which more than a single score is produced are suggested as two approaches to better direct assessment.

INTRODUCTION

Over the years writing skill has been appraised through two approaches, direct assessment and indirect assessment. Direct assessments are those in which a sample of an examinee's writing is obtained under controlled conditions and then evaluated by one or more judges, usually English teachers trained in making judgments about writing skill. Indirect assessments are so termed because an estimate of probable skill in writing is made through observations of specific kinds of knowledge about writing, such as grammar and sentence structure, although more advanced skills can also be observed. These indirect assessments are commonly made by means of multiple-choice questions. Thus, direct assessments tend to be associated with writing samples and indirect assessments with multiple-choice questions. Later in this review, the distinction between direct and indirect assessments of writing skill will be reconsidered because the usual distinction may be more simplistic than it needs to be. For the moment, however, direct assessments will be thought of as writing samples evaluated by one or more judges. Indirect assessments will not be covered in this review.

Diederich (1974) probably captured better than anyone else the reasoning behind the widespread use of writing samples for the assessment of writing skill.

> As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student's writing, especially if the writing is done under test conditions in which one can be sure that each sample is the student's own unaided work. People who uphold the view that essays are the only valid test of writing ability are fond of using the analogy that, whenever we want to find out whether young people can swim, we have them jump into a pool and swim. (p. 1)

From this perspective, if one wants to know if any given individual can perform any given task, a test of performance in that task is what is needed. Coffman (1971a) presented the same kind of argument for the academic context:

> The only way to assess the extent to which a student has mastered a field is to present him with questions or problems in the field and see how he performs. The scholar performs by speaking or writing. The essay examination constitutes a sample of scholarly performance; hence, it provides a direct measure of educational achievement. (p. 273)

The logic of these kinds of arguments is so cogent that despite more than a half a century of criticism by educational measurement specialists, the essay remains a principal means of evaluation in courses of instruction of all types. In recent years, in fact, the essay has gained more and more advocates as evidence of a decline in writing skills among high school and college students accrues with each day. Faced with this, it is difficult to deny that students need more exposure to writing whether in the form of instruction or examination.

Also related to direct assessment are issues of national impact—the message that is implicitly sent to students and teachers by direct assessment used on a wide scale: If large numbers of students are required to produce compositions for assessments important for graduation, certification, or admission to higher levels of education, then students will be encouraged to learn composition skills and teachers to teach them.

Nonetheless, the history of direct writing skill assessment is a bleak one. As far back as 1880 it was recognized that the essay examination was beset with the curse of unreliability (see Huddleston 1954; Follman and Anderson 1967). One of the first demonstrations of the reliability problem occurred in the 1920s when it was shown that the score a student received on a College Board examination could depend more on which reader read his or her paper, or on when the examination was taken, than on what was actually written (Hopkins 1921).

The reliability problem is perhaps best illustrated by a simple example. In 1961 a study was conducted at the Educational Testing Service in which 300 essays written by college freshmen were rated by 53 readers representing several professional fields (French 1962). Each rater used a nine-point scale. The results showed that none of the 300 essays received less than five of the nine possible ratings, 23 percent of the essays received seven different ratings, 37 percent received eight different ratings, and 34 percent received all possible ratings. It was clear from this study that the score received was to a large degree dependent upon which expert happened to be doing the scoring.

The severity of the reliability problem noted was accentuated by the realization that readers represented only *one* source of error. Perhaps greater errors in a direct assessment are introduced by the limited sampling of topics on which students can write. Furthermore, additional errors are introduced by a tendency for errors to be correlated (because readers are influenced similarly by extraneous factors such as essay length, handwriting quality, and neatness) and by interactions among the different sources of error. The sources of reader error are many. A study by Shepard (1929) showed dramatic variations in the scores received by identical essay responses differing only in penmanship. Of course penmanship is probably less important today, but there is some evidence that it can still affect the score assigned to an essay examination (Markham 1976). In another early study, Traxler and Anderson (1935) showed that two independent scores made by experienced readers of essay examinations agreed fairly well for one essay topic but not for a second topic. It was also observed that the grades assigned to essays tended to be influenced by the grades given to the papers immediately preceding. Stalnaker (1936) noted, in this regard, that

> A "C" paper may be graded "B" if it is read after an illiterate theme, but if it follows an "A" paper, if such can be found, it seems to be of "D" caliber.

The overall impact of the reliability problem manifests itself when one attempts to correlate judgmental scores of essays with external criteria for purposes of validation. More often than not, correlations of judgmental scores with other measures are lower than would be expected, and this is usually caused by the low reliability of the judgmental scores.

Reliability will be revisited in a later section of this review, but it is first of value to review some of the different types of writing tasks commonly used in direct assessments and the methods used for evaluating them—since the specific task and evaluation procedure can influence reliability.

TYPES OF DIRECT ASSESSMENT

This section is intended only as a brief summary of various types of direct assessment as background for subsequent discussions of reliability, validity, and other issues that at times are influenced by the type of assessment. More complete (and more precise) treatments of types of assessment described here, as well as other types, are given in a number of writings by members of the English teaching profession (see, for example, Cooper 1977; Lloyd-Jones 1977; Myers 1980; Odell 1981). The types of direct assessments commonly used may be classified with respect to task types and the method of evaluation used. At times the evaluation procedure is closely linked with the task, as in primary-trait scoring. Most scoring methods, however, can be applied to more than one specific task, though modifications may be necessary as the tasks vary.

Task Types

Task types are infinite in their variety, since they vary not only with the topic to be addressed but with the specific kind of prompt or stimulus used, the audience to be addressed, and the purpose intended. Prompts may be written, aural, or pictorial. The audience and purpose may be only implicit, as when a student writes something to be evaluated by his or her teacher or an anonymous teacher or group of teachers. A task may allow consultation of reference works, such as dictionaries, and time for revision, editing, and rewriting. Or, it may be a brief, impromptu task, which allows no consultation of reference works and no time for rewriting. Following are brief descriptions of some well-known types of writing tasks.

Letter

An examinee might be asked to write a letter of some type: to a friend, to the editor of a newspaper, to a potential employer, to a company complaining about a product or service, and so on.

Narrative

An autobiographical account, a description of a vacation or other experience, or a historical description of some other type would all be narratives. These narratives could, of course, also be written in the form of a letter, and narratives can be either real or imaginary.

Descriptive

Although a narrative is usually descriptive, the term implies the description of a series of events. A piece of writing may be simply the description of some object, how it looks, how it works, or some other aspect of it, or some other kind of description.

Argumentative

In this type of task, the examinee is usually asked to take a position on some issue and argue persuasively for that position using evidence from his or her own personal experience or reading. It is probably the most common task type used because it requires the integration of several different writing skills. Sometimes this type of task is referred to as an "expository-argumentative" task.

Expressive

Rather than argue persuasively, the task may be only to express one's opinion on some issue or event. While expository in nature, this kind of task is usually distinguished from a persuasive or argumentative exposition.

Role-Playing

One may be asked to assume a role in some situation and then to write something (such as a letter or a memorandum) for some specific purpose. Examples would include responding to an irate customer as a customer relations official, or writing a memorandum to a superior or a subordinate in an organization. For role-playing tasks, the audience and purpose are usually quite clear.

Precis or Abstract

A real-life task of some importance is that of synthesizing a large body of information for transmittal to an audience different from that intended in the original piece. Scientists abstract complex scientific investigations for nonspecialists. Diplomats abstract current information about specific countries, at times originally written in other languages, for use by others. Lawyers synthesize case histories having legal precedents in making arguments. Therefore, a useful task is to ask students to read something and then to prepare a brief precis or abstract of it.

Diary Entry

This could be similar to any of the preceding tasks, but the fact that it is written for personal use would probably change its tone.

Literary Analysis

This is a common task used in literature courses and in the more difficult English examinations

Revision or Editing

Any of the tasks above might be the subject of a task requiring revision or editing.

Evaluation Methods

Having obtained a response to one or more of the stimuli represented in the task types discussed in the preceding section, one can then usually choose among a number of different methods for evaluation of the response. As noted earlier, some evaluation methods are closely tied to the stimulus, namely, primary-trait methods. Thus, the task may predetermine the evaluation method. Among the several different approaches to evaluation, some are more widely used than others. The descriptions that follow, it should be cautioned, do not represent a consensus of opinion on the meaning of terms. Rather, they are an attempt to describe briefly methods about which there is often much disagreement.

Holistic Scoring

According to Cooper (1977), in holistic scoring "the rater takes a piece of writing and either (1) matches it with another piece in a graded series of pieces, or (2) scores it for the prominence of certain features important to that kind of writing, or (3) assigns it a letter or number grade. The placing, scoring, or grading occurs quickly, impressionistically, after the rater has practiced the procedure with other raters." Holistic scoring is at times conducted using scoring guides, or rubrics. Some practitioners of holistic scoring distinguish it from impressionistic scoring, since the latter is viewed as a haphazard, noncontrolled, and unmonitored procedure. Holistic scoring is the most widely used evaluation procedure.

Focused Holistic Scoring

This method is essentially the same as holistic scoring except that scores are produced for more than a single dimension of the writing sample being evaluated. For example, one might score for content and mechanics, or for some other specific aspects. The scoring might be done for each dimension after a single reading, or it might be done for each separately so as to minimize influences of one focus on the other. The number of focuses must of course be limited; otherwise, the procedure tends to be more like an analytical procedure. As in holistic scoring, no counts or enumerations of any type are used. Scoring rubrics for each of the dimensions focused on, however, may be used.

Analytic Scoring

This evaluation procedure is perhaps best exemplified by that associated with Diederich (1974). The Diederich procedure is based on a factor analysis of writing samples scored by experts representing several different academic disciplines. The factors derived were ideas, organization, wording, flavor, and mechanics. In some versions of the method, mechanics is further divided into usage, punctuation, spelling, and handwriting. Each factor is rated on a scale from 5 (high) to 1 (low), and two of the scales (ideas and organization) receive a double weighting. Thus it is possible to obtain a score as high as 50, or as low as 10. Other analytic procedures are described by Cooper (1977), Odell (1981), and Follman and Anderson (1967).

Atomistic Scoring

Somewhat akin to analytic scoring are methods in which detailed enumerations are made of quite a number of different features of a piece of writing. While certainly "analytic" in many senses, it is useful to distinguish atomistic scoring from analytic scoring, as described here, because it is very different with respect to the detail required. One example of an atomistic scoring procedure was described by Moss (1982). In this procedure, the total number of errors was counted in each of four categories: spelling, capitalization, punctuation, and expression. To develop a score from these counts, the total number of errors was divided by an index of paper length so as to avoid inappropriate penalties for writing more.

Primary-Trait Scoring

Mullis (1980) explains that the rationale of primary-trait scoring "is that writing is done in terms of an audience and can be judged in view of its effects upon the audience." The primary, or most important, trait of a piece of writing will be the approach used by the writer to reach the audience intended. The primary trait of a set of directions, for example, "would be an unambiguous, sequential, and logical progression of instructions," according to Mullis. Another example given by Mullis is a piece of political campaign literature intended to persuade a reader to vote for a particular candidate. "A successful campaign paper will have certain persuasive traits that an unsuccessful one will not have, and these traits will differ from those necessary for a successful set of directions," Mullis notes. For any given task, the scoring directions must be prepared beforehand, and they are usable only with that specific task.

Syntactic Scoring

Hunt (1977) has popularized a method of gauging syntactic maturity which is most often associated with the term, *T-unit.*" A *T-unit* is defined by Hunt as a "single main clause plus whatever other subordinate clauses or nonclauses are attached to, or embedded within, that one main clause." In other words a T-unit is a single main clause and whatever else goes with it. The T-unit is used, rather than the sentence, because it is empirically useful in describing the changes that occur in the syntax of writ^rs as they mature.

Communicative Effectiveness

In a sense similar in objectives to primary-trait scoring, this method of measuring the quality of prose is also concerned with the effects it has on an audience. But, operationally, the method is very different from primary-trait scoring. Hirsch and Harrington (1981) describe the theoretical basis for this new method and some of its advantages over traditional methods of scoring. The method is also similar in some ways to recent approaches being taken by cognitive psychologists, in which the theory and structures of reading comprehension research are applied to the analysis of text (see, for example, Bracewell et al. 1982; Bruce et al. 1982; Fredericksen 1983). Usually, an objective index of communicative effectiveness, such as reading speed or comprehension, is derived for the assessment.

Automated Scoring

Another new method of evaluation that is of considerable interest is that done by computer. Frase et al. (1981) and MacDonald et al. (1982) describe a computer-based system developed at Bell Laboratories that is presently operational. A more sophisticated parsing system is under development by IBM (see Heidorn et al. 1982). These methods will be discussed in more detail in a later section of this review.

RELIABILITY OF DIRECT ASSESSMENTS

Numerous research investigations have demonstrated that direct assessments of writing skill, as usually conducted, tend to yield low reliabilities. The sources of error are several, but most analyses have focused on two primary sources: rater inconsistency and sampling bias. Rater inconsistency occurs not only between raters but with the same rater from one occasion to the next—even when the same writing sample is being scored. Rater variability consists of three different components (Coffman 1971a): First, raters differ with respect to leniency. Some may tend to score high and others low; thus, the level of score obtained by any individual examinee depends upon the rater or raters assigned to score the responses of that examinee. Second, raters differ in the degree to which they have a central tendency, an inclination to score near the average. Third, different raters have different values that many times lead them to assign grossly different scores to the same response.

While less research has been concerned with the problem of sampling error, it is probable that sampling is also a serious source of error in direct writing assessments. A highly reliable writing assessment will require more than one writing sample, and each sample will be independent from all other samples. Such independence does not occur, of course, when several tasks are required that relate to the same topic stimulus. The most reliable assessment will occur when all of the responses are scored independently by different raters. The more the number of independent responses and the more the number of independent ratings of each response, the greater will be the reliability of the assessment. Unfortunately, it has not proved to be economically feasible to conduct large-scale writing assessments using multiple writing samples and multiple independent ratings. For the same reason, there have been few research investigations of multiple samples scored independently by multiple readers. Table 1 presents a summary of 24 research studies in which reliability estimates were reported for direct assessments of writing skill. These studies are summarized with respect to a number of factors that may have influenced the magnitude of the estimates reported. A consideration of these factors is useful as an introduction to the reliability estimation for direct assessments.

Factors Influencing Reliability Estimates

Table 1 is limited to studies reporting reliability estimates for direct assessments of junior high, high school, and college populations. However, quite a variety of social, ethnic, and ability groups is represented. The population sampled can influence reliability estimates if it is restricted in range of ability, but how such influences operate is not always clear. It is usually assumed that restrictions in range will attenuate estimates, but the actual effects are dependent on other aspects of the population distributions as well. The number of cases used for the estimate affects its stability. The larger the number of cases, the more stable will be the estimate. Reliability is also influenced by the type of writing tasks used and the amount of time allowed for response, but little evidence is available concerning the effects of task type and timing on reliability. The most common type of writing sample is the brief, persuasive, or argumentative essay in which some position is to be taken on an issue presented and a thesis developed to support that position

Table 1. Studies Reporting Reliability Estimates for Direct Assessments

Study	*Population Description*	*Subsamples*	*Cases*	*Task*		*Scoring*	
				Type	*Timing (Minutes)*	*Method*	*Range*
1. Akeju (1972)	West African 18-year olds	None	100	Not described	Not given	Not described	Not given
2. Breland and Gaynor (1979)	College freshmen	Four colleges	2,000	Persuasive	20	Holistic	1-6
3. Breland (1983)	College applicants	Black	200	Persuasive	20	Atomistic	20-100
		White	200			Analytic	3-15
		Hispanic-Y	200			Holistic	1-4
		Hispanic-N	200				
4. Clemson (1978)	High school students	None	567	Letter	20	Holistic	1-4
5. Coffman (1966)	High school students	None	646	Five types	20-40	Holistic	1-3
6. Coffman (1971a)	Adv. Placement students	None	495	Not described	Not given	Holistic	1-9
7. Coffman (1971b)	Hypothetical	None	25			Holistic	1-9
8. Conry and Jeroski (1980)	Canadian students	8th grade	382	Narrative	Not given	Holistic	1-9
			397	Expository		Holistic	1-9
			88	Narrative		Analytic	Varied
			90	Expository		Analytic	Varied
		12th grade	371	Narrative	Not given	Holistic	1-9
			382	Narrative		Holistic	1-9
			70	Narrative		Analytic	Varied
			70	Expository		Analytic	Varied
9. Coward (1952)	Foreign Service applicants	None	100	Four types	45	Holistic Analytic	1-10
10. ETS (1982)	High school students	None	86,039	Persuasive	20	Holistic	1-4
11. Finlayson (1951)	English 12-year olds	None	197	Two choices	60	Holistic	1-20
12. Godshalk et al. (1966)	High school students	None	646	Five types	20-40	Holistic	1-3
13. Follman and Anderson (1967)	Hypothetical	None	10	Expository Persuasive	Not given	Various	1-5
14. Hackman and Johnson (1977)	College freshmen	None	173	Persuasive	40	Holistic Analytic	1-5
15. Huddleston (1954)	College entrants	None	129	Persuasive	20	Analytic	?
			138	Interpretative			
			136	Expository			
16. Michael et al. (1980)	College students	A	100	Descriptive	30	Holistic	1-4
		B	100				
17. Moss et al. (1982)	High school students	7th grade	40	Letter	Not given	Holistic	1-4
		10th grade	94	Descriptive		Atomistic	0-20
19. Myers et al. (1966)	College applicants	None	125	Not described	20	Holistic	1-4
20. Powills et al. (1979)	Junior High students	7th grade	80		30	Holistic	1-4
		8th grade	135				
21. Quellmalz et al. (1982)	High school students	None	200	Expository	Not given	Analytic	1-4
22. Steele (1979)	College freshmen	Sample 1	65	Letters	20	Analytic	1-5
		Sample 2	50				
23. Weiss and Jackson (1982)	College students	None	224	Descriptive	40	Atomistic	0-64
			224	Persuasive	40	Holistic	1-6
		Posttest	123	Persuasive	20	Holistic	1-6
24. Werts et al. (1980)	College students	None	234	Persuasive	20	Holistic	1-6

using examples, facts, or other evidence. The second most common type of sample is a narrative essay (at times called a descriptive or narrative-descriptive essay). Any of those tasks can require the writer to consider a specific audience or purpose. More commonly, however, neither the audience nor the purpose is specified. Very few assessments of this type offer the examinee a choice of topics. The time allowed for the writing tasks in Table 1 varied from 20 minutes to 2 hours with 20-minutes being the most common.

The method of evaluation of writing samples can influence reliability. Essentially, three principal approaches to scoring of writing samples are represented in Table 1. Holistic scoring is the most common, but what is termed holistic scoring may vary from one study to the next. The other two types of scoring represented are analytic scoring and atomistic scoring. The distinction between these two is not always cle ir, but in this review analytic scoring refers to the development of several subscores which are either interpreted separately or combined to produce a total score. Atomistic scoring refers to a very detailed count of errors or a detailed scoring of many aspects of a sample. Any scoring with as many as 20 subscores has been considered here to be atomistic, even if the authors called it analytic.

Scoring scales differ somewhat, and these also can affect reliability. The most common scale has been the 1 (low) to 4 (high) scale often used for holistic scoring. Some observers believe (for example, Coffman 1971a, 1971b) that a greater scale range produces better reliabilities. A field test comparing a 1-3 scale with a 1-4 scale by Godshalk et al. (1966) suggested some improvement in reliability with the 1-4 scale, but Coffman (1971b) indicated a preference for an even greater range in scores. Large scale-ranges can be simply many points on a holistic scale, or they can be developed through analytic and atomistic scoring as in the Breland (1983) 3-15 range scale based on three subscales or the Moss et al. (1982) 0-20 atomistic scale.

Once scores have been assigned, they may or may not be adjudicated. Adjudication usually involves engaging an additional reader to resolve a scoring discrepancy between two other readers. Since highly discrepant scores are eliminated through adjudication, reliabilities increase. Two final procedural differences in direct assessments have to do with the total number of readers engaged and the physical context of their engagement. The more readers there are, the more difficult training and instruction is. Consequently, it is usually expected that reliabilities will be less for a large group of readers than for a very small operation. For example, if only two readers are used, and if they are carefully instructed and monitored, one would not expect much difference in their judgments. The two readers may also represent the same educational setting, such as an English department, and thus the likelihood of agreement may be quite high.

The other procedural difference has to do with the setting in which the scores are generated. The most common setting is the conference setting in which readers are assembled at some central facility and supervised in some way as they read. Another approach used less often is what might be called the "remote" method in which readers are not assembled but are mailed samples with written instructions on scoring. At times, readers may be assembled initially for instruction, but the actual reading is conducted in their individual homes or offices and the materials returned through the mail.

The reliability estimates reported in the studies of Table 1 were generated through different statistical procedures. Often, a simple correlation between reader scores on a single topic is reported. At other times test-retest, alternate forms, and other types of correlations are reported. Coffman (1971a, 1971b) asserts that correlations at times tend to overestimate reliabilities because they do not take into account mean differences among scores. Analysis of variance procedures are preferred, he observes. Similar estimates are generated through confirmatory factor analysis procedures, but these depend on the specific model postulated for the analysis.

All of the above differences in the ways direct assessments are conducted and analyzed often combine to produce unpredictable influences on reliability estimates reported in the literature. In an attempt to gain some sense of the magnitude of reliabilities that one might expect in a given situation, estimates reported in the literature have been assembled and identified as much as possible with respect to procedures. A basic distinction made in assembling these estimates has been between reading reliability estimates and score reliability estimates.

Reading Reliability Estimates

Reading reliability reflects error variance attributable to the inconsistencies among readers, but it does not reflect sampling error (the error introduced by providing only a limited opportunity to compose) or other sources of error. Reading reliability estimates will thus be inflated and cannot be used as an estimate of score reliability. Nevertheless, it is often useful to obtain an estimate of reading reliability as a gauge of the consistency of readers. When only one writing sample has been scored, it is not possible to estimate accurately anything but reading reliability. A comparison of reading reliability estimates obtained in a number of research investigations is presented in Table 2. Estimates are grouped with respect to the number of tasks scored and the number of ratings per task obtained.

Overall median estimates of .64, .70, and .78 were computed and are given at the bottom of Table 2 for three common situations. Note that relatively low estimates were reported in the Coffman (1966) paper. The estimates in Table 2 range from a low of .39 (for one task rated by one reader) to a high of 88 (for three tasks rated by five readers). In two other papers, Coffman (1971a, 1971b) observes that the range of scores assigned, the number of readers, and the method of estimate used will all affect

Table 2: Reading Reliability Estimates Reported for Direct Assessments

				One Task					*Two Tasks*					*Three Tasks*				
				Ratings per Task					*Ratings per Task*					*Ratings per Task*				
Study	*Estimate Number*	*Scoring Method*	*Scale*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
Akeju (1972)		Not described		.72	.84	.88												
Breland (1983)	1	Atomistic	1-20	.40[a]	.57[a]													
	2	Analytic	3-15	.67[a]	.80[a]													
	3	Holistic	1-4	.54[a]	.70[a]													
Coffman (1966)		Holistic	1-3	.39	.56	.65	.72	.76	.51	.68	.76	.81	.84	.60	.75	.82	.86	.88
Coffman (1971a)		Holistic	1-4		.70													
Conry and Jeroski (1980)		Holistic	1-9			.74[b]												
Coward (1952)	1	Holistic	1-9	.54[c]	.69[c]													
	2	Analytic		.70[c]	.82[c]													
ETS (1982)		Holistic	1-4		.71													
Finlayson (1951)		Holistic	1-20	.71	.83	.88	.91		.80	.89	.92	.94						
Hackman and Johnson (1977)		Holistic	1-5	.61														
Huddleston (1954)		Analytic		.60[d]														
Michael et al. (1980)		Holistic	1-4	.66[e]	.80[f]													
Moss et al. (1982)	1	Holistic	1-4							.86								
	2	Atomistic	0-20							.89								
Myers et al. (1966)		Holistic	1-4	.41	.58	.67	.73											
Powills et al. (1979)		Holistic	1-4			.81[g]												
Steele (1979)	1	Holistic	1-5												.84			
	2	Analytic	3-15												.90			
Traxler and Anderson (1935)	1	Analytic	1-10		.94													
	2	Analytic	1-10		.84													
Weiss and Jackson (1982)	1	Atomistic	0-64	.71	.55													
	2	Holistic	1-6	.80[h]	.66[h]													
Estimate Medians				.64	.70	.78												

[a] Average over 4 samples.

[b] Average over 2 samples; narrative and expository tasks.

[c] Average over 4 tasks.

[d] Average over 3 tasks and 3 samples.

[e] Average over 8 conditions.

[f] Average over 8 conditions.

[g] Average over 8 conditions.

[h] Average over 2 conditions.

estimates. With respect to range of scores, the suggestion was made that the greater the range, the greater the variance obtainable, and thus the greater the reliability estimate. Table 2 supports such a speculation, since the low estimates reported by Coffman (1966) were based on a score range of only 1 (low) to 3 (high). Myers et al. (1966) used similar methods with a 1-4 scale and obtained estimates similar to those reported by Coffman (1966). In a hypothetical set of data, Coffman (1971b) demonstrated that two ratings of the same 25 papers correlated .87 when a 15-point scale was used but only .77 when a 5-point scale was used. These correlations, which represent reliability estimates for a single task and one rater, change also when the 15-point scale is cut in different places.

As noted previously, when the number of readers used is large, it is more difficult to achieve consistency than when the number is small (because it is easier to train and instruct a small number). None of the studies in Table 2 examined this issue specifically, but the magnitude of estimates is to some degree associated with numbers of readers, where such information is available. The Coffman (1966) estimates, for example, are based on ratings by 25 different raters; Finlayson (1951), in contrast, used only six raters. Estimates based on product-moment correlations will also tend to be higher than those based on analysis of variance, because one set of scores may have a different mean than another, and differences in means are not reflected in a product-moment correlation. A comparison of the two methods was made by Coffman (1971b) using his hypothetical set of 25 essays. For the 15-point scale, the reading reliability was .87 for the correlational method and .85 for the analysis of variance method. No comparison was made for the 5-point scale. The investigation of Michael et al. (1980) summarized in Table 2 also computed reliability estimates based on both methods, though the main object of the study was to compare expert and lay readers. The two types of estimates were quite close with the exception of one comprison where the analysis of variance estimate was somewhat lower.

A fourth influence believed by many to be important is the length of the essay. In Table 2, the longest writing time reported in any of the studies is the 60-minute papers of Finlayson (1951). The reading reliability estimates are relatively high (.71 to .94), but these might be attributable to the large range of scores (1-20), the use of the analysis of variance method of assessment, the use of only six raters, or the combination of all three of these factors. A comparison of essay length (or time allowed) is possible within the Coffman (1966) study and within the Weiss and Jackson (1982) study. In the Coffman estimates, the suggestion is that essay length is unimportant because the 20-minute essays were estimated to have about the same reading reliabilities as the 40-minute essays. In the Weiss and Jackson study, a 40-minute essay had a slightly higher reading reliability estimate (.68) than did the 20-minute essay (.63). It is not clear from the studies listed in Table 2, therefore, whether reading reliability is influenced by the length of the essay or the time allowed to write it.

A fifth influence suggested by Coffman and others on direct assessment reliabilities is the method of scoring. Three of the studies in Table 2 allow for a comparison of scoring methods. Coward (1952) compared scores on responses to four different tasks that were scored both analytically and holistically on a 1–9 scale. The analytical scoring involved the rating and weighing of several components, although the actual range of scores developed was not given. The reading reliability estimates were higher for analytic scoring for each of the four tasks analyzed. Weiss and Jackson (1982) used both holistic and atomistic scoring methods, and both holistic scorings yielded higher reading reliability estimates than did the atomistic scoring.

In my own work (Breland 1983), I have conducted all three types of scoring on the same set of 20-minute essays. An atomistic scoring was conducted through a 20-element checklist in which scorers checked specific attributes of essays on a 5-point scale. The scores on each checklist item were combined into an equally weighted sum to produce a score range from 20 to 100. An analytic scoring was accomplished by a different set of raters using a three-facet skill rating, each on a 5-point scale. The three facets were discourse quality, syntactic quality, and lexical quality, and were based on an analysis of the 20-element checklist. The analytic score was based on an equally weighted sum of the three-skill facets. Holistic scorings of the same essays were also made on two different occasions by two different sets of readers using a 1-4 scale. The results of scoring the same essays three different ways leads one to the conclusion that holistic scoring yields higher reliabilities than detailed atomistic scoring, and it is also a great deal less tedious. On the other hand, it indicates that a limited amount of analysis, such as in the three-facet scoring, can produce reading reliabilities higher than those obtained with holistic scoring.

The analytic ratings, of course, required more reading time, but costs were minimized by conducting the reading through the mail rather than in a conference setting. This difference in mail versus conference reading suggests one final influence on the reliability of readings. In a conference setting, readers can discuss their ratings and be supervised by table leaders and a chief reader. These influences have demonstrated through countless readings to result in better reliabilities of scores. But it is possible that carefully worded instructions sent through the mail can also result in improved reading reliabilities. The suggestion of the results from Table 2 is that carefully written instructions, when combined with analytic scoring procedures, result in improved reliabilities. Whether holistic scoring conducted in a similar way would yield even higher reliabilities is not known, but Table 2 indicates that analytic scoring tends generally to produce the highest reading reliabilities when a single task is being scored.

Table 3. Score Reliability Estimates Reported for Direct Assessments

				One Task				Two Tasks				Three Tasks			
	Estimate Number	Scoring		Ratings per Task				Ratings per Task				Ratings per Task			
Study		Method	Scale	1	2	3	4	1	2	3	4	1	2	3	4
Breland and Gaynor (1979)		Holistic	1-6		.51				.51						
Clemson (1978)		Holistic	1-4		.55				.55						
Coffman (1966)		Holistic	1-3	.26	.38	.44	.49	.42	.55	.62	.66	.52	.65	.71	.74
Finlayson (1951)		Holistic	1-20	.69	.78	.82	.84	.82	.88	.90	.91				
Moss et al. (1982)	1	Holistic	1-4						.46						
	2	Atomistic	0-20						.73						
Quellmalz (1982)	1	Analytic	1-4										.61		
	2	Analytic	1-4										.83		
Steele (1979)	1	Holistic	0-4		.43	.49			.58	.62			.65	.70	
	2	Holistic	0-4		.58				.73				.76		
	3	Analytic	3-15										.82		
Traxler and Anderson (1935)		Analytic	1-10						.60						
Werts et al. (1980)					.44										
Estimate medians					.53				.66				.70		
Spearman-Brown estimates									.69				.76		

Score Reliability Estimates

Table 3 provides a summary of 10 studies reporting estimates of score reliabilities, estimates that include not only reader inaccuracies but also error variance associated with sampling. To develop such estimates, more than a single task and more than a single reading are required.* The most frequent type of estimate reported, as Table 3 shows, is that for two ratings per task—whether one, two, or three tasks were rated. For these cases, medians of the estimates are given at the bottom of the table. The median estimates for two and three tasks, respectively, are slightly less from what would be computed by the Spearman-Brown formula using the .53 median estimate for a single task as a base. This could mean that the .53 estimate is too high, and that the estimates for two and three tasks are too low. The low (.38) estimate made by Coffman was based on an extension from a 5-task, 5-reading analysis of variance and, additionally, is based on a 1-3 score scale—which probably attenuated the base estimate. The next higher figure of .58 reported by Steele (1979) is based on unusually explicit instructions and numerous prescored samples—advantages readers usually don't have. Thus, the .53 median estimate for the score reliability obtained when one task is scored by two readers seems reasonable.

For three tasks and two ratings per task, a Spearman-Brown estimate of .76 is higher than the median estimate of .70. The Steele (1979) generalizability coefficient estimate was .65 for a 3-task, 2-rating situation, but it may have been low because rating instructions were in the process of development. After rating instructions were improved, the generalizability coefficient increased to .76—the same as the Spearman-Brown estimate.

A few studies have reported reliability estimates for numbers of tasks or ratings in excess of those given in Table 3. These are of interest because they give some indication of what accuracy one might expect if resources were available to conduct such assessments. Table 4 gives estimates reported in four studies. The Finlayson (1951) estimates for two tasks and six raters exceed the score reliability attained by many objective tests (and thus appear extreme).

Coffman (1966), using empirical estimates as a base, produced an extended matrix of reading and score

Table 4. Reported Reliability Estimates Based on Multiple Tasks or Ratings in Excess of Three

		Reliability Estimate	
	Tasks/Raters per Task	Reading	Score
Akeju (1972)	1 / 7	.95	
Coffman (1966)	5 / 5	.92	.84
Diederich and Link (1967)	4 / 2		.80
Finlayson (1951)	2 / 6	.96	.93
Steele (1979)	6 / 2		.75
	6 / 3		.79
	9 / 2		.79
	9 / 3		.83

*Note that multiple tasks may consist of multiple topics in the same discourse mode, a single topic in different discourse modes, or multiple topics in different discourse modes.

Table 5. Past Estimates of Score and Reading Reliabilities for Sets of Short Essays Read Holistically on a 1-3 Scale

		Number of Tasks				
Number of Ratings per Task	*Type of Reliability*	*1*	*2*	*3*	*4*	*5*
1	Score	.26	.41	.52	.59	.64
	Reading	.38	.51	.60	.66	.70
2	Score	.38	.55	.65	.71	.75
	Reading	.56	.68	.75	.79	.82
3	Score	.44	.62	.71	.76	.80
	Reading	.65	.76	.82	.85	.88
4	Score	.49	.66	.74	.79	.82
	Reading	.72	.81	.86	.88	.90
5	Score	.52	.68	.76	.81	.84*
	Reading	.76	.84	.88	.91	.92*
∞	Content	.68	.81	.86	.89	.91

Source: Adapted from Coffman (1966).

*Based on empirical data

to examine this matrix of estimates and to summarize the procedures Coffman used to generate it. The procedure begins with the 5-task, 5-rating cell based on empirical data and these assumptions:

1. The essay tasks are random samples from a pool of tasks; consequently, the relationships among score reliabilities as tasks vary in number are governed by the Spearman-Brown formula.
2. The raters are selected at random and randomly assigned to essays. Under these conditions it is also assumed that the Spearman-Brown holds for reading reliabilities as the number of readings varies.
3. The relationship between reading and score reliabilities is governed by the concept of "content" reliability (Gulliksen 1950, 211–214), in which content reliability remains constant as the number of readings changes. Content reliability is computed as the ratio of the score to reading reliability.

Using these assumptions, it is possible to start at the 5-task, 5-rating cell (based on empirical data) and complete the entire matrix as shown in Table 5. The further one proceeds from the empirical base, of course, the less confidence one has in the estimates made. In the 1-task, 1-rating cell, for example, the estimates would be expected to be less accurate. Unfortunately, it is at the low end of the matrix (few tasks and few ratings) where most assessments are made. As a result, it would be of value to have better estimates for those situations. Moreover, since the Coffman (1966) estimates were based on an extreme scoring scale (only 1-3 points), they are not generally applicable. One approach to better estimates would be to use the Coffman procedure, but to use as a base empirical evidence more generally applicable and to start at the opposite end of the matrix (few tasks and few raters). Median estimates from Tables 2 and 3 can be used as an empirical base. For the 1-task, 2-rating cell, good median estimates are available for both reading and score reliabilities. For the 1-task, 1-rating cell and for the 1-task, 3-rating cell, Table 2 provides reasonably stable reading reliabilities.

Table 6 shows the matrix of reading and score reliabilities developed using the Coffman procedure, the indicated empirical bases, and some of Coffman's assumptions. The second assumption, that the Spearman-Brown

Table 6. New Estimates of Score and Reading Reliabilities for Various Combinations of Tasks and Ratings per Task

		Number of Tasks				
Number of Ratings per Task		*1*	*2*	*3*	*4*	*5*
1	Score reliability	.48	.65	.74	.78	.83
	Reading reliability	.64*	.76	.82	.85	.88
2	Score reliability	.53*	.70	.76	.81	.85
	Reading reliability	.70*	.81	.85	.88	.90
3	Score reliability	.59	.75	.81	.84	.88
	Reading reliability	.78*	.87	.90	.91	.94
∞	Content reliability	.76	.86	.90	.92	.94

Note: The Spearman-Brown formula is: $r_{nn} = \frac{nr_{tt}}{1 + (n-1)\,r_{tt}}$

Where, r_{nn} = the estimated coefficient
r_{tt} = the original coefficient
n = the number of times a test is lengthened

*Based on the empirical data of Tables 2 and 3.

can be used to increase reading reliabilities as the number of ratings increases, was not used. Such a table of estimates can be only a rough guide to the magnitude of reliabilities one might expect in a given situation, of course. More precise estimates would recognize the specific effects on reliability noted previously, namely, scoring scale range, number of readers to be trained, and other factors. Additionally, the greater the sampling from the various stimulus and discourse modes the greater the reliability one would expect. A comparison of Tables 5 and 6 suggests that Coffman's extended estimates were somewhat lower than would usually be obtained, and that his estimates nearer to his empirical base were slightly low.

Reliabilities of Analytic Subscales

Several of the studies summarized in the previous sections also examined analytic subscales. In Table 7, six studies are summarized in which reliabilities were reported either for separate analytic subscales or for an overall score derived

Table 7. Reliabilities Reported for Analytic Subscales

			Reading Reliability Estimates				
	Number of Tasks	*Ratings per Task*	*#1*	*#2*	*#3*	*#4*	*Score Reliability Estimates*
Breland (1983)							
Discourse quality	1	2	.69	.74[a]	.—	.—	
Syntactic quality	1	2	.70	.75[a]			
Lexical quality	1	2	.71	.74[a]			
Total of subscales	1	6	.78	.82[a]			
ECT Holistic Score	1	2		.76[a]			
Conry and Jeroski (1980)							
Organization	1	3	.32[b]	.55[c]	.56[d]	.66[e]	
Sentence structure	1	3	.47	.62	.63	.51	
Spelling	1	3	.46	.53	.61	.71	
Handwriting	1	3	.47	.55	.51	.65	
Vocabulary	1	3	.59	.57	.76	.58	
Punctuation	1	3	.28	.52			
Diederich and Link (1967)							
Ideas	4	2					.80+[f]
Organization	4	2					
Wording	4	2					
Flavor	4	2					
Usage	4	2					
Punctuation	4	2					
Spelling	4	2					
Handwriting	4	2					
Hackman and Johnson (1977)							
Mechanics (subsentence level)	1	2		.83[a]			
Mechanics (sentence level)	1	2		.81[a]			
Organization	1	2		.64[a]			
Thought	1	2		.66[a]			
Style	1	2		.70[a]			
Overall quality	1	2		.61[a]			
Quellmalz, Capell, and Chou (1982)							
Focus	3	2					
Organization	3	2					
Support	3	2					
Mechanics	3	2					
Steele (1979)							
Language	3	2					.83
Organization	3	2					.74
Audience	3	2					.48
Total analytic	9	2					.82
Holistic	3	2					.76

[a] Adjudicated scores
[b] 12th grade, narrative
[c] 12th grade, expository
[d] 8th grade, narrative
[e] 8th grade, expository
[f] Overall reliability of composite of analytic scores

from the analytic subscales. The best known of these analytic scoring schemes is that of Diederich and Link (1967). The construction and use of these subscales are described in Diederich (1974), and the factor analysis from which they were derived is reported by French (1962). A total score is derived from the eight scales by rating each on a 1 (low) to 5 (high) scale, doubling the weight for ideas and organization, and summing. Thus the total score can range from 10 to 50. Diederich and Link (1967) report that this cumulative total of eight ratings, when applied independently to four different papers, results in a score reliability of .80 or more. The average reading time per paper is about 5 minutes.

The analytic subscales of Conry and Jeroski (1980), Hackman and Johnson (1977), and Quellmalz et al. (1982) are somewhat similar to the Diederich subscales. All have organization as one subscale, and all have mechanics—either as a subscale or as represented by specific aspects of mechanics. The Steele (1979) subscales are different in that they don't attend to mechanics at all, except as it relates to language. The stimulus was aural (taped) rather than written, and the examinee was required to consider audience and purpose as important. Each element was rated on a 0-4 scale. The use of audience as a subscale is of particular interest because such a scale allows for an evaluation of skills relating to audience issues and (implicitly) issues of purpose. But the score reliability obtained for the audience subscale (.48) was disappointing, suggesting that such a factor is difficult to score.

These analytic approaches tend to be limited also because they focus only on parts of the total domain of interest. Since there are numerous aspects of writing skill, and since these vary from one mode of discourse to the next, it is usually assumed that only a few aspects can be rated. And when only a few characteristics are rated, there is always the possibility that something important may have been overlooked or that one element may receive more weight than it merits. Of course, these limitations of analytic scoring—as well as the added time it takes—are the principal arguments for holistic scoring.

The Breland (1983) scales represent a compromise between analytic scoring as it is usually conducted and holistic scoring. While empirically based, these scales do not represent an extraction of factors as in the Diederich approach. Such factor analytic approaches are limited because (1) they are appropriate only for the particular discourse mode used for the factor analysis, and (2) they do not cover the entire domain of skills, as does holistic scoring. As a compromise, the Breland (1983) scales might be more aptly labeled "focused holistic scales." That is, they focus on three distinct qualities of writing, but in doing so they do not exclude any specific characteristics. They represent a dividing up of holistic scoring into three domains. Because nothing is excluded, the scales can be applied to samples from any mode of discourse—provided that each subscale is appropriately defined. For example, in an argumentative mode of discourse, the scale "discourse quality" would include an evaluation of the degree to which supporting evidence is used, but in a narrative-descriptive mode, use of supporting evidence would not be evaluated as a part of discourse quality because no argument is being made.

Summary of Reliability Evidence

The reliability of direct assessments of writing skill is limited primarily by measurement errors resulting from reader inconsistencies, content sampling biases, and interactions between these two sources of error. Reliability estimates found in the literature are influenced by the population studied, the number of cases examined, task type, number of tasks, number of readers, time allowed, scoring method used, and scoring range. The most important influences appear to be number of tasks, number of raters, scoring method, and scoring range. Considering only number of tasks and number of ratings per task, it can be expected that score reliabilities will range from about .50 (for one task and one rater) to about .90 (for five tasks and three ratings per task). Higher scoring ranges, up to about 15 judgmental points, seem to generate slightly higher reliabilities. Analytic scoring methods with a limited set of scales may produce higher reliabilities than holistic scoring, though detailed analysis using many scales (atomistic scoring) appear to yield the lowest reliabilities.

VALIDITY OF DIRECT ASSESSMENTS

Validity is often considered with respect to several specific procedures used in the process of examining measures: concurrent validation, predictive validation, incremental validation, validation of subscores, content validation, and construct validation. With the exception of the last two procedures, the methods used are essentially correlational methods. That is, a criterion of some type is correlated with the measure being examined. In content validation, a systematic examination of the content of a test is made to determine the degree to which it samples the skill purported to be measured. Construct validation requires an examination of the degree to which an assessment measures some theoretical construct, or trait. Construct validation involves the gradual accumulation of evidence from a number of sources including correlational evidence, internal consistency, the influence of instructional interventions, and any other available sources. The following sections report evidence for direct assessments for various types of validity.

Concurrent Validity

Table 8 summarizes five studies in which some direct measure of writing skill was correlated with a criterion measure at or about the same time. The most common criterion used in these studies was the high school GPA, but concurrent correlations are also shown for high school and college grades in English composition courses, for high school instructors' ratings of writing ability, and for more reliable direct assessments of the same type being validated.

The validity evidence reported by Coffman (1966) is different conceptually from that of the other studies in Table 8. The criterion variable was the sum of scores obtained from four different essay tasks, each scored independently by four different raters. As a result, the criterion was based on 16 independent judgments and had a score reliability estimated at .79. This is a relatively high reliability for direct assessment; moreover, the single essay being examined for validity was similar to the criterion essays and was scored in the same way. The correlation of .56 obtained is therefore not surprising nor is the fact that it is the highest of any of the correlations in Table 8.

The earliest study of this type reviewed was that of Huddleston (1954). For the 763 high school students studied two criterion variables—average high school English grade and an instructor's rating of their writing ability—were accessible. An essay score was the total of two judgments (content and style) of a sample of writing (approximately 150 words) made by each of two English teachers. This essay score was found to correlate .43 and .41 respectively with high school English grades and high school instructors' ratings of writing ability.

The concurrent validity comparisons in the Breland (1977) study were based on the criteria of high school rank (self-reported), high school English grades (self-reported), college freshman English grades (fall), and college freshman english grades (spring). The relationships between the essay pretests (administered in college English courses) and both high school rank and high school grades was .37. The relationship with college grade was much less, .23. The smaller correlation with college grades may have been a result of instructional influences, or to the probably lower reliability of English grades as compared to high school rank. In any event, some correlation with course grades would be expected because the essays were written toward the end of courses.

The Hackman and Johnson (1977) study, reported in Table 8, used high school GPA as the criterion and a holistic

Table 8. Studies Reporting Concurrent Correlations with Direct Measures of Writing Skill

Study and Setting	*N*	*Direct Measure of Writing Skill*	*Criterion Measure*	*Correlation*
Breland (1977)	799	Fall essay pretest	High school rank	.37
College	756	Fall essay pretest	Last high school English grade	.37
freshmen	878	Fall essay pretest	Fall English grade	.23
	491	Spring essay posttest	Spring English grade	.23
Breland (1983)	800	ECT holistic score	Last high school English grade	.20[a]
College applicants			High school rank	.18[a]
Coffman (1966) High school students	296	One essay scored by two readers	Four essays scored by four readers	.56
Hackman and Johnson (1977) Yale freshmen	36	Fall essay pretest	High school GPA	.20
Huddleston (1954)	763	Essay score	High school English grades	.43
High school students	763	Essay score	Instructors rating of writing ability	.41
Michael et al. (1980) College juniors	100 (first sample)	30-minute essay (expert readers)	Cumulative college GPA	.40
		30-minute essay (lay readers)		.36
	100 (second sample)	30-minute essay (expert readers)		.05
		30-minute essay (lay readers)		.06
Michael and Shaffer (1978)	687	45-minute essay	High school GPA	.15
High school students	656	In-class essay		.17
Median correlation				.23

[a] Median over four samples

score on a 40-minute essay read independently by two readers. The relatively low correlation of .20 may be related to restriction of range, because all subjects had been admitted to Yale University. Most had very good high school records.

The Michael and Shaffer (1978) investigation also used high school GPA as the criterion. The validity correlations reported, .15 and .17, are similar to the .20 figure reported by Hackman and Johnson for Yale students, even though the California State University and Colleges (CSUC) sample was not restricted in its range of abilities.

In the Michael et al. (1980) study, two random samples of approximately 100 college juniors each wrote 30-minute essays on two different topics. Each response was rated by both English professors (experts) and by professors in other departments (lay readers). Two of each type of reader read each essay, and the total score was obtained by adding the two ratings. The criterion measure was the cumulative GPA of each student up to the time of the investigation. For the first sample, the observed correlations between reader scores and GPA were better (.41 and .40) than those for the second sample (.05 and .06). Small differences in reliabilities of ratings favored the expert readers, but these differences were not considered important ones. The main differences between the first sample and the second sample data were in the writing tasks, though the details of the tasks used were not reported. It was suggested that the specific topic of an essay, or the specific writing task required, may have a substantial bearing on the validity of an assessment.

Predictive Validity

While the concurrent correlations just reviewed are predictive in a sense, the usual interest is in examining how well a measure predicts some event which occurs at a later time. In the case of writing skills, therefore, we want to demonstrate a relationship, for example, between a precourse test and a course grade, between a preadmission test and GPA after admission, or between writing skill as assessed at one time and writing skill as assessed at a later time. Table 9 presents results from four studies that have reported such relationships.

The Breland (1977) and Michael and Shaffer (1978) studies, reviewed earlier for concurrent correlations, also examined data on student English course grades and on writing samples collected toward the end of courses. The Werts et al. (1980) article represented a refinement of the same data of the Breland (1977) study through analyses of a complete but smaller data sample. As in the concurrent correlations of Table 8, the relationships between writing sample scores obtained at different times are higher than relationships between writing sample scores and later course grades. The direct measures correlate with each other about at the level of their score reliability (about .50 in this case), but they are not highly predictive of performance either in English courses or overall.

Incremental Validity

Because of the expense of direct assessments of writing skill, a central issue over the years has been whether or not an essay adds significantly to the measurement accuracy provided by other available measures—the high school record, objective test scores, or other information. Despite the importance of this issue, it has not often been examined. Table 10 gives the results from five studies that have in some way provided useful evidence.

The Breland and Gaynor (1979) study considered the effect of adding an essay when already available were high school rank (self-reported), last high school English grade (self-reported), SAT-verbal score, and TSWE score. Two criteria were used: freshman English composition course grade and a postcourse essay assessment consisting of the sum of scores received on essays written toward the end of both the fall and spring semesters. The grade criterion was examined within each of four colleges; the essay criterion was examined for all four colleges combined. Significant beta weights were obtained for the essay pretest in all four colleges combined when the essay criterion was used. The

Table 9. Studies Reporting Predictive Correlations for Direct Measures of Writing Skill

Study and Setting	*N*	*Predictive Measure*	*Criterion Measure*	*Correlation*
Breland (1977)	886	Fall essay pretest	Fall English grade	.28
Four colleges	400	Fall essay pretest	Spring English grade	.26
	904	Fall essay pretest	Fall essay posttest	.52
	316	Fall essay pretest	Spring essay posttest	.51
Michael and Shaffer (1978)	1, 36	EPT essay	Fall GPA	.21
California State	637	EPT essay	Fall English grade	.31
University, Northridge	657	In-class essay	Fall GPA	.25
	604		Fall English grade	.32
Werts et al. (1980)	234	Fall essay pretest	Fall essay posttest	.56
		Fall essay pretest	Spring essay posttest	.57

Table 10. Studies Reporting Incremental Validity Evidence for Direct Measures

Study	*N*	*Criterion*	*Predictors*	*r*	*beta*	*R*	*Incremental R (direct)*
Breland and Gaynor (1979	76	Freshman English	HS rank	—	.10		
College freshmen	(College A)	course grades	HS English grade	—	.17	.39	.04
			SAT-V	—	.00		
			TSWE	—	.10		
			Essay pretest	—	.20		
	160		HS rank	—	.04	.43	.04
	(College B)		HS English grade	—	.28		
			SAT-V	—	.00		
			TSWE	-	.05		
			Essay pretest	—	.22		
	204		HS rank	—	.20	.51	.03
	(College C)		HS English grade	—	.00		
			SAT-V	—	.25		
			TSWE	—	.03		
			Essay pretest	—	.22		
	135		HS rank	—	.25	.50	.02
	(College D)		HS English grade	—	.13		
			SAT-V	—	.00		
			TSWE	—	.13		
			Essay pretest	—	.19		
	213	Postcourse essay	HS rank	—	.11	.76	.05
	(Four	assessment	HS English grade	—	.09		
	colleges)		SAT-V	—	.16		
			TSWE	—	.22		
			Essay pretest	—	.38		
Checketts and Christensen (1974)							
CLEP examinees	123	Freshman English GPA	CLEP objective	—	—	.53	.06
			CLEP essay	—	—		
Godshalk et al. (1966)							
High school students	237	Four brief essays,	PSAT-V sentence	.69	.28	.77	.02[b]
		each read 5 times	Correction prose groups	.67	.27		
			Essay A (2 readings)	.56	.13		
				.55	.26		
	254		PSAT-V sentence	.63	.20	.75	.03[b]
			Correction prose groups	.68	.36		
			Essay B (2 readings)	.56	.15		
				.49	.23		
Huddleston (1954)							
High school students	420	Average English grade	Objective English	.60	.18	.80	—
			Essay-content	.26	.02		
			Essay-style	.39	.10		
			Paragraph A	.29	.03		
			Paragraph B	.33	.08		
			Verbal test	.77	.58		
		Instructor's rating	Objective English	.58	.16	.79	—
		of writing ability	Essay content	.22	-.03		
			Essay-style	.39	.13		
			Paragraph A	.26	.00		
			Paragraph B	.33	.09		
			Verbal test	.76	.60		
	763	Average English grade	Objective English	.34	—	.56	.07[a]
			Two essay ratings	.43	—		
		Instructor's rating	Two paragraph ratings	.34	—	.56	.05[a]
		of writing ability	Two essay ratings	.41	—		

Note: A dash (—) indicates information not reported.

[a] Includes two essays and two paragraph ratings.

[b] This increment is based on a comparison with prediction by four objective tests. Actually, one objective test was replaced by an essay in conducting the study. Consequently, the increment attributable to the essay is slightly larger than the figure reported here, but the precise amount is unknown.

Table 10. **Studies Reporting Incremental Validity Evidence for Direct Measures (Continued)**

Study	*N*	*Criterion*	*Predictors*	*r*	*beta*	*R*	*Incremental R (direct)*
Michael and Shaffer (1978) College freshmen	1583	Fall GPA	HS GPA	—	.25	.38	—
			EPT-reading	—	.11		
			EPT-essay	—	.09		
			EPT-sent. constr.	—	ns		
			EPT-logic & org.	—	.06		
	637	Freshman English grade	HS GPA	—	.23	.48	—
			EPT-reading	—	.15		
			EPT-essay	—	.12		
			EPT-sent. constr.	—	.18		
			EPT-logic & org.	—	ns		

Note: A dash (—) indicates information not reported.

[a] Includes two essays and two paragraph ratings.

[b] This increment is based on a comparison with prediction by four objective tests. Actually, one objective test was replaced by an essay in conducting the study. Consequently, the increment attributable to the essay is slightly larger than the figure reported here, but the precise amount is unknown.

average increment in the multiple correlation, attributable to the essay, was about .04.

Checketts and Christensen (1974) studied the CLEP objective and essay components and obtained an increment in the multiple correlation predicting a freshman English average of .06 owing to the essay. The CLEP essay and objective components are each 90 minutes in length—so the results are not precisely comparable to the more common 20-minute essay and somewhat shorter objective component. But the similarity of the .06 increment to the .04 increment indicated in the Breland and Gaynor study would suggest that not a great deal is gained by the longer essay.

The Godshalk et al., (1966) study has been cited on a number of occasions in this report. The incremental validity evidence reported in Table 10 was developed in a special field trial in which four of the five essays used were criteria and the fifth was a predictor. Two different essay topics were used as predictors, Essay A and Essay B. The criterion thus excluded either Essay A or Essay B. As noted in Table 10, the incremental *R* observable in the Godshalk et al. study was the difference between the *R* obtainable from four objective predictor tests and the *R* obtained when one of the four objective tests was replaced by an essay test. Thus the incremental *R* shown is attenuated by some unknown amount. Another possible comparison is between an objective test prediction using three objective tests of composition (but excluding the PSAT-verbal) and the Table 10 multiple *R*s. Such a comparison tends to artificially inflate the increment, but the values obtained are .05 and .04 respectively for essays A and B. The true increment lies between these figures and those shown in Table 10.

The Huddleston (1954) study reported that a verbal test (essentially the SAT-verbal) accounted for practically all of the variance in both of these criteria—average high school English grade and high school instructors' ratings of writing ability. A multiple correlation of .80 was obtained for the prediction of average high school English grades from an essay (rated for both content and style), two paragraph revision exercises, an objective test of English, and the verbal test. But the verbal test alone correlated .77 with the criterion, indicating that all other variables including the essay test, the paragraph revision exercises, and the objective test of English added little (.03) to the prediction. A similar result was obtained when the criterion was instructor rating of writing ability. The essay style rating contributed more to the prediction than the content rating, suggesting that content was less reliably assessed.

The final study of Table 10, that of Michael and Shaffer (1978), also used two criteria. The first criterion was fall semester GPA and the second, grades in a freshman English course. Significant beta weights were obtained for the 40-minute EPT Essay (scored by two readers) for both criteria. Incremental multiple correlations comparable to other studies in Table 10 were not reported, but some were. For example, the summation of the EPT composition components (sentence construction, logic and organization, and the essay) predicted fall semester GPA with an $r = .29$, whereas sentence construction correlated .27 and logic and organization .26 with the same criterion. For predicting grades, the summation of the three composition scores produced a correlation of .41, whereas sentence construction and logic and organization correlated respectively .38 and .33 with the criterion.

Validity of Analytic Subscores

Recent interest in diagnosis calls for an examination of validity evidence reported for analytic subscores in direct assessments. Although analytic scales are often used in developing scores for direct assessments, data are not often reported for them. Some reliability data for analytic subscales were described previously in Table 7. Table 11 summarizes three investigations in which some kind of correlational validity evidence was reported for an analytic subscore. The studies by Hackman and Johnson (1977) and Huddleston (1954) are in some senses similar because of the high school grade criterion and the types of subscales used. In both *style* appears to be a more valid subscore than *content* (*thought* in Hackman and Johnson). However, *grammar* in the Huddleston study had the highest validity ($r = .49$ with instructor rating of writing ability). The generally lower correlations in the Hackman and Johnson study are probably attributable to the select sample (Yale freshmen) being studied.

The Breland (1983) data also show slightly higher validities for grammatical types of ratings as opposed to higher order skills. For both criteria, a syntactic quality rating and a lexical quality rating yielded higher correlations than a rating on discourse quality. The discourse quality rating reflected qualities similar to the organization, thought, and content subscores reported for other studies in Table 11. Despite the importance that most observers, including members of the English teaching profession, place on discourse, thought, content, organization, and similar qualities, the validity evidence shown in Table 11 favors the more mundane skills.

Construct Validity

Quellmalz et al. (1982) have recently revived issues of construct validity in writing skill assessment. One construct validity issue was the long-standing question of whether direct and indirect assessments both measure a unitary trait that is not easily divisible. Most past research has concluded that direct and indirect assessments are highly correlated, even if it could not be demonstrated conclusively that they measured the same underlying trait (Huddleston 1954; Breland and Gaynor 1979; Coffman 1966; Werts et al. 1980). While Quellmalz et al. were not able to answer the question unequivocally, their results indicated that indirect assessments, as well as different types of direct assessments, measure different skill constructs. In particular, discourse mode (for example, expository, narrative) and response mode (production vs. recognition) were suggested as influences on the assessment. The study also compared analytic judgments of essays with objective assessments of

Table 11. Validity Evidence for Analytic Subscores

Study and Setting	*N*	*Subscore*	*Criterion Measure*	*Correlation*
Breland (1983)	800	Discourse quality	Last high school	.19 (.20)[a]
Random samples		Syntactic quality	English grade	.26
of ECT-takers		Lexical quality		.24
		Analytic total		.26
		Discourse quality	High school rank	.17 (.18)[a]
		Syntactic quality		.21
		Lexical quality		.20
		Analytic total		.22
Hackman and Johnson	173	Mechanics (subsentence)	High school grade average	.20
(1977)		Mechanics (sentence)		.22
Yale college freshmen		Organization		.19
		Thought		.19
		Style		.27
Huddleston (1954)	294	Punctuation	High school English grades	.25
High school students		Idiom		.21
		Grammar		.33
		Sentence structure		.33
		Punctuation	Instructor rating of	.29
		Idiom	writing ability	.22
		Grammar		.49
		Sentence structure		.36
	763	Content	High school English grades	.28
		Style		.40
		Content	Instructors' ratings of	.24
		Style	writing ability	.39

[a] The discourse ratings were similar in emphases to the holistic ratings made for the ECT administration. Correlations between the criterion and the ECT holistic ratings are given in parentheses.

parallel skills—focus, organization, support, and mechanics. Their analyses indicated that focus and organization defined a single factor (termed *coherence*), but that support and mechanics were distinct factors measurable by both direct and indirect methods.

Content Validity

No analyses of the content validity of direct assessments were encountered in the present review. The analyses of Quellmalz et al. (1982) touched on content validity, however, since different modes of discourse were examined. In that study, students who scored high on narrative tasks were not the same students who scored high on expository tasks. These results suggest that content sampling is important in direct assessments. Beyond the influence of discourse mode, the specific topic of the direct assessment may have additional influences. All students do not have equivalent knowledge about all topics. Direct assessments in which a single topic and a single discourse mode are used clearly are limited in content validity.

Summary of Validity Evidence

Evidence in support of the validity of direct assessments of writing skills is available from several perspectives. Concurrent correlations with high school rank, high school English grades, instructor rating of writing ability, and college GPA all showed statistically significant relationships, though these correlations were at times relatively low. Predictive correlations with college English grades and GPA were similar in magnitude, although also significant statistically. Incremental validity evidence was reported in a number of studies, showing that direct assessments of writing skill contribute information beyond that available through previous academic records and other kinds of test scores. In those few investigations reporting validity evidence for direct assessments of writing subskills, ratings of grammatical skills tended to yield slightly higher validity coefficients than ratings of content, discourse quality, or thought. The only type of validity evidence not located for direct assessments was evidence of content validity. Since only one writing task was often employed, content sampling from the domain of all possible writing tasks was of course severely limited.

TECHNOLOGICAL DEVELOPMENTS

Recent technological developments in text processing may afford an opportunity to improve direct assessments of writing skill. There is hope that the present impasse between the unreliability of the usual assessments and the labor intensiveness of more reliable and valid assessments can be broken by appropriate applications of technology. Past solutions to this dilemma have relied on multiple-choice assessments as a source of reliability and brief judgmental assessments as a source of validity. Few accept such a combination as the ultimate solution. Most multiple-choice assessments cover only a narrow range of the writing skill domain, and most judgmental assessments are made on one sample written in one mode of discourse. The limitations of current direct assessments are almost always a consequence of the labor intensiveness of better direct assessments.

The use of technology in writing assessment is not a new idea. In an extensive project conducted for the U.S. Office of Education, Page and Paulus (see Page 1966, 1968a, 1968b for summarizations of this work) developed techniques for scoring essays and for providing instructional feedback to students through computer analysis of essays. Indices were developed that predicted judgmental scores through a procedure adapted from Diederich's (1974) analytic scoring procedure. The computer was shown to be about as good a predictor of human judgments as human judges themselves. In view of the time that has now passed, however, the optimism expressed by Page (1966, 238) was clearly excessive: "We *will* soon be grading essays by computer, and this development *will* have astonishing impact on the educational world." [Page's emphasis] This statement was in error—at least with respect to the word *soon*.

One of the reasons Page's work did not catch on was the English profession's negative response (see, for example, Macrorie 1969). Although it has been pointed out that some of the negative reactions to Page's work miss the point (Slotnick and Knapp 1971; Slotnick 1972), that what is being studied are the cognitive processes of experienced English teachers, these assertions have not sufficed to revive the idea. The limitations of the technology of the late sixties and its consequent lack of availability also caused the idea of computer assessment of writing to founder at that time. Recent strides in microchip technology and widespread acceptance and use of text processing procedures have changed the context in which technology operates. Despite this changed context, most English teachers and most examinees would probably never accept a computer's judgment of the quality of a piece of writing. On the other hand, descriptions, counts, and other computer-generated information that is useful but not evaluative would likely be more acceptable.

An example of such descriptive information is provided by the Writer's Workbench program developed at Bell Laboratories (Frase 1980; Frase et al. 1981). The Writer's Workbench consists of a growing set of computer aids for editing and reformatting written documents. In addition to simple programs that check spelling and punctuation, included also are more complex routines that flag poor diction, weak phrases, and others that compute readability indices, compute the total number of unique words used, and compare a written piece with some standard piece written by a well-known writer. Frase et al.

(1981) have also written about the ethics of imperfect measures, such as readability indices. Because of the limitations of imperfect measures, they recommend the use of multiple measures, the use of relative rather than absolute evaluations, and the treatment of imperfect measures as information rather than decisions. Noting the failure of Page's idea to grade essays by computer, it is suggested that humans will never relinquish human judgment to imperfect measures and that this fact must be recognized by those who develop imperfect measures of writing skill.

A much more sophisticated text-critiquing system, EPISTLE, is currently under development at IBM (Heidorn et al. 1982). The EPISTLE system is more sophisticated than the Writer's Workbench because it uses a parser that breaks down sentences into component parts of speech and relates the form, function, and syntax of each part. By contrast, the Writer's Workbench is only a collection of programs that identify characteristics of writing. A parse tree of a sentence can show for example, that the distance between a subject and verb is too great. EPISTLE also performs paragraph-level critiques such as noting that there are too many passive sentences or too many compound or complex sentences. Heidorn et al. emphasize, however, that EPISTLE is still in the experimental stage.

There are also writing computer assessment activities under way in the academic setting. One well-developed computer-assisted instructional program is JOURNALISM (Bishop 1974). JOURNALISM performs stylistic analysis by reporting variety in sentence length and overuse of articles, passives, adjectives, and adverbs. It also checks spelling and keeps students' records of progress. Another academic approach to the writing assessment problem is that of Finn (1977). Finn's approach is to focus only on word choices and to relate those to standard frequency counts to develop an index of writing maturity. Moe (1980) describes programs that count words and word strings of various types, analyze sentences, and estimate readability.

A fundamental notion, that of an automated dictionary, was brought to the attention of the National Institute of Education in 1978, and later a conference was held (Miller 1979). Since that time software companies have developed automated dictionaries that function in consort with proofreading programs. These kinds of developments are likely to proceed rapidly if recent history is any guide.

SUMMARY AND CONCLUSIONS

The history of direct writing skill assessment is dominated by the issue of reliability. Specifically, the issue is the limited reliability of the usual kind of direct assessment in which an examinee produces a sample of writing on some topic during a limited time period, and that sample is then evaluated by one or more judges. As simple and straightforward as such procedures seem on the surface, the fact is that they are not simple at all. Much has been written about the inconsistency of the judgments of writing samples by English teachers and others. But there has been little examination of other kinds of limitations in the usual writing assessments. One important limitation not often examined has to do with the degree of content sampling usually conducted.

The sampling domain for direct assessments reflects all possible types of stimuli (written, pictorial, aural, for example) and all possible modes of discourse (narrative, expressive, argumentative, for example). For each combination of s. nulus and discourse mode, different contexts for writing occur. How much time is allowed for the writing? What reference materials, if any, are allowed? What is the purpose of the writing? Who is the audience? When one adds the context variables to the different stimulus types and the different modes of discourse, the domain from which any particular writing sample is drawn is extensive indeed. Because the usual writing sample represents only one kind of stimulus, only one mode of discourse, and only one context, it is a small sample of the possible domain of tasks that might be used to assess writing skill. Since some examinees are likely to perform better at some tasks than at others, the use of only a limited sample from the domain will result in errors in the assessment. These errors, in addition to the errors introduced by reader inconsistency, make reliable direct assessment difficult to attain.

Reliabilities of essay assessments can be made acceptable, of course, through the use of expensive multiple-topic, multiple-mode, and multiple-reader procedures—as the evidence presented shows. Consequently, there is nothing inherently unreliable about the general approach. It is probably true, nevertheless, that student behavior in producing writing samples is less consistent than it is for more structured tests. There are more choices to make, more consequences of poor choices, and there is less control over the order of responses. As a result, it is difficult to attain very high reliabilities when these inconsistencies are coupled with those of readers making judgments of the samples.

It has been effectively argued (Coffman 1966) that direct assessments of writing skill can be valid even if reliability is often a problem. To the degree that they relate to actual performance in English composition courses or to more extensive assessment of writing performance, direct assessments are valid measures. And substantial relationships with course performance have been reported. Moreover, direct assessments have been shown to contribute, incrementally, beyond the prediction possible using past academic performance and objective test scores. Therefore, it is difficult to argue that direct assessments of writing skill are not valid. Such validity could be increased, however, by improvements in the reliabilities of direct assessments.

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.

An important but seldom examined validity issue is concerned with the purposes of testing. If the purpose is to rank students, the direct assessments with holistic scoring are clearly valid for that purpose. But if the interest is in specific strengths and weaknesses of a student's writing for use as instructional feedback, a holistically scored essay is at best a blunt instrument. Analytic scoring may not be much better when the writing samples obtained represent only a very small proportion of the domain of possible samples. Therefore, the validity of direct assessments for diagnostic and instructional purposes can easily be questioned despite the obvious instructional utility of commentary on one's writing.

Validity in direct assessments has also been questioned with respect to issues of test bias, but no evidence on this issue was available. A specific question is whether judges discriminate against minorities and others who speak dialects and other languages. Hoover and Politzer (1981), for example, observe that impressionistic judgments of the writing of speakers of dialects may be biased because the judge may react primarily to less important subskills (such as punctuation and grammar) and fail to note that other more important goals of the essay were achieved. The rating of subskills is suggested to minimize bias effects. Such subskill rating would increase the time required by judges.

Recent analyses suggest technology offers some promise as a means of relieving the labor intensiveness of direct assessment. Implementation of technology is not without problems, however, because some procedure is first needed for entering the sample into the computer and because the types of analyses that can be performed by computer are limited. Obviously, word processing is not quite the same thing as composition. Nonetheless, some aspects of good and bad writing can probably be evaluated by an appropriately programmed word processor.

One must conclude, first, that writing skill is inherently difficult to assess accurately. While direct assessment accuracy is limited by rater inconsistencies and domain sampling problems, the indirect assessment of writing skill has other limitations. A second conclusion that is unavoidable is that assessment is labor intensive, expensive, and cumbersome. A means has yet to be devised that significantly relieves this efficiency problem, though computers may represent a potential long-term solution. Faced with the present dilemma of either excessively high costs or low reliability, a solution is not easily found. Worse, some assessments in current use have high costs *and* low reliability.

It may be that writing is simply too complex a skill to be measured completely. An approach that avoids some difficulties is to focus on specific support skills that are usually necessary but not sufficient for effective writing. Knowledge of the rules of syntax, lexical knowledge, and spelling of course come to mind. But it is probably also possible to assess better than we now do other more advanced skills like organizational skills, coherence skills, transition skills, and skills of revision and editing.

Interestingly, it is this approach toward the assessment of specific skills that an English professor recently arrived at after facing some of the same problems described above. Matalene (1982) chronicles the experience of an English professor who became director of freshmen English at a large state university. After struggling with the complex political issues surrounding an exit examination, she decided to develop a test of her own—with the assistance of English department faculty members. An early step was a survey of English professors and teaching assistants.

With the survey as the basis, a revision and editing test was developed which consists of 30 items divided into two parts: the first 15 questions deal with units larger than a sentence, the last 15 questions are on how to improve sentences or groups of sentences. The test is printed with the entire essay on one page of the test booklet. Questions ask students to

- discover thesis and topic sentences;
- judge level of language, voice, coherence, logic, and diction
- discern methods of development, errors of logic, unstated assumptions, sentence variety, patterns of errors, effectiveness of examples; and
- offer suggestions for revision.

Following administrative and computer scoring of the test, each teacher receives a printout of his or her class which shows each student's answer to each question. After an extensive trial period, the test has now been made a requirement for completion of freshman English.

What seems important in this example is that the test developed is not a test of writing skill per se but a revision and editing test, even though everyone (and certainly every English teacher) knows that there is more to writing than revising and editing. To the English professors and teaching assistants in this one university, however, these were the *most important* support skills. And a successful measure of these support skills was developed and is now in use.

The example test described is not a direct assessment of writing skill, nor is it an indirect assessment of writing skill. It is a direct assessment of revision and editing skills. Similar direct assessments of other writing support skills would also seem to be possible. Direct assessments of written organization skills, of thesis statements, of methods of thesis development, and of the use of supporting evidence would also seem possible. Thus, the direct assessment of writing support skills represents one possible approach to the dilemma described earlier.

Research into the assessment of these higher-level support skills is recommended. Also recommended is research of the following types:

1. The development of a comprehensive criterion measure based on multiple writing samples written in different modes of discourse with each carefully evaluated by multiple judges on more than a single dimension. While in some senses similar to the Godshalk et al. (1966) study, an effort to develop a new writing criterion would benefit from more recent research on writing skill development. Furthermore, the new criterion could be used to evaluate new assessments of writing support skills. The collection of data on writing support skills using instruments such as those of Matalene (1982) and others and the analysis of such skills in relation to the overall variance in a comprehensive criterion would be especially useful.

2. The conduct of confirmatory factor analyses as well as other kinds of analyses to examine the construct validity of the measures available as contrasted with the validity of new prototype measures.

3. The analysis of judgmental assessments in conjunction with automated assessments to determine in what ways these two approaches might be *combined* to optimize efficiency, reliability, and validity.

4. The exploration of more efficient means for obtaining human judgments of written products. Such efficiency may be obtainable through the mail (particularly electronic mail) if appropriate quality control procedures are implemented at the same time.

5. Since practicality usually dictates that only limited samples of an examinee's writing be taken, it would be important to examine what specific *kinds* of tasks elicit the most reliable and valid information. While persuasive/argumentative tasks may be preferred by English teachers, for example, they may be so difficult as to preclude much writing by many students. A comparative validity examination of task types would be valuable.

6. Equating of direct assessments of writing is inherently difficult because tasks vary in difficulty. This problem is usually handled through the use of multiple-choice measures as anchors. If the examinee is allowed a choice of topics, the problem of equating is even more difficult. A useful investigation would explore equating issues as they relate to task types, choice of tasks by the examinee, and optimum methods for weighting direct and indirect components.

7. Bias in judgments of essays may be influenced by methods used, as has been suggested by Hoover and Politzer (1982). An examination of holistic as opposed to analytic ratings for different dialect and linguistic groups would provide a better understanding of this issue.

REFERENCES

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

Bishop, R. L. 1974. Computing in the Teaching of Journalistic Skills. *On-Line* 3 (3), 5–12.

Bracewell, R. J., Frederiksen, C. H., and Frederiksen, J. D. 1982. Cognitive Processes in Composing and Comprehending. *Educational Psychologist* 17 (3).

Breland, H. M. 1977. *A Study of College English Placement and the Test of Standard Written English* (College Board Research and Development Report RDR-76-77-4). Princeton, N.J.: The College Entrance Examination Board.

Breland, H. M. 1983. Linear Models of Writing Assessments. ETS Research Report, Educational Testing Service, Princeton, N.J.

Breland, H.M. and Gaynor, J.L. 1979. A Comparison of Direct and Indirect Assessments of Writing Skill. *Journal of Educational Measurement* 16, 119–128.

Bruce, B., Collins, A., Rubin, A.D., and Gentner, D. 1982. Three Perspectives on Writing. *Educational Psychologist* 17 (3).

Checketts, K.T., and Christensen, M.G. 1974. The Validity of Awarding Credit by Examination in English Composition. *Educational and Psychological Measurement* 34, 357–361.

Clemson, E. 1978. *A Study of the Basic Skills Assessment Direct and Indirect Measures of Writing Ability*. Princeton, N.J.: Basic Skills Assessment Program, Educational Testing Service.

Coffman, W.E. 1966. On the Validity of Essay Tests of Achievement. *Journal of Educational Measurement* 3, 151–156.

Coffman, W.E. 1971a. Essay Examinations. In *Educational Measurement*. 2d ed. R.L. Thorndike, ed. Washington, D.C.: American Council on Education.

Coffman, W.E. 1971b. On the Reliability of Ratings of Essay Examinations in English. *Research in the Testing of English* 5 (1).

Conry, R., and Jeroski, S. 1980. The British Columbia Assessment of Written Expression. Report prepared for the Learning Assessment Branch of the Ministry of Education.

Cooper, C.R. 1977. Holistic Evaluation of Writing. In *Evaluating Writing: Describing, Measuring, Judging*. C.R. Cooper and L. Odell, eds. Urbana, Ill. National Council of Teachers of English.

Coward, A.F. 1952. A Comparison of Two Methods of Grading English Compositions. *Journal of Educational Research* 46, 81–93.

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

Diederich, P.B., and Link, F.R. 1967. Cooperative Evaluation in English. In *Evaluation as Feedback and Guide*. F.T. Wilhelms, ed. Washington, D.C.: Association for Supervision and Curriculum Development, 181–231.

Educational Testing Service. 1982. *Testing Analysis of College Board Achievement Examinations English Composition Test with Essay 3DBE Literature Test 3DAC December 1981 Administration* (Report No. SR-82-68).

Finlayson, D.S. 1951. The Reliability of the Marking of Essays. *British Journal of Educational Psychology* 21, 126–134.

Finn, P.J. 1977. Computer-Aided Description of Mature Word Choices in Writing. In *Evaluating Writing: Describing, Measuring, Judging*. C.R. Cooper and L. Odell, eds. Urbana, Ill.: National Council of Teachers of English.

Follman, J.C., and Anderson, J.A. 1967. An Investigation of the Reliability of Five Procedures for Grading English Themes. *Research in Teaching of English* 1, 190–200.

Frase, L.T. 1980. *Writer's Workbench: Computer Supports for Writing and Text Design*. Paper presented at the annual American Educational Research Association Meeting, Boston.

Frase, L.T. 1981. Ethics of Imperfect Measures. *IEEE Transactions on Professional Communication*, Vol. PC-24, No. 3.

Frase, L.T., Gingrich, P.S., and Keenan, S.A. 1981. *Computer Content Analysis and Writing Instruction*. Paper presented at the annual American Educational Research Association Meeting, Los Angeles.

Frase, L.T., MacDonald, N.H., Gingrich, P.S., Keenan, S.A., and Collymore, J.L. 1981. Computer Aids for Text Assessment and Writing Instruction. *NSPI Journal*, 21–25.

Frederiksen, C.H., Frederiksen, J.B., and Bracewell, R. 1983. *Discourse Analysis of Children's Written and Oral Production*. Paper presented at the annual American Educational Research Association Meeting, Montreal.

French, J.W. 1961. *Schools of Thought in Judging Excellence of English Themes*. Reprint from the Proceedings of Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1962.

Godshalk, F.I., Swineford, F., and Coffman, W.E. 1966. *The Measurement of Writing Ability*. New York: College Entrance Examination Board.

Gulliksen, H. 1950. *Theory of Mental Tests*. New York: Wiley.

Hackman, J.D., and Johnson, P. 1977. How Well Do Freshmen Write? Implications for Placement and Pedagogy. *College and University*. 81–99.

Heidorn, G.E., Jensen, K., Miller, L.A., Byrd, R.J., and Chodorow, M.S. 1982. The EPISTLE Text-Critiquing System. *IBM Systems Journal* 21 (3).

Hirsch, E.D., and Harrington, D.P. 1981. Measuring the Communicative Effectiveness of Prose. In *Writing: The Nature, Development, and Teaching of Written Communication*, Vol. 2. C.H. Frederiksen and J.F. Dominic, eds. Hillsdale, N.J.: Lawrence Earlbaum Associates.

Hoover, M.R., and Politzer, R.L. 1981. Bias in Composition Tests with Suggestions for a Culturally Appropriate Assessment Technique. In *Writing: The Nature, Development, and Teaching of Written Communication, Vol. 1, Variation in Writing: Functional and Linguistic-Culture Differences*. M.F. Whiteman, ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hopkins, T.L. 1921. The Marking System of the College Entrance Examination Board. *Harvard Monographs in Education, Series 1, No. 2*. Cambridge, Mass.: The Graduate School of Education, Harvard University.

Huddleston, E. 1954. Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Testing Techniques. *Journal of Experimental Psychology* 22, 165–213.

Hunt, K. 1977. Early Blooming and Late Blooming Syntactic Structures. In *Evaluating Writing: Describing, Measuring, Judging*. C.R. Cooper and L. Odell, eds. Urbana, Ill.: National Council of Teachers of English.

Kay, P.M., and Sabban, Y. 1983. *The Language Skills Structure of a Group of College Level Remedial Writing Students: Further Evidence on the Distinction Between Essay and Objective Tests*. Paper presented at the annual American Educational Research Association meeting, Montreal.

Lloyd-Jones, R. 1977. Primary Trait Scoring. In *Evaluating Writing: Describing, Measuring, Judging*. C.R. Cooper and L. Odell, eds. Urbana, Ill.: National Council of Teachers of English.

MacDonald, N.H., Frase, L.T., Gingrich, P.S., and Keenan, S.A. 1982. The Writer's Workbench: Computer Aids for Text Analysis. *Educational Psychologist* 17, (3).

Macrorie, K. 1969. Roundtable Review. *Research in the Teaching of English* 3, 228–236.

Markham, L.R. 1976. Influences of Handwriting Quality on Teacher Evaluation of Written Work. *American Educational Research Journal* 13, 227–283.

Matalene, C.B. 1982. Objective Testing: Politics, Problems, Possibilities. *College English* 44 (4).

Michael, W.B., and Shaffer, P. 1978. The Comparative Validity of the California State University and Colleges English Placement Test (CSUC-EPT) in the Prediction of Fall Semester Grade-Point Average and English Course Grades of First Semester Entering Freshmen. *Educational and Psychological Measurement* 38.

Michael, W.B., Cooper, T., Shaffer, P., and Wallis, E. 1980. A Comparison of the Reliability and Validity of Ratings of Student Performance on Essay Examinations by Professors of English and by Professors in Other Disciplines. *Educational and Psychological Measurement* 40, 83–195.

Miller, G.A. 1979. Automated Dictionaries, Reading and Writing. Chairman's Report of a Conference on Educational Uses of Word Processors with Dictionaries, National Institute of Education.

Moe, A.J. 1980. Analyzing Text with Computers. *Educational Technology* (July) 29–31.

Moss, P.A., Cole, N.S., and Khampalikit, C. 1982. A Comparison of Procedures to Assess Written Language Skills at Grades 4, 7, and 10. *Journal of Educational Measurement* 19 (1), 37–47.

Mullis, I.V.S. 1980. *Using the Primary Trait System for Evaluating Writing* (Report No. 10-W-51). Denver, Col.: National Assessment of Educational Progress, Education Commission of the States.

Myers, M. 1980. *A Procedure for Writing Assessment and Holistic Scoring*. Urbana, Ill.: National Council of Teachers of English.

Myers, A.E., McConville, C.B., and Coffman, W.E. 1966. Simplex Solution in the Grading of Essay Tests. *Educational and Psychological Measurement* 26, 41–54.

Noyes, E.S., Sale, W.M., and Stalnaker, J.M. 1945. *Report on the First Six Tests in English Composition*. New York: College Entrance Examination Board.

Odell, L. 1981. Defining and Assessing Competence in Writing. In *The Nature and Measurement of Competency in English*. C.R. Cooper, ed. Urbana, Ill.: NCTE.

Page, E.B. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan* 47, 238–243.

Page, E.B. 1968a. *The Analysis of Essays by Computer* (Final Report, U.S. Office of Education Project 6-1318). Storrs, Conn.: University of Connecticut.

Page, E.B. 1968b. The Use of the Computer in Analyzing Student Essays. *International Review of Education* 14, 210–225.

Powills, J.A., Bowers, R., and Conlan, G. 1979. *Holistic Essay Scoring: An Application of a Model for the Evaluation of Writing Ability and the Measurement of Growth in Writing Ability Over Time*. Paper presented at the annual American Educational Research Association meeting, San Francisco.

Quellmalz, E.S., Capell, F.J., and Chou, C. 1982. Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement* 19 (4).

Sheppard, E.M. 1929. The Effect of Quality of Penmanship on Grades. *Journal of Educational Research* 19, 102–105.

Slotnick, H.B. 1972. Toward a Theory of Computer Essay Grading. *Journal of Educational Measurement* 9, 253–263.

Slotnick, H.B., and Knapp, J.V. 1971. Essay Grading by Computer: A Laboratory Phenomenon? *English Journal* 60, 75–77.

Stalnaker, J.M. 1936. The Problem of the English Examination. *Educational Record* 17 (Suppl. 10).

Steele, J. 1979. *The Assessment of Writing Proficiency Via Qualitative Ratings of Writing Samples*. Paper presented at the annual Meeting of the National Council on Measurement in Education.

Traxler, A.E., and Anderson, H.A. 1935. Reliability of an Essay Test in English. *School Review*. 43, 534–540.

Weiss, D., and Jackson, R. 1982. *The Validity of the Descriptive Tests of Language Skills* (Unpublished report). Princeton, N.J.: Educational Testing Service.

Werts, C.E., Breland, H.M., Grandy, J., and Rock, D.R. 1980. Using Longitudinal Data to Estimate Reliability in the Presence of Correlated Measurement Errors. *Educational and Psychological Measurement* 40, 19–29.