DOCUMENT RESUME

ED 196 038 CS 206 045

AUTHOR Spandel, Vicki: Stiggins, Richard J.
TITLE Direct Measures of Writing Skill: Issues and Applications.
INSTITUTION Northwest Regional Education Lab., Portland, Oreg. Clearinghouse for Applied Performance Testing.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ISBN-0-89354-629-4
PUB DATE Jan 80
GRANT OB-NIE-G-78-0206
NOTE 70p.
AVAILABLE FROM Northwest Regional Educational Laboratory, 300 S. W. Sixth Ave., Portland, OR 97204 ($3.75)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education: *Evaluation Criteria: *Evaluation Methods: Writing (Composition): *Writing Instruction: *Writing Skills
IDENTIFIERS *Writing Assessment

ABSTRACT

This monograph provides educators with the fundamental knowledge needed to develop and use direct assessments of student writing proficiency. The first chapter offers a brief review of the current status of writing assessment in the United States, focusing on emerging interest in the topic over the past decade. The second chapter presents an overview of writing assessment procedures touching on (1) differences between direct and indirect tests of writing proficiency, (2) considerations in improving the quality of the assessment, (3) strategies for exercise development, and (4) alternative approaches to scoring. The concluding chapter contains a discussion of approaches for conducting writing assessments in various educational contexts. Alternative testing methods are linked to various testing purposes, and strategies are outlined for optimizing the match between the two. (FL)

Direct Measures of Writing Skill:
Issues and Applications

Vicki Spandel
Richard J. Stiggins

Clearinghouse for Applied Performance Testing

NORTHWEST REGIONAL EDUCATIONAL LABORATORY
710 S.W. 2nd Avenue
Portland, Oregon 97204

ISBN 0-89354-829-4

January 1980

This work is published by the Clearinghouse for Applied Performance Testing (CAPT) of the Northwest Regional Educational Laboratory, a private nonprofit corporation. The work contained herein has been developed under grant OB-NIE-G-78-0206 with the National Institute of Education (NIE), Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position of the National Institute of Education, and no official endorsement by the Institute should be inferred.

TABLE OF CONTENTS

PREFACE

The Clearinghouse for Applied Performance Testing (CAPT) is pleased to provide you with the first in a series of monographs exploring the use of applied performance tests in selected content areas. The measurement of writing ability was selected as the initial monograph topic because in recent years educators have expressed steadily increasing concern with students' writing abilities. That concern has precipitated a vigorous search for tests that accurately reflect students' competence in writing. Various approaches to assessing writing proficiency have resulted, including direct assessment via writing samples and indirect assessment via objective tests. *This monograph deals specifically with direct assessment alternatives.*

Until only a few years ago, experience in designing and carrying out direct assessment of writing was the province of relatively few organizations—principally the Educational Testing Service and the National Assessment of Educational Progress. Now, techniques pioneered by those agencies are being adopted by school districts, state education agencies, and postsecondary institutions. Acknowledging the desirability of directly assessing writing proficiency, many educators are seeking information on alternative approaches to writing assessment.

This monograph was prepared to provide some of that information. The monograph offers the interested educator the fundamental knowledge needed to develop and use educationally sound assessments of writing proficiency. It does not, however, present step-by-step instructions on how to measure writing skill. Rather, the document describes strategies for planning and conducting writing assessment, then provides references and contact persons capable of providing additional, more precise information on how to assess writing skill.

The intended audience is the educator interested in (1) the developmental state of procedures for measuring writing proficiency and/or (2) the measurement issues to be addressed in such an assessment. The discussion of procedures and issues herein is appropriate for educators serving in (1) public education, either in the classroom or in administra-

tion, (2) state educational agencies, or (3) postsecondary education. For the reader who wants more information on the general status of writing assessment methods, this monograph may suffice. However, for those who plan to develop and conduct an assessment of writing skills, more in-depth study will be required.

The monograph begins with a brief review of the current status of writing assessment in American education, focusing on emerging interest in the topic over the last decade. Chapter 2 presents an overview of writing assessment procedures, touching on (1) differences between direct and indirect tests of writing proficiency, (2) considerations in maximizing the quality of the assessment, (3) strategies for exercise development, and (4) alternative approaches to scoring. The monograph concludes in Chapter 3 with a discussion of approaches for conducting writing assessments in various educational contexts. Alternative testing methods are linked to various testing purposes, and strategies are outlined for optimizing the match between the two.

CAPT will be developing further informational pieces and training materials on writing assessment. Your comments and questions on this monograph and suggestions for future materials are encouraged.

Beverly L. Anderson
CAPT Director

ACKNOWLEDGMENTS

Many persons contributed their time and skills to the preparation of this monograph. In particular, the authors wish to thank the chief contributors, Stephen Slater and Beverly Anderson, for their assistance in conceptualizing the document and providing background information on selected sections.

Thanks also to the following reviewers, whose invaluable critiques guided us through several revisions: Richard Brickley, Thomas Corcoran, Marjorie Kirrie, Connie Kravas, Christine McGuire, Dean Nafziger, Don Ochs, Judy Shoemaker and Frank Womer.

In addition, we are grateful for the excellent secretarial and production assistance provided by Laura Hopkins, Ken Jordan and Gervaise McCoy of the NWREL Assessment and Measurement Program; Perry Colton and Cathy Winters of the Marketing Department; and Archie Matthew and Warren Schlegel of the Media Center.

Special thanks to Vicki Fredrick of the Wisconsin Department of Public Instruction and to the many state representatives from throughout the country for sharing the information that appears in the appendix. We especially appreciate the courtesy of those who agreed to have their names included in our list of contact persons. Thank you, everyone.

Vicki Spandel and Rick Stiggins,
Authors

CHAPTER I: A Status Report on Writing Assessment

Declining writing skills—no longer news to educators anywhere—made headlines just a few years ago. In 1969 and 1974, the National Assessment of Educational Progress (NAEP) conducted a nationwide study to describe students' writing abilities. Results of that longitudinal study, published by NAEP in 1975, revealed that, for 13- and 17-year-olds, those five years had witnessed an increase in writing problems: awkwardness, run-on sentences, incoherent paragraphs, rambling prose, inappropriate language and inadequate sentence structure. NAEP's findings received widespread publicity, increased public awareness of a serious educational problem, and—along with other evidence of declining academic abilities among students—spurred renewed demands for educational accountability.

In revealing the scope of the problem, the NAEP assessments represented an important step in measuring students' writing skills. However, the results were too general to give educators practical guidance in planning. Because the NAEP writing assessment focused on broad dimensions of writing skill summarized over large samples of students, it set the stage for a barrage of new questions. If curricula were to be improved, if students were to be taught to write better, national summaries of performance would not suffice. Educators could address the writing problem realistically only through more precise answers to such questions as: How well can each individual student write? What specific writing skills

are lacking? What kinds of writing tasks give students the most difficulty?

Such questions continue to guide efforts to devise assessment procedures capable of assessing a student's writing skills precisely. Educators from all levels have participated in this developmental effort. Their varied approaches remain the source of intense debate about how best to measure writing skill. Each new assessment effort draws the attention of those who confront similar problems and seek to benefit from their colleagues' experiences. Some of those efforts are described herein.

A National Survey

In 1979, the Wisconsin Pupil Assessment Program conducted a nationwide survey to determine the current status of statewide writing assessments and provide information for others to use in planning and conducting writing assessments. A total of 18 states reported writing assessments in progress and an additional five reported writing assessments in the planning stages. The final report of survey results (Fredrick, 1979) includes a brief overview of the assessment procedures used in each state. That overview is reproduced in Table 1. Fredrick also provides (1) a summary of each state's assessment procedures, as well as (2) a summary of the administration, scoring and reporting procedures, and (3) a review of writing objectives assessed, assessment-related costs, problems encountered and recommendations for conducting effective writing assessments. The detailed profiles prepared by Fredrick have been abstracted for the reader in the appendix of this monograph.

Postsecondary Developmental Activities

The assessment of writing proficiency has been a major focus of research and development in postsecondary education as well. In response to requests from postsecondary institutions for more information on student writing skill, the College Board, in conjunction with Educational Testing Service, developed and offered as an optional component of the Scholastic Aptitude Test (SAT), a 20-minute assessment of writing proficiency using writing samples (Breland, 1977). This represented an expansion of assessment procedures used for years by the College Board as part of its Advanced Placement Test in English Composition. Similarly, the

Table 1

Overview of Statewide Writing Assessment Programs*

STATE	Grade(s) or Age(s) Tested	Type of Test	Type of Scoring
California	Grade 12	Writing samples	Holistic—Scale 1-9
Hawaii	Grades 4, 8, and 11	Writing samples	Primary, secondary, and tertiary trait
Idaho	Grade 9	Writing samples	Holistic—Scale 1-5
Louisiana	Grades 4, 8, and 11	Multiple-choice, Writing samples	Machine scoring, Primary and secondary trait
Maine	Grades 8 and 11	Multiple-choice, Writing samples	Machine scoring, Primary trait
Massachusetts	9-year-olds 17-year-olds	Writing samples, Multiple-choice	Holistic—Scale 1-8, Machine scoring
Missouri	Grade 8	Writing samples	Teacher evaluation of proficiency
New Hampshire	Grades 5 and 9	Writing samples	Holistic—Scale 1-4
New Mexico	Grades 10, 11 and 12	Writing samples, Multiple-choice	Teacher evaluation of proficiency
New York	Grades 9, 11 and 12	Writing samples	Holistic—Scale 1-4
Ohio	Grade 8 (1977) Grade 12 (1978)	Writing samples, Multiple choice	Primary & secondary traits and mechanics, Machine scoring
Oregon	Grades 4, 7 and 11	Writing samples, Multiple-choice	Holistic—Scale 1-4, Machine scoring

*V. Fredrick. 1979. Reproduced by permission of the publisher.

Table 1 (Continued)

STATE	Grade(s) or Age(s) Tested	Type of Test	Type of Scoring
Pennsylvania	Grades 5, 8 and 11	Multiple-choice	Machine scoring
Rhode Island	Grade 11	Writing samples, Multiple-choice	Analytical, Machine scoring
Texas	9-year-olds 13-year-olds 17-year-olds	Writing samples, Multiple-choice	Primary trait, Machine scoring
Vermont	All students in all grades	Writing samples	Teacher evaluation of proficiency
Washington	Grade 8 (1976) Grade 11 (1977)	Writing samples, Multiple-choice	Primary trait, Machine scoring
Wisconsin	Grades 4, 8 and 11	Writing samples, Multiple-choice	Primary trait, Holistic— Scale 1-8, Machine scoring

American College Testing Program (ACT), as part of its new College Outcome Measures Project (Steele, 1979), developed a direct, writing sample-based assessment of writing proficiency. At the state level, the California State University System and Colleges has recently instituted mandatory English proficiency and placement testing programs, including direct writing assessment for all entering freshmen.

Still—No "Best" Answer

These are but a few of the many instances in which writing assessment is being successfully conducted on national, state, and local levels. The remainder of this monograph describes (1) some of the procedures used in various assessment contexts and (2) key measurement issues in the testing of writing skill.

The assessment of writing skill is a very complex task, because of the broad range of potentially relevant writing competencies and the difficulties in setting standards of accept-

able performance. *There is not now, nor will there ever be, a single best way to assess writing skill. Each individual educational assessment and writing circumstance presents unique problems to the developer and user of writing tests. Therefore, great care must be taken in selecting the approach and the methods to be used in each writing assessment. Methods used in one context to measure one set of relevant writing skills should not be generalized to other writing contexts without very careful consideration of writing circumstances.*

CHAPTER II: An Overview of Direct Writing Assessment Procedures

The most effective way to clarify what is meant by "direct" assessment of writing proficiency is first to differentiate it from indirect assessment, then to discuss direct assessment test development procedures.

Direct versus Indirect Assessment

There are at least two methods for gathering useful information about writing proficiency. One is to gather samples of writing and to evaluate them according to prespecified criteria. This is the direct approach. The second is to construct objective tests measuring some of the language usage skills important to effective writing. This is the indirect approach. Though each is capable of yielding sound information about writing proficiency, they employ totally different measurement methods to achieve different goals. These differences become more clear when one examines the purpose for each type of assessment.

When resources and expertise are available, it is generally acknowledged that the best way to assess writing skill is through the direct assessment approach—that is, by having students write. The purpose of direct assessment is to simulate—under controlled conditions—commonly encountered writing circumstances and to evaluate examinee performance within those circumstances. Resources and expertise must be available to (1) specify the writing skills to be assessed, (2) develop writing exercises, (3) train those who are

to evaluate the writing samples, and (4) conduct at least two independent readings of each writing sample. If these conditions are satisfied, writing exercises provide the most sound, appropriate alternative for generating valid and reliable information about writing skill.

When the costs and complexities of direct assessment make it impracticable, the indirect approach might be considered. Useful information can be gathered about a student's language proficiency by using objective—usually multiple choice—tests. Examples of such tests are the English Usage Test of the ACT Assessment Program and the SAT Test of Standard Written English. The purpose of these tests is not to measure writing skill per se (ACT, 1978; CEEB, 1975). Rather, it is to measure the students' understanding of the basic elements and conventions of standard English usage. These prerequisites of effective writing represent necessary but not sufficient ingredients in writing skill.

The ACT and SAT tests are similar. The examinee is presented with a prose passage and is asked to identify and correct problems in usage. The ACT test incorporates extended passages. Items focus on the following skills: punctuation, grammar, sentence structure, style, diction, logic and organization. The SAT test presents single sentences; items focus predominantly on grammar and sentence structure. Each has been shown to yield consistent (reliable) scores (.85 to .90) and to be moderately to highly correlated (.60 to .70) with writing proficiency measured via the direct approach (Huntley, Schmeiser, and Stiggins, 1978; Breland and Gaynor, 1979).

As with direct measures, indirect measures like the ACT and SAT tests should be used only when rcliability and validity (discussed later) can be assured. If a published test is used, the publisher should be asked to present evidence of score reliability and validity. In judging validity, test users should give careful attention to the difference in purpose between direct and indirect measurcs.

In summary, direct assessment of writing skill via writing samples simulates real life writing circumstances; writing must be evaluated by trained judges. Indirect assessment, on the other hand, measures knowledge of language usage to dctermine whether students have mastered the prerequisites of effective writing; responses are usually machine scorable. Though the results of indirect assessment can be shown to be

related to the results of direct assessment, under no circumstances should indirect assessment be considered a substitute for direct assessment.

In the case of direct assessment, the cost of test administration and scoring are high, while with indirect assessment the costs of test development (or purchase) are high. All things considered, indirect assessment is generally less costly. In selecting a measurement approach, therefore, the user must carefully examine the tradeoffs: quality of the resulting information (favoring direct assessment) vs. cost of assessment (favoring indirect assessment). If an indirect measure is to be developed locally for local use, resources should be available for (1) specifying skills to be assessed, (2) constructing exercises, and (3) scoring the tests. Extensive experience in writing objective test items on language usage is a necessity. For this reason alone, the user is advised to exhaust the list of available published tests before opting for local indirect writing test development.

Ensuring High Quality Assessment

Two key considerations in determining the quality of writing assessment are the reliability and validity of the scores generated by the assessment. The exercise development and scoring procedures outlined in the following two sections of this chapter have been developed and refined specifically to ensure score reliability and validity. However, before describing those procedures, it may be useful to explain reliability and validity as they relate to direct writing assessment.

Reliability. To be useful for educational decisions, tests must yield scores that are consistent or reliable. When scores are unreliable, the assessment results can lead to erroneous conclusions or decisions. In writing assessment, score inconsistency can take any of several forms.

For example, suppose a writing assessment were administered to the same students twice, the second administration following a two- to three-week interval. And suppose that even though no writing instruction took place, the scores obtained the second time were totally different from those achieved the first time for nearly every examinee. The examiner would not know which score (if either) to depend on as the true reflection of the students' proficiency. Or suppose two writing exercises were developed to measure exactly the

same skills and yet when both were administered to a student, the exercises resulted in totally different estimates of proficiency. Again, the examiner would not know which score was the better indicator of proficiency. Or, from a third perspective, suppose two judges read and evaluated a writing sample from the same student and drew totally different conclusions regarding the student's proficiency. In this case, as with the others, the examiner would not know which judgment to rely on. These three examples show how unreliability can manifest itself in the assessment of writing skill with writing samples.

When scores are unstable over time, differ across ostensibly equivalent writing exercises and/or differ across independent evaluations of proficiency, there is reason to question the usefulness of the assessment procedures. However, when the procedures employed yield scores that are stable over time, across exercises and across independent evaluators, those scores can be confidently used for educational decisions. The test developer is responsible for (1) employing assessment development procedures that maximize score reliability, and (2) presenting systematic evidence of score reliability for review by users.

Three factors are important in developing reliable tests. First, the writing skills to be measured must be clearly and concisely defined by writing experts. Only then is it possible to (1) demonstrate to users, exercise developers, and others precisely what skills are to be assessed; (2) judge exercise appropriateness; and (3) inform judges about the criteria for acceptable performance.

Second, there must be a clear and unambiguous link between the skills to be tested and the exercises developed. This interrelationship ensures that exercises give the competent writer the stimulus and opportunity to demonstrate whatever skill(s) the user wants to measure.

And third, judges must be carefully trained to conduct the evaluation according to prespecified criteria and agreed upon standards. If these three guidelines are followed, chances are that scores will be consistent over time, across exercises, and across raters. If scores are found to be inconsistent, assessment procedures should be re-examined in light of these guidelines and revised accordingly.

Validity. Even if a developer of a direct writing assessment

is successful in achieving score stability through careful skill identification, exercise development and evaluator training, the writing assessment developmental task is only partly completed. Attention must also be given to the validity of the assessment scores. The validity of a score depends on (1) the test used to generate that score, and (2) the intended purpose for that score. Intended purpose can be identified in a variety of ways, each of which can be considered a dimension of validity. Cronbach (1971) has identified a number of such dimensions that can be applied to the direct assessment of writing proficiency. For example, a test may be designed to measure a specific set of writing skills. If review of that test by qualified experts reveals that the exercises do indeed cover those skills, then the test is said to cover the intended content validly. It has achieved its content coverage purpose.

From a different but related perspective, a test that plays a significant role in educational decision making (e.g., provides a basis for placement or selection) should inspire confidence among users. The exercises must appear to assess truly important skills. If this face validity is missing, the test will not be used—regardless of the actual appropriateness of the exercises. It is important that the exercises seem appropriate even to the least sophisticated of the intended users.

There are other ways of revealing whether a test is achieving its intended purpose. For example, a test of writing proficiency is only one of many potential indicators of writing skill. If a test is valid, then scores should be consistent with (or reflect the same level of proficiency as) other indicators of writing skill: for example, performance on job-related or real-world writing tasks, amount of formal training in writing, grades received in writing courses, and/or scores achieved in other objective or writing sample-based tests of writing skill. To the extent that the writing assessment developer is able to show that performance on a newly developed writing assessment is consistent with performance on other writing-related tasks, the assessment has achieved its goal of reflecting writing proficiency.

Test purpose largely determines the requirements for documenting validity. For example, a direct writing assessment may be very general, or it may be narrowly focused to be precise and diagnostic. Suppose, for instance, that one wished to measure students' letter writing skills. A general exercise might present the student with these directions:

> Pretend that you are applying for a job as a salesperson with Acme, Inc. Write a letter to Acme explaining your interest and qualifications.

Because these instructions are very broad, responses can only be judged on general merit. Raters will likely consider such factors as word choice, sentence structure, organization, mechanics—in short, the kinds of things one would consider in judging any piece of writing. And the result will be a general profile of overall student writing performance. But suppose one wished to measure students' performance on explicit letter writing skills, in order to diagnose individual students' strengths and weaknesses. This would call for some modification in the item so that it might read as follows:

> Pretend that you are applying for a job as a salesperson with Acme, Inc. Write a business letter addressed to Ms. Jones, Sales Manager of Acme. 2525 Main. Huntsville, New York 20201. Explain your interest and qualifications. Attempt to convince Acme that you're the best person for the job. Use proper business letter form.

These specific directions will allow responses to be judged according to explicit criteria: students' ability to be convincing and use proper business letter format. Responses to the first item could not be scored in this manner because the intended audience, purpose and expected letter format were not specified in the instructions. In summary, if diagnostic information is desired, items must be carefully structured to elicit the appropriate type of response. Evidence of success in achieving the desired level of precision should be included in validation research.

The purpose for testing may also be considered in terms of the specific educational decision in question. That is, a test may be intended to rank order examinees in terms of proficiency for selecting the most able for further training or the least able for remediation. Or the assessment may be intended to provide information for mastery/nonmastery decisions with regard to specific writing objectives. Because these are different purposes, the assessment strategies used to achieve them will differ. It is up to the developer to determine the usefulness and appropriateness of assessment procedures for meeting each specific decision-oriented purpose.

The essential point is that validity is a reflection of success

in achieving the testing purpose. As with reliability, the test developer has two primary responsibilities: to maximize validity through careful test development and to report evidence of validity for users. Strategies for maximizing validity are similar to those for maximizing reliability. The writing skills to be assessed should be clearly and unambiguously defined. Both the skills and exercises developed to reflect those skills should carefully be reviewed by subject experts to ensure appropriateness. And once the test is administered and scored, scores should be related to other relevant writing proficiency indicators to be sure the assessment is focused on the desired dimensions of writing skill.

Subsequent CAPT publications will deal in greater detail with procedures for documenting reliability and validity of writing assessment. However, in the interim, interested readers are urged to refer to such standard measurement textbooks as Ebel (1978), Mehrens and Lehmann (1977), Sax (1974) and Thorndike (1971) for more information.

Developing Exercises

In the discussion that follows, a writing exercise is considered to comprise all stimulus materials and instructions used to define the writing task. Developing exercises for direct assessment of writing involves five carefully conducted steps. The first two steps are crucial for any writing assessment: (1) assessment planning and (2) exercise development. The remaining three steps, while very important, are not always implemented, depending on the resources available and the seriousness of the decisions to be made. These are (3) test specification and exercise review, (4) exercise pretesting and (5) final revision. Each of these five developmental steps is discussed in detail in the following paragraphs.

Assessment planning. The ultimate quality of any assessment is influenced more by the thoroughness and detail of its original blueprint than by any other factor. Several very important test design questions must be thoroughly considered. If each is not individually considered, the chances of creating a valid and reliable assessment—especially a writing assessment—are greatly reduced.

The first planning question concerns purpose. The sole reason for conducting any educational assessment is to provide information to facilitate some educational decision.

Therefore, the primary step in writing assessment planning is to state precisely the specific educational decision to be influenced by the resulting scores. Potential decisions include (1) diagnosing individual student proficiency in specific writing skill areas; (2) rank ordering examinees with regard to general writing proficiency for selection or placement; and (3) assessing specific or general writing proficiency to evaluate the impact of an instructional program. (Additional decisions will be presented later.) Specific assessment strategies vary according to purpose. Therefore, the decision(s) to be facilitated must be clearly specified at the outset.

Second, test developers must determine the specific form of writing to be produced (e.g., essay, business letter, fiction), the audience to be addressed, and the purpose to be served in addressing that audience. Any given student's level of proficiency will vary as a function of writing form.

A third planning step calls for identifying evaluation criteria (to be used in judging writing skills) and levels or standards of acceptable performance for each criterion chosen. For example, organization, style, tone and sense of audience are typical *criteria*. In order to judge performance, however, evaluators need guidelines or *standards* for determining good, poor or mediocre organization, style and so on. Both criteria and standards are closely related to the purpose for assessment and the form of writing called for. For a broad assessment, it is only necessary to stipulate the general dimensions of writing skill; for a diagnostic assessment, the precise writing skills to be evaluated must be identified.

In summary, the writing assessment blueprint must include (1) the educational decision(s) to be facilitated, (2) the writing context (purpose, audience and type of writing to be required) and (3) the specific criteria (skills) and standards of performance. If any element is missing, it will be difficult—if not impossible—to construct writing exercises that give students an opportunity to demonstrate proficiency.

Exercise development. Once planning is completed, the developmental goal becomes quite apparent: the design exercises that provide the competent student with the necessary stimulus and writing conditions to demonstrate his/her level of competency. In other words, the writing tasks must inform students of the purpose for the writing, the audience to be addressed and the type of writing expected (necessary condi-

tions), while at the same time allowing students the latitude (e.g., sufficient exercises and time) to demonstrate their capabilities. It should be apparent that unless careful planning has preceded this step, appropriate exercise development will be difficult at best.

Here are some specific guidelines to be observed in constructing writing exercises: First, the exercise developer should recognize the impossibility of covering all possible instances of relevant writing. A realistic objective is to construct and include in the assessment an appropriate sample of relevant exercises. Based on student performance on that sample, one can generalize about expected performance in parallel contexts. To insure the appropriateness of these generalizations, however, samples must be carefully selected. For example, if one wishes to know whether students can write expository prose for an academic audience, one exercise is probably not enough; two or three similar exercises may be necessary to ensure that the sample is sufficiently representative. At the same time, ability to construct other forms for other audiences—e.g., an entertaining piece of fiction for young children—is irrelevant to the testing purpose at hand.

To use another example, suppose the purpose of an assessment is to determine mastery of a single clearly focused writing objective: ability to present map directions effectively in written form. Enough examples of student performance should be gathered to ensure that addition of another exercise would not significantly alter any conclusions about student performance. In other words, exercises must be clearly focused and sufficient in number.

The reader may recognize that this issue of skill sampling is related to both reliability and validity, as described earlier. For example, it is important to provide enough samples of student writing to allow for stable scores (reliability), and to fairly and adequately sample the skill domain the test is intended to cover (validity).

Certainly the key question in all writing assessment is: How much writing is enough? There is no hard and fast answer. The number of exercises required and the length of those exercises are functions of the range of skills to be evaluated and the level of precision at which those skills are defined. Broader assessments covering many skills generally require more samples than precisely focused, narrow assess-

ments. Recent research on this topic (Steele, 1979 and Breland, 1977) offers some guidance. The Steele research involved a broad assessment of end-of-college writing proficiency via three 20- to 30-minute writing exercises. Analysis of score consistency revealed that the use of only one or two exercises yielded unreliable scores. However, the use of all three exercises raised score consistency to an acceptable level. Further, the study revealed that the addition of more exercises beyond the original three would not significantly increase reliability. These results were supported by Breland's research which revealed that, in a similar college-level assessment, a single 20-minute exercise was incapable of yielding consistent scores.

Braddock, Lloyd-Jones and Schoer (1963) offer guidance from a different perspective as to the amount of writing needed to judge proficiency:

> Even if the investigator is primarily interested in nothing but grammar and mechanics, he should afford time for the writers to plan their central ideas, organization, and supporting details; otherwise their sentence structure and mechanics will be produced under artificial circumstaoces. Furthermore, the writers ordinarily should have time to edit and proofread their work after they have come to the end of their papers. . . . *Investigators should consider permitting primary grade children to take as much as 20 to 30 minutes, intermediate graders as much as 35 to 50 minutes, junior high school students 50 to 75 minutes, high school students 70 to 90 minutes, and college students two hours (to demonstrate proficiency).* [Emphasis added.]

Exercises should frame a clear and concise writing task so that students fully understand what is required—whether or not they can fulfill the requirements. Time pressure is undesirable; it is an artificial imposition that may not replicate the circumstances in which real life writing occurs. Items should offer the writer a realistic, sensible challenge so as to maintain interest. Varied stimulus materials (written, auditory, or visual) should be used. Most important, examinees must be given time to think, organize, write, reread and revise.

Some writing assessments have attached great importance to revision. As Rivas (1977) notes:

> Rewriting skills are often considered to be the essence of

> good writing. All of us can express ourselves in some form, however ambiguous or inappropriate, but a good writer knows how to revise such preliminary statements so that they become less ambiguous and more appropriate.

Part of NAEP's 1974 writing assessment called for writing and rewriting the same copy in an attempt to get at revision (Rivas, 1977). Students were asked to write a class report about the moon, given certain facts. They were given 15 minutes to write the first draft, using a pencil. Upon finishing, they were given 13 minutes to revise the first draft, using a blue pen so that any changes would stand out clearly. They were told to make any changes they wished, including crossing out words or rewriting if necessary: rewriting was not required, however. Papers were scored for overall organization (based on the quality of the revision), and were categorized to indicate the kinds of revisions attempted: cosmetic (improved legibility), mechanical, grammatical, transitional, informational, holistic (complete rewriting), and so on. Though some educators might feel the test was not a true measure of revision skills (many students, for reasons unknown, attempted no revision), the NAEP moon test represents at least a step toward development of a proper revision test.

Clearly, attention must be given to editing and revision as part of any writing assessment, whether by providing sufficient time and opportunity for the examinees to revise on their own, or by providing specific instructions to revise, as NAEP did. If extensive revision (beyond proofreading for spelling and other mechanical errors) is desired, it will be necessary to construct the assessment to allow students time for proper reflection—just as in a real-life writing situation. It will not be sufficient merely to give students an additional five or ten minutes at the end of a writing exercise to "fix things up." A better approach might be to allow students time to write one day, time to revise on a subsequent day. This kind of provision may increase administration time and costs. However, it will also provide a more relevant (i.e., true to real life) test of revision skills than one-session assessment.

Review of specifications and exercises. Whenever possible, the writing and assessment personnel responsible for

assessment specifications and writing exercises should present their work to an independent group of writing and measurement specialists for review and formative evaluation. This review should cover—

1. The purpose for the assessment (decision to be made).
2. The definition of the assessment context (form of writing, audience and reason for writing).
3. The criteria (skills to be assessed) and standards of acceptable performance.
4. Relevance of exercises in terms of skills to be assessed.
5. Representativeness of exercises in terms of the domain of possible exercises.
6. Sufficiency of the exercises in providing students with the opportunity, in terms of time and tasks, to demonstrate proficiency.
7. Clarity and conciseness of prescribed writing tasks.
8. Level of interest and challenge conveyed in stimulus materials and writing instruction.
9. Adequacy of instructions and opportunity for revision, if that is a desired part of the assessment.

As the importance of an educational decision and/or as the number of students to be included in the writing assessment increases, the importance of independent review increases also. Thus, review is less critical with small-scale, local or classroom assessments than with large-scale assessments on which selection decisions are often based.

Exercise pretesting. Whenever possible, exercises should be administered to a sample of students prior to actual full-scale administration so that potential problems can be identified and corrected. Pretesting procedures should closely approximate actual administration in terms of type (though not number) of pretest students, conditions (e.g., facilities, time limits, methods for providing directions) and scoring procedures. Developers should then independently evaluate results, attending to (1) the level of proficiency demonstrated (and whether that level seems to fluctuate from exercise to exercise), (2) the nature of the responses produced

(in terms of quality, appropriateness, length and enthusiasm), (3) the consistency of ratings across independent evaluations, and (4) the apparent clarity of instructions to students. Exercises that appear to yield inconsistent or repeatedly low quality results can be identified and the reasons for apparent problems discussed. Often, exercises can be adjusted. As with independent exercise review, the importance of pretesting increases with the scope and importance of the assessment.

Final exercise revision. The final step in exercise development is to revise exercises on the basis of the review and pretest results. As final revisions are made, developers should continue to ensure reliability and validity of scores through careful use of test specifications, exercise development and preparation for scoring.

Procedures for Scoring Writing Samples

Many forms of objective tests can be machine scored. Writing tests that rely on writing samples, however, require individual hand scoring by qualified persons trained to apply agreed upon criteria and performance standards. Several different methods have been devised for scoring writing samples depending on the assessment purpose. The most appropriate method in any given situation depends upon what information one wishes to gain through scoring, how that information will be used, and what resources are available. Some scoring methods are more complicated—and therefore more costly—than others. The purpose of this section is to present a comparative overview of the general advantages and disadvantages inherent in each of five approaches: holistic scoring, analytical scoring, primary trait scoring, scoring for mechanics and grammar, and T-unit analysis.

Holistic scoring. In holistic scoring, raters review a paper for an overall or "whole" impression. Specific factors such as grammar, usage, style, tone and vocabulary undoubtedly affect the rater's response, but none of these considerations is directly addressed. As with all rating methods, raters must be carefully trained to conduct the evaluation. The purpose of training is to minimize (at least temporarily) the effects of individual biases by helping raters internalize an agreed upon set of scoring standards. It is generally recommended

that raters be experienced in language arts, familiar with pertinent terminology and practiced in rating student papers at the level for which they will be scoring. Consistency—both among raters and among scores assigned by a single rater—is very important in holistic scoring. Initial training takes about half a day, but it is also necessary to build in time for "refresher" sessions throughout the course of any scoring activity.

Papers are rated on a numerical scale. NAEP has used both 4-point and 8-point scales. Four-point scales are most common. An even-numbered scale is recommended because it eliminates the convenience of a mid-point "dumping ground" for borderline papers.

Prior to actual scoring, the trainer and the most qualified or experienced raters review a subset of the papers to be scored in order to identify "range finders." These are papers that are representative of all the papers at a given scoring level. With a four-point scale, for example, there would be range finders for the 4, 3, 2 and 1 levels. Range finder papers must be so typical of papers at a given level that virtually all readers agree on the assigned score. This is vital because range finders are used in training, and later used as models to assist raters during scoring. Trainers and their assistants may have to read dozens of papers in order to find the "typical" range finder papers with which everyone is satisfied. For training purposes, it is advisable to have at least two (preferably more) range finders at each level.

Trainers do not work from any predetermined set of criteria in identifying range finders. They may, of course, discuss their findings and observations during the process. But it is important to realize that in holistic scoring, there is no preconceived notion of the "ideal" paper. A paper assigned a score of 4 will simply be a relatively high quality paper within a given group; it may or may not be an excellent paper in its own right. As Brown (1977) notes, "It is possible that all of the papers at the top of the score are horribly written. They may be better than the rest, but still may be unacceptable to most teachers of composition." If one has in mind some specific criterion of performance that students must meet, holistic scoring will not be appropriate. Scoring levels are set from within, irrespective of external standards.

Despite personal preferences, the holistic approach quickly produces marked consistency among raters—in virtually

any group. This may be partly the result of peer pressure. But more likely it suggests that language arts people can agree—though the bases for their conclusions may differ—on what constitutes a relatively good and a relatively poor paper. Interrater reliability (that is, agreement between any two raters) can be expected to run from about .60 to .80 (Diederich, 1974). It may be higher in a few cases, depending upon the background of the raters and the amount of training time allowed (so that raters can internalize the system).

All papers should be read by at least two raters to minimize the chance of error resulting from rater fatigue, prejudice or other extraneous factors. ACT has achieved an interrater reliability of .75 using two raters and three writing samples (ACT, 1979). Increasing the number of raters beyond two does not seem to enhance score reliability (Steele, 1979).

Scores may be added or averaged across raters to determine a final score. Disagreements of more than one rating point should be resolved by a third reader or through discussion by the disagreeing raters. Such disagreements can typically be expected to occur in fewer than 5 percent of all cases if careful assessment planning and rater training is conducted.

Holistic scoring is rapid and efficient. Depending on the length of student responses, experienced raters can usually go through 30 to 40 papers per hour (though inexperienced raters cannot be expected to match this rate). Six hours of scoring per day is considered about maximum to maintain high reliability. Scoring is intensive work; short hours with frequent break periods yield the best results.

Because scoring levels are never defined, holistic scoring does not permit the reporting of specifics on student performance. After reading hundreds of papers, however, raters typically have a supremely clear notion of what factors influenced them to assign particular scores. For reporting purposes they may translate those observations into level definitions. Suppose, for example, that students were asked to write a job application letter. One might then say that a "typical" 4 paper used proper business letter format, used vocabulary and tone appropriate to the occasion, described the student's qualifications in a way that reflected a clear understanding of job requirements (as presented in the item), and reflected consistently good sentence structure, correct mechanics, and so on. Such a definition would not necessarily

apply in total to every 4 paper, but would certainly capture the essence of papers at that level and help make results meaningful to parents and other audiences. Presentation of such definitions in conjunction with sample student papers can be an extremely effective reporting technique.

Analytical scoring. Analytical scoring involves isolating one or more characteristics of writing and scoring them individually. Analytical scoring is most appropriate if one wants to measure (and report) students' ability to deal with one or more specific conventions of writing: punctuation, organization, syntax, usage, creativity, sense of audience, and so on. Criteria of this sort must be explicit and complete, and must be well understood by all raters. Except for the setting of criteria, training procedures are similar to those for holistic scoring.

Analytical scoring provides data on specific aspects of student writing performance. But does it really reveal whether, in general, students write well? The answer depends on (1) whether enough traits are analyzed to provide a comprehensive picture, and (2) whether those traits analyzed are significant—that is, whether they actually contribute to good writing. In an effort to identify those characteristics that seem most to influence a reader's judgment about the quality of a piece of writing, Diederich (1974) performed a content analysis on a sample of student essays scored holistically. Marginal comments were invited (as would not be the case in a traditional holistic session), and later tallied to isolate those factors that seemed to influence experienced raters' scores most. Here, in order of significance, are the factors Diederich isolated through that study:

1. Ideas
2. Mechanics (including usage, punctuation and spelling)
3. Organization
4. Wording
5. Flavor (or style)

Of course, individual examiners may identify other traits they wish to score. However, this list of traits permits a reasonably comprehensive analysis of writing.

Factor-by-factor analysis of writing elements is more time consuming than holistic scoring. Depending on how many factors one looks at, it requires two to three times as long (or more) to rate a paper analytically as it does holistically.

Analytical rating has been criticized because there is some indication it produces a "halo" effect: that is, students who are rated high on one trait will tend to be rated high on all traits. Page (1968) explains,

> A constant danger in multi-trait ratings is that they may reflect little more than some general halo effect, and that the presumed differential traits will really not be meaningful. . . . We find (in our research) a very large halo, or tendency for ratings to agree with each other.

Despite these disadvantages, however, analytical scoring has one great advantage: it provides potential for trait-by-trait analysis of students' writing proficiency.

Primary trait scoring. Primary trait scoring is similar to analytical scoring in that it focuses on a specific characteristic (or characteristics) of a given piece of writing. However, while analytical scoring attempts to isolate those characteristics important to any piece of writing in any situation, primary trait analysis is rhetorically and situationally specific. The most important—or primary—trait(s) in a letter to the editor will not likely be the same as that (those) in a set of directions for assembling a bicycle.

The primary trait system is based on the premise that all writing is done in terms of an audience, and that successful writing will have the desired effect upon that audience. For example, a good mystery story will excite and entertain the reader; a good letter of application will get the interview. In a scoring situation, of course, papers must be judged on the likelihood of their producing the desired response.

Because they are situation-specific, primary traits differ from item to item, depending on the nature of the assignment. Suppose a student were asked to give directions for driving from his/her home to school. The primary trait might then be sequential organization, for any clear, unambiguous set of directions would necessarily be well organized with details presented in proper order. As Mullis (1974) points out, "Successful papers will have that [primary] trait; unsuccessful papers will not—regardless of how well written they may

be in other respects."

Raters determine that some traits are essential to success in a given assignment. However, additional traits that contribute but are not necessarily essential to the success of a paper are termed "secondary" traits and may also be included in the evaluation, if they can be clearly defined and exemplified for raters. Scores may be weighted to show the relative importance of various traits, if desired, then totalled to indicate the overall quality of the paper.

The first step in primary trait scoring is to determine which trait or traits will be scored. The second is to develop a scoring guide to aid raters in assigning scores. To illustrate, consider the following guide developed by NAEP for scoring "letters to the principal on solving a problem in school." It was determined that a good letter would identify the problem, present a solution, and explain how that solution would improve the school. Here are NAEP's criterion levels:

1. Respondents do not identify a problem or give no evidence that the problem can be solved or is worth solving.
2. Respondents identify a problem and either tell how to solve it or tell how the school would be improved if it were solved.
3. Respondents identify a problem, explain how to solve the problem, *and* tell how the school would be improved if the problem were solved.
4. Respondents include the elements of a "3" paper. In addition, the elements are expanded and presented in a systematic structure that reflects the steps necessary to solve the problem (Mullis, 1974).

Range finder papers may be used in addition to the scoring guide.

All raters should be familiar with the rationale underlying the primary trait system, and with the level definitions to be used in scoring. Raters must accept the fact that they will be looking for specific, well-defined traits, and be cautious about allowing extraneous criteria to influence scoring. NAEP recommends that raters prescore (for practice) at least 10 sample papers at each level during training in order to become comfortable with applying the criteria (Mullis, 1974).

As with analytical scoring, defining criterion levels is the

most time consuming step. It may be necessary to "test" numerous definitions on sample papers in order to come up with a set that works. Herein lies a strong argument for keeping the list of traits to be scored brief. On an average, count on a day of trial and error, discussion and debate for each trait to be defined. This may sound time consuming, but the quality and clarity of the final definitions, and the ease with which they can be applied, will readily justify the time spent.

Like analytical scoring, primary trait scoring can allow the reporting of student performance with respect to specific characteristics: e.g., organization, awareness of audience. For this reason, primary trait scoring is greatly favored over holistic scoring in contexts where more precise information is needed. But this advantage should be carefully weighed against the time and effort required to set up a workable primary trait scoring system. Aside from adopting already written criteria (e.g., from NAEP), there are no known shortcuts.

Scoring language usage and mechanics. Of the types of scoring mentioned thus far, the scoring of writing mechanics is the most time consuming, and the most complex approach for which to provide training. This realization often comes as a great surprise to inexperienced raters, who may look on mechanics as a rather cut and dried affair—until faced with the prospect of setting up a scoring system.

The fact is, the standards of appropriate usage are subject to continual change through popular usage. So rapid has that change become now that even usage textbooks sometimes reflect different notions of what is appropriate. For the sake of consistency in scoring mechanics, it is necessary that a fairly comprehensive guide be developed. It is possible, of course, to use a standard reference—an English handbook—for this purpose. But raters must agree to abide by the document, and if there are too many areas of disagreement, it may be simpler to design their own. Whatever the decision, it is imperative that everyone agree to score according to the rules of the guide, regardless of personal preference. Otherwise, the inconsistency will render the scores useless.

Several other decisions must be made as well:

1. Whether to count errors of commission and errors of omission equally.

2. Whether to require formal usage, or to base guide rules on informal usage.

3. Whether to count errors involving concepts or rules with which students may not be familiar.

4. Whether to count every identifiable error or to focus on specific areas for easier reporting of results.

In addition, raters must establish a workable rating scale. If they choose to retain a 4-point scale, for example, it will be necessary to determine how many errors will be allowed in a 4 paper, how many in a 3 and so on.

One additional step necessary in scoring writing mechanics is obtaining an accurate word count for each paper. Errors can then be tabulated per 100 words. Analyzing errors in this way does not penalize those who write long responses, or give unfair advantage to those who write very little.

Test administrators should be cautioned about scoring mechanics as one trait within a primary trait system. As the foregoing discussion indicates, it is far more time consuming to score than other traits, and demands a number of special considerations. Therefore, test administrators should weigh carefully the advantages and disadvantages of such a combined approach.

Educators considering using the direct assessment approach to evaluate mechanics should remember that understanding of such usage elements as punctuation, grammar, diction, and sentence structure can be very efficiently, validly and reliably assessed using available indirect assessment measures. For mechanics or usage assessment, very careful consideration should be given to the objective test because it forces examinees to demonstrate explicit ability to deal effectively with the precise elements being tested. If a writing sample is used to assess these elements, examinees will typically avoid language constructions which they are unable to use effectively. Further, inconsistencies in usage patterns will make comparisons among examinees, on the basis of mechanics, difficult if not impossible. Such comparisons are generally possible with objective usage tests. In addition, because a writing sample taps but a small, arbitrary portion of an examinee's proficiency in writing mechanics, results cannot appropriately be used in diagnosis, whereas objective

test results may be quite suitable for this purpose.

T-unit analysis. The concept of T-unit analysis was introduced in the 60s, and has gained popularity ever since as a means of measuring writing sophistication. A T-unit may be thought of as an independent clause plus whatever subordinate clauses or phrases accompany it. In simple terms, a T-unit is the smallest group of words in a piece of writing that could be punctuated as a sentence (T stands for "terminable"). Consider the following passage:

> I yelled at my cat Manfred and he ran away, but he came home when he got hungry.

This passage has only one terminal mark of punctuation as written, but actually contains three T-units:

- I yelled at my cat Manfred
- and he ran away,
- but he came home when he got hungry.

Each of these T-units is an independent clause that could be punctuated as a sentence. Note that T-unit analysis is independent of punctuation: a writer may or may not punctuate T-units as sentences.

Studies have shown that T-unit length tends to increase with the age and skill of the writer* (Hunt, 1977). In addition, it has been demonstrated that with increased skill, writers can incorporate a greater number of distinct concepts into a single T-unit. Consider the following example, using six short sentences, each of which consists of one T-unit, abstracted from a longer piece:

1. Aluminum is a metal.
2. It is abundant.
3. It has many uses.
4. It comes from bauxite.
6. Bauxite looks like clay.

*There are notable exceptions: therefore, this tendency cannot be applied as a general rule. Highly experienced, sophisticated writers may consistently use short T-units. Conversely, the use of lengthy T-units does not of itself render one a skillful writer.

Table 2

A Comparison of Scoring Methods for Direct Writing Assessment

DESCRIPTOR	HOLISTIC	ANALYTICAL
GENERAL CAPABILITIES	Comprehensive, general picture of student performance; writing viewed as a unified coherent whole. Applicable to any writing task.	Thorough, trait by trait analysis of writing; provides comprehensive picture of performance *if* enough traits are analyzed; traits are those important to *any* piece of writing in *any* situation (e.g., organization, wording, mechanics).
RELIABILITY	High reliability if standards are carefully established and raters are carefully trained.	High reliability if criteria and standards are well defined, and careful training is conducted.
PREPARATION TIME	Up to one day per item to identify range finder (model) papers; up to one-half day to train readers using 4-point scale; full day to train with 8-point scale.	One full day to identify traits; one day per trait to develop scoring criteria (unless traits and criteria are borrowed from another source); one to two days to review results of pilot test and refine traits and criteria as necessary; one-half day to train raters.
READERS	Qualified language arts personnel recommended; high reliability can be achieved with non-language arts readers given sufficient training.	Qualified language arts personnel recommended.
SCORING TIME	One to two minutes per paper (experienced readers may read faster).	One to two minutes per paper per trait.
CLASSROOM USE	May be adapted for use in class.	May be adapted for use in class.
REPORTING	Allows reporting on students' overall writing skill.	Allows reporting of student performance on wide range of generalizable traits (i.e., the qualities considered important to all good writing).
GROUP/ SAMPLE SIZE*	Primarily usable with a larger sample; with a small sample, responses may be difficult to scale.	Best with smaller samples; extensive scoring time *may* make costs prohibitive with larger groups.

*These are very general guidelines. Due to the nature of the scoring-cost/amount-of-information trade-off across scoring methods, readers are urged to seek the technical assistance of a qualified writing assessment specialist if there is a question regarding the appropriate use of available scoring resources.

PRIMARY TRAIT	WRITING MECHANICS	T-UNIT ANALYSIS
Highly focused analysis of situation-specific primary trait (and possibly secondary traits): provides specific information on a narrowly defined writing task (e.g., ability to recount details in chronological order).	Can provide either a general or a specific profile of the student's ability to use mechanics properly.	Provides a measure of syntactical sophistication.
High reliability if criteria and standards are well defined, and careful training is conducted.	High reliability if given sufficient training time and authoritative, complete, acceptable guidelines (e.g., an English handbook).	High reliability provided trained and experienced raters are used.
One full day to identify traits; one day per trait to develop scoring criteria (unless traits and criteria are borrowed from another source); one to two days to review results of pilot test and refine traits or criteria as necessary; one-half day to train raters.	One to two days to set up a scoring system (unless borrowed from another source). Minimum of one day to internalize the scoring system and practice scoring.	Half day to full day, depending on raters' previous experience.
Qualified language arts personnel recommended; non-language arts staff may be able to score some traits.	Qualified language arts personnel recommended.	Raters *must* be experienced language arts personnel, preferably those already familiar with the concept of T-unit analysis.
One to two minutes per paper per trait.	Five minutes or more per paper, depending on number of criteria.	Varies greatly, depending on raters' skill.
May be adapted for use in class.	May be adapted for use in class.	May be adapted for use in class.
Allows reporting of student performance on one or more situation-specific traits important to a particular task.	Allows reporting of group or individual data on students' general strengths or weaknesses in mechanics.	Allows group or individual reporting on syntactical sophistication.
Generally more cost-effective with smaller samples, depending on the number of traits to be scored (with one trait, sample size is not an issue).	Best with smaller samples; extensive scoring time *may* make costs prohibitive with larger groups.	Best with smaller samples; extensive scoring time *may* make costs prohibitive with larger groups.

Here's how a fourth grader rewrote the passage:

> Aluminum is a metal and it is abundant. It has many uses and it comes from bauxite. Bauxite is an ore and looks like clay. (6 sentences to 5 T-units)

The revision of a typical eighth grader:

> Aluminum is an abundant metal, has many uses, and comes from bauxite. Bauxite is an ore that looks like clay. (6 sentences into 2 T-units)

And finally, the revision of a skilled adult, a professional writer:

> Aluminum, an abundant metal with many uses, comes from bauxite, a claylike ore. (6 sentences into 1 T-unit)

T-unit analysis and review of conversions (from simple sentences into T-units) provide a good measure of sentence maturity and of a student's ability to consolidate multiple thoughts.

Sophisticated, condensed writing has undeniable appeal. T-unit analysis used in conjunction with holistic scoring is likely to reveal that the highest scored papers (i.e., those that appealed most to readers) were in fact those with the most sophisticated use of T-units.

T-unit analysis is still in the experimental stages. It is time consuming and costly to conduct. Moreover, it can only be done by highly trained language arts specialists. Further research and use may, however, reveal more widespread applicability than has so far been anticipated. Two interesting footnotes: syntactical maturity is apparently reflected in oral speech as well as in writing, and such maturity can be enhanced through a sentence combining curriculum (Hunt, 1977).

A Comparison of Scoring Methods

Table 2 offers a comparative overview of the scoring procedures discussed in this section, focusing on several key descriptors.

Chapter III: Adapting Writing Assessment to Specific Purposes

Educational tests have only one function: to facilitate educational decision making. A test should not be administered, therefore, until the decision or decisions that rest on the results of that test have been clearly articulated. This applies to all tests, including writing tests.

In many educational contexts, writing tests can be and are being used effectively. For example, tests can play a role in instructional management decisions. Such decisions include (1) the diagnosis of individual learner strengths and weaknesses for instructional planning, (2) the placement of students into the next most appropriate level of instruction, and (3) educational and vocational planning as part of student guidance and counseling.

Tests can also be administered at key points in an educational program to check student development in order to (1) screen the admission to an advanced or remedial program, or (2) certify minimum proficiency (e.g., for high school graduation).

And finally, tests can be used for program evaluation purposes such as in (1) large-scale survey assessment, (2) formative program evaluation, and (3) summative program evaluation.

In the discussion that follows, each of these eight contexts is described in terms of the decision to be made, the primary decision makers, and the type of writing skill information needed to make the decision. Decision makers include stu-

dents, parents, teachers, administrators (including specific project or program administrators, as well as building-, district- and state-level administrators), guidance counselors, and the public (including taxpayers and elected officials).

Using Tests to Manage Instruction

Diagnosis. Teachers often use tests and other performance indicators to track each student's level of development, thereby determining where that student is in the instructional sequence, and anticipating the next appropriate level of instruction. Diagnostic data gathered via direct writing assessment can help individualize instruction by simplifying student grouping or instructional scheduling decisions. In addition, diagnostic writing skill data gathered over time may provide a basis for grading or communicating progress to parents.

Placement. Decision makers such as teachers and educational administrators must place each student at the level of instruction best suited to his/her skills. Typically, they use such performance indicators as writing skill tests, previous courses completed, and grades to rank order students along a continuum of writing skill development, then place them in the appropriate course.

Guidance and Counseling. In deciding their future educational or vocational activities, students need to know how their writing skill compares to that of other students with whom they could compete. Performance indicators like writing tests can help provide such information. Writing tests can indicate the probability that a given student will find success and satisfaction in a program or professional position for which writing skill is a prerequisite. More specifically, normative test data can help students, their parents and their guidance counselors answer students' typical questions: Should I pursue advanced training in a postsecondary educational program in which writing is a key element? In which school or job am I most likely to be successful? Though test scores should never serve as the sole basis for answering such questions, they can play a valuable role.

Using Tests to Screen Students

Selection. It is not uncommon to have more candidates than program openings. When this happens, teachers, counselors and administrators must select students for admis-

sion. Performance indicators such as writing tests can be used to rank order examinees to facilitate selection. Selection decisions most often affect those at either end of the skill continuum. That is, more able students are selected for inclusion in advanced writing programs, while less able students are selected for remedial writing programs.

Certification. Tests tailored to a specified certification domain are often used to verify and document a student's mastery of specific knowledge or skills. For example, teachers might use writing tests to certify mastery of beginning writing skills for purposes of grading or promotion. Or district and state administrators might use minimum writing competency tests as criteria for high school graduation. Both examples show how certification may be accomplished through testing.

Using Tests to Evaluate Programs

Survey Assessment. Survey assessment refers to the collection of group achievement data to determine general educational development (e.g., in writing). Data may be gathered by administering a writing test to a carefully selected random sample of students in the target population. Survey assessment is often cyclical, thus allowing for the examination of trends in writing skill development over time. Decision makers include (1) building-, district- or state-level administrators who allocate resources for special instructional needs pinpointed by the assessment, or (2) the public, which makes value judgments regarding perceived and reported levels of student writing skill development.

Formative Evaluation. In the context of formative program evaluation, program administrators and teachers attempt to determine which components of instruction are functioning as intended and which need further refinement. They may test students on each of the intermediate and final outcomes of a writing program, for example. Assessment for formative evaluation may also involve multiple test administrations to determine the effectiveness of ongoing modifications in a writing program.

Summative Evaluation. Summative evaluation reveals a program's overall merit, suggesting whether that program should be continued or terminated. Tests designed to assess students' performance on final learning outcomes are an important part of such an evaluation. Teachers, program,

building or district administrators, and the public (including the board of education) may be involved in summative evaluation decisions. As with survey assessment and formative evaluation, multiple test administrations are common. Tests may be given prior to as well as following instruction, with retention testing after a given time interval.

Selecting Examinees as a Function of Purpose

In the three program evaluation contexts just cited (survey assessment, formative evaluation, and summative evaluation), testing costs can be significantly reduced through random sampling. If the student population is very large, then data summarized across a carefully selected random subset of students will reflect group performance every bit as accurately as if every student were tested—often at a fraction of the cost. It is not within the scope of this paper to present all the important considerations in sampling, as each specific educational situation is unique. The intent is to point out the potential financial advantage of sampling and to urge its consideration.

It should be apparent that sampling is not feasible with instructional management or student screening decisions because in these contexts, individual student data are necessary.

Developing Exercises as a Function of Purpose

Generally, the process for developing writing assessment exercises remains constant across all eight educational assessment contexts. Careful planning is essential in all cases, and attention must always be given to designing exercises that give the examinee sufficient opportunity (in terms of time, appropriate stimulus and range of tasks) to demonstrate proficiency. Further, in all cases, the type of audience and purpose for communication should be made clear to the student. In addition, exercises should frame challenging tasks based on varied and directly relevant stimulus materials. And finally, in all cases, clear and concise instructions are essential.

A few factors vary according to context and the nature of the decisions to be made. As a general rule, the specificity of an exercise (i.e., level of detail in instructions) should increase along with the specificity of the skills to be assessed. In other words, exercises to be used in broad survey assessment need

not be quite so focused as exercises to be used in, say, a diagnostic test.

The amount of writing required might also vary, depending on the decisions to be made. For example, it might be possible to rank order students in terms of general writing proficiency (via holistic scoring) on the basis of three or four general, relatively short writing samples. However, it would probably be very difficult to use those same three or four short writing samples to reliably and validly determine whether a student had mastered 10 to 15 specific, independent writing skills. Generally, the more precise and numerous the criteria and standards of acceptable performance, the more writing needed to evaluate performance.

And finally, exercises developed for use in a large-scale statewide assessment or where important selection decisions are pending *must* be (1) independently reviewed by writing and assessment experts and (2) pretested. Pretesting and review are less critical with writing assessment exercises used in instructional classroom management.

Selecting Scoring Procedures as a Function of Purpose

Selection of scoring procedures is, in effect, part of assessment planning, since this decision is influenced by the purpose for the assessment and criteria to be used in judging writing proficiency. Though it is possible to conceptualize instances within each of the eight educational assessment contexts in which any given scoring approach could be employed, the actual scoring approach most commonly used will vary by context.

To illustrate, diagnosis of individual student strengths and weaknesses demands the level of specificity provided through analytical, primary trait or mechanics scoring. Placement and guidance, on the other hand, may only require holistic ratings because the objective of assessment is simply to rank order students on a continuum of writing skill.

Consider measurement of student status. While selection may require a holistic ranking of students, certification may be done through holistic ratings or analytical or primary trait scoring, depending on the specificity of the minimum competencies to be certified.

Holistic scoring procedures are well suited to the relatively broad, unfocused nature of large-scale survey assessment. However, analytical scoring may serve as well if the desire for

Table 3

Writing Assessment Procedures as a Function of Assessment Context

	Assessment Context		Assessment Procedure	
Context	Decision to be made	Decision makers	Examinees assessed	Exercise specificity
Diagnosis	Determine and track ednca tional development	Teacher Stndent	Individnal	Specific
Placement	Match level of student development to level of instruction	Teacher Connselor	Individnal	General
Guidance	Rank order for edncational planning decisions	Administrator Connselor Teacher Parent Stndent	Individnal	General
Selection	Rank order examinees for selection into instruction	Administrator Connselor Teacher	Individnal	General
Certification	Determine mastery of specific competencies	Teacher Stndent	Individnal	Specific
Snrvey Assessment	Policy decision re: status of stndent edncational development	Administrators Public	Sample	General
Formative Evalnation	Determine components of instructional program in need of revision	Program Developer Teacher	Sample	Depends on program objectives
Summative Evaluation	Program continnation	Administrator	Sample	Depends on program objectives

	Assessment Procedure				
Context	Holistic	Analytical	Primary trait	Mechanics	T-unit
Diagnosis			X	X	X
Placement	X	X			
Guidance	X	X			
Selection	X	X			
Certification		X	X	X	X
Survey Assessment	X	X			
Formative Evaluation		X	X	X	
Summative Evaluation	X	X			

individual data justifies the additional time required.

Scoring procedures for formative evaluation depend on the specificity of the enabling and terminal objectives that guide instruction. If overall writing proficiency is the focus of the program, analytical scoring may be selected. However, if instruction focuses on situation-specific rhetorical skills, primary trait scoring may be most appropriate. Similarly, emphasis on mechanics indicates selection of a corresponding scoring approach. In most instances, formative evaluation demands scoring procedures more specific than holistic.

With summative evaluation, holistic assessment may provide sufficient data to judge program viability. However, if stated program goals subdivide writing skill into component parts, analytical scoring may be appropriate. Instructional programs in writing seldom focus on a single rhetorical circumstance. Rather, they deal with writing of many types, for many purposes. Therefore, primary trait scoring will have limited value in this context.

Ensuring Efficient, Effective, and High Quality Assessment

The keys to successful direct writing assessment are careful planning, thoughtful and creative exercise development, and consistent application of performance criteria during scoring. If these factors are given meticulous attention, the assessment will yield data that are (1) sufficiently precise to support necessary decisions, (2) reliable, (3) valid for the intended purpose, and (4) maximally cost-effective.

The preceding discussion is intended to acquaint the interested educator with available assessment strategies and to highlight some of the issues involved in selecting a scoring procedure appropriate for a specific context. Table 3 provides an overall summary of the key points made in that discussion.

The reader is encouraged to refer to the list of REFERENCES *following this section and to the* APPENDIX, *which names contact persons in many states who can offer further information on writing assessment approaches and contingencies. In addition,* CAPT *welcomes further inquiries regarding writing assessment.*

REFERENCES

American College Testing Program. Alternative strategies for the assessment of writing proficiency. Iowa City, IA: Author. 1979.

American College Testing Program. Content of the tests of the ACT Assessment. Iowa City, IA: Author. 1978.

Braddock, R., Lloyd-Jones, R., and Schoer, L. *Research in written composition*. Urbana, IL: National Council of Teachers of English. 1963.

Breland, H. M. *A Study of college English placement and the Test of Standard Written English*. Princeton, NJ: Educational Testing Service. 1977.

Breland, Hunter M. and Judith L. Gaynor. A comparison of direct and indirect assessments of writing skills. *Journal of Educational Measurement*, Summer 1979. *16* (2).

College Entrance Examination Board. The Test of Standard Written English: A preliminary report. Princeton, NJ: Educational Testing Service. 1975.

Cronbach, L. J. Test validity. In R. L. Thorndike, *Educational Measurement*. Washington, DC: American Council on Education. 1971.

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

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, NJ: Holt Rinehart Winston. 1973.

Fredrick, V. Writing assessment research report: A national survey. Monograph published by the Wisconsin Department of Public Instruction, Madison, WI. 1979.

Hunt, K. W. Early blooming and late blooming syntactic structures. In C. Cooper & L. Odell (Eds.), *Evaluating writing*. Urbana, IL: National Council of Teachers of English. 1977.

Huntley, R. M., Schmeiser, C. & Stiggins, R. The assessment of rhetorical proficiency: The role of objective tests and writing samples. Paper presented at the annual meeting of the National Council on Measurement in Education. 1979.

Mehrens, W. A. and Lehman, I. J. *Measurement and evaluation in education and psychology*. New York: Holt Rinehart Winston. 1973.

Mullis, I. The primary trait system for scoring writing tasks. Denver: National Assessment of Educational Progress. 1974.

National Assessment of Educational Progress. *Writing mechanics 1969-1974: A capsule description of changes in writing mechanics*. Denver: Author. 1975.

Page, E. B. *The analysis of essays by computer* (Final Report, U.S. Office of Education Project 6-1318). Storrs, CT: University of

Connecticut. 1968.
Rivas, F. *Write/rewrite: An assessment of revision skills* (Writing Report No. 05-W-04). Denver: National Assessment of Educational Progress, 1977.
Sax, G. *Principles of educational measurement and evaluation*. Belmont, CA: Wadsworth, 1974.
Steele, J. 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, 1979.
Thorndike, R.L. *Educational measurement*. Washington, DC: American Council on Education, 1971.

ADDITIONAL READINGS

Apstein, B. Deficiencies of CLEP writing examinations. *College Composition and Communication*, 1975, *26*, 350-355.

Brown, Rexford. What we know now and how we could know more about writing ability in America. Paper presented at the NIE Conference on Writing, Los Angeles, June 1977.

Chase, C. I. The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement*, 1979, *16*, 39-42.

Coffman, W.E. On the reliability of ratings of essay examinations in English. *Research in the Teaching of English*, 1971, 5(1), 24-37.

Clemson, E. A study of the basic skills assessment: direct and indirect measures of writing ability. Princeton: Educational Testing Service, 1978.

Cooper, C. Measuring growth in writing. *English Journal*, 1975, *64*(3), 111-120.

Cooper, C. and L. Odell (Eds.). *Evaluating writing*. Urbana, IL: National Council of Teachers of English, 1977.

Della-Piana, G., Odell, L., Cooper, C. & Endo, G. The writing skills decline: So what? *Educational Technology, 1976, 16*(7), 30-39.

Diederich, P. B. How to measure growth in writing ability. *English Journal*, 1966, *55* 435-449.

Fowles, M. E. Manual for scoring the writing sample. Princeton: Educational Testing Service, 1978.

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

Isaac, S. & Michael, W. B. *Handbook in research and evaluation*. San Diego: Robert R. Knapp, 1971.

Jewell, R. M., et al. *The effectiveness of college-level instruction in freshman composition* (Final Report, U.S. Office of Education Project 2188). Cedar Falls, IA: University of Northern Iowa, 1969.

Markham, L. R. Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 1976, *13*, 277-283.

Maxwell, John C. National assessment of writing: Useless and uninteresting? *English Journal*, 1973, *62*, 1254-1257.

McColly, W. What does educational research say about the judging of writing ability? *Journal of Educational Research*, 1970, *64*(4), 148-56.

Mullis, I. Scoring writing mechanics. Denver: National Assessment of Educational Progress, 1974.

Paule, L. Holistic scoring. Unpublished paper, Portland, OR: Northwest Regional Educational Laboratory, 1978.

APPENDIX
Profiles of Statewide Writing Assessments

CALIFORNIA, 1975

Overall Goals:	To determine the general writing skills of twelfth graders.
Specific Components Tested:	Writer's overall skills in punctuation, diction, usage and sentence sense.
Students Tested:	4000 high school seniors in 28 schools throughout California.
Testing Strategy:	40-minute writing sample on one of five randomly assigned topics.
Scoring Procedures:	Papers were scored holistically using a 9-point scale. Two independent ratings were combined to determine the final score (2-18). Only blank papers received a zero. Significant discrepancies were resolved by a third reading. A representative sample of 750 papers were subsequently scored analytically to identify common strengths and weaknesses.
Reporting Results:	Results for the holistic scoring were reported to schools as number of essays per score and percent of essays per score. Al-

so, correlations between essay scores and scores on the written expression section of the *Survey of Basic Skills: Grade 12* (multiple choice items testing fundamental linguistic skills) were calculated.

Contact Person:

Beth Breneman, Consultant
Office of Program Evaluation & Research
California State Department of Education
721 Capitol Mall
Sacramento, CA 95814
Phone: (916) 322-2200

HAWAII, May 1979

Overall Goals:

To determine how well Hawaii's students could write in response to specific writing objectives.

Specific Components Tested:

1. **Expressing feelings:** To express personal feelings clearly and vividly.
2. **Giving information:** To give clear, accurate and complete information to others.
3. **Promoting ideas:** To present a convincing argument.
4. **Entertaining:** To use language artfully to move the reader into the imaginary world of the writer.

Students Tested:

A sample of fourth, eighth and eleventh graders statewide.

Testing Strategy:

Writing sample.

Scoring Approach:

Papers were scored using a primary trait system based on the 4-point primary trait scale developed by NAEP, in which

1 = Absence of the trait
2 = Presence of the trait
3 = Adequate expression of the trait
4 = Excellent expression of the trait

	Secondary and tertiary traits were also scored. Four three-member teams—one for each grade level—scored papers. Two members of each team read and scored each exercise; the third team member acted as judge and final arbitrator in case of substantial differences.
Reporting Strategy:	Alternative reporting strategies are being studied and have not been finalized at the time of this writing.
Contact Person:	Ronald L. Johnson, Administrator Evaluation Section State of Hawaii Department of Education P.O. Box 2360 Honolulu, HI 96804 Phone: (808) 548-6911

IDAHO, 1979

Overall Goals:	To provide a statewide profile of students' writing skills. The writing assessment is one component of the statewide Proficiency Testing Program.
Specific Components Tested:	General writing proficiency as demonstrated by ability to organize and present a persuasive argument in written form (specifically, a letter to the principal).
Students Tested:	Ninth graders statewide. (Those who do not pass will be given an opportunity to retake the test twice a year until they achieve a satisfactory score. Though a passing score is not required for graduation, those who do pass will receive special commendation on their diplomas.)
Testing Strategy:	Writing Sample
Scoring Approach:	Papers were scored holistically on a five-point scale, with 3 designated as a passing score. All papers were read at least twice; discrepant papers and 2/3 splits were given a third reading.

Reporting Strategy:	Individual student score reports sent to participating districts, with information on assessment procedures and performance levels; statewide summary of results disseminated to participating districts; follow-up workshops held for participating districts to deal with results.
Contact Person:	Ms. Driek Zirinski Language Arts Consultant Proficiency Testing Department of Education State of Idaho Boise, ID 83720 Phone: (208) 384-3301

LOUISIANA, 1979-80

Overall Goals:	The assessment was conducted in response to legislative mandate. Its purpose was to determine students' general writing skills, as measured according to minimum standards set by the state.
Specific Components Tested:	Skills within two domains were tested: 1. Types and Forms of Writing • Description • Narration • Exposition • Persuasion 2. Writing Skills • Handwriting • Spelling • Capitalization • Punctuation • Language Structure • Organization • Proofreading
Students Tested:	All fourth, eighth and eleventh graders (including special education students who are tested separately by special education teachers).

Testing Strategy:	Multiple choice and writing sample.
Scoring Approach:	Machine scoring for multiple choice; primary (and secondary) trait scoring for writing sample. (Only a sample of papers were scored.)
Reporting Strategy:	A general report was prepared on the results of the writing sample; results were reported at the state level only. However, for the objective portion of the test, each participating teacher received an individual report on each child in his/her class.
Contact Person:	Donna N. Shows Administration Officer State of Louisiana Dept. of Education P.O. Box 44064 Baton Rouge, LA 70804 Phone: (504) 342-1148

MAINE

Overall Goals:	Maine's writing assessment was conducted in response to a legislative mandate to assess basic skills: reading, writing and math. The test is not part of a graduation requirement, though individual districts have the option to initiate such a requirement if they so choose.
Specific Components Tested:	Ability to write in a social or business context.
Students Tested:	Eighth and eleventh graders statewide.
Testing Strategy:	Multiple choice and writing sample.
Scoring Approaches:	Machine scoring for multiple choice; primary trait scoring for writing sample.
Reporting Strategy:	Results were reported for each item showing comparisons with previous assessments, national assessments, and Northeastern regional percentages. Results were also reported according to school size and student's sex.

Contact Person:	Horace P. Maxcy, Jr. Educational Planner State of Maine Dept. of Educational/Cultural Services Augusta, ME 04333

MASSACHUSETTS, 1975-76

Overall Goals:	Assess writing proficiency in various specific contexts.
Specific Components Tested:	1. Writing to communicate adequately in a social situation 2. Writing to communicate adequately in a business or vocational situation 3. Writing to communicate adequately in a scholastic situation 4. Mechanics
Students Tested:	A representative statewide sample of 9- and 17-year-olds.
Testing Strategy:	Multiple choice and writing sample.
Scoring Approaches:	Multiple choice questions were machine scored. Writing samples were scored holistically on an 8-point scale. Each paper was read at least twice; discrepant papers were given a third reading. Scores were summed to produce a final score of 2-16. In addition, spelling and mechanics were considered in scoring responses to Objective 3: Writing to communicate adequately in a scholastic situation.
Reporting Strategy:	For those items borrowed from NAEP scores were reported in comparison to national percentages and Northeast regional percentages. Scores were also considered in relation to educational region within the state and type of community in which the school was located.

Other reporting variables included student's sex, mother's education, father's education, occupation of head of household, type of high school program, future plans, attitude toward school, and friendliness of school.

Contact Person:

Matthew H. Towle
The Commonwealth of Massachusetts
Department of Education
31 St. James Avenue
Boston, MA 02116
Phone: (617) 727-0190

MISSOURI

Overall Goals:

Writing is assessed as part of the Basic Essential Skills Test (BEST) to identify students that are having problems with basic skills while there is still ample time for remediation.

Specific Components Tested:

The BEST includes two objectives that measure writing skills:

1. The student will demonstrate the ability to write with complete sentences, acceptable sentence structure, acceptable grammatical construction, and correct spelling and punctuation.

   Criteria for evaluating performance on this objective:

   - Legible handwriting
   - Correct spelling
   - Correct punctuation
   - Correct sentence structure
   - Good paragraphing

2. The student will demonstrate the ability to complete a business form correctly and neatly.

   Criteria for evaluating performance on this objective:
   - Legible handwriting

- Demonstrated ability to follow directions (e.g., student prints in capital letters if instructed to do so)
- All required information included in correct space

Students Tested:	All eighth graders in Missouri public school districts.
Testing Strategy:	Writing samples.
Scoring Approach:	Teacher evaluation of proficiency.
Reporting Strategy:	The Department of Elementary and Secondary Education provides suggested criteria for evaluating "satisfactory" student performance, and a suggested format (essentially a checkoff chart) for reporting satisfactory performance. However, districts are free to develop their own criteria and reporting procedures.
Contact Person:	Charles Foster, Director Pupil Personnel Services Missouri Department of Elementary and Secondary Education P.O. Box 480 Jefferson City, MO 65102 Phone: (314) 751-3545

NEW HAMPSHIRE, 1978

Overall Goals:	Generate a statewide profile of student narrative writing proficiency.
Specific Components Tested:	1. Prewriting, including collection and organization of ideas 2. Writing, including drafting narrative text 3. Revision of drafted text
Students Tested:	Random samples of fifth and ninth grades.
Testing Strategy:	Writing samples.

Scoring Approaches:	Holistic scoring with the objective of describing student proficiency at each of four proficiency levels.
Reporting Strategy:	1. Written report of assessment results prepared for legislators and district superintendents 2. Workshop convened to review results with district representatives 3. All assessment exercises released for district use
Contact Person:	Joanne Baker Consultant, English and Language Arts New Hampshire Dept. of Education 64 N. Main Street Concord. NH 03301 Phone: (603) 271-3747

NEW MEXICO, Annual

Overall Goals:	The Writing Skills Appraisal is part of the High School Proficiency Examination required by the New Mexico Basic Skills Plan. The purpose of the test is to ensure that students possess the skills they will need to function successfully as adults.
Specific Components Tested:	Four writing tasks are assessed: 1. Abbreviated message 2. Business letter 3. Description 4. Comparison/Contrast To determine successful completion of each task, performance on the following skills was considered: 1. Legibility 2. Spelling 3. Language mechanics 4. Appropriate language 5. Sentence construction 6. Paragraph construction

7. Cohesiveness and transition
8. Appropriate organization
9. Letter format
10. Ordering

Students Tested: All students in grades 10, 11 and 12.

Testing Strategy: Writing samples and multiple choice. (Note: Specific exercises for each of the four generic tasks are designed by individual school districts and are not part of a statewide writing assessment.)

Scoring Approach: Teacher evaluation of proficiency.

Reporting Strategy: Final verification (whether or not the student passes) must be entered on the student's transcript the year of or the year preceding graduation. Students who perform successfully receive a state "proficiency endorsement" on their diplomas.

Contact Person:
Michael Glover
Elementary/Secondary Education Office
New Mexico State Dept. of Education
Education Bulding, Capitol Complex
Santa Fe, NM 87503
Phone: (505) 827-5391

NEW YORK,* Annual

Overall Goals: To assure the early identification of students who need special help and to assure that students have acquired an adequate competence before receiving a high school diploma.

Specific Components Tested: The writing test consists of three tasks:

1. A business letter registering a complaint and requesting corrective action

2. A report based upon data supplied

*Program being revised for 1979-80.

3. A statement of about 200 words that will persuade a specific audience

Students Tested: All students in grade 9, retested if necessary in grades 11 and 12.

Testing Strategy: Writing sample.

Scoring Approach: Papers are scored holistically on a 4-point scale. Raters are given guidelines for rating papers: excellent (4), very good (3), minimally acceptable (2), and very poor (1). In particular, raters are told to emphasize content, organization and development, and mechanics.

Each of the student's three papers is read by a different rater. The three scores are summed to determine a final score.

Reporting Strategy: Student and parents informed of pass/fail results; pass/fail percentage reported to districts; and statewide results made available to the public.

Contact Person:
Charles Chen
Bureau of English Education
State Department of Education
Washington Avenue
Albany, NY 12234
Phone: (518) 474-5917

OHIO,* 1977 and 1978

Overall Goals: To assess general strengths and weaknesses in specific subject areas.

Specific Components Tested: The eighth grade test covered three objectives:

The student will demonstrate an ability to write—

1. To reveal personal feelings and ideas through free expression.

*Note: As of July 1979, Ohio's writing assessment program has been phased out.

2. To communicate adequately in a social situation.

3. To communicate adequately in a business situation.

The twelfth grade test covered two objectives:

The student will demonstrate an ability to—

1. Plan, write and edit a communication adequately.

2. Write to communicate adequately in a business situation.

Students Tested: Eighth graders in 1977; twelfth graders in 1978.

Testing Strategy: Writing sample and multiple choice.

Scoring Approach: Machine scoring for multiple choice; primary and secondary trait and mechanics scoring for writing samples.

Reporting Strategy: The Department prepared an executive summary and technical report which were issued to all schools involved in the assessment. Results were reported on a statewide basis by sex, race, socioeconomic level and type of district.

Contact Person:
Jim Payton
Educational Consultant
Ohio Department of Education
Columbus, OH 43215
Phone: (614) 466-3641

OREGON, 1978

Overall Goals: To provide a general profile of student performance statewide.

Specific Components Tested: General writing skills as evidenced through performance on a business letter.

	friendly letter, or how-to-do-it narrative. Fourth and seventh graders also responded to objective items designed to measure—
	1. Writing conventions.
	2. Grammar.
	3. Organization.
Students Tested:	A representative statewide sample of fourth, seventh and eleventh graders.
Testing Strategy:	Writing sample and multiple choice.
Scoring Approach:	Multiple choice items were machine scored; writing samples were scored holistically using a four-point scale.
Reporting Strategy:	Press releases were issued shortly following the assessment. In addition, a report detailing results was prepared by the Department and made available to educators, legislators and interested members of the general public. Individual tests were returned to parents; however, no individual, school or district data were reported.
Contact Person:	Barbara Cole, Coordinator Oregon Statewide Assessment Oregon Department of Education 700 Pringle Parkway SE Salem, OR 97310 Phone: (503) 378-2923

PENNSYLVANIA, 1978-79

Overall Goals:	To gain a general picture of skills deemed necessary to produce coherent written material.
Specific Components Tested:	Three separate tests were administered, one to each grade level tested. Most skills assessed were common to all three grade levels. However, there were slight differences.

The fifth grade writing assessment measured skill areas, including—

1. Punctuation.
2. Use of regular and irregular verbs.
3. Transforming sentences with no change in meaning.
4. Choosing appropriate language for a given purpose.
5. Choosing opening or topic sentences.

The eighth grade writing assessment measured skill areas including—

1. Placing modifiers.
2. Paraphrasing sentences.
3. Determining relevance of ideas to an essay.
4. Choosing appropriate language for a given purpose.
5. Making a transition to a new paragraph.

The eleventh grade writing assessment measured skill areas including—

1. Combining sentences with clarity.
2. Choosing appropriate language for a given purpose.
3. Choosing a sentence to develop a topic.
4. Including critical information in a message.

Students Tested: Approximately 30,000 students at each grade level were tested. There were three forms of the fifth grade test, four of the eighth grade test and four of the eleventh

grade test. Multiple matrix sampling ensured that approximately an equal number of students from each building received each form.

Testing Strategy: Multiple choice.

Scoring Approach: Machine scoring.

Reporting Strategy: Results were reported to schools in several different formats:

1. General Summary, listing each goal area, the number of students who were given scores for each goal area, the average (mean) of the raw student scores, percentile rank statewide of the school in each goal area, and the mean raw score range predicted for the school.
2. Percentile Bands by Goals, showing the prediction band by school for each goal area.
3. Condition Variables, showing information from administrative records and responses from teacher and student questionnaires.
4. Summary of Criterion Referenced Information for Each Goal Area.
5. Item Frequency Analysis, listing the percentage of responses for each item.

Contact Person: Richard I. Kohr, Research Associate
Bureau of Research and Evaluation
Pennsylvania Department of Education
333 Market Street
Harrisburg, PA 17126
Phone: (717) 787-4234

RHODE ISLAND, Annual

Overall Goals: To provide data on writing skills as part of statewide objective referenced testing program.

Specific Components Tested:	Typical topics include completing a resume, expressing personal ideas and values, and generating written criteria to judge a work of art.
Students Tested:	Eleventh graders.
Testing Strategy:	Multiple choice plus writing sample.
Scoring Approach:	Machine scoring for multiple choice; analytical scoring for writing sample.
Reporting Strategy:	Scores are reported in terms of the average percentage of items correctly completed.
Contact Person:	Martha C. Highsmith, Consultant Statewide Assessment Program Rhode Island Department of Education 199 Promenade Street Providence, RI 02908 Phone (401) 277-3126

TEXAS, 1978

Overall Goals:	To obtain a statewide profile of students' writing performance, and to compare the performance of Texas students with that of students nationwide (as assessed by NAEP).
Specific Components Tested:	Items were borrowed from NAEP and were designed to measure (1) performance on specific writing tasks, (2) mechanics of written expression, and (3) recognizing appropriate writing and valuing written communication. Students in each age group were asked to respond to two kinds of writing assignments: explanatory or persuasive letters and a descriptive essay. In addition, nine-year-olds were given an exercise in expressive writing. Time was adequate for simple corrections, but no additional time was provided for editing or revising.
Students Tested:	A representative statewide sample of 9-, 13- and 17-year-olds.

Testing Strategy:	Writing samples to measure objectives 1 and 2; multiple choice items to measure objective 3.
Scoring Approach:	Multiple choice items were machine scored. Writing samples were hand scored using the primary trait system and criteria developed by NAEP.
Reporting Strategy:	Generally, results on writing samples were reported as percentage of students scoring at each of four carefully defined criterion levels; comparisons with national percentages (provided by NAEP) were also offered. In addition, performance was reported relative to the following variables: (1) family income status, (2) ethnicity, (3) size and type of school district, (4) per pupil expenditure, (5) student's sex, and (6) language spoken in the home.
Contact Person:	Keith L. Cruse, Division Director Educational Assessment Texas Education Agency 201 East 11th Street Austin, TX 78701 Phone: (512) 475-2066

VERMONT, Annual

Overall Goals:	The Vermont Basic Competency Program is designed to give every Vermont student an opportunity to learn to read, write, listen, speak, compute and reason. In 1977-78, a statewide assessment was conducted to determine which specified competencies in each of these areas Vermont students had mastered (competencies were those identified by Vermonters as necessary to successful functioning in today's society). By 1981, mastery of all competencies will be one requirement for graduation from high school.
Specific Competencies Tested:	Eight competencies have been identified for writing:

1. The student will write all required material, including signature, legibly in manuscript and cursive.

2. Given a list of commonly misspelled words, the student will spell them with 80 percent accuracy.

3. Given material to copy, the student will do so with no errors or omissions.

4. Given directions to write a message related to his/her own interests and environment, the student will write a message that will be clear to the receiver and will contain no more than two grammatical errors.

5. Given forms such as application blanks and order forms, the student will complete them correctly and neatly with no omission of essential information.

6. Given directions to write a friendly letter, to fold it correctly and to address the envelope, the student will do so using correct form and having no more than two errors in grammar or punctuation.

7. Given directions to write a business letter, to fold it and to address the envelope, the student will do so with no errors in form, grammar or punctuation.

8. Given directions to select a topic of interest or importance to him/her—including personal opinion—and to write in complete sentences, to use the dictionary as needed to check spelling, and to proofread carefully, the student will write one page of organized material with a total of no more than five errors in grammar, usage, spelling and punctuation.

Students Tested: All students in the state.

Testing Strategy: Writing samples.

Scoring Approach:	Teacher evaluation of proficiency.
Reporting Strategy:	Teachers enter results in classroom records; schools enter results on permanent records; commissioner issues statewide proficiency report.
Contact Person:	Ms. Pat Austin English/Language Arts Consultant State of Vermont Dept. of Education Montpelier, VT 05602 Phone: (802) 828-3111

WASHINGTON, November 1976 and May 1977

Overall Goals:	Both assessments used NAEP objectives and items that (1) related as much as possible to current curricular trends and emphases in Washington, (2) measured significant or worthwhile skills, knowledge or understanding, (3) related both to in-school and out-of-school applications and requirements, and (4) offered the opportunity to compare Washington's eighth and eleventh graders' performance with that of comparable populations nationally and regionally (a legislative requirement).
Specific Components Tested:	1. Writing to reveal personal feelings and ideas through free expression 2. Writing in response to societal demands and obligations
Students Tested:	Representative statewide samples comprising approximately 1600 eighth graders in 1976 and 1500 eleventh graders in 1977.
Testing Strategy:	Writing samples and multiple choice.
Scoring Procedures:	Multiple choice items were machine scored; writing samples were hand scored using primary trait procedures and criteria developed by NAEP.
Reporting Strategy:	Results were reported in comparison to na-

	tional and western regional scores provided by NAEP. Results were also reported in relation to type of community, sex of student and age group.
Contact Person:	Gordon B. Ensign, Jr., Supervisor Testing and Evaluation State Superintendent of Public Instruction Old Capitol Building Olympia, WA 98504 Phone: (206) 753-3449

WISCONSIN. March 1980

Overall Goals:	To obtain a statewide profile of how well Wisconsin public school students demonstrate expected skills and knowledge and perform compared to the rest of the nation.
Specific Components Tested:	Fourth graders are assessed in the areas of social writing and scholastic writing. They are asked to: 1. Write a friendly letter (social) 2. Write a set of directions (social) 3. Write a descriptive essay (scholastic) 4. Write an expository essay (scholastic) Eighth graders are assessed in the areas of social writing, business/vocational writing and scholastic writing. They are asked to: 1. Write a set of directions (social) 2. Write a telephone message (social) 3. Write a business letter (bus./voc.) 4. Fill out an application blank (bus./voc.) 5. Write a descriptive essay (scholastic) 6. Write an expository essay (scholastic)

	Eleventh graders are assessed in the areas of social writing, business/vocational writing and scholastic writing. They are asked to: 1. Write a set of directions (social) 2. Write a telephone message (social) 3. Write a job application letter (bus./voc.) 4. Fill out an application blank (bus./voc.) 5. Write a descriptive essay (scholastic) 6. Write a persuasive essay (scholastic)
Students Tested:	Representative statewide samples composed of approximately 4800 students at each grade level tested (grades 4, 8 and 11).
Testing Strategy:	Writing samples and multiple choice. (The Comprehensive Test of Basic Skills in the area of language arts is given only to the students who wrote the holistically scored scholastic writing exercises.)
Scoring Procedures:	Multiple choice items (CTBS) are machine scored. Social writing samples, business/vocational writing samples and the persuasive essay are scored by the primary trait method. All other samples are scored holistically, and the scores on the holistically scored exercises are correlated with the scores on the CTBS. The holistic scale is 1-8. Each exercise is read twice, and the two scores are added together to produce a final score of 2-16. On both the primary trait and holistic exercises, third readers resolve discrepancies.
Reporting Strategy:	Results on the CTBS are reported in comparison to national norms. Holistic and primary-trait scores are reported as percent of samples receiving each score. A correlation study between CTBS scores

and the holistic scores is reported.

Contact Person:

Vicki Fredrick, Assessment Specialist
Pupil Assessment Program, Room 227B
Wisconsin Dept. of Public Instruction
126 Langdon Street
Madison, WI 53702
Phone: (608) 267-7268