A Comparison of the Information Provided by Essay, Multiple-Choice, and Free-Response Writing Tests

Terry A. Ackerman
American College Testing Program

Philip L. Smith
University of Wisconsin-Milwaukee

This study investigated the similarity of information that is provided by direct and indirect methods of writing assessment. The skills required by each of these techniques provide a framework for a cognitive model of writing skills from which these procedures can be compared. It is suggested that practitioners interested in reliably measuring all aspects of the proposed writing process continuum, as characterized by this cognitive model, use both indirect and direct methods. *Index terms: Confirmatory factor analysis, Essay tests, Free-response tests, Multiple-choice tests, Writing assessment, Writing processes.*

An issue in ability and achievement testing that has gained increased attention in recent psychometric literature is the relative value of information gained from objective (multiple-choice) and free-response (including essay) testing methodologies. Critics of objective testing claim that the information that can be gained from multiple-choice tests is limited relative to that gained from free-response options. Others claim that the scoring reliability and validity of essay and other free-response formats is so poor as to outweigh any such advantage.

Research on the comparability of the construct validity of the two general classes of measures has been somewhat sparse and varied in terms of conclusions. Early work comparing multiple-choice tests with constructed-response tests (Davis & Fifer, 1959; Heim & Wats, 1967; Vernon, 1962) generally indicated that tests employing different formats cannot be expected to have the same means, standard deviations, and correlations with criterion variables. Some of these differences can be assumed to be due to changes in the scale of measurement and amount of error variance associated with each format. Thus the results of this earlier work do not necessarily imply lack of construct equivalence.

Traub and Fisher (1977) recognized these problems and employed methodology that equated scale parameters and error variances on three response formats for both verbal and quantitative measures. Two of these formats were multiple-choice and constructed-response. Using confirmatory factor analysis (CFA), Traub and Fisher found little evidence of a format effect for the mathematical reasoning items, and only weak evidence that the free-response and multiple-choice items were measuring different constructs for verbal comprehension items.

Ward, Frederiksen, and Carlson (1980), also using CFA, compared machine-scored and constructed-response forms of a test of ability to formulate scientific hypotheses. While their data were somewhat restricted and their analysis was more concerned with correlations of the resulting scores with personality and other cognitive variables, the Ward et al. results indicated that the two formats measure, at least in part, different constructs. In an additional study, Ward (1982) concluded that

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 2, June 1988, pp. 117–128

0146-6216/88/020117-12$1.85

for verbal aptitude items, various item formats produce much the same information and are essentially equivalent in terms of both the technical adequacy of the resulting measures and the construct interpretations of the resulting scores.

With the exception of a few isolated instances, the results of these studies seem to suggest that item format has little to do with the construct measured. Nonetheless, the debate regarding the equivalence of various item formats continues within various subject matter areas. Expressions of concern over multiple-choice versus free-response assessment formats have probably been greatest in the area of writing assessment. The present study used CFA to examine differences in student writing abilities using three response formats: multiple-choice, constructed-response, and essay.

While it might be expected that subject matter experts and psychometricians would agree that writing ability is best assessed using a direct method (e.g., an essay or free-response item) rather than an indirect method such as multiple-choice, the literature in this area is far from unequivocal. A study by Godshalk, Swineford, and Coffman (1966), using both essay and multiple-choice assessments of writing for 11th and 12th grade students, concluded that (1) "When objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove highly valid," and (2) "The most efficient *predictor* of a reliable direct measure of writing ability is one which includes both essay questions . . . and objective questions" (p. 41).

Although research has cited the predictive or concurrent validity of indirect writing measures (Breland, 1977; Coffman, 1971), these measures are regarded by a number of researchers to be weak in content, construct, and "ecological" validity (Braddock, Lloyd-Jones, & Schoer, 1963; Cooper & Odell, 1977). In essence, the critics of indirect assessment of writing, although they do concede that these measures probably adequately measure comprehension and editing abilities, claim that the tests make little or no attempt to measure unity, content, or organization because the examinee does no actual writing.

There is empirical evidence to suggest that indirect and direct measures of writing ability may, in fact, be measuring different types of abilities. Most attempts to investigate the relationship between scores resulting from varieties of direct and indirect writing assessment methodologies have been primarily correlational in nature, and have focused on rather limited definitions of the two assessment strategies. Several studies have produced results that show moderate correlations (ranging from .3 to .6) between standardized tests of language skills and holistically scored essays at both the high school and college levels (Breland, Conlon, & Rogosa, 1976; Breland & Gaynor, 1979; Coffman, 1966; Hogan & Mishler, 1980; Moss, Cole, & Khampalikit, 1982). One general conclusion that can be drawn from the results of these studies is that the two assessment methods may be assessing dissimilar skills and therefore are not of equal value in evaluating skills thought to comprise writing proficiency.

Although moderate correlations between direct and indirect methods of writing assessment are routinely reported in the literature, little objective evidence exists pertaining to why these correlations are not larger, or what unique information is provided by each of the two assessment methods. Although it has been argued that the most efficient and objective method of writing skill assessment is through the use of indirect methods (Breland, 1977), language arts experts and cognitive theorists point out that such methods fail to effectively assess many important aspects of the writing process (Braddock et al., 1963).

Despite this controversy, most users of both assessment formats assume that each measures the same construct(s). The purpose of the present investigation was to provide information regarding the unique skills and/or abilities measured by each approach. Using a conceptual framework for the writing process proposed by Hayes and Flower (1980), CFA was used with data obtained from 219 examinees on a multiple-choice, free-response, and essay assessment of writing ability.

The Conceptual Model

The basis for the analysis used in this study is a cognitive model of writing behavior first pro-

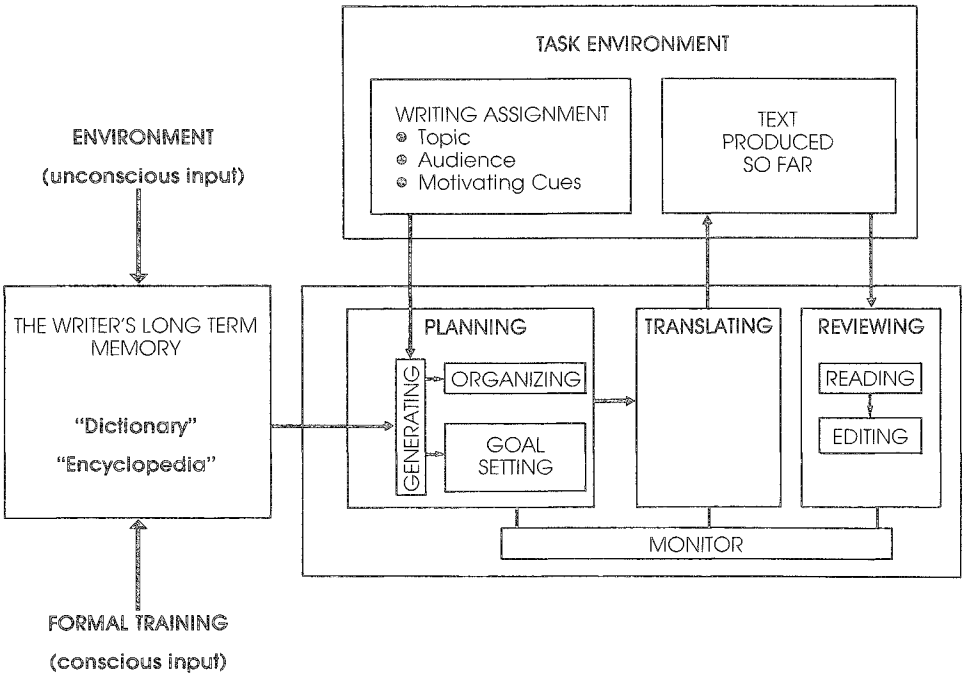
posed by Hayes and Flower (1980). This model provides a loose framework within which the writing process can be examined. Using this model, it can be hypothesized that a scoring discrepancy between direct and indirect methods of writing assessment should exist because of differences in the types of tasks required of the examinee. Specifically, discrepancies would be due largely to the *procedural* processes of knowledge generation and organization that are required in direct methods, but these processes would be less of a factor in indirect methods of writing assessment. Further, it can be hypothesized that the indirect method of assessment could be modified to include some of these procedural components and thus make the scores resulting from the two methods more comparable.

As a starting point in this analysis, features of an elaborated model of writing proposed by Hayes and Flower (1980) will be used to highlight the distinction between the tasks required of direct and indirect writing assessment strategies. Using protocol analysis, Hayes and Flower proposed a model of the writing process which identifies the procedures involved in creating written text. The model (Figure 1) includes three basic components: planning, translating, and reviewing. The model's components are both interactive and iterative. The main components are interdependent.

The *planning* component is composed of three subskills: generating ideas, organizing ideas, and goal setting. The *translating* process transforms the generated ideas into grammatically complete sentences. The function of the *reviewing* process is to monitor what has been written to check for mechanical errors and to improve the general quality of the text. Subsequent research has examined and verified the role of each of these components in the writing task. (For additional details on this model, see Ackerman, 1985.)

One elaboration of the Hayes and Flower model has been made. The writer's long-term memory has been more explicitly defined. Specifically, the long-term memory is believed to be composed of

Figure 1
A Modified Version of the Hayes and Flower (1980) Writing Model



two basic components: a dictionary and an encyclopedia (Clark & Clark, 1977) which are thought to be constantly changing and building through at least two identifiable sources. One source is acquired language: A person unconsciously increases his or her knowledge of language and the world through interaction with the environment (e.g., conversation and listening). The second source is formal instruction (i.e., the educational system), by which a person makes a conscious effort to acquire knowledge.

It is important to note that in defining the memory component of the model in this manner, formal training in grammar need not be a prerequisite for the writing process. Research has shown that students can write essays without being aware of the rules of grammar which they are applying. Research by Michaels (1981) and Scollon and Scollon (1981) has shown that through the learning of "story grammars," coherent and cohesive discourse can be learned before and during the writing process.

Using the Hayes and Flower (1980) model of writing as a basis, the model in Figure 2 illustrates the procedures or skills required of the examinee in a multiple-choice test. This modified model is proposed as a framework for examining the component differences between the processes involved in direct and indirect assessment of writing using CFA.

Although somewhat similar to the Hayes and Flower model, the modified model contains no components for planning or translating. The task of responding to a multiple-choice item is hypothesized to involve the reading and encoding of the content of the stem and alternatives into active memory, as illustrated in Figure 2. Once in active memory, the examinee edits the stem by comparing the alternatives with the schema of grammatical rules the examinee has previously learned, either through formal training or environmental interaction. Once a match occurs, the examinee selects the proper letter or number pertaining to the correct alternative.

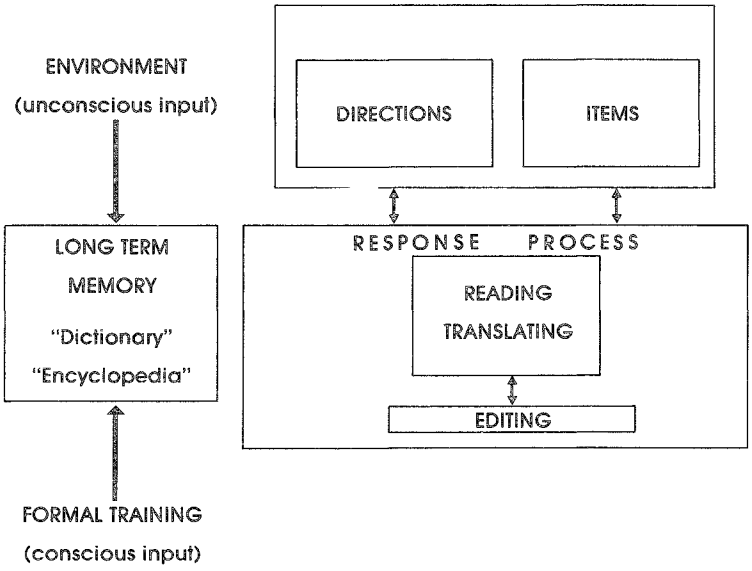
The reviewing process is the primary component of the objective test model. The multiple-choice format does not require the examinee to generate any text. In most multiple-choice tests the examinee is not required to link the test items in any coherent or cohesive fashion. Each of the items must be considered as an independent entity. (However, it may be argued that students respond not to the individual items but to response patterns or sets of previous items.)

The goal of the multiple-choice task is stated in the test directions or explained by the test administrator. There is no input required of the examinee with respect to the types of goals on which he or she may focus. These goals include identifying errors in usage, spelling, punctuation, or grammar. The examinee need not worry about tailoring the response to meet the needs of the reader or scorer. Another difference between the essay and multiple-choice models is that the latter does not require the rhetorical organization demanded by the various modes of written discourse (e.g., narration, persuasion, argumentation).

To summarize, the difference between writing an essay and answering a series of multiple-choice items can be considered in terms of the types and numbers of procedural skills that each task demands. The objective test item model requires only the editing and reading skills (i.e., primarily declarative knowledge) to select an appropriate solution. The writing task, as described above, demands the procedures of setting goals, generating information, organizing this information, imposing a grammatical framework on it, and then reviewing it for possible errors in meaning or structure; thus the task requires both declarative and procedural knowledge.

In an effort to explore the common information provided by various methods of writing assessment, the present study employed a design based on these two models. It was hypothesized that the procedural components of organizing, goal setting, and translating are unique to the "direct" writing task. By manipulating these processes, it should be possible to identify the specific factors associated with various assessment paradigms through CFA. Specifically, this study investigated whether indirect methods of writing assessment measure only the rudimentary skills of writing, which rep-

Figure 2
The Multiple-Choice Model for Responding to a Standardized Test Item



resent primarily *declarative* knowledge (i.e., the initial rules that are first learned in writing development).

This study proceeded from the expectation that the factor structure of the indirect method of assessment would differ from that of the direct method, because the direct method's scoring uniquely measures the integration of these subskills into *procedural* knowledge, that is, an ability that represents the final stage in the development of writing ability. Also of interest in this study was the question of whether the format of the indirect assessment could be modified to include the measurement of at least some of the aspects of procedural knowledge that most objective tests currently fail to measure.

Method

Examinees

The examinees used in the experiment were 219 10th-grade students from a parochial high school in southeastern Wisconsin. Students were randomly selected from traditional English classes.

Instruments

Three instruments were used in the study to elicit varying degrees of procedural knowledge on the part of the students. The first was a multiple-choice (MC) standardized achievement test published by Scott, Foresman. Two subtests, Language and Writing, from the Comprehensive Assessment Program were used to obtain six measures of writing ability: Spelling (SP), Capitalization/Punctuation (CP), Correct Expression (CE), Usage (US), Paragraph Development (PD), and Paragraph Structure (PS). The CE subtest measures the appropriate use of pronouns, verb form, and verb tense. The US subtest measures the knowledge of commonly misused word forms such as lay/lie, among/between, and stationary/stationery. PD tested the students' knowledge of paragraph transition, and the PS subtest questioned the examinee about the relationship among sentences within paragraphs.

A second measure was an essay task. The essay prompt asked students to "choose one or two ways in which you feel that television is of benefit to individuals, to families, or even to all of society.

Tell exactly what you consider the benefits of television to be and tell how television can improve or help people."

The third instrument was a free-response (FR) version of the standardized test. This instrument was identical to the standardized test except that it required the student to *generate* the correct answer instead of selecting the correct alternative. For example, in the FR test the student was asked not only to identify which word in each item was misspelled, but also to generate the correct spelling.

Procedure

Data collection took place over a five-week period in the fall of 1983. A total of 219 students were administered the Comprehensive Assessment Program battery. The Writing and Language portions of the MC test required approximately 50 minutes of testing time.

Approximately two weeks later, 192 students wrote essays. Students were given 50 minutes for this task. Classroom teachers had attended one organizational session to ensure uniform administration throughout the classrooms.

Two weeks after the essay task, students were administered the transformed test within their respective classrooms. Students were allowed 50 minutes to complete the test. Instructors indicated that the time period was sufficient; most students finished the test before the end of the 50-minute session.

The MC test was machine-scored. The essays were graded by six English teachers, none of whom taught at the target school. Three of the English teachers graded the essays holistically, using a 1 (poor) to 6 (excellent) scoring scheme. The other three readers graded the essays analytically, following the same set of objectives that were represented in the standardized subtests. All readers attended two training sessions in which examples of different types of errors or different levels of quality were clearly identified to ensure parallel grading. (Unknown to the readers, 11 of the essays were randomly chosen and duplicated and then distributed throughout the 192 essays. This was done to allow an estimate of intra-reader reliability.)

The FR test was hand-scored by a seventh English teacher. Each FR test was graded on the same six measures of writing ability as measured by the standardized test. Only the PS subtest required significant modification of the scoring criteria. In this subtest, students were required to arrange six sentences into a logical and meaningful order. One point was given for each correct identification of the first, second, third, and concluding sentences.

In addition, 1 point was awarded if the student could identify which of the given sentences did not belong in the paragraph. Unlike the MC test, the FR test was scored to give credit for each correct pairing of sentences. Thus it was possible to receive a maximum score of 8 on the FR test and only 5 on the MC test.

Analysis and Results

Reliability estimates were computed for each of the subscores from the MC test and the FR test, and for the analytic subscores from the essay. Internal consistency estimates for the standardized test ranged from .31 to .60; those for the FR test ranged from .71 to .88. Generalizability coefficients for the essay scores based on six readers ranged from .26 to .83.

CFA was used to determine whether the factor structure of the data collected in this analysis reasonably reflected that which would be predicted by the theoretical structure proposed earlier. Thus, factors were specified which represented the similar and unique procedures involved in each of the three types of writing assessment.

All analyses were done using LISREL IV (Jöreskog & Sörbom, 1978). Using LISREL, hypotheses about trait and method influences on observed scores corresponding to specified factor models can be estimated using the covariance structure. With the conceptual model used here, target factors can be specified to include specific variables, to be free of variables, or to be constrained. An overall chi-square test of the model's fit can be computed to determine the degree of match between the specified factor structure and the structure of the data.

The first analysis examined the fit of the MC and FR test subscore structures to the hypothetical model.

Because the FR version of the test had been modified to include a generation component corresponding to the procedural aspect thought to be required in this test, some confirmation was needed to determine whether this component existed. To verify this, an eight-factor CFA was imposed on all 12 subtest scores for the MC and FR tests. The first six factors were specified as trait factors (e.g., SP, CP, etc.). The seventh factor was hypothesized to be a "recognition (editing) factor" (REC) on which all 12 subtests were allowed to load, because all subtests were hypothesized by the model to contain this procedure. The eighth factor was targeted to be the "generation factor" (GEN) and only the six FR test subscores were free to load on it. The REC and GEN factor intercorrelations were fixed so they would not correlate with each other or with the six trait factors. The CFA solution is presented in Tables 1 and 2.

The $\chi^2(df = 24, N = 192)$ goodness of fit yielded a nonsignificant value of 20.4 ($p = .69$), indicating a reasonably good fit. Each of the MC subscores had a sizable weight on the recognition factor, with the exception of the PS subscore (.20). Only the subscores for the SP, CP, and CE subtests of the FR test weighed significantly on this factor. The generation factor is clearly dominated by the US subscore (.80), providing some indication that the subscore measure belongs to the generation factor specified in the three-factor solution. Intercorrelations of the trait factors (upper triangle of Table 2) suggest that measures for CP, CE, and PD could be combined into one or two factors.

To provide further evidence of the generation component, an additional CFA was performed after the analytic essay scores were added to the correlation matrix. Specifically, it was speculated that although the FR test contained a generation component, the essay test results would include a generation component not available in the objective test format. A nine-factor solution was imposed on the data. The first six factors were again specified to be pure trait factors. The seventh factor was specified to be the recognition factor on which all three methods would be free to load. The eighth factor was specified to be the generation factor on which both the FR and the essay scores were free to load. The last factor was the organization factor (ORG) on which only essay subscores would be free to load.

The results are shown in Tables 2 and 3. A very good fit of the model to the data was achieved: $\chi^2(df = 84, N = 192) = 71.48$ ($p = .83$). It can be seen in Table 3 that factor 7 (REC) is dominated by the MC and FR loadings (with the exception of the loading for the PS subscore for the MC test).

Table 1
Trait/Procedure Confirmatory Factor Loadings and Error Variances (θ_ε) for the Multiple-Choice and Free-Response Test Subscores

Test and Subscore	Factor								θ_ε
	SP	CP	CE	US	PD	PS	REC	GEN	
Multiple-Choice									
SP	.65	*	*	*	*	*	.61	*	.19
CP	*	.37	*	*	*	*	.57	*	.53
CE	*	*	.63	*	*	*	.58	*	.25
US	*	*	*	.40	*	*	.73	*	.30
PD	*	*	*	*	.44	*	.52	*	.53
PS	*	*	*	*	*	.53	.20	*	.66
Free-Response									
SP	.43	*	*	*	*	*	.69	.09	.31
CP	*	.44	*	*	*	*	.59	.12	.43
CE	*	*	.12	*	*	*	.71	.26	.39
US	*	*	*	.90	*	*	.33	.80	.56
PD	*	*	*	*	.51	*	.43	.05	.54
PS	*	*	*	*	*	.61	.30	.12	.51

*Factor loading fixed to 0.0.

Table 2
CFA Factor Intercorrelations for the Factor Matrices in Table 1 (Above the Diagonal) and Table 3 (Below the Diagonal)

	Factor								
Factor	SP	CP	CE	US	PD	PS	REC	GEN	ORG
SP	--	.01	−.09	.18	.24	.16	*	*	NA
CP	.17	--	.36	.16	.81	.25	*	*	NA
CE	.13	.56	--	.37	.78	.25	*	*	NA
US	.32	.21	.27	--	.41	.15	*	*	NA
PD	−.18	−.45	−.29	−.05	--	.62	*	*	NA
PS	−.02	−.25	.15	−.14	−.22	--	*	*	NA
REC	*	*	*	*	*	*	--	*	NA
GEN	*	*	*	*	*	*	*	--	NA
ORG	*	*	*	*	*	*	*	*	--

*Factor loading fixed to 0.0.

The REC loadings ranged from .44 (PS) to .74 (PD) for the FR test. No single subtest dominates the generation factor (factor 8), with the CE subscore of the FR test having the highest loading of .35. The last factor, representing the organizational component, is clearly dominated by the subscores believed to characterize organization: PD and PS (.71 and .42 respectively).

The factor intercorrelation matrix (lower triangle of Table 2) suggests that each of the trait factors could be considered to be unique with very low intercorrelations (the only exception being the .55 correlation between the CP and the CE factors).

Further verification of the presence of the procedural continuum can be seen by reanalyzing the individual correlation matrices of the 18 subscores. The correlational information is summarized in Table 4. Because the MC and FR tests are closer on the generation factor continuum than the MC and essay tests, it is expected that the correlations between the two indirect measures should be higher than the correlations between the MC and essay tests. Likewise, the correlation between the FR and essay tests should be higher than the correlation between the essay and MC tests. That is, the closer any two methods are on the continuum, the higher the correlation between the individual traits. This pattern is followed in almost all cases. The only exception is that the MC/essay total score correlation (.20) is slightly higher than the FR/essay total correlation (.15).

Previous studies have demonstrated that only moderate correlations exist between indirect and direct measures of writing assessment. Quellmalz, Capell, and Chou (1982) found that factors containing only loadings for essay subscores were independent of factors characterized by multiple-choice score weightings. To verify this finding, an additional CFA was done using the standardized and transformed test and essay subscores. In this analysis, correlations between the trait factors and the test format factors were constrained to be 0. Specifically, only the spelling subtests were permitted to load on the SP factor. The same was true for the other five trait factors. A seventh factor was created so that only the six MC test subtest scores could load on it. There were two other factors, one for the FR subtest and one for the essay subscores. Intercorrelations between method factors were also fixed to be 0. This indicates that the subscore loadings for each method factor are accounting for variance that is unique to that factor. Results of the CFA are shown in Tables 5 and 6.

The fit of the hypothesized factor structure to the combined group data was extremely good: $\chi^2(df = 102, N = 192) = 64.92$ ($p = .98$). The trait factors weigh most heavily on the FR test except for the CE and US subscores. The essay trait loadings are smallest for each of the six traits. The MC test method factor is dominated by the negative weightings on the SP and the CP subscores, −.25 and −.43, respectively. The essay factor loadings are highest for the two organizational subscores: PD (.60) and PS (.79).

The intercorrelation matrix shows three very large coefficients for the CP and CE factors (.95); the CP

Table 3
Trait/Procedure Confirmatory Factor Analysis Loadings and Error Variances (θ_ε) for the Multiple-Choice, Free-Response, and Analytical Essay Scores

Test and Subscore	Factors									θ_ε
	SP	CP	CE	US	PD	PS	REC	GEN	ORG	
Multiple-Choice										
SP	.52	*	*	*	*	*	.52	*	*	.44
CP	*	.33	*	*	*	*	.56	*	*	.57
CE	*	*	.21	*	*	*	.66	*	*	.53
US	*	*	*	.84	*	*	.62	*	*	.00
PD	*	*	*	*	−.19	*	.67	*	*	.54
PS	*	*	*	*	*	.44	.27	*	*	.73
Free-Response										
SP	.73	*	*	*	*	*	.57	.13	*	.11
CP	*	.35	*	*	*	*	.64	.02	*	.45
CE	*	*	.73	*	*	*	.52	.35	*	.17
US	*	*	*	.37	*	*	.49	.18	*	.57
PD	*	*	*	*	−.04	*	.74	−.24	*	.46
PS	*	*	*	*	*	.64	.44	.21	*	.35
Essay										
SP	.27	*	*	*	*	*	.32	.11	−.27	.74
CP	*	.01	*	*	*	*	.43	.08	−.11	.79
CE	*	*	.05	*	*	*	.32	.25	−.27	.77
US	*	*	*	.08	*	*	.31	.33	−.34	.66
PD	*	*	*	*	.59	*	.30	.29	.71	.07
PS	*	*	*	*	*	−.21	.07	.22	.42	.92

*Factor loading fixed to 0.0.

and PD factors (.88); and the CE and PD factors (.94). These correlations suggest that a better fit might be achieved if these factors were combined, and indicate that similar skills are being required by the three subtests to some degree.

These factor loadings represent standardized regression coefficients, with each factor representing a linear combination. The squares of the coefficients represent the proportion of unit variance that can be assigned to trait, method, and error. By examining the error variances, it is possible to determine which trait was most accurately measured by which method.

Discussion

Several limitations of this study should be noted. First, the data were collected from a single 10th-grade population and thus the effects of any developmental stages of writing could not be addressed (i.e., it is uncertain whether the results are developmentally dependent). A second concern is that only one essay measure was used for the direct assessment. However, it is speculated that if more essay data were obtained, not only would the direct measure be more reliable, but similar factor structures would result as well.

Table 4
Correlations Between Multiple-Choice (MC), Free-Response (FR), and Essay Test Subscores and Total Score

Test Pair	Subscore						
	SP	CP	CE	US	PD	PS	Total
MC/FR	.69	.48	.41	.37	.41	.77	.77
FR/Essay	.41	.31	.33	.26	.19	.01	.01
MC/Essay	.31	.22	.24	.22	−.04	−.08	.19

Table 5
CFA Factor Loadings and Error Variances (θ_ε)
for the Multiple-Choice, Free-Response, and Analytical Essay Scores

Test and Subscore	Factors									θ_ε
	SP	CP	CE	US	PD	PS	MC	FR	E	
Multiple-Choice										
SP	.74	*	*	*	*	*	−.25	*	*	.38
CP	*	.61	*	*	*	*	−.43	*	*	.43
CE	*	*	.72	*	*	*	.19	*	*	.43
US	*	*	*	.87	*	*	−.13	*	*	.22
PD	*	*	*	*	.66	*	.09	*	*	.54
PS	*	*	*	*	*	.43	−.14	*	*	.78
Free-Response										
SP	.90	*	*	*	*	*	*	−.13	*	.15
CP	*	.70	*	*	*	*	*	.08	*	.49
CE	*	*	.70	*	*	*	*	−.15	*	.48
US	*	*	*	.71	*	*	*	−.15	*	.49
PD	*	*	*	*	.68	*	*	.76	*	.03
PS	*	*	*	*	*	.91	*	−.17	*	.12
Essay										
SP	.45	*	*	*	*	*	*	*	−.14	.77
CP	*	.46	*	*	*	*	*	*	−.06	.78
CE	*	*	.34	*	*	*	*	*	−.14	.86
US	*	*	*	.37	*	*	*	*	−.17	.83
PD	*	*	*	*	.20	*	*	*	.60	.59
PS	*	*	*	*	*	.01	*	*	.79	.37

*Factor loading fixed to 0.0.

Unlike most of the studies reviewed, results of this study suggest, at least in the area of writing assessment, that the construct being measured is a function of the format of the test. That is, scores obtained from direct and indirect methods of writing assessment provide different information. Specifically, the skill of generating topic knowledge is more accurately assessed with an essay task. However, results from this study also suggest that objective testing formats can be modified to measure some of the procedural components contained in the writing task without sacrificing the advantage of faster, easier scoring schemes. The CFA procedures suggest that the free-response format measured the ability to organize coherent paragraphs (i.e., PD) much better than did the multiple-choice format. This should provide encouragement to test constructors to develop standardized test item formats that can facilitate the measurement of procedural components of writing in a more efficient format than the essay task provides.

While differences between direct and indirect methods of assessment exhibited in the results of this study are exactly those that might have been predicted by language arts experts, the study contains confirmatory evidence on the nature of these differences. Specifically, while it is reasonable to assume that more generation components would be present in the essay task, it is interesting to note from these data that the essay scores are almost totally dominated by these tasks. That is, variance in the essay score structure is heavily dominated by higher-order generation components such as paragraph development and paragraph structure. The moderate correlations previously observed between direct and indirect assessment procedures can be explained by the fact that direct methods contain more types of skills than indirect. These results suggest that the overlap may not be totally eclipsed; rather, indirect methods can be characterized by declarative components of text generation.

This difference suggests that practitioners interested in reliably measuring all aspects of the writing process characterized by this continuum may require the use of both indirect *and* direct methods

Table 6
Intercorrelations of the Subscore and Format CFA Factor Loadings

	Factor								
Factor	SP	CP	CE	US	PD	PS	MC	FR	E
SP	1.00								
CD	.66	1.00							
CE	.65	.95	1.00						
US	.68	.70	.85	1.00					
PD	.71	.88	.94	.69	1.00				
PS	.29	.40	.45	.29	.58	1.00			
MC	*	*	*	*	*	*	1.00		
FR	*	*	*	*	*	*	*	1.00	
Essay	*	*	*	*	*	*	*	*	1.00

*Factor loading fixed to 0.0.

of assessment. When instructors wish to measure only individual components of writing skill (i.e., organization), the type of knowledge required by this skill should suggest the method of assessment. The findings of this study indicate that direct methods provide a better measure of procedural-type writing skills, and that indirect methods should be preferred for measuring declarative-type writing skills. However, indirect methods can be modified to assess the aspects of writing which require organizational skills.

References

Ackerman, T. A. (1985). An investigation of the assessment of writing ability from a declarative/procedural perspective. (Doctoral dissertation, University of Wisconsin-Milwaukee, 1984). *Dissertation Abstracts International, 45,* 12.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). *Research in written composition.* Champaign IL: National Council of Teachers of English.

Breland, H. M. (1977). *A study of college English placement and the test of standard written English.* Princeton NJ: Educational Testing Service.

Breland, H. M., Conlon, C. G., & Rogosa, D. A. (1976). *Preliminary study of the test of standard written English.* Princeton NJ: Educational Testing Service.

Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessment of writing skill. *Journal of Educational Measurement, 16,* 119–128.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3,* 151–156.

Coffman, W. E. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5,* 24–36.

Cooper, C. R., & Odell, L. (1977). Consideration of sound in the composing process of published writers. *Research in the Teaching of English, 10,* 103–115.

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 19,* 159–170.

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

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In E. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing.* Hillsdale NJ: Erlbaum.

Heim, A. W., & Wats, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology, 37,* 339–346.

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

Jöreskog, K. G., & Sörbom, D. (1978). *LISREL IV: A general computer program for the estimation of linear structural equation systems by maximum likelihood methods.* Uppsala, Sweden: University of Uppsala, Department of Statistics.

Michaels, S. (1981). Sharing time: Children's normative styles and differential access to literacy. *Language in Society, 10,* 423–442.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement, 19,* 37–47.

Quellmalz, E. S., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19,* 241–258.

Scollon, R., & Scollon, S. B. K. (1981). Narrative literacy and face in interethnic communication. In R. Freedle (Ed.), *Advances in discourse process*. Norwood NJ: Ablex Press.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355–369.

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22,* 269–286.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11–29.

Author's Address

Send requests for reprints or further information to Terry A. Ackerman, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.