Domain-Referenced Specifications for Writing Proficiency*

Edys S. Quellmalz

CSE Report No. 127

January 1979

*Paper presented at the annual meeting of the American Educational Research and Association, Toronto, 1978.

Measurement and Methodology Program
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

This Project was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

How familiar now is the cry that American youth is woefully unable to produce logical, clearly written communications. Small wonder, given the state of affairs in writing instruction and assessment for at least the past decade. The field of composition, like its sister field, reading, has long been suffering an identity crisis. In composition the suffering is intensified by a schizophrenic split between an atomistic focus on easily understood, taught, and measured skills in mechanics and usage and such intellectually titillating rhetorical devices as voice and playful language.

Unfortunately for all of us concerned about writing competence, the question is as often "What is writing?" as "What is good writing?" The response to these questions, as to analogous questions, "What is science?" or "What is good math?", is "It depends." Delineation of "What is writing?" depends on such factors as the age of the learners and the perceived value and relevance of types of writing for those learners. "What is good writing?" depends on what is written, for what purpose, by whom, and for whom.

In view of the growing concern about fundamental competencies in all basic skills areas, the Center for the Study of Evaluation (CSE) is investigating methods for designing domain-referenced achievement tests. Interest in developing valid measures of writing competence stems both from concern about current inadequacies in existing approaches to composition assessment and from hope that technological advancements in the evaluation of written production will also contribute to methods for assessing constructed responses in other subject areas. One of the research strands in the CSE

domain-referenced testing project has sharpened the focus of the question "What is writing?" to "What is basic, fundamental writing competence?" What skills are presumably necessary for students to "get out and get along" (exit skills, vocation-related skills) or to "get in and go on" (college admission, college survival writing skills). Within this focus, the Center is initiating a research program designed to examine how to extend current methodology for specifying objective, domain-referenced achievement tests to largely unresolved problems in assessing constructed responses. In this paper I will describe some of the measurement problems involved in constructing valid, stable domain-referenced measures of basic writing competence, and some preliminaty studies investigating promising approaches.

Domain-referenced measurement

First, it might be useful to describe briefly the purpose and characteristics of domain-referenced tests. The hallmark of domain-referenced measurement is its clear characterization of the domain or class of content and responses to be measured. A set of design specifications guide construction of a homogeneous pool of test items that will measure a precisely circumscribed competency domain. Proponents of domain-referenced tests (DRTs) concur that domain specifications for competency-based assessment should include (1) a description of the outcomes that the test is intended to measure, (2) a precise definition of the content range and response requirements of the test items, and (3) a set of rules to guide item generation. In addition to these elements, the Center proposes that a domain-referenced test must meet the following criteria:

1. Meaningfulness/Publicness: Ecological validity. The test tasks should represent writing topics, levels of difficulty, and standards of judgment that are typical of both students' present school situation and future work contexts.

2. Parsimony: Maximum generalizability. Performance on a task should predict performance across a range of related tasks.

3. Instructional relationship: Diagnostic value. The dimensions of the task to be manipulated must have clear implications for instructional practice, for example, placement of students or selection of teaching strategies.

Is actual writing necessary?

Given DRTs heavy reliance on the descriptive power of the tests specifications, traditional notions of content validity raise questions concerning the representativeness and ecological validity of tasks posed to assess composition skill. Certainly the most salient threat to the validity of composition assessment instruments is that most do not collect samples of actual writing at all. A look at norm-referenced tests currently available in writing (such as CTB/McGraw Hill's Education Skills Test, 1971, or the Houghton Mifflin College English Placement Tests, 1969) uncovers measures which contain multiple-choice items dealing with organization of ideas, sequencing topics, mechanics and usage. Even the few criterion-referenced tests produced in composition, notably the 1972 Composition Skills tests published by the Instructional Objectives Exchange, provide only selected response items.

Test publishers have abandoned constructed responses for a number of reasons. Economics, the time and expense required to score written samples, is no doubt the most important reason. A major problem also has been the reliability and stability of constructed response measures, arising primarily when scoring criteria are ambiguous. An even more prominent reason cited for

excluding constructed responses has been the body of data indicating the high correlation between scores on selected and constructed items (Coffman, 1971). Test developers have apparently concluded that selected response modes measure skills just as well, if not better, than constructed responses.

In reply, the continued cry of subject matter experts has been that total reliance on selected response modes to measure writing achievement violates construct and content validity. Across subject matters, the reaction could be summarized as "life is not a multiple choice." Few curriculum theorists are satisfied with recognition/identification levels of response as outcomes in any subject matter expertise. This position is particularly the case in the area of composition, where the truly valid measure is actually student production of a composition (Braddock, 1963). While the requirement for production is accepted prima facie in other areas calling for student performance, e.g., music, drama, art (Fitzpatrick & Morrison, 1971), it is interesting that in the area of composition, a basic skill, test publishers tenaciously adhere to the validity of student comprehension and critique as substitutes for actual written production.

Are selected response modes for measuring composition competencies appropriate? The plausible answer is "no," not as valid measures of the terminal skills. Findings from the psychology of learning have repeatedly demonstrated that the constructed response imposes different requirements upon the learner than does the selected response (Bourne, 1966). The constructed response is more difficult as measured by error rate and increased response latency. Recent theory and research in information processing and

models of memory also may suggest that the process of input verification (selection) is a different process from the access and integration of information from multiple schemata (Anderson, 1975; Atkinson & Schffrin, 1967).

It is likely however that selected response items can be useful measures of simpler, enroute or component competencies required for the production of a composition. Writers in the area of instructional design (Gagné, 1975) assert that the ability to recognize an exemplar of a concept is a prerequisite skill to the ability to produce an example. These contentions imply that selected response items measuring recognition of examples of basic elements of a composition can be valuable from the standpoint of parsimony of measurement and diagnosis for planning of instruction. They will not suffice, however, as the only form of measurement. Also, in light of the legendary problem of subjectivity and variability of scoring writing samples, selected responses can provide an objective base for decisions about student achievement of basic enroute competencies. It has been proposed that a mixture of selected and constructed response items in criterion-referenced tests can provide a powerful information base for instructional decisions (Snidman & Quellmalz, 1975). I will assume that most educators would agree that actual writing samples are required as valid measures of writing ability.

What should be written?

A second validity issue in composition assessment is what the student is asked to write. Our initial focus will be on the major modes of discourse and their associated structural elements. In the area of writing

there do exist recognized delineations of the major forms of discourse, or genre, and the components of them that comprise criteria for "well written" work (Kinneavy, 1971). That there is a common set of concepts or schema at the text level (Anderson, 1975) has been asserted and characterized from Aristotle to Ajay (1974). The problem has been that tests frequently fail to tap the range of these genres and skills, to consider methods for collecting writing samples on component skills, or to define criteria in terms of teachable elements. While some writers may dismiss the appropriateness of these basic schema as formula writing, it seems reasonable that these schema must be mastered as fundamental building blocks. It is anticipated that by employing conventionally recognized structural and stylistic elements of discourse to describe writing tasks and criteria in domain specifications will facilitate collection of evaluation information with practical classroom relevance. That such a characterization of writing can be employed by practitioners is evidenced already by the established success of the SWRL/Ginn Composition Skills Program and by such admirably detailed specifications for criterion-referenced tests as those used in the Shawnee Mission Schools (Roberts & Wolfe).

We cannot agree with the approach used in norm-referenced tests which treats a single piece of writing as representative of skill in all discourse. In fact one anticipated outcome of our studies will be to corroborate the position that writing tasks for different genres do indeed tap different skills and that student competencies in one mode will not necessarily predict performance in another mode.

In terms of economy of testing, the tasks that are most important to measure will depend on both the purpose of the test and the level of decision to be made. At the instructional level one might be interested in all genres. At the program, district, or state level the test designer may have to focus on those modes of discourse with ecological validity, i.e., that the student actually needs to continue in school or use in a job situation. Thus some tests may collect writing samples of exposition, persuasion, or description in preference to literary self-expression. Basic skills oriented tests produced at CSE will focus more on what Lloyd-Jones calls other-oriented, social-effectiveness discourse than on self-expressive discourse (Lloyd-Jones, 1977). Writing tasks will call for informing, explaining, or persuading rather than the "imaginative expression of feeling" included on the National Assessment exercise described by Lloyd-Jones (1977).

How should the writing task be described to the examinee?

How the writing task is described to the learner can have profound effects on the written product. The problem of selecting a topic appropriate for the mode of discourse and to the background of the learner has been described extensively elsewhere (Braddock, 1963; Coffman, 1971). Of interest in some of our research will be the influence of specifying the important elements the writer should include and the effects of clearly describing the purpose and audience of the writing task. These elements will in turn be used as criteria for judging the written response. By prompting the writer with the essential elements to include, we may

eliminate some of the problems associated with tests measuring performance rather than competence. One might speculate that differing levels of specificity of instructions to writers might selectively direct the content and organization of their production, just as pre-reading objectives and questions affect selective attention on text comprehension studies (Rothkopf & Billington, 1975).

What scoring criteria should be used?

To generate specifications for domain-referenced tests, the scoring criteria must be clearly delineated. In contrast to topic and item specific criteria used often in norm-referenced tests or described by Lloyd-Jones (1977) in his discussion of primary-trait analysis, it would seem reasonable to judge responses along dimensions that are generalizable to other writing assignments in the same genre and that are both sensitive to instruction and amenable to instruction. Thus basic structural and stylistic elements appropriate to the particular mode of discourse such as "concrete imagery" or "spatial organization" would be used rather than "creativity" or "imaginative projection." Indeed those elements used as criteria for judging a well-written essay should be elements amenable to instructional intervention.

What should be the length of the writing sample?

In those tests that do call for writing samples, the tests seem to call for complete essays. Some CSE research will explore the relative feasibility and utility of collecting shorter samples such as one, two, or three

paragraphs. Shorter writing samples would allow collection of more than one sample for a skill or skills during a test administration. Also, shorter samples measuring skills also assessed on a full (2-5 paragraph) essay would provide a means for establishing the stability of performance on the skill within the test.

How many writing samples should and can be gathered?

The variability of performance from one writing task to the next has also been extensively reported (Coffman, 1971; Cooper, 1977). This is often described as the "writer variable," fluctuation in writer mood, fatigue, and knowledge of topic from one task to another. Mastery of a writing competency which applies to a class of writing tasks, however, should be more stable. By describing the content and response parameters of the writing task in the domain specifications, it should be possible to elicit more consistent performance on writing competencies that presumably generalize across writing tasks within a mode of discourse. If a student has indeed mastered the use of time order as an organizational technique, for instance, performance should be relatively similar on two narrative writing tasks requiring chronological organization. Also, if time order is a skill of particular interest, short writing tasks of a paragraph in length may be sufficient indices of it as a separate skill. One would not expect too disparate a performance on a similar writing task (homogeneous item drawn from the same domain) given at a different administration time. Studies will vary the number of samples gathered on a skill both within and between administration times.

What scoring procedures should be used?

The different methods of holistic and analytic scoring procedures have also been extensively described (Cooper, 1977). Comparisons of the differential decisions generated by these methods will also be explored. It is likely that different scoring procedures would be differentially appropriate for varying levels of decision making. Holistic methods, based, perhaps on criteria such as those described before, might suffice for gross placement decisions (e.g., remedial vs. freshman composition) while analytic scales might be more appropriate for skill diagnosis or for documenting performance problems where life-impact decisions are to be made (e.g., getting out, getting in).

The predictive validity of alternative scoring procedures and the criteria included is a related issue. Which of the scoring criteria are most important for making particular decisions and for distinguishing among different competency levels? CSE studies will attempt to cross-reference writing performance with external indices of writing competence whenever possible.

How can these writing measures be sensitive to instruction?

To reiterate, CSE's position is that tests must have relationships to instruction. Seemingly obvious, this criteria has long been violated by tests presenting students with ill-defined, vacuous writing tasks and employing sophisticated, esoteric, or just unteachable criteria such as "creativity" or "projected self." Ideally, writing tests should be sensitive to learners' instructional history. This, of course, requires knowledge

of that history. When students have engaged in a particular program such as the SWRL Composition Skills Program, clearly relevant tests can be designed. The task is more difficult, but possible, if information is based on a district or state's general curriculum or course of study.

A test may not be sensitive to instruction, however, if learners have not had much instruction. This, I suspect is most often the case with today's secondary students. The test task and criteria, then, should at least be amenable to instruction and clearly employ elements for which there is some consensus in the field regarding their fundamental importance, generalizability, and utility.

What about other dimensions of writing?

Our initial work is on writing products and their basic structural and stylist elements. Characterizations of writing processes or stages, from conceptualization to planning, writing and editing, imply differently constructed measures. Given the exploratory state of much current research into the writing process and writing instruction, evaluation methods for these areas will be targets of future work.

In some studies in our series we also hope to attend to the information processing requirements of different response modes, genres, and genre elements by analyzing student writing samples.

Designing tests of writing competency

By attending to these measurement questions posed for assessing composition skill, we hope to refine techniques for developing writing tests

referenced to clearly defined competency domains. In the course of our studies we will assess a range of writing competencies to determine if there are different skills involved in different modes of discourse. We expect to document that the design of the evaluation instrument and its concommitant scoring procedures can profoundly influence decisions made about learner's writing ability.

Empirical studies

CSE is conducting three pilot studies which attempt to answer some of the measurement questions discussed above. The first study, examines the differential effects on high school students' performance of varying writing task requirements. The study will investigate relationships between subjects' analytic and impressionistic rating scale scores on expository writing tasks and between analytic scores and scores on analogous objective items. The study will also examine the relationship of writing performance to external indices of writing competence (e.g., teacher judgment, self-report, grades in English classes) and to subjects' instructional history.

The effects of differing response criteria on the assessment of writing competence are investigated in a second study. Its particular research questions are:

1. Which scoring systems are the best predictors of criterion group membership?

2. What is the pattern of criterion group performance on different scoring systems?

3. How does the performance of different groups of writers compare on the same scoring system?

4. Which scoring systems are most generalizable across writing groups?

5. Which scoring systems discriminate most effectively among criterion groups?

The six criterion groups are composed of low or high high school juniors, low or high college students, and teachers or business people who write on the job. Students will produce two writing samples of three paragraphs, two hundred words on two occasions one week apart. Topics will be expository and evaluated according to criteria such as development by reason, by examples, or by analogy.

A third study will investigate the relationship of students' instructional history to their performance on a descriptive writing task. In particular the study will collect student reports about instructional principles employed in the classroom (task description, use of instructional time on: writing activities, methods of discourse, and particular skills; and practice and feedback).

REFERENCES

Ajay, H. A survey of style manuals. Technical note. SWRL Research and Development. TN 2-73-44. Los Alamitos, CA., 1974.

Anderson, R. C. The notion of schemata and the educational enterprise. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, N. J. Laurence Erlbaum Associates, 1977.

Atkinson, R. C., & Shiffrin, R. M. Human memory: A proposed system. In K. Spense & C Spence (Eds.), The psychology of motivation: Vol. 2. New York: Academic Press, 1967.

Bourne, L. J. Human conceptual behavior. Boston, Mass: Allyn & Bacon, 1966.

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

Coffman, W. E. Essay exams. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cooper, C. Holistic evaluation of writing. In C. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. State University of New York at Buffalo, 1977.

Fitzpatrick R., & Morrison, J. Performance and product evaluation. In R. L. Thorndike (Ed.) Education measurement (2nd ed.). Washington: American Council on Education, 1971.

Gagne, R., & Brigg, L. J. Principles of instructional design. New York: Holt, Rinehart & Winston, 1974.

Kinneavy, J. L. A theory of discourse: The aims of discourse. Englewood Cliffs, N. J. Prentice Hall, 1971.

Lloyd-Jones, R. Primary tract scoring. In C. Cooper & L. Odell (Eds,), Evaluating writing. Buffalo: State University of New York at Buffalo, 1977.

Rothkopf, E. Z., & Bellington, M. J. A two-factor model of the effect of goal-descriptive directions on learning from text. Journal of Educational Psychology, 1975, 67, 692-704.

Roberts, D., & Wolfe, D. Sequencing & Keying of Unified Studies, Test specifications for criterion-referenced testing, achievement awareness record for language arts. Shawnee Mission Schools. ED 116193.

Snidman, N. S., & Quellmalz, E. S. Issues in criterion-referenced test construction. Paper presented at the Annual Meeting of the California Educational Research Association, San Diego, CA., 1976.