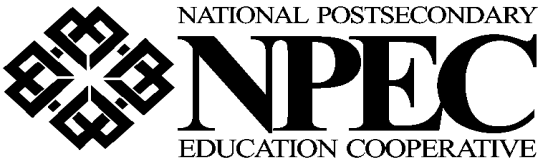
The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing

National Postsecondary Education Cooperative
Student Outcomes Pilot Working Group:
Cognitive and Intellectual Development



The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing

National Postsecondary Education Cooperative
Student Outcomes Pilot Working Group:
Cognitive and Intellectual Development

Prepared for the National Postsecondary Education Cooperative (NPEC) and its Student Outcomes Pilot Working Group by T. Dary Erwin, Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, under the sponsorship of the National Center for Education Statistics (NCES), U.S. Department of Education.

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
1990 K Street
Washington, DC 20006

March 2000

The NCES World Wide Web Home Page address is
http://nces.ed.gov

Suggested Citation

U.S. Department of Education, National Center for Education Statistics. *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing,* NCES 2000--172, prepared by T. Dary Erwin for the Council of the National Postsecondary Education Cooperative Student Outcomes Pilot Working Group: Cognitive and Intellectual Development. Washington, DC: U.S. Government Printing Office, 2000.

Contact
Nancy Borkow
202–502-7311

National Postsecondary Education Cooperative
Student Outcomes Pilot Working Group:
Cognitive and Intellectual Development

Members:

Toni Larson, Executive Director, Independent Higher Education of Colorado (Chair)

Paul Brinkman, Director, Planning and Policy Studies, University of Utah

Sal Corrallo, Consultant, Potomac Management Services

Eric Dey, Associate Professor, Higher Education, University of Michigan

Mary Golladay, Program Director, Education and Human Resources Group, National Science Foundation

Henry Hector, Executive Director, Alabama Commission on Higher Education

Jacqueline King, Director, Federal Policy Analysis, American Council on Education

Cheryl Lovell, Assistant Professor, University of Denver

Meredith Ludwig, Director, Postsecondary Statistics Support, ESSI

Vivian Makosky, Consultant

Larry Mayes, Coordinator, Institutional Research, University of Kentucky Community College System

Margaret Miller, President, AAHE

John Muffo, Director, Academic Assessment Program, Virginia Polytechnic Institute and State University

Michael Pavel, Assistant Professor, Washington State University

Gregory Smith, Vice President, Information Resources and Planning, Community College of Denver

Consultant to the Working Group:

T. Dary Erwin, Center for Assessment and Research Studies, James Madison University

NPEC Staff:

Brenda Albright, Consultant to NPEC

Melodie Christal, Director, SHEEO/NCES Communication Network, State Higher Education Executive Officers

Roslyn Korb, Program Director, Postsecondary and Library Cooperative Programs, NCES

Nancy Borkow, NPEC Project Director, NCES

Robert Wallhaus, Consultant to NPEC

TABLE OF CONTENTS

Table of Contents—Continued

List of Tables

Table of Contents—Continued

Table

PREFACE

The National Postsecondary Education Cooperative (NPEC) was authorized by Congress in 1994. It charged the National Center for Education Statistics to establish a national postsecondary cooperative to promote comparable and uniform information and data at the federal, state, and institutional levels. In accordance with this charge, the projects supported by the Cooperative do not necessarily represent a federal interest, but may represent a state or institutional interest. Such is the case with this Sourcebook. While there is no federal mandate to assess the cognitive outcomes of postsecondary education, some states and many institutions have identified cognitive assessment as a way of examining the outcomes of their educational programs. This project was undertaken to facilitate these efforts.

In a climate of accelerating costs and greater requirements for high-quality services, policymakers are attempting to understand the value of higher education and are demanding greater accountability from institutions. Concurrently, accreditation agencies are requiring assessment of student outcomes as an integral part of the accreditation process. Increasingly, colleges and universities are being asked for more direct measures of student outcomes. How much did students learn? Did they learn the "right things"? Did they complete college prepared for employment? And postsecondary education is increasingly asking itself: What information really answers these questions? How do we measure what was learned? Can institutions that have different missions or that deliver instruction using different learning modes respond in a comparable way?

The National Postsecondary Education Cooperative (NPEC), in its first council meeting (held in the fall of 1995), identified the assessment of student outcomes as a high priority. The NPEC Steering Committee appointed two working groups, Student Outcomes from a Policy Perspective and Student Outcomes from a Data Perspective, to explore the nature of data on student outcomes and their usefulness in policymaking. The exploratory framework developed by the policy working group is presented in the paper *Student Outcomes Information for Policy-Making* (Terenzini 1997) (see http://nces.ed.gov/pubs97/97991.pdf). Recommendations for changes to current data collection, analysis, and reporting on student outcomes are included in the paper *Enhancing the Quality and Use of Student Outcomes Data* (Gray and Grace 1997) (see http://nces.ed.gov/pubs97/97992.pdf). Based on the work undertaken for these reports, both working groups endorsed a pilot study of the Terenzini framework and future research on outcomes data and methodological problems.

In 1997, a new working group was formed to review the framework proposed by Terenzini vis-a-vis existing measures for selected student outcomes. The working group divided into two subgroups. One group focused on cognitive outcomes, and the other concentrated on preparation for employment outcomes. The cognitive outcomes group produced two products authored by T. Dary Erwin, a consultant to the working group: *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing*; and *The NPEC Sourcebook on Assessment, Volume 2: Selected Institutions Utilizing Assessment Results*. Both publications can be viewed on the NPEC Web site at http://nces.ed.gov/npec/ under "Products."

The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing is a compendium of information about tests used to assess the three skills. Volume 1 is a tool for people who are seeking comparative data about the policy-relevance of specific student outcomes measured in these areas. The interactive version of Volume 1 (see http://nces.ed.gov/npec/evaltests/) allows users to specify their area(s) of interest and create a customized search of assessment measures within the three domain areas.

Volume 1 should be regarded as a work in progress and has certain limitations. First, it focuses on three kinds of student outcomes: critical thinking, problem solving, and writing. The Student Outcomes Working Group recognizes that there are many more outcome variables and measures that are of interest to postsecondary education constituents. Second, Volume 1 describes tests that are designed, for the most part, to measure cognitive variables for traditional students. It does not describe more "nontraditional" methods such as portfolios and competencies. Similarly, the tests themselves are not assessed with nontraditional settings in mind. Finally, the evaluations of the tests found in this volume are based mainly on the way the developers of the tests represent them in their materials and, in some cases, on material available through third-party test reviews. Each prospective user of any of the tests must evaluate the test's appropriateness for the user's own particular circumstances. Different needs, motivations, and focuses affect the utilization of the various assessments.

The tests described in Volume 1 are those that the consultant to the group was able to identify through careful searching and consideration. Some tests may have been inadvertently missed. Also, the comments in the book are not to be taken as a recommendation or condemnation of any test, but rather as a description. The descriptive process used is unique to NPEC and was developed for the purpose of the Student Outcomes Working Group project. We intend to update this volume on an as needed basis. Updates will be available at the NPEC web site: http://nces.ed.gov/npec/evaltests/.

The NPEC Sourcebook on Assessment, Volume 1 is a companion volume to *The NPEC Sourcebook on Assessment, Volume 2*. Volume 2 provides eight case studies of institutions that have addressed policy-related issues through the use of the assessment methods presented in Volume 1.

Your comments on Volume 1 are always welcome. We are particularly interested in your suggestions concerning student outcomes variables and measures, potentially useful products, and other projects that might be appropriately linked with future NPEC student outcomes efforts. Please e-mail your suggestions to Nancy Borkow (Nancy_Borkow@ed.gov), the NPEC Project Director at the National Center for Education Statistics.

Toni Larson, Chair
NPEC Student Outcomes Pilot Working Group:
Cognitive and Intellectual Development

1. GENERAL AND SPECIFIC ISSUES IN SELECTING ASSESSMENTS

1.1 Introduction

The educational goals for the year 2000, announced by the President of the United States and state governors in 1990, included the abilities to think critically, solve problems, and communicate. In a national response to the educational goals, a list of communication and critical thinking skills was obtained from a study of 500 faculty, employers, and policymakers who were asked to identify the skills that these groups believe college graduates should achieve (Jones et al. 1995). To address these national concerns, there is a need to provide evidence of attainment of these essential skills in general education. Providing the assessment results of general education gives proof of "return" to policymakers, as general education assessment enables collection of all students' performance, regardless of individual major. A variety of assessment methods have been developed to measure attainment of these skills. This report will present definitions of critical thinking, problem solving, and writing, along with a detailed review of assessment methods currently available.

In addition to specific information pertaining to critical thinking, problem solving, and writing, there are general issues pertaining to the assessment of these skills. Definitions of the particular conceptual and methodological criteria that play a key role in evaluating and selecting assessments for use in higher education are outlined in the first section. More specifically, issues to be examined in this section include the following: relevance to policy issues, utility for guiding specified policy objectives, applicability to multiple stakeholder groups, interpretability, credibility, fairness, scope of the data generated, availability or accessibility for specified/diversified purposes, measurability considerations, and cost. In the second section, the test format (multiple-choice vs. performance-based), which impacts the type of data generated and the resultant inferences that are justified, will be reviewed. The last section gives a detailed description of methodological concerns, such as reliability, validity, and method design. Because of the many factors to consider when undertaking a testing project, an assessment specialist who can create a comprehensive testing plan that accounts for conceptual and methodological issues as well as other factors relevant to the outcomes should be consulted. Due to the limitations in length of this report, only conceptual and methodological considerations will be discussed, but readers should take note that there are variables not explained in this report that greatly impact test selection (i.e., student motivation, the sample chosen, or the assessment design).

1.2 Selection of Assessment Methods: Specific and General Considerations

With the development of critical thinking, problem solving, and writing skills being increasingly recognized as integral goals of undergraduate education, a number of different measures have been designed across the country. Selection of an appropriate instrument or strategy for evaluating students' competencies in these areas often depends on whether the assessment is formative or summative in nature. In formative evaluation the goal is to provide feedback, with the aim of improving teaching, learning, and the curricula; to identify individual students' academic strengths and weaknesses; or to assist institutions with appropriate placement of individual students based on their particular learning needs. Summative evaluation, on the other hand, tends to be used to make decisions regarding allocation of funds and to aid in decisionmaking at the program level (e.g., personnel, certification, etc.). Data are derived from a summative assessment chiefly for accountability purposes and can therefore be used to meet the demands of accrediting bodies, and state and federal agencies.

Once an institution identifies the specific purpose of its assessment and defines the particular critical thinking, problem solving, or writing skills it is interested in measuring, selection of the appropriate test becomes much easier. In some cases, there is not a measure that adequately examines the

forms of student achievement that have been the focus of curriculum objectives, producing a need to develop a test locally. When the type of assessment falls into the formative category, often only outcome data derived from locally developed tests provide enough congruence with the learning objectives and curriculum aims, in addition to yielding a sufficient quantity of information, to guide decisionmaking. This is certainly not always the case, and oftentimes an institution will find a commercially produced test that samples content and/or skill areas that were emphasized in their programs in addition to providing detailed student reports. When an assessment is conducted for external purposes, typically the widely recognized, commercially produced assessments are preferred. Unfortunately, if measures are selected for this reason only, institutions may end up with a measure that is not valid for use with their unique student population or particular programs. For example, an innovative general education program that emphasizes the development of critical thinking in the context of writing instruction might focus on students learning to write essays reflecting substantial critical thinking and integration of ideas. If the students are tested with a multiple-choice writing assessment, emphasizing mechanics and editing, the degree to which the program has met its objectives would not be legitimately measured.

Conceptual Considerations

Regardless of the specific objectives associated with a given assessment approach, a number of conceptual considerations should enter into the decision to use a particular measure. First, if the outcome data will be used for making a decision regarding an important policy issue, how **relevant** is the outcome to the particular issue at hand? For example, if an assessment is conducted to determine those writing skills needed for college graduates to function effectively in the business world, the context of an essay test should probably include products such as writing letters and formal reports rather than completing a literary analysis of a poem.

A second critical conceptual issue relates to **utility**, or the potential of data generated from a particular measure to guide action directed toward achieving a policy objective. For instance, a policy objective might involve provision of resources based on institutions' sensitivity to the learning needs of students from demographically diverse backgrounds. It would be difficult to convince funding agencies that students' individual needs are being diagnosed and addressed with a measure that is culturally biased in favor of white middle-class students. Ewell and Jones (1993) noted that indirect measures often help individual colleges and universities improve instruction, but such measures tend to be less effective in terms of providing a clear focus of energy for mobilizing public support for national improvement. They base this judgment on the fact that data originating from many different types of institutions cannot be usefully combined into a single summary statistic without substantial distortion and loss of validity.

Sell (1989) has offered several suggestions for enhancing the utilization of assessment information. These include the following: (1) attending to institutional characteristics and readiness to change in the design and implementation of assessment strategies; (2) ensuring the data are valid, reliable, and credible; (3) providing information in a concise and timely manner; (4) involving potential audiences (users) in the process; and (5) providing extensive feedback and consultation regarding recommended changes.

Applicability of assessment measures relates to the extent to which information on a particular outcome measure meets the needs of multiple stakeholder groups. In other words, to what extent will data generated from a critical thinking, problem solving, or writing assessment yield information that can be used by multiple groups, such as faculty and administrators who wish to improve programs, or government officials and prospective employers who desire documentation of skill level achievement or attainment?

A fourth critical conceptual issue pertains to the **interpretability** of the test information. Will the outcome data be provided in a format that is comprehensible to individuals with different backgrounds? Data generated must be readily consumable, or individuals trained to interpret outcome

data need to be available to translate score data into a form that can be readily understood by decisionmakers who will use the data.

Credibility, which refers to how believable the information generated by a particular outcome is for policymakers, represents a fifth dimension of outcomes that should be incorporated into the selection process. Credibility is a multidimensional quality, with some overlap with the other dimensions. Credibility is established based on the amount of time, energy, and expertise that goes into a particular measure; the psychometric qualities associated with a test; the ease of interpretation of the materials and results; the amount of detail provided pertaining to student outcomes; and the cultural fairness of the test. Moreover, the credibility of outcome data is perhaps most closely tied to the degree to which the assessment information is conceptually related to the actual skills deemed important. Credibility, hence, is a part of validity, in that the validation process involves justifying or supporting the types of inferences drawn from data, which includes issues of fairness, the evaluation of psychometric properties of a test, and most importantly the interpretation of information (Messick 1981). Information pertaining to credibility will often be found through validation of test results (i.e., how congruent is test performance to the identified skills). Generally speaking, the results obtained with direct assessments have become more accepted as credible measures of learning to think critically, solve problems, and write effectively than nonperformance-based assessments, such as reports of student satisfaction or descriptions of student academic activities.

Although **cultural fairness** is an important element in the overall credibility of a measure, it also constitutes a primary conceptual consideration. The information yielded by a particular assessment approach should not be biased or misleading in favor of particular groups. Bias can be subtle, requiring extensive analysis of item content and analysis of performance by students with comparable abilities, who differ only in terms of group association, to ensure fairness. A measurement analysis, Differential Item Functioning (DIF), allows for the control of ability level so that bias can be detected. In this way, cultural fairness is a measurement issue.

Methodological Considerations

In addition to the preceding conceptual considerations, several methodological criteria should be examined when critical thinking, problem solving, and writing assessments are selected. First, the **scope** of the data needed should be considered. If "census-type" data drawn from all students in attendance at all institutions in a particular locale are needed, then researchers should opt for measures that can be efficiently administered and scored in addition to measures that assess skills and content which are universally covered across curricula. However, if the scope of data needed is more restricted (of the "knowledge-base" type), with examinees selected via sampling strategies requiring fewer participants (perhaps drawn from particular institutions or regions), then measures designed to assess more highly specified curriculum-based skills can be used. Moss (1994) noted that there tends to be an inverse relationship between the number of students that can be tested and the complexity, depth, and breadth of outcome information that can be provided due to budgetary considerations. For the purposes of accountability, it is not necessary to assess every student to derive valid estimates of system performance, and a much wider range of outcome data can be generated when careful sampling is conducted.

Availability of appropriate outcome measures represents a second methodological consideration. This refers to issues revolving around the availability of existing measures, the feasibility of developing new measures, and the logistics of using specified measures (both of the commercially available and locally developed variety). For instance, do the facilities and personnel exist for analysis and storage of data? Can the data be readily collected and the results disseminated without too much difficulty? Are the competencies and abilities of the individuals involved consistent with the tasks involved? Is the selected measurement strategy feasible with existing funds? How does the cost of one outcome measure compare to the cost of another?

Measurability refers to how the outcome is operationally defined and measured, including the methodological soundness of the chosen measures. A number of different approaches to assessing the constructs of critical thinking, problem solving, and writing ability are available in the literature; however, individuals involved in any particular assessment must arrive at a definition that is specific enough to be translated into definitive assessment objectives. In addition to construct definitions, reliability and validity of an assessment instrument must be carefully scrutinized to match the appropriate assessment test with the test givers' objectives. There is a critical validity issue with particular relevance to direct measures of ability. Although direct assessments may possess high content validity, it is important that they are not considered "exempt from the need to marshal evidence in support of their use" (Powers, Fowles, and Willard 1994). For example, it is essential to establish a clear link between performance on a particular direct writing assessment and demonstrated writing on both concurrent (such as grades in a writing class) and future performances (demonstrating competence in graduate courses requiring writing or on-the-job writing tasks). Although the inferential leaps between authentic measures of abilities and actual tasks encountered in coursework or elsewhere are substantially reduced when direct measures are used, the need to provide validation of a test for a particular use remains the same (Powers, Fowles, and Willard 1994).

Multiple-Choice Measures

Assessment of critical thinking, problem solving, and writing in higher education has traditionally taken two forms: direct (constructed response) and indirect (multiple-choice) measurement. Indirect assessments involve an estimate of the examinee's probable skill level based on observations of knowledge about skill level (i.e., to assess writing, one would observe vocabulary, grammar, sentence structure, etc.). Indirect assessments are exemplified by many of the standardized, commercially available tests. Perhaps the most frequently cited advantage of multiple-choice tests is the high reliability estimates often associated with them. Indirect assessments also tend to possess higher predictive validity with a variety of outcome measures, such as college GPA or scores on other standardized tests. An additional advantage is ease of scoring. Scoring is less time consuming and costly because computers can be readily used. Enhanced political leverage associated with outcomes derived from indirect assessments due to the extensive development process and general familiarity associated with commercially designed tests represent two other benefits.

One of the commonly cited disadvantages of indirect assessment involves the time and resources needed to develop and revise the tests. Further, many have argued that indirect assessments dramatically under-represent the construct. For instance, when writing or critical thinking is defined as a process, multiple-choice tests do not adequately represent the definition. Inferences about the processes students use to arrive at the correct choice on a multiple-choice test are often made, but scrutinized for their accuracy. Ewell and Jones (1993) point out that conclusions drawn from indirect indicators are highly inferential even when the data are presented from multiple measures. White (1993) contends that many indirect assessments fail to assess higher-order thinking skills. Finally, allegations of bias based on gender, race, and language have been leveled against specific multiple-choice tests, and there is some evidence suggesting that the selected response format may generally favor certain groups more than the constructed format or essay-type test (Koenig and Mitchell 1988; White and Thomas 1981). However, general conclusions such as this should be viewed very cautiously, as the majority of available critical thinking, problem solving, and writing assessments have not been systematically examined for evidence of bias.

Essay Tests

Direct assessments involve evaluation of a sample of an examinee's skill obtained under controlled or real life conditions by one or more judges, and are most frequently associated with the timed essay format. The specific types of essay assessments may be classified in terms of the types of tasks

employed and/or the scoring method implemented. Breland (1983) identified nine different types of tasks employed in direct measures of writing. Each of these will be described briefly. An examinee may be directed to write a **letter** to a friend, a potential employer, a politician, or an editor. Another type of essay prompt, termed a **narrative**, requires the student to write a personal account of an experience or convey the details of a particular story or historical event. Narratives can be real or imaginary. The **descriptive** format requires that the writer describe an object, place, or person, with the goal of creating a vivid image or impression in the reader's mind. An **argumentative** prompt (also referred to as a persuasive task) instructs the examinee to adopt a position on an issue and present a persuasive argument in favor of the chosen side using relevant information obtained through personal experience and/or reading. For an **expressive** task, the examinee simply conveys his or her own personal opinion on a particular issue or event. With a **role-playing** prompt, the student is asked to assume a role in some situation and write a response to a given situation. A **precis or abstract** requires a summary or synthesis of a large body of information. The purpose of a **diary entry** is personal usage necessitating an informal tone, and finally, a **literary analysis** requires interpretation of a passage or other literary work.

Several benefits of essay tests in general have been touted, including the following: (1) enhanced construct validity; (2) reduced racial bias; (3) faculty involvement in development and scoring, leading to more awareness of the central role of critical thinking, problem solving, and writing in the college curriculum; and (4) the flexibility to assess a wider range of skills than is feasible with the multiple-choice format. Although essay tests have earned increasing support from faculty, administrators, and test development experts in recent years, many professionals who are committed to the process model of writing object strongly to the timed essay as it precludes revision. Many adherents of a process definition of writing believe that revision represents the most critical part of the process, and when revision skills are not measured, an essential component of the construct is neglected. A disadvantage of critical thinking essay tests is that the ability to write is often entangled with the measurement of critical thinking ability. Essay tests have also been criticized because they are routinely conducted in artificial settings, provide only a small sample of the universe of writing, and have compromised reliability.

Although this report will focus on specific assessment instruments and measurement issues surrounding each test, there will be no discussion of implementation issues at the state or university level. This information, although beyond the scope of this report, is still pivotal in selecting an assessment test. For instance, sample size, time of testing, the audience, and assessment design (pre/post-testing) are just a few examples of variables that greatly affect assessment outcomes. Such factors and many others should be reviewed with an assessment specialist before a measure is chosen. In addition to implementation issues, there are methodological and conceptual considerations that should steer the test selection process. Many of the considerations overlap, as in the cases of credibility and validity or cultural fairness and measurability. Therefore, the methodological and conceptual considerations are not independent issues, but parts of a whole that create a comprehensive and rigorous test selection process.

1.3 Test Properties

One of the methodological considerations in test selection involves the psychometric properties of a test. The test tables or templates provide a condensed review of studies that address the psychometric qualities of critical thinking, problem solving, and writing tests. The first column indicates the test name, author(s), publisher, date of publication, testing time, and cost. Any special comments or notes about the tests are at the bottom of this column. The second column gives the name(s) of the reported scores. Often tests have a total score and then several subtest scores. Whether or not subtest scores can be reported independently varies from test to test. The Definition column includes critical thinking, problem solving, or writing as defined by the author. It is important to note that the test items should match the definition given by the author(s). The next column, Reliability, involves the consistency of scores across a test. The statistics reported under this column will be addressed further in the report. Method Design combines both reliability and validity issues concerning the internal structure of a test.

Next is the Validity column, which gives information about studies that have implemented the tests. Readers should especially take note of studies conducted independently of test authors. The last column, Correlation with Other Measures, is a form of validity, and is given a separate section, due to the amount of information found for most tests. A review of correlations can be found under the heading, Validity. The following section is meant as a brief review of statistical procedures. For a more extensive explanation of reliability, validity, correlations, and method design issues, see Crocker and Algina (1986), Felt and Brennan (1989), or Cole and Moss (1989).

Reliability

Reliability is an estimate of test takers' performance consistency internally, across time, test forms, and raters (when applicable). Tests are not reliable in and of themselves, but the scores generated from the tests can be reliable. This means that across varying populations, reliability estimates may change. Important factors to consider when interpreting reliability estimates are the following: longer tests tend to be more reliable, reliability fluctuates with test takers, speeded tests can change the reliability estimate, homogeneity of test taker ability lowers the reliability, different levels of skill may be measured with different levels of accuracy, and longer time intervals for test-retest reliability lower the reliability estimate. With these factors in mind, different types of reliability estimates will be reviewed. Generally, reliability estimates above .70 indicate an acceptable level, although values in the .80 and above are more commonly accepted reliabilities.

Internal consistency can be measured using several methods. Coefficient Alpha, Split-half, KR-20, and inter-rater reliability are the four methods reported in the context of the test reviews. Internal consistency is another term for a test of item homogeneity. Item homogeneity indicates that content and item quality are consistent throughout the test. This reliability coefficient ranges from 0 to 1.0, representing the degree of relationship among items on a test. A test with homogenous or more related items will produce higher reliability coefficients (values closer to 1.0).

The most often used estimate of internal consistency is **Alpha**, indicated as "internal consistency" on the templates. For instance, the California Critical Thinking Dispositions Inventory (Facione and Facione 1992) has internal consistency coefficients ranging from .75 to .96, indicating that the items are highly related. The KR-20, another reliability estimate reported in the templates, can be interpreted in the same manner as Alpha. The Critical Thinking Test of the CAAP (American College Testing Program 1989) has a KR-20 value of .81–.82, indicating that it is a reliable measure with homogeneous items.

Split-half reliability estimates represent another internal consistency method. The most often used method of split-half reliability involves using the even numbers to create one half-test and the odd numbers to compose the second half-test. In addition, test content can determine the division of items on a test. The same students are given each half-test and the scores are correlated, giving a coefficient of equivalence. As an overall reliability measure, the split-half reliability will give an underestimate of total test reliability, due to fewer items. The utility of the estimate is that item homogeneity is tested. In the case of the Watson-Glaser Critical Thinking Appraisal (Watson and Glaser 1980), the split-half reliability estimates ranged from .69 to .85, indicating item homogeneity and a reliable measure.

Inter-rater reliabilities are estimated to find the consistency of scores across raters. The Reflective Judgement Interview (King and Kitchener 1994) was found to have an inter-rater reliability of .97 (Mines et al. 1990), indicating that across raters there was high consistency in scores. Although this measure gives some indication of consistency, it only considers consistency across raters. What if items affect the performance of individuals? Some items may be harder or easier for students and raters; therefore, inter-rater reliability is a limited reliability estimate for performance assessment. The Generalizability coefficient discussed later is a more extensive estimate of reliability. Related to inter-rater reliability is inter-rater agreement. Inter-rater agreement is not a reliability estimate, but rather an

item-by-item percentage of agreement across raters. The inter-rater agreement percentages reflect the degree of similarity in ratings for each item.

Another estimate of reliability is **test-retest reliability**, which assesses test consistency over time. The same form of a test is given at different occasions that can vary from hours to days to weeks, or even years. The time interval may depend on factors such as content of the test or developmental and maturational considerations. The test-retest reliability estimate is often called the coefficient of stability, since it addresses test score stability over time. The Problem Solving Inventory (Heppner 1982) has been tested across various time intervals, with more reliable estimates found for shorter time intervals: .83–.89 across 2 weeks, .77–.81 across 3 weeks and .44–.65 across 2 years (Heppner and Peterson 1982a; Ritchey, Carscaddon, and Morgan 1984).

To test the consistency of two forms purported to be identical, **alternate forms reliability** is calculated. This method involves two versions of a test given to the same subjects on the same testing occasion. A correlation between the scores on each form indicates the alternate forms reliability, also called the coefficient of equivalence. The higher the correlation between the two sets of scores, the more equivalent the forms are considered. If two forms exist, alternate forms reliability is recommended. The Tasks in Critical Thinking tests have alternate forms with reliability across the varying skills (not the tasks) ranging from .17 to .90 (Educational Testing Service and the College Board 1989). These values indicate that some of the skills assessed by the tasks are reliable, while others fall in an unacceptable range. The Watson-Glaser Critical Thinking Appraisal reports an alternate forms reliability of .75, moderately supporting the use of the separate forms as identical. Subscales that are internally correlated with one another is another form of alternative reliability, which is reported under the Method Design section.

The **Generalizability coefficient** estimates the consistency of scores while accounting for more than one variable at a time (error). Instead of conducting a separate internal consistency study and an inter-rater reliability study, the two studies can be done at one time using a Generalizability study. A Generalizability study creates a G coefficient that can be interpreted as a reliability estimate. The Tasks in Critical Thinking (Educational Testing Service and the College Board 1989) have G coefficients ranging from .57 to .65, indicating that across raters and items, students' scores are only moderately reliable.

Method Design

There are several methods used to support the structure of a test. The structure of a test includes the item representations on subtests and the test, along with the relationship of the subtests to one another. More developed tests will use procedures such as factor analysis and differential item analysis. Most tests will report item-total correlations or discrimination indices as support for the structure of the test.

Factor analysis is a method that identifies the underlying constructs or factors among items. Each subtest is created from a set of items, which theoretically should correlate with one another, since they are purported to measure the same concept. By applying factor analysis, the relationships among the items can be understood. Factor loadings indicate the amount of relationship or contributing power an item has within a subtest or test. Therefore, higher factor loadings indicate items that are more strongly related. Optimally, factor analysis results should parallel the hypothesized structure of the test. For instance, support for the three subtest structure of the Problem Solving Inventory (Heppner 1982) was found using factor analysis (Heppner 1988; Chynoweth 1987; Heppner and Peterson 1982a).

Another method used to validate test design is item total correlations. These correlations reveal how well each item correlates with the total score. The larger the item total correlation, the more the item contributes to the subscale or test. Values below .10 indicate an item does not measure the same construct as other items on the test, while negative items indicate an inverse relationship among items and the total. An analysis of the item total correlations for the California Critical Thinking Skills Test

(CCTST) (Facione 1990a) revealed that many of the items did not correlate well with the total test or respective subtests. For instance, 10 out of the 34 items on the total test had values below .10 (Jacobs 1995), indicating little relationship between these items and the total test. Erwin (1997) further supported Jacobs' results, finding that 7 out of 34 of the items on the CCTST had item total correlations below .10.

Validation of test design can also be supported with item discrimination indexes. Item discrimination indexes are a measure of the difference in item responses between high and low scorers. They range from 0 to 1.00, with values closer to 1.00 indicating higher discrimination. Greater item discrimination indexes suggest a test that is sensitive to differences in ability. The Cornell Critical Thinking Test (Ennis, Millman, and Tomko 1985) had indexes ranging from .20 to .24, suggesting moderate discrimination among high and low scorers.

Fairness, related to bias in testing, is usually focused on differences among test takers based on variables such as inclusion in a group. For instance, are there unintended differences between males and females on critical thinking tests? This is the typical argument in defining whether a test is "fair." What is not considered in this argument is whether a difference in ability level actually exists across gender. Males or females may have a naturally higher competency level in critical thinking. In this case, it is important to know if items are fair indicators of ability across groups (gender, ethnicity, etc), not just whether groups score differently on items.

Differential item analysis (DIF) allows for the control of ability level, so that differences found in scores are attributed to a variable other than ability. When items exhibit DIF they are considered "unfair," meaning that individuals from one group are more likely to answer the item correctly than individuals from another group, even when ability levels are the same. Traditionally DIF is performed across groups such as gender and ethnicity. For instance, the Cornell Critical Thinking Test has four items that exhibit gender DIF. Three of the items were more likely to be answered correctly by males compared to females with similar critical thinking ability levels. Content analysis of the items revealed some hypotheses for the differing scores. Two of the items that males had a better chance of answering correctly pertained to stockcars, a subject perhaps more interesting to males than females. Whether the content contributed to the differences found, it is clear that males and females of similar ability levels do not have a fair chance at getting these items correct. By applying gender DIF analysis, ability levels were controlled and a true bias in the test could be found.

Validity

Validity involves "building a case" that a test is related to the construct it is intended to measure. There are three types of validity: content, criterion, and construct. The most important type of validation is construct validity, because it encompasses both content and criterion validity. Therefore, inferences made from test scores that have only content or criterion validation are not considered valid until construct validity is addressed. When reviewing validity studies in the templates, the external validation studies or studies conducted by those other than the test author should be given more consideration. External validation studies reveal the amount of use and exposure of the test and can be considered unbiased toward the outcomes of the study.

Content validity deals with the conceptualization of the constructs. Is the content of the test representative of the construct (critical thinking or writing) it purports to measure? Does the test represent the test developer's definition? Is there a discrepancy between the test developer's definition and the test user's definition? Do experts judge the test to measure the constructs adequately and appropriately? Tests that are conceptualized from theory have stronger content validity over tests that have no theoretical backing. The CCTST (Facione 1990a) is a good example of a test with strong content validation. The test was conceptualized from a definition of critical thinking developed by the American Philosophical Association and the California State University system.

A second type of validation involves whether a test can be used to infer standing on another test or variable. This is called **criterion validity**. Criterion validity can be measured as predictive (i.e., how well one score predicts scores on another test), or as concurrent (i.e., how well one's current standing on a given measure can be predicted from another measure). Typically variables such as class standing, GPA, grades, SAT scores, and other relevant tests are used in criterion validation studies. If, for instance, SAT scores did accurately predict critical thinking test scores, then it could be inferred that the critical thinking test and the SAT test are measuring similar abilities. A study by Mines et al. (1990) revealed that one subscale of the Cornell Critical Thinking Test (CCTT) (Ennis, Millman, and Tomko 1985) and three subscales of the Watson Glaser Critical Thinking Appraisal (WGCTA) (Watson and Glaser 1980) could accurately predict 50 percent of students' Reflective Judgement Interview scores (King and Kitchener 1994). The high level of prediction highlights that tests often measure the same construct, even if authors profess their tests to be based on different constructs. In general, more studies are needed relating critical thinking, problem solving, and writing to other criteria such as job performance or citizenship.

Construct validity involves content and criterion validity. Construct validity specifically addresses the questions of whether the test measures the trait, attribute, or mental process it is purported to measure, and whether the scores should be used to describe test takers. Two methods of construct validation are correlation studies (convergent and divergent validity) and outcome analysis. To understand correlation studies, a brief review of correlations will be given. The **correlation coefficient** represents the amount of relationship between two variables and ranges from -1.00 to 0 to 1.00, with values closest to 1.00 and -1.00 indicating a strong relationship. A correlation coefficient from .10 to .20 represents a small relationship, and values from .30 to .50 indicate moderate relationships between tests. A negative correlation, or inverse relationship, indicates that as one variable increases the other decreases. Some correlations are corrected for attenuation, which means corrected for unreliability. Measurement of variables always involves "error." By removing the error, a perfect correlation between two variables can be calculated. For instance, the correlation between the WGCTA and CCTT is .71, and when corrected for attenuation the correlation is .94, indicating that the lack of reliability in the two tests is accounting for the lower correlation.

Convergent and divergent validity involves finding the relationship of the critical thinking, problem solving, or writing test to other tests that measure similar and opposite constructs. The column Correlation with Other Measures on the templates represents convergent and divergent validity. To interpret correlations with other measures, one needs to understand the content behind the measures, and how they should logically be related. Two similarly conceptualized writing tests correlated with one another should produce moderate correlations around .40 to .60, since some overlap of content is expected. High correlation values could be considered indicators of a strong relationship, suggesting that individual tests may be measuring the same construct. Many critical thinking tests come under scrutiny as being measures of verbal ability. This criticism can be tested using correlation studies comparing critical thinking scores with SAT verbal scores or other verbal tests. The CCTT (Ennis, Millman, and Tomko 1985) scores were correlated with SAT verbal scores ($r = .36, .44$), revealing that test scores were related to a moderate degree (Ennis, Millman, and Tomko 1985; Frisby 1992). Higher correlation values between critical thinking tests and verbal ability measures suggest that critical thinking test scores might actually be tapping into verbal ability.

The last method of construct validity is to conduct experimental studies analyzing outcomes. If students take a critical thinking, problem solving, or writing course, the hypothesized outcome is that students would exhibit a gain in the appropriate skill from pre- to post-testing and would score higher compared to students who did not take the proposed course. These studies add substantial support to tests as measures of critical thinking, problem solving, and writing. Although significant differences across pre- and post-testing give an indication of change, the degree of change is not known. To calculate the degree of change, an effect size is used. Effect sizes are the standardized difference between the treatment groups (those who received skill training) and the control groups (those who did not receive skill training). By standardizing the group differences, comparisons can be made from one study to the next. An effect size of .50 indicates half a standard deviation difference between groups. For instance, the CAAP was reported to have an effect size of .41 for full-time students versus part-time students,

indicating a .41 standard deviation increase for students enrolled full-time. Effect sizes should be interpreted in light of the degree of change that is expected or desired.

The reliability and validity of a test cover an immense amount of information regarding the consistency and usefulness of scores. As a first step in the review process, it should be noted that reliability must be established before validity issues are addressed. If scores are not consistent, then the inferences made will also be inconsistent. Once reliability is determined, the content of a test, most specifically the definition and domains covered by the test, should be examined for fit with the purpose of testing. Any outcome information regarding the content and inferences made from the test should help to guide the content review. Correlations with other measures can also help to clarify the tests' relationships with other well-known variables. Perhaps the most important information comes from studies that investigate gains in ability not only across time, but across treatment. For instance, individuals receiving intense instruction in writing should out-perform those who do not receive training. If a test detects the differences in writing ability between these two groups, then the test is supported as a measure of writing. Overall, the review process is tedious and involved. Each test must be considered based on the merits of its structure, content, score consistency, and inferential potential, in addition to how these elements fit with the purpose of testing and the outcomes desired.

2. CRITICAL THINKING AND PROBLEM SOLVING

2.1 Introduction

Critical thinking and problem solving have been identified as essential skills for college students. Many colleges across the nation have begun to teach courses based on these pertinent skills. For instance, Chaffee (1991) authored a book *Thinking Critically*, which can be used as a curriculum guide. Although the importance of students demonstrating these skills has been determined, defining these terms and finding appropriate assessment methods are complex and involved tasks. In a national report on higher education, Jones et al. (1997, pp. 20–21) and Jones et al. (1995, p. 15) give comprehensive definitions of problem solving and critical thinking, making distinctions between the two terms. With a consensus among 500 policymakers, employers, and educators, the following definitions were created. **Problem solving** is defined as a step-by-step process of defining the problem, searching for information, and testing hypotheses with the understanding that there are a limited number of solutions. The goal of problem solving is to find and implement a solution, usually to a well-defined and well-structured problem. **Critical thinking** is a broader term describing reasoning in an open-ended manner, with an unlimited number of solutions. The critical thinking process involves constructing the situation and supporting the reasoning behind a solution. Traditionally, critical thinking and problem solving have been associated with different fields: critical thinking is rooted in the behavioral sciences, whereas problem solving is associated with the math and science disciplines. Although a distinction is made between the two concepts, in real life situations the terms critical thinking and problem solving are often used interchangeably. In addition, assessment tests frequently overlap or measure both skills. In keeping with the Jones et al. (1995, 1997) definitions, this report will analyze critical thinking and problem solving separately, yet attempt to integrate the two skills when appropriate.

2.2 Definition of Critical Thinking

A comprehensive definition of critical thinking, the product of studies by Jones et al. (1995, 1997) can be found in tables 2–8. Critical thinking is defined in seven major categories: Interpretation, Analysis, Evaluation, Inference, Presenting Arguments, Reflection, and Dispositions. Within each of these categories are skills and subskills that concretely define critical thinking. As a content review of critical thinking assessment methods, comparisons were made for each test across the comprehensive definition of critical thinking. If test content addresses a skill, then the test acronym appears next to that skill. The following table indicates the tests and acronyms used. Tests were chosen for review based on several factors: (1) the ability to measure college students' critical thinking skills and/or critical thinking dispositions, and (2) broad scale availability to colleges and universities.

Table 1—Test acronyms

Acronym	Test Name
A. PROFILE	Academic Profile
CAAP	Collegiate Assessment of Academic Proficiency
CCTDI	California Critical Thinking Dispositions Inventory
CTAB	CAAP Critical Thinking Assessment Battery
CCTST	California Critical Thinking Skills Test
CCTT	Cornell Critical Thinking Test

Acronym	Test Name
COMP	College Outcomes Measures Program – Objective Test
ETS TASKS	ETS Tasks in Critical Thinking
MID	Measure of Intellectual Development
PSI	Problem Solving Inventory
RJI	Reflective Judgement Inventory
WGCTA	Watson Glaser Critical Thinking Appraisal

Several methods were used to match the test content with the definition of critical thinking. For the Academic Profile, CAAP, CCTDI, CTAB, CCTST, COMP, and ETS Tasks, the definitions created by the author(s) were used as a guide in determining content on the test. For the CCTT, PSI, and WGCTA, the tests were reviewed to determine the content, due to the lack of specific skills or definitions given by the author(s) in the test manual. The RJI and MID, which are based on stages, were analyzed in light of the information that would be needed to separate individuals at different stages. It should also be noted that the PSI measures perceptions of critical thinking skills; therefore, if the PSI is indicated to measure a skill in the tables, it should be interpreted as measuring perception of that skill. Caution should be used in interpreting tables 2–8, due to the subjective process used to compare tests and definitions.

Table 2—Interpretation skills measured by critical thinking tests

Interpretation	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
<u>Categorization</u>												
1. Formulate categories, distinctions, or frameworks to organize information in such a manner to aid comprehension.					*		*	*				
2. Translate information from one medium to another to aid comprehension without altering the intended meaning.					*			*				
3. Make comparisons; note similarities and differences between or among informational items.					*			*				
4. Classify and group data, findings, and opinions on the basis of attributes or a given criterion.					*			*				
<u>Detecting Indirect Persuasion</u>												
1. Detect the use of strong emotional language or imagery that is intended to trigger a response in an audience.					*	*		*				
2. Detect the use of leading questions that are biased towards eliciting a preferred response.						*		*				
3. Detect "if, then" statements based on the false assumption that if the antecedent is true, so must be the consequence.					*	*						*

Table 2—Interpretation skills measured by critical thinking tests—Continued

Interpretation	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
4. Recognize the use of misleading language.						*		*				
5. Detect instances where irrelevant topics or considerations are brought into an argument that diverts attention from the original issues.					*	*		*				*
6. Recognize the use of slanted definitions or comparisons that express a bias for or against a position.					*	*	*	*				
<u>Clarifying Meaning</u> 1. Recognize confusing, vague, or ambiguous language that requires clarification to increase comprehension.		*		*		*		*				*
2. Ask relevant and penetrating questions to clarify facts, concepts, and relationships.												
3. Identify and seek additional resources, such as resources in print, which can help clarify communication.							*	*				
4. Develop analogies and other forms of comparisons to clarify meaning.								*				
5. Recognize contradictions and inconsistencies in written and verbal language, data, images, or symbols.					*	*						*

Table 2—Interpretation skills measured by critical thinking tests—Continued

Interpretation	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
6. Provide an example that helps to explain something or removes a troublesome ambiguity.							*		*			

Table 3—Analysis skills measured by critical thinking tests

Analysis	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
<u>Examining Ideas and Purpose</u> 1. Recognize the relationship between the purpose of a communication and the problems or issues that must be resolved in achieving that purpose.												
2. Assess the constraints of the practical applications of an idea.												
3. Identify the ideas presented and assess the interests, attitudes, or views contained in those ideas.								*				
4. Identify the stated, implied, or undeclared purpose(s) of a communication.								*				

Table 3—Analysis skills measured by critical thinking tests—Continued

Detecting and Analyzing Arguments												
1. Examine a communication and determine whether or not it expresses a reason(s) in support or in opposition to some conclusion, opinion, or point of view.	*	*		*	*	*		*				*
2. Identify the main conclusions of an argument.	*	*		*	*	*		*				*
3. Determine if the conclusion is supported with reasons and identify those that are stated or implied.	*	*		*	*	*		*				*
4. Identify the background information provided to explain reasons that support a conclusion.	*	*		*	*	*		*				*
5. Identify the unstated assumptions of an argument.	*	*		*	*	*						*

Table 4—Evaluation skills measured by critical thinking tests

Evaluation	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
1. Assess the importance of an argument and determine if it merits attention.					*			*				*
2. Evaluate an argument in terms of its reasonability and practicality.		*		*	*	*		*				*
3. Evaluate the credibility, accuracy, and reliability of sources of information.		*		*	*	*		*				*
4. Determine if an argument rests on false, biased, or doubtful assumptions.		*		*	*	*	*	*				*
5. Assess statistical information used as evidence to support an argument.		*		*	*	*						*
6. Assess how well an argument anticipates possible objectives and offers, when appropriate, alternative positions.					*			*				
7. Determine how new data might lead to the further confirmation or questioning of a conclusion.					*	*						
8. Determine and evaluate the strength of an analogy used to warrant a claim or consolation.								*				

Table 4—Evaluation skills measured by critical thinking tests—Continued

9. Determine if conclusions based on empirical observations were derived from a sufficiently large and representative sample.						*						
10. Determine if an argument makes sense.					*	*	*	*				*
11. Assess bias, narrowness, and contradictions when they occur in the person' point of view.		*		*	*	*						*
12. Assess degree to which the language, terminology and concepts employed in an argument are used in a clear and consistent manner.		*		*	*	*						*
13. Determine what stated or unstated values or standards of conduct are upheld by an argument and assess their appropriateness to the given context.					*	*						*
14. Judge the consistency of supporting reasons, including their relevancy to a conclusion and their adequacy to support a conclusion.	*	*			*	*	*	*				*
15. Determine and judge the strength of an argument in which an event(s) is claimed to be the results of another event(s) (causal reasoning).	*	*			*	*						*

Table 5—Inference skills measured by critical thinking tests

Inference Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
<u>Collecting and Questioning Evidence</u> 1. Determine what is the most significant aspect of a problem or issue that needs to be addressed, prior to collecting evidence.					*		*	*				*
2. Formulate a plan for locating information to aid in determining if a given opinion is more or less reasonable than a competing opinion.							*	*				
3. Combine disparate pieces of information whose connection is not obvious, but when combined offer insight into a problem or issues.												
4. Judge what background information would be useful to have when attempting to develop a persuasive argument in support of one's opinion.								*				
5. Determine if one has sufficient evidence to form a conclusion.					*	*						*
<u>Developing Alternative Hypotheses</u> 1. Seek the opinion of others in identifying and considering alternatives.												

Table 5—Inference skills measured by critical thinking tests—Continued

Inference Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
2. List alternatives and consider their pros and cons, including their plausibility and practicality, when making decisions or solving problems.								*		*	*	
3. Project alternative hypotheses regarding an event, and develop a variety of different plans to achieve some goal.							*	*		*		
4. Recognize the need to isolate and control variables in order to make strong causal claims when testing hypotheses.						*						
5. Seek evidence to confirm or disconfirm alternatives.					*	*	*			*		
6. Assess the risks and benefits of each alternative in deciding between them.								*		*		
7. After evaluating the alternatives generated, develop, when appropriate, a new alternative that combines the best qualities and avoids the disadvantages of previous alternatives.												

Table 5—Inference skills measured by critical thinking tests—Continued

Inference Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
Drawing Conclusions												
1. Assess how the tendency to act in ways to generate results that are consistent with one's expectations could be responsible for experimental results and everyday observations.					*		*					*
2. Reason well with divergent points of view, especially with those with which one disagrees, in formulating an opinion on an issue or problem.											*	
3. Develop and use criteria for making judgments that are reliable, intellectually strong, and relevant to the situation at hand.					*	*	*	*			*	*
4. Apply appropriate statistical inference techniques to confirm or disconfirm a hypothesis in experiments.					*	*		*				*
5. Use multiple strategies in solving problems including means-ends analysis, working backward, analogies, brainstorming, and trial and error.					*							
6. Seek various independent sources of evidence, rather than a single source of evidence, to provide support for a conclusion.							*		*			

Table 5—Inference skills measured by critical thinking tests—Continued

Inference Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
7. Note uniformities or regularities in a given set of facts, and construct a generalization that would apply to all these and similar instances.						*						
8. Employ graphs, diagrams, hierarchical trees, matrices, and models as solution aids.					*	*	*	*				

Table 6—Presenting arguments skills measured by critical thinking tests

Presenting Arguments Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
1. Present supporting reasons and evidence for their conclusion(s) which address the concerns of the audience.				*			*					
2. Negotiate fairly and persuasively.				*			*		*			
3. Present an argument succinctly in such a way as to convey the crucial point of issue.				*			*	*	*			
4. Cite relevant evidence and experiences to support their position.				*			*	*	*			
5. Formulate accurately and consider alternative positions and opposing points of view, noting and evaluating evidence and key assumptions on both sides.				*				*		*		

Table 6—Presenting arguments skills measured by critical thinking tests—Continued

Presenting Arguments Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
6. Illustrate their central concepts with significant examples and show how these concepts and examples apply in real situations.				*			*		*			

Table 7—Reflection skills measured by critical thinking tests

Reflection Skills	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
1. Apply the skills of their own analysis and evaluation to their arguments to confirm and/or correct their reasoning and results.			*				*					
2. Critically examine and evaluate their vested interests, beliefs, and assumptions in supporting an argument or judgment.							*					
3. Make revisions in arguments and findings when self-examination reveals inadequacies.			*				*				*	

Table 8—Dispositions measured by critical thinking tests

Dispositions	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
1. Be curious and inquire about how and why things work.			*							*		
2. Be organized, orderly, and focused in inquiry or in thinking.			*					*		*		

Table 8—Dispositions measured by critical thinking tests—Continued

Dispositions	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CT A
3. Willingly persevere and persist at a complex task.			*							*		
4. Be flexible and creative in seeking solutions.								*		*		
5. Be inclined to arrive at a reasonable decision in situations where there is more than one plausible solution.			*					*		*	*	
6. Apply insights from cultures other than their own.								*				
7. Exhibit honesty in facing up to their prejudices, biases, or tendency to consider a problem solely from their viewpoint.			*									
8. Monitor their understanding of a situation and progress toward goals.							*			*		
9. Find ways to collaborate with others to reach consensus on a problem or issues.												
10. Be intellectually careful and precise.			*					*		*		
11. Value the application of reason and the use of evidence.			*							*		
12. Be open-minded; strive to understand and consider divergent points of view.			*							*	*	

Table 8—Dispositions measured by critical thinking tests—Continued

Dispositions	A. Profile	CAAP	CCTDI	CTAB	CCTST	CCTT	COMP	ETS TASKS	MID	PSI	RJI	WG CTA
13. Be fair-minded; seek truth and be impartial, even if the findings of an inquiry may not support one's preconceived opinions.			*								*	
14. Willingly self-correct and learn from errors made no matter who calls them to our attention.												

In reviewing tables 2–8, it should be noted that no single test measures every aspect of critical thinking. In fact, even with all of the tests combined, all critical thinking skills are not assessed. Although in comparison to the Jones et al. definition, a comprehensive test is not available, many tests are still adequate measures of some critical thinking skills. Analysis of these particular tests can be found in the test templates at the end of this section.

2.3 Definition of Problem Solving

The ability to solve problems has been defined through a consensus of college and university faculty members, employers, and policymakers. The resulting definition produced by Jones et al. (1997) will be used as a base for examining the scope of problem-solving assessments reviewed within this report. Problem solving is defined as understanding the problem, being able to obtain background knowledge, generating possible solutions, identifying and evaluating constraints, choosing a solution, functioning within a problem-solving group, evaluating the process, and exhibiting problem-solving dispositions. Only three tests were identified as addressing problem-solving skills: ACT College Outcomes Measures Program (COMP) problem-solving subscale, the ETS Tasks in Critical Thinking; and the Problem Solving Inventory (PSI). The PSI, when compared to the Jones et al. definition, was not found to assess any of the skills; therefore, only the COMP and ETS tests were included in the comparison. The full definition follows in table 9. Again, the process used to determine if tests measured a skill was subjective and based on the authors' claims; therefore, the results presented in table 9 should be interpreted cautiously. The test templates at the end of this section include in-depth reviews of the problem-solving tests.

From the definition table, it is evident that there is not an adequate measure of problem-solving skills and that the most comprehensive measure is the ETS Tasks in Critical Thinking. These tasks are purported to measure critical thinking, yet also address many of the skills of problem solving. This brings to light the issue that there is considerable overlap in critical thinking and problem solving. For instance, the ability to state a problem; evaluate factors surrounding the problem; create, implement, and adjust solutions as needed; and analyze the process and fit of a solution—as well as having an active inclination towards thinking, solving problems, and being creative—are all skills necessary for both problem solving and critical thinking. Therefore, the clear distinctions between problem solving and critical thinking exhibited in the definition by Jones et al. may prove difficult to assess and tease apart in application.

Perhaps the most important element in measuring critical thinking or problem solving at the college level is the choice of a clear, comprehensive definition to steer the assessment process. If, for instance, the purpose of testing is to assess effectiveness in a general education program, then the definition should match the curriculum objectives identified and resemble the students' classroom experiences. Once a firm definition is determined and the purpose of testing is known, conceptual and methodological considerations can be evaluated. Test users should understand the limitations of particular tests to assess a broad range of skills and incorporate these limitations into the assessment plan. The test format, multiple-choice or constructed response, is another consideration affecting the types of inferences that can be made and the data generated. In essence, there are many complex issues to evaluate; therefore, it is recommended that an assessment specialist always be contacted and included in the testing process.

Table 9—Problem-solving skills as measured by the COMP and ETS Tasks in Critical Thinking

Problem-Solving Skills	**COMP**	**ETS Tasks**
Understanding the Problem		
Recognize the problem exists.	*	*
Determine which facts are known in a problem situation and which are uncertain.		*
Summarize the problem to facilitate comprehension and communication of the problem.	*	*
Identify different points of view inherent in the representation of the problem.		*
Identify the physical and organizational environment of the problem.		*
Describe the values that have a bearing on the problem.		
Identify time constraints associated with solving the problem.		
Identify personal biases inherent in any representation of the problem.	*	*
Obtaining Background Knowledge		
Determine if they have the background information to solve the problem.		*
Apply general principles and strategies that can be used in the solution of other problems.	*	*
Use visual imagery to help memorize and recall information.		
Identify what additional information is required and where it can be obtained.	*	*
Develop and organize knowledge around the fundamental principles associated with a particular discipline.		*
Develop and organize knowledge around the fundamental principles associated across functions or disciplines.		*
Use systematic logic to accomplish their goals.	*	*
Evaluate arguments and evidence so that competing alternatives can be assessed for their relative strengths.		*
Organize related information into clusters.		*

Table 9—Problem-solving skills as measured by the COMP and ETS Tasks in Critical Thinking —Continued

Problem-Solving Skills	**COMP**	**ETS Tasks**
Recognize patterns or relationships in large amounts of information.		*
Use analogies and metaphors to explain a problem.		
Identify persons or groups who may be solving similar problems.		
Obtaining Background Knowledge—Continued Identify time constraints related to problem solving.		
Identify financial constraints related to problem solving.		
Use clear, concise communication to describe a problem.	*	*
Generate Possible Solutions Think creative ideas.		*
List several methods that might be used to achieve the goal of the problem.	*	*
Be flexible and original when using experiences to generate possible solutions.		
Use brainstorming to help generate solutions.		
Divide problems into manageable components.		*
Isolate one variable at a time to determine if that variable is the cause of the problem.		
Develop criteria that will measure success of solutions.	*	*
Determine if cost of considering additional alternatives is greater than the likely benefit.		
Measure progress toward a solution.		
Identifying and Evaluating Constraints List the factors that might limit problem-solving efforts.		
Question credibility of one's own assumptions.		*
Recognize constraints related to possible solutions.		
Apply consistent evaluative criteria to various solutions.	*	*
Utilize creative and original thinking to evaluate constraints.		
Choosing a Solution Reflect upon possible alternatives before choosing a solution.	*	*
Use established criteria to evaluate and prioritize solutions.	*	*
Draw on data from known effective solutions of similar problems.		*
Evaluate possible solutions for both positive and negative consequences.		*
Choosing a Solution—Continued Explore a wide range of alternatives.	*	*

Table 9—Problem-solving skills as measured by the COMP and ETS Tasks in Critical Thinking —Continued

Problem-Solving Skills	**COMP**	**ETS Tasks**
Form a reasoned plan for testing alternatives.	*	*
Work to reduce the number of alternatives from which they choose a solution.		
Analyze alternatives to determine if most effective options have been selected.	*	*
Identify deficiencies associated with solutions and how they may be resolved.		
Explain and justify why a particular solution was chosen.		
Prioritize the sequence of steps in a solution.		
Group Problem Solving Identify and explain their thought processes to others.		
Be patient and tolerant of differences.		
Understand there may be many possible solutions to a problem.		
Use discussion strategies to examine a problem.		
Channel disagreement toward resolution.		
Fully explore the merits of innovation.		
Pay attention to feelings of all group members.		
Identify and manage conflict.		
Identify individuals who need to be involved in problem solving process.		
Search for aids of methods to reach agreement.		
Integrate diverse viewpoints.		
Stimulate creativity rather than conformity.		
Listen carefully to other's ideas.		
Understand and communicate risks associated with alternative solutions.		
Work on collaborative projects as a member of a team.		
Evaluation Choose solutions that contain provisions for continuous improvement.		
Seek alternative solutions if goals aren't achieved.		
Determine and review steps in implementation.	*	
Seek support for solutions.		
Revise and refine solutions during implementation.		*
Determine if their solutions integrate well with other solutions.		

Table 9—Problem-solving skills as measured by the COMP and ETS Tasks in Critical Thinking —Continued

Problem-Solving Skills	**COMP**	**ETS Tasks**
Dispositions		
Learn from errors.		
Work within constraints.		
Actively seek information.		
Take responsible risks.		
Remain adaptable and flexible when implementing solutions.		
Think creatively.		*
Search outside their expertise for solutions.		*

TEMPLATES — CRITICAL THINKING AND PROBLEM SOLVING

Critical Thinking Methods

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
The Academic Profile (A. Profile) Long form 144 items Short form 36 items **Author** Educational Testing Service **Publisher** Educational Testing Service Higher Education Assessment Princeton, NJ 08541 609–951–1509 **Date** 1986 **Testing Time** Long form 2.5 hrs. Short form 40 min. **Cost** $300 institutional fee Long form $15.50 Short form $9.95 *Entire test must be given* *All information from author*	Total Critical thinking subscore defined (6 more subscores available)	7 subscores Humanities: recognize cogent interpretation of a poem, distinguish between rhetoric and argumentation, draw reasonable conclusions, recognize elements of a humanities selection that strengthen or weaken the argument presented Social Sciences: recognize assumptions made in a piece of social science writing, recognize th best hypothesis to account for info presented in a social science passage, recognize info that strengthens or weakens arguments made in such a passage Natural Sciences: recognize the best hypothesis to explain scientific phenomenon, interpret relationships between variables in a passage, draw valid conclusions based on passage statements, recognize info that strengthens or weakens arguments in passage	.94 internal consistency for total .74–.85 internal consistency for subscores .74 internal consistency for critical thinking subscore .80 internal consistency of total short form .90 total alternate forms .77 critical thinking alternate forms .80 KR-20 for critical thinking subtest (Banta and Pike 1989)	Freshman inter-correlations of CT subscore w/ Humanities .78 Social Sciences .79 Natural Sciences .79 Reading .72 Writing .64 Math .52 Junior/senior inter-correlations of CT subscore w/ Humanities .84 Social Sciences .87 Natural Sciences .86 Reading .78 Writing .73 Math .52 Factor analysis supported 3 factors: reading/critical thinking, math, writing; CT factor correlated w/ Math .72 Writing .85 Intercorrelations of CT subscore w/ Reading .80/1.0 Writing .75/.99 Math .69/.89 (second correlation corrected for attenuation) (Marr 1995)	Critical thinking scores significantly different across major field, GPA, and core curriculum, but not for class level (Marr 1995) Content addresses consensus from American Association of Colleges "Integrity in the Core Curriculum"; content reviewed by ETS faculty, college-level assessment professionals, and senior faculty 68% of students' proficiency levels change across the various skills 67% of variance in critical thinking subtest scores accounted for by ACT scores, coursework not strong predictor of critical thinking scores (Banta and Pike 1989)	COMP objective test total .68 Subscores .15–.57 Percentage of total core curriculum completed w/ critical thinking subscore .17 (Marr 1995) COMP objective test total .64 (Banta and Pike 1989)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
California Critical Thinking Dispositions Inventory (CCTDI) 75 Likert scale items **Authors** Peter A. Facione and Noreen C. Facione **Publisher** California Academic Press 217 La Cruz Ave. Millbrae, CA 94043 **Date** 1992 **Testing Time** 15–20 minutes **Cost** $205/pack of 225 *Not a measure of critical thinking ability or skills*	Total Truth-seeking Open-mindedness Analyticity Systematically Confidence Inquisitiveness Cognitive maturity	All subscores Eager for knowledge and courageous to ask questions, even if knowledge fails to support or undermines preconceptions, beliefs, or self interests Tolerant of different views and self-monitoring for bias Prizing application of reason/evidence, alert to problematic situations, anticipate consequences Being organized, orderly, focused, and diligent in inquiry Trusting one's own reasoning process Curious/eager to acquire knowledge even if applications not immediate Prudence in making, suspending or revising judgment; awareness of multiple solutions	.80 internal consistency (Koehler and Neer 1996) .90–.91 total .71–.80 subscale internal consistency (Facione 1992)	Factor analysis 62/75 items loaded on the 7 subscales (Koehler and Neer 1996) No categorization format given for items (Callahan 1995; Ochoa 1995) Range of factor loadings for 7 subscales .029–.693 Item-total correlations for each subscale Truth .167–.467 Open .205–.573 Anal. .272–.510 Syst. .269–.568 Conf. .393–.569 Inquist. .317–.627 Maturity .175–.597 (Facione, Facione, and and Giancarlo 1992)	Content derived from American Philosophical Association committee, prompts screened by college-level CT educators (Facione, Facione, and Giancarlo 1992) No gender, ethnic, or geographical location information in manual (Ochoa 1995) No difference for overall means across gender (Facione, Sanchez, Facione, and Gainen 1995)	Age .18 Trait Argumentative Scale .43 (Koehler and Neer 1996) CCTST .66, .67 (authors) WGCTA .10 (Erwin 1996) WGCTA .17 (Erwin 1997)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
California Critical Thinking Skills Test (CCTST) Forms A and B 34-item multiple-choice **Author** Peter Facione **Publisher** California Academic Press 217 La Cruz Ave. Millbrea, CA 94043 **Date** 1990–1992 **Testing Time** 45 minutes **Cost** $225/Pack of 200 *Not for use with non-native, non-English speaking students*	Total Analysis: items 1–9 (includes interpretation) Inference: items 14–24 Evaluation: items 10–13 items 25–35 (includes explanation) Deductive reasoning: items 1, 2, 5, 6, 11–19, 22, 23, 30 Inductive reasoning: items 25, 27–29, 31–35	All subscores Categorization Decoding sentences Clarifying meaning Examining ideas Detecting arguments Analyzing arguments Querying evidence Conjecturing alternatives Drawing conclusions Assessing claims Assessing arguments Stating results Justifying procedures Presenting arguments Syllogisms Proofs in math Argument's conclusion follows from truth of its premises	Total Form A/B KR-20 .70–.71 (Facione and Facione 1992) Form A and B *respectively* Total .56, .59 Induction .42, .35 Deduction .50, .53 Analysis .04, .16 Evaluation .45, .33 Inference .36, .42 (Jacobs 1995) Form A .58–.59 internal consistency (Erwin)	Number of corrected item-total correlations below .1 for Forms A and B respectively (Total # items) Total (34) 10, 10 Induction (14) 5, 7 Deduction (16) 5, 2 Analysis (9) 9, 6 Evaluation (14) 3, 8 Inference (11) 3, 2 (Jacobs 1995) Principal component analysis did not support item classification (Jacobs 1995) -.08 to .34 item correlations with total, 7 out of 34 items correlated from -.08 to .09 with total (Erwin)	Content derived from American Philosophical Association committee and objectives of the California State University system Differences in CT across gender after critical thinking course—differences not found when SAT scores and GPA controlled Blacks and whites show significant improvement in CT skills after CT course, yet Hispanics and Asians show no gains Differences found for academic majors across critical thinking courses (all above, Facione and Facione 1992) Effect sizes for critical thinking courses .22–.33 (Erwin) Effect sizes .22–.44 for critical thinking course (Pike 1997) SAT—V, SAT—M, GPA, H.S. GPA accounted for 41% of variance in CCTST scores (Jacobs 1995)	SAT—V .55–.62 SAT—M .44–.48 Nelson-Denny .49 Age -.006 College GPA .20–.29 (Facione and Facione 1992) CCTT .56 WGCTA .50 SAT—V .45 SAT—M .37 (Freshmen, N = 131) (Erwin 1996) SAT—V .52–.59 SAT—M .55–.62 (Jacobs 1995) WGCTA .50 (Erwin 1997)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Collegiate Assessment of Academic Proficiency (CAAP) **Critical Thinking Test (CTT)** 32 multiple-choice items Same 32 items as the CTAB **Author** American College Testing Program **Publisher** American College Testing Program Iowa City, IA **Date** 1988 **Testing Time** 40 minutes **Cost** $285 for first order plus $8.80 per student (critical thinking section only) *For use with end-of-the-year sophomores*	Critical thinking total	Measure the ability to clarify, analyze, evaluate, and extend arguments Analysis of the elements of an argument 20 items Evaluation of an argument 6 items Extension of an argument 6 items	Total KR-20 .81–.82 (ACT Program 1989) Form A: KR-20 .30 (all female 2-yr. institution) KR-20 .79–87 (for all other 2- and 4-yr. public/private institutions) Form B: KR-20 .77–.84 (for all 2- and 4-yr. public/private institutions) (ACT Program 1991) Forms 88A/88B respectively: Total .82, .78 (freshman) Total .87, .82 (sophomores) (Lehigh Community College 1992) KR-20 critical thinking subtest .53 (Pike 1989)	Critical thinking inter-correlations with subscores, corrected for attenuation, respectively Form A: Writing skills .66, .75 Reading .70, .84 Math .53, .66 Form B: Writing skills .72, .86 Reading .77, .91 Math .48, .60 Median inter-correlations among Forms A/B: Writing skills .57 Reading .60 Math .36 Mean item discrimination indices from freshman to sophomore across 2- and 4-yr. public/private institutions Form A: .47–.58 (one exception, .27 for all female 2-yr. private institutions) Form B: .45–.54 (ACT Program 1991) Critical thinking subtest .85 factor loading w/ reading and writing (Pike 1989)	ACT encourages local validation Content determined by panel of subject experts (ACT Program 1991) Intended to measure group, not individual change first-year students at a 4-year college full-time (24 hrs. or more per year) attained a higher critical thinking score than part-time (6 hrs. or less per year) students; effect size advantage .41 (Pascarella et al. 1996) Jr. English GPA predicted by CT scores, .32 .9 Mean difference from freshman-sophomore longitudinal study 1.7 mean difference from cross-sectional study Beyond precollege experiences, college experience explained 7–17% of variance in first-year critical thinking gains ACT total scores account for 30% of variance in critical thinking subtest scores (Pike 1989)	Sophomore GPA .34 Jr. Cum. GPA .35 (Lehigh Community College 1992) WGCTA .75 (McMillan 1986)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
College Outcome Measures Program (COMP) Objective Test 60 multiple-choice items: simulation activities with excerpts from TV documentaries, radio newscasts, commentaries, magazine articles, music, and art 2 correct responses, 2 distractors—points subtracted for incorrect response **Author** American College Testing Program **Publisher** American College Testing Iowa City, IA **Date** 1976 **Testing Time** 2.5 hours **Cost** $6–$17/per test	Total Communicating Solving problems Clarifying values Functioning within social institutions Using science and technology Using the arts	All subscores Send and receive info. in a variety of modes, within a variety of settings, and for a variety of purposes Analyze a variety of problems, select or create solutions, and implement solutions Identify one's personal values and values of others, understand how personal values develop, analyze implications of decisions made on personally held values Identify, analyze, and understand social institutions, impacts of self and others Identify, analyze, and understand tech., impacts of self and others Identify, analyze, and understand art, impacts of self and others	Alternate forms reliability for Objective test (forms 9/6, 10/5, 11/9) Total .83, .86, .86 Communicating .66, .70, .76 Solving problems .69, .70, .72 Clarifying values .65, .73, .71 .84 internal consistency .63–.68 subscores G study forms 9/10 Total .86–.97 Subscores .71–.96 (values vary across sample size) (ACT Program 1990) Alternate forms reliability .70 Subscales .53–.68 (Forrest and Steele 1982) KR-20 problem-solving subtest .51, G coefficient .61 (Pike 1989)	High ceiling: 6% of nation's high scorers get 67% correct Subscale correlations Fresh. .43–.55 Seniors .48–.53 (ACT Program 1990) Subscale correlations Fresh. .51–.58 Seniors .54–.57 (Forrest and Steele 1982) Solving problems with other COMP subscales .50–.71 (Sibert 1989) Single factor supported by factor analysis (Banta and Pike 1989) 13 items (54%) on solving problems exhibited race DIF (blacks and whites), favoring whites most often (Pike 1989b)	Content reviewed by ACT staff, senior college faculty, and consultants Faculty rated problem-solving subtest as 100% content coverage for college outcomes (Pike 1989) For solving problems subtest means from freshman (72.0) to senior (74.5–76.5) increase; mean difficulty from freshman (50%) to senior (55.2–59.4%); no gender differences Preprofessional Skills English score + social sciences ACT score account for 45% of variance in problem-solving scores (Sibert 1989) 8.9 gain in mean scores for institutions that have 46% of degree gen. ed. requirements 3.9 gain in mean scores for institutions that have 31% of degree gen. ed. requirements ACT (academic ability) accounts for 20% of variance in problem-solving scores (Pike 1989) Student scores higher for subtests related to major (Forrest and Steele 1982)	COMP UAP area tests correlated with objective test, .47–.59 Preprofessional Skills Test .36–.56 National Teacher Exams .53–.62 Major GPA .33 Cumulative GPA .35 ACT total .58 (Sibert 1989) CAAP subscores .24–.65 A. Profile Critical Thinking subtest w/ problem solving .42 Total ACT score .46 (Banta and Pike 1989) Senior GPA .32 Amount of reading .14 Seniors mean ACT score and mean gains -.34 SAT total .66–.68 GRE subscores w/ communication, solving problems, clarifying values subscores respectively Verbal .66, .53, .62 Quant. .54, .22, .34 Anal. .67, .48, .57 GMAT subscores w/ communication, solving problems, clarifying values subscores respectively Verbal .49, .54, .57 Quant. .45, .13, .31 Total .60, .28, .48

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Cornell Critical Thinking Test (CCTT) 50-item multiple-choice Level Z: grade 13 and above **Authors** Robert Ennis and Jason Millman **Publisher** Critical Thinking Press and Software P.O. Box 448 Pacific Grove, CA 93950–0448 **Date** 1971, 1982 **Testing Time** 50 minutes **Cost** $16.95/pack of 10	Total	Deduction (items 1–10) Semantics (items 11 and 21) Credibility (items 22–25) Induction—judging conclusion (items 26–38) Induction (items 39–42) Definition and assumption identification (items 43–46)	.50–.77 split half internal consistency (Ennis, Millman, and Tomko 1985) .74–.80 split half internal consistency (Frisby 1992) .70 internal consistency (Mines et al. 1990) .58 internal consistency freshman .72 internal consistency sophomores	Discrimination indices .20–.24 Legitimate low-scoring test takers and those who "guessed" produced scores in the same range (Frisby 1992) -.17–.43 item correlations with total 11 out of 52 items correlations range from -.17 to .08 Gender DIF analysis found 3 items favor males, while 1 item favors females	Review of items and keyed responses by Illinois Critical Thinking Project members (authors) Cross-sectional study from freshman to seniors showed significant CT improvement (Mines et al. 1990) Validity study contains sample group and data collection procedure deemed consistent with test purpose; possible test bias/lack of cross validation (Modjeski and Michael 1983) Differences found across ability levels (Frisby 1992) Subtest scores increased across reflective judgment stages Detecting ambiguous arguments and 3 WGCTA subtests accounted for 50% of variance in RJI stages (Mines et al. 1990) Contradictory findings: Study 1—No differences found across CT course; Study 2—Significantly higher gains for students who took critical thinking course vs. no critical thinking course (Langer and Chiszar 1993)	SAT—V .36 SAT—M .51 Rokeach Dogmatism Scale -.41, -.37 WGCTA .48, .79 Logical Reasoning Test, part II, Form A .25 Test of Critical Thinking, Form G .44 RJI .62 (authors, all above) GPA .32–.38 Graduate units .34–.41 (Garret and Wulf 1978) WGCTA .48 CCTST .56 SAT—V .48 SAT—M .36 (Erwin) SAT Writing .42 SAT Verbal .44 LSAT .48 (Frisby 1992) MMPI (ego-related subscales) Men .21–.25 Women .31–.38 WGCTA .71, .54, .94 RJI .46, .27, .59 (for WGCTA and RJI: correlation, w/ academic ability controlled, corrected for attenuation, respectively) ACT .62 (King, Wood, and Mines et al. 1990; Mines et al. 1990)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Critical Thinking Assessment Battery (CTAB) **Author** American College Testing Program **Publisher** American College Testing Program Iowa City, IA **Date** 1997 **Testing Time** 2.5 hours **Cost** $15 (pilot) *All info. from author*	Critical thinking (32 multiple-choice items—total score) Applied reasoning (3 essays and 15 double multiple-choice questions—total; social, scientific, and artistic reasoning subscores) Engagement in reasoning and communicating (15 ranked sets of questions—total score) Persuasive writing (3 essays—same essays rated for applied reasoning score—total score; audience, organization, language subscores)	Assesses skills in clarifying, analyzing, evaluating, and extending arguments Assesses skills in analyzing problems, generating logical and reasonable approaches to solve and implement solutions, reflecting consistent value orientations Inventories past involvement in community/social contexts, requiring application of problem-solving and communicating skills Assesses skills in written communication, including making contact with a relevant audience, organizing a persuasive message that develops a number of relevant ideas, and using language to present ideas clearly and effectively			No validity studies done as of 3/21/97 Pilot testing was planned for fall 1997 and winter 1998 Test takers will be rated as Level 1, Level 2, or Level 3 (Level 3 = high degree of competence) Validation studies will be done on these criterion-referenced levels of proficiency for CTAB components during pilot testing Content validity of CATB's four components supported by the inclusion of: Paul's elements of reasoning/ intellectual standards NSF/NAEP problem-solving steps Bloom's cognitive levels of thinking Torrance's criteria for creative thinking	

Description:

Part I (three essay responses to role-playing tasks)—Assesses skills in analyzing problems and generating logical and reasonable approaches to solve and implement solutions, reflecting consistent value orientations.

Part II (utilizing the same essays produced for part I)—Provides a performance assessment of skills in written communication including making contact with a relevant audience, organizing a persuasive message that develops a number of relevant ideas, and using language to present ideas clearly and effectively.

Part III (32 multiple-choice questions)—Assesses skills in clarifying, analyzing, evaluating, and extending arguments.

Part IV (15 innovative double multiple-choice items)—Measures applied skills in reasoning and decisionmaking.

Part V (16 ranked self-report items and optional short written responses)—Inventories past involvement in community/social contexts, requiring application of problem solving and skills.

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Measure of Intellectual Development (MID) Single essay (2 forms) **Author** William S. Moore **Publisher** Center for the Study of Intellectual Development 1505 Farwell Ct. NW Olympia, WA 98502 **Date** 1988 **Testing Time** 20–30 minutes **Cost** $15 (pilot)	Position 2 Position 3 Position 4 Position 5	Dualistic thought, content-oriented, high level of external control Some ownership of thought, methods become authority, fairness important Realization of many alternatives, independent thinker, active in the learning process, flexibility and learning from others Diversity assumed, meta-thought, seeking knowledge, search for the truth, realization of no absolute truth	Rater agreement 51.2% within 1/3 of position agreement 93.6% (Mentkowski no date available) Expert rater agreement correlation .45, .53 Correlations w/ dominant position .76, .80 (Moore 1990)		Scoring based on Perry scheme of intellectual and ethical development, test first developed by Knefelkamp, Widick, and Stroad (1976) (author) Dualist treatment gain .85 Relativist treatment gain .79 (Knefelkamp, Widick, and Stroad 1976) Treatment group gain .85 vs. control groups .42, .12 (Stephenson and Hunt 1977) Longitudinal study, from freshman to senior year, increase in mean score, no difference across gender (Moore 1990) *All studies cited from Moore 1990*	DIT (measure of moral reasoning) .45 DIT .13 Sentence completion task (ego development) .30 (Wertheimer 1980) MER (measure of epistemological reflection) .13 Interview ratings for Perry scheme .74, .77 (Knefelkamp and Sleptiza 1976)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
ETS Tasks in Critical Thinking Nine essay/short answer tasks: three each in humanities, social science, and natural science **Author** New Jersey Faculty **Publisher** Educational Testing Service Higher Education Assessment Princeton, NJ 08541 609–951–1509 **Date** 1989 **Testing Time** 90 minutes **Cost** $12 each *No individual student scores*	Inquiry Analysis Communication	Plan a search; use various methods of observation and discovery; comprehend and extract; sort and evaluate Formulate hypotheses and strategies; apply techniques, roles, and models to solve problems; demonstrate breadth, flexibility, and creativity; evaluate assumptions, evidence, and reasoning; find relationships and draw conclusions Organize the presentation; write effectively; communicate quantitative or visual information	To be determined by users (manual) Inter-rater reliability G coefficients Task 15 .65 Task 19 .57 Task 22 .61 ETS raters vs. local raters across tasks .67–.95 correlations (Erwin and Sebrell) All author reliabilities based on NJ GIS assessment (tasks and MC items) .80 and .87 mean inter-rater reliabilities for pilot tests Alternate forms reliability across skills Planning .17 Gathering info. .66 Evaluating assumptions .20 Finding relationships .69 Analyzing info. .57 Quant. reasoning .90 (ETS and the College Board 1990)	Intertask correlations 15/19 .22 15/22 .19 19/22 .22 Interskill correlations based on tasks 15, 19, 22 .23–.30 Inquiry .10–.23 Analysis -.03–.43 Comm. (Erwin and Sebrell)	Content based on and reviewed by NJ faculty, ETS, and College Board; original test was New Jersey General Intellectual Skills Assessment (GIS) Bias in scoring guide due to people in the discipline related to the task creating guide; tests all mimic classroom tasks; essay-writing performance affecting CT performance (Scriven 1991) To be determined by users (manual)	SAT—V .32 SAT—M .21 Local logic test .27 With individual tasks SAT—V .16–.47 SAT—M .03–.39 (Erwin and Sebrell)

Tasks in Critical Thinking Scoring Rubrics

Core scoring method—Analysis and inquiry

1 Not proficient—A response was attempted but students scoring at this level either did not understand the questions or their explanations were erroneous, illogical, totally unrelated to the requirements.

2 Limited proficiency—The basic requirements were not met, and responses were very brief, inappropriate, and/or incorrect. Responses were vaguely expressed or inaccurate.

3 Some proficiency—Student understood the question, yet the basic requirements were not met. Responses were vague, incomplete, and/or inappropriate.

4 Fully proficient—The **Core Score** means that the questions were understood and the responses were correct and complete. Students met all basic requirements.

5 Exceeds requirements—Students met all the basic requirements and provided some expansion or extension—citing evidence, providing additional information, or in some other way going beyond what was required.

6 Superior performance—All basic requirements were met and expanded upon; in addition, students presented ideas, interpretations, relationships, or examples that showed originality and insight.

Holistic: Communication

1 Not proficient—A paper demonstrating incompetence. It is seriously flawed by very poor organization, very thin development, and/or usage and syntactic errors so severe that meaning is somewhat obscured.

2 Limited proficiency—A paper flawed by weaknesses such as failure to develop the required assignment, poor organization, thin development, using little or inappropriate detail to support ideas, and/or displaying frequent errors in grammar, diction, and sentence structure.

3 Some proficiency—A slightly less than adequate paper that addresses the writing task in a vague or unclear way, shows inadequate organization or development, and/or has an accumulation of errors in grammar, diction, or sentence structure.

4 Fully proficient—This is an adequate paper with only occasional errors or lapses in quality. It is organized and somewhat developed and uses examples to support ideas. It shows a basic command of, and adequate facility in, use of language.

5 Exceeds requirements—A very strong paper with only occasional errors or lapses in quality. It is generally well organized and developed, displaying facility in language, range of vocabulary, and some variety in sentence structure.

6 Superior performance—A superior paper that is well organized and developed, using appropriate examples to support ideas. It displays facility in language, range of vocabulary, and variety in sentence structure.

OT Off topic, this designation is used for responses that were completely off the assigned topic.

Omit No response was attempted

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Problem Solving Inventory (PSI) Forms A and B 35 Likert statements **Author** P. Paul Heppner **Publisher** Consulting Psychologist Press 3803 E. Bayshore Rd. Palo Alto, CA 94303 **Date** 1982, Form A 1988, Form B **Testing Time** 15 minutes **Cost** 1990, $14.50/per 25 tests *Low scores indicate positive problem-solving abilities*	Total 32 items Problem-solving confidence (PSC) 11 items Approach-avoidance (AA) 16 items Personal control (PC) 5 items	General index of problem-solving appraisal Self-assurance while engaging in problem-solving activities Tendency of individuals to approach or avoid problem-solving activities Extent to which individual believes he or she is in control of emotions and behavior in problem-solving activities	All Form A reliabilities .72–.90 internal consistency for total and subscales .83–.89 2-wk. Test-retest (Heppner and Peterson 1982a) .77–.81 3-wk. Test-retest reliability (Ritchey, Carscaddon, and Morgan 1984) .44–.65 2-yr. test-retest reliability	Several factor analyses give support for 3-factor model on Form A (Chynoweth 1987 cited in Heppner 1988; Heppner and Peterson 1982a) Factor loadings for subscales Confidence .42–.75 Approach-avoid .30–.71 Control .42–.71 Congruence coefficients indicate overlap in factors .96–.99 (Heppner and Peterson 1982a) Interscale correlations PSC/PC .46–.53 PSC/AA .39–.51 PC/AA .40–.48 (Elliott et al. 1995)	Based on 5-stage problem-solving model; differences found after problem-solving training compared to controls; cross-validation of normative data (Heppner and Peterson 1982a) No differences across academic levels No group differences after motivation course, yet students who successfully completed course perceived improved CT ability (Chynoweth, Blankinship, and Parker 1986) Blind judges correctly rated 83% students as high- and low-scorers based on interviews (Heppner and Anderson 1985) Increases in clients' problem-solving ability after problem solving (effect size change = 2.49) vs. problem focused (effect size change = .46) vs. no therapy (Nezu 1986) Positive PSI scores predict greater positive and lower negative affect (Elliott et al. 1995) Low PSI scores associated with tendency to enjoy cog. activities, fewer dysfunctional thoughts, stronger self concepts, lower irrational belief scores, and positive coping skills (Heppner and Peterson 1982b) Masculinity (16.2%) and maleness (20.3%) predictors of PSI scores (Brems and Johnson 1989)	Social desirability scale -.16 Rotter I-E scale .61 SCAT-II .13 MCET(writing ability) -.08 MMPT (algebra) .08 H.S. rank .06 Self-rating scales problem solving -.46 satisfaction w/ problem solving -.42 (Heppner and Peterson 1982a) SAT—V -.19 SAT—M -.31 Test anxiety scale .20–.35 (Blankstein, Flett, and Batten 1989) State-trait personality inventory .47 (Carscaddon, Poston, and Sachs 1988) PST (index of distress) PSC/PST .21 PC/PST .22 AA/PST .03 PANAS (trait affect) positive -.28– -.40 negative .17–.38 (Elliott et al. 1995)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Reflective Judgment Interview (RJI) 4 intellectual problems with contradictory views along with a series of standardized probe questions **Authors** Patricia King and Karen Kitchener **Publisher** **Date** 1983 **Testing Time** 45–60 minutes **Cost**	Stage 1 Stage 2 Stage 3 Stage 4 Stage 5 Stage 6 Stage 7	Absolutism Dogmatism Knowledge uncertain; beliefs based on whim Accept uncertainty of knowledge; skeptically argue Subjective evaluations of reality; objective reality does not exist Objectively compare claims; beliefs are plausible opinions Uncertainty part of objective reality; knowledge consequence of critical inquiry and evaluation	Internal consistency .89 (Mines et al. 1990) .75 (Brabeck 1983) .85 (King, Wood, and Mines et al. 1990) .96 (King and Kitchener 1994) Inter-rater reliability .97 (Mines et al. 1990; King, Wod, and Mines et al. 1990) Rater agreement .76 (Brabeck 1983) .90 (Mines et al. 1990; King, Wood, and Mines et al. 1990)	.35–.47 correlation between all pairs of dilemmas .52–.59 dilemma-total correlations-subjects subscores across dilemmas correlated (Brabeck 1983)	Based on reflective judgement stage theory (authors) WGCTA and CCTT had increasing linear pattern across RJI scores; Differences for seniors vs. grad. students with ability controlled; men scored higher even w/ ability controlled (King, Wood, and Mines et al. 1990) Differences in cross-sectional freshman–graduate students controlling for ability (Mines et al. 1990) Increase in scores from H.S. seniors to sophomores to college seniors, w/ ability controlled college seniors different than others; high WGCTA scorers had greater variability on RJI stages than low WGCTA scorers (Brabeck 1983)	WGCTA .40 SAT—V .53 (Brabeck 1983) ACT .44 CCTT .46, .27, .59 WGCTA .46, .27, .51 (for CCTT and WGCTA: correlation, w/ academic ability controlled, corrected for attenuation, respectively) (Mines et al. 1990; King, Wood, and Mines et al. 1990)

Name	Scores	Definition	Reliability	Method Design	Validity	Correlation With Other Measures
Watson-Glaser Critical Thinking Appraisal (WGCTA) Forms A and B (YM and ZM forms—older versions) 80 multiple-choice items **Authors** Goodwin Watson and Edward M. Glaser **Publisher** The Psychological Corp. 555 Academic Ct. San Antonio, TX 78204–2498 **Date** 1980 **Testing Time** 40 minutes **Cost** $97/pack 25	Total	Inference: Discriminating among degrees of truth or falsity of inferences drawn from given data Recognition of assumptions: Recognizing unstated assumptions or presuppositions in given statements or assertions. Deduction: Determining whether certain conclusions necessarily follow from information in given statements or premises. Interpretation: Weighing evidence and deciding if generalizations or conclusions based on the given data are warranted Evaluation of arguments: Distinguishing between arguments that are strong and relevant and those that are weak or irrelevant to a particular question at issue	.69–.85 split half .75 alternate forms .73 test-retest over 3 mo. (author) .70s split half internal consistency (Sternberg 1983, cited in King, Wood, and Mines et al. 1990) .82 internal consistency (Mines et al. 1990) .76 internal consistency (Brabeck 1983) .78 internal consistency (Taube 1995) .87 internal consistency .54–.80 subscale internal consistency freshman	Form A subtest intercorrelations .29–.50 (Brabeck 1983) Intercorrelations based on nursing students Forms A and B .45–.69 (authors) With recognition of assumptions excluded, 3 out of 4 factors loaded with test of divergent thinking, but not convergent thinking (Fontana et al. 1983) Confirmatory factor analysis supported WGCTA as ability factor with SAT scores and GPA (Taube 1995) .01–.48 item correlations with total; 6 out of 80 item correlations range .01–.09; 4 items exhibited DIFF, 2 items favored females, 2 items favored males	Manual contains validity evidence for suggested inferences; sample and data collection consistent with test use; universe of sampled performance defined; possible test bias/lack of cross-validation studies, Forms YM, ZM (Modjeski and Michael 1983) Content based on definition of Dressel and Mayhew (1954) (authors) Increase in scores across RJI stages; 3 WGCTA subtests and 1 CCTT subtest accounted for 50% of variance in RJI stages (Mines et al. 1990) 3 out of 8 studies found differences for CT across CT courses versus non-CT courses (McMillan 1987) Successful prediction of women's performance in physics courses, but not men's (McCammon, Goldman, and Wuensch 1988) Differences for college versus non-college students, effect size .44 (Pascarella 1989) Differences not found across nursing program (Saucier,1995) Differences in CT across grades (A>B>C) for freshman courses (Gadzella et al. 1996) Lower CT for med. students who took extra time to complete courses or changed their curricula (Scott and Markett 1994)	SAT—V .37–.69 SAT—M .29–.48 ACT, composite .65 ACT: Math .30, English .21 CA. Achievement test, reading .64 (author) CCTDI .10, .17; CCTST .50 CCTT .48 SAT—V .48, .35 SAT—M .36, .25 (Erwin 1996) RJI .40 (Brabeck 1983) CCTT .71, .54, .94 RJI .46, .27, 51 (for CCTT and RJI: r, academic ability controlled, corrected for attenuation respectively) ACT .59 (Mines et al. 1990) (King, Wood, and Mines et al. 1990) Math anxiety rating scale -.30 Arithmetic Skills Test .36 Primary Mental Abilities Test .44 (McCammon, Golden, and Wuensch 1988) WG: Form A/YM .78 WG: Form B/ZM .69 (Berger 1985) MCAT scores Reading .57 Quantitative .40 Age -.23 (Scott and Markett 1994) CLEV (dualism) .33 SAT—V .43 SAT—M .39 GPA .30 (Taube 1995)

3. WRITING

3.1 Introduction

An effective and meaningful evaluation of postsecondary writing assessments is predicated upon a comprehensive understanding of the definition of writing competency. Therefore, the writing part of this sourcebook begins with an overview of existing approaches to the definition of writing. This preliminary segment also contains a table highlighting the writing skill components measured by several existing postsecondary writing tests. In the second section, descriptions of different types of formats used to assess writing competency—both directly and indirectly—are provided, with consideration of the advantages and disadvantages of each method. This section closes with a discussion of computerized writing assessment and an exploration of some global issues relevant to all postsecondary writing assessment efforts. Finally, to further aid individuals in the selection of a useful writing assessment, details of each existing measure (scoring, author/publisher, testing time, date, cost, specific purposes, current users, details related to the utility, and psychometric properties, as well as the scale definition and rubrics) are displayed in the context of a comprehensive chart.

3.2 Definition of Writing

Although writing is clearly a form of communication that connotes activity and change, attempts to define writing often focus on the products (essays, formal reports, letters, scripts for speeches, step-by-step instructions, etc.) or the content of what has been conveyed to whom. When writing is defined only as a product, elaboration of the construct tends to entail specification of whether particular elements, such as proper grammar, variety in sentence structure, organization, etc., are present (suggestive of higher quality writing) or absent (indicative of lower quality writing). Attention is given to describing exactly what is generated and detailing the skill proficiencies needed to produce a given end-product. Although educators, researchers, and theorists in the writing field tend to prefer a process-oriented conceptualization of writing, research suggests that employers in industry are more interested in defining writing competence with reference to products (Jones et al. 1995). Section 3.4 (see below) provides a brief summary of the history of process theory in writing assessment.

In a report on national assessment of college student learning, Jones et al. (1995) provided a comprehensive definition of writing, which in addition to including several subcomponents of the process, delineates critical aspects of written products. The general categories of key elements composing the construct of writing produced by these authors include awareness and knowledge of audience, purpose of writing, prewriting activities, organizing, drafting, collaborating, revising, features of written products, and types of written products. These researchers developed this definition based on an extensive review of relevant literature and feedback from a large sample of college and university faculty members, employers, and policymakers representative of all geographic regions in the United States. Stakeholders were asked to rate the importance of achieving competency on numerous writing skills upon completion of a college education. Jones et al. found that in every area of writing there were certain skills that each respondent group believed were essential for college graduates to master in order to facilitate effective functioning as employees and citizens. However, there were areas of contention as well. For example, employers and policymakers placed less emphasis on the importance of the revision process, tending to expect their graduates to be able to produce high-quality documents on the first attempt. In addition, employers found the ability to use visual aids, tables, and graphs as more important than did faculty members; and faculty members attached more importance to being able to write abstracts and evaluations. The resulting definition produced by Jones et al., which only includes skills that were universally endorsed by all three groups, is distinct from other definitions in that it is based on a *consensus* derived

empirically from groups that possess very different interests regarding the development of writing skill competency through undergraduate training. The Jones et al. definition will, therefore, be used as a base for examining the scope of the writing assessments to be surveyed herein.

Table 10 provides a detailed list of all of the subcomponents addressed in the definition, in addition to an indication of which currently available measures assess particular components. Only multiple-choice and essay tests are included in the table, because the rubrics used with most portfolio measures tend to only address very global dimensions of writing quality.

Table 10—Dimensions of writing reflected in assessment methods

Multiple-Choice Tests

Components	CLEP	SAT-II	AP-Eng. Comp.	CAAP	A. Profile	COMPASS	TASP	CLAST
Awareness and Knowledge of Audience 1. Consider how an audience will use the document. 2. Choose words that their audience can understand. 3. Understand the relationship between the audience and the subject material. 4. Address audiences whose cultural and communication norms may differ from those of the writer. 5. Clearly understand their audiences' values, attitudes, goals, and needs. 6. Understand the relationship between the audience and themselves. Other dimensions are covered generally.							*	
Purpose of Writing 1. State their purpose(s) to their audience. 2. Use vocabulary appropriate to their subject and purpose(s). 3. Arrange words within sentences to fit the intended purpose(s) and audiences. 4. Make appropriate use of creative techniques of humor and eloquence when approaching a writing task. 5. Draw on their individual creativity and imagination to engage their audience. Other dimensions are covered generally.	*	*					*	*

Table 10—Dimensions of writing reflected in assessment methods—Continued

Multiple-Choice Tests

Components	CLEP	SAT-II	AP-Eng. Comp.	CAAP	A. Profile	COMPASS	TASP	CLAST
Prewriting Activities								
1. Discuss their piece of writing with someone to clarify what they wish to say.								
2. Research their subject.								
3. Identify problems to be solved that their topic suggests.								
Other dimensions are covered generally.								
Organization								
1. Organize the material for more than one audience.								
2. Include clear statements of the main ideas.	*							
3. Demonstrate their method of organization to their audience(s) by using informative headings.								
4. Write informative headings that match their audiences' questions.								
5. Maintain coherence within sentence.	*	*		*	*			
6. Maintain coherence among sentences, paragraphs, and sections of a piece of writing.	*			*	*			
7. Develop patterns or organization for their ideas.								
8. Use knowledge of potential audience expectations and values to shape a test.								
9. Create and use an organizational plan.								
10. Organize their writing in order to emphasize the most important ideas and information within sentences and larger units such as paragraphs.	*							
11. Cluster similar ideas.								
12. Provide a context for the document in the introduction.	*							
13. Set up signposts such as table of contents, indexes, and side tabs.								
14. Demonstrate patterns of reasoning in their writing.				*	*	*	*	
Other dimensions are covered generally.								

Table 10—Dimensions of writing reflected in assessment methods—Continued

Multiple-Choice Tests

Components	CLEP	SAT-II	AP-Eng. Comp.	CAAP	A. Profile	COMPASS	TASP	CLAST
Drafting								
1. Avoid common grammatical errors of standard written English.								
2. Quote accurately.								
3. Establish and maintain a focus.								
4. Write effective introductions and conclusions.								
5. Write effectively under pressure and meet deadlines.								
6. Make general and specific revisions while they write their drafts.								
7. Move between reading and revising of their drafts to emphasize key points.								
8. Refine the notion of audience(s) as they write.								
Other dimensions are covered generally.								
Collaborating								
1. Collaborate with others during reading and writing in a given situation.								
Other dimensions are covered generally.								
Revising								
1. Correct grammar problems.					*			
2. Revise to improve word choice.								
3. Select, add, substitute, or delete information for a specified audience.								
4. Reduce awkward phrasing and vague language.					*			
Other dimensions are covered generally.								
Features of Written Products								
1. Use active or passive voice where appropriate.	*							
2. Use language their audience understands.								
3. Define or explain technical terms.				*			*	
4. Use concise language.								
5. Use correct grammar, syntax (word order), punctuation, and spelling.	*	*		*	*	*	*	*
6. Use correct reference forms.							*	
7. Use the specific language conventions of their academic discipline or professional area.						*	*	*
Other dimensions are covered generally.								

Table 10—Dimensions of writing reflected in assessment methods—Continued

Multiple-Choice Tests

Components	CLEP	SAT-II	AP-Eng. Comp.	CAAP	A. Profile	COMPASS	TASP	CLAST
Written Products								
1. Write memoranda.								
2. Write letters.								
3. Write formal reports.								
4. Write summaries of meetings.								
5. Write scripts for speeches/presentations.								
6. Complete pre-printed forms that require written responses.								
7. Write step-by-step instructions.								
8. Write journal articles.								
9. Write policy statements.								
Other dimensions are covered generally.								
Other								
1. Style.				*				
2. Avoidance of figurative language.					*			
3. Shifts in construction.						*		
4. Analyzing rhetoric.			*					
5. Ambiguity/wordiness.	*	*						
6. Insightful support for ideas.	*							
7. Point of view exemplified.	*							
8. Maintenance of a consistent tone.								
9. Effective opening and closing.								
10. Avoidance of generalizations, cliches.								
11. Awareness, insight into complexities of prompt.								
12. Separating relevant from irrelevant information.								
13. Depth, complexity of thought.								
14. Sentence variety.	*							

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Local Essay Tests					Commercial Essay Tests		
Components	TASP	CLAST	SEEW	IIEP	NJCBSPT	SMSU	College Base	Praxis I
Awareness and Knowledge of Audience								
1. Consider how an audience will use the document.								
2. Choose words that their audience can understand.	*							
3. Understand the relationship between the audience and the subject material.	*							
4. Address audiences whose cultural and communication norms may differ form those of the writer.								
5. Clearly understand their audiences' values, attitudes, goals, and needs.								
6. Understand the relationship between the audience and themselves.								
Other dimensions are covered generally.	*							
Purpose of Writing								
1. State their purpose(s) to their audience.		*		*	*	*		*
2. Use vocabulary appropriate to their subject and purpose(s).	*	*					*	
3. Arrange words within sentences to fit the intended purpose(s) and audiences.	*	*						
4. Make appropriate use of creative techniques of humor and eloquence when approaching a writing task.								
5. Draw on their individual creativity and imagination to engage their audience.	*	*						
Other dimensions are covered generally.			*	*	*	*		*
Prewriting Activities								
1. Discuss their piece of writing with someone to clarify what they wish to say.								
2. Research their subject.								
3. Identify problems to be solved that their topic suggests.								
Other dimensions are covered generally.								

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Local Essay Tests					Commercial Essay Tests		
Components	TASP	CLAST	SEEW	IIEP	NJCBSPT	SMSU	College Base	Praxis I
Organization								
1. Organize the material for more than one audience.			*		*	*		
2. Include clear statements of the main ideas.								
3. Demonstrate their method of organization to their audience(s) by using informative headings.								
4. Write informative headings that match their audiences' questions.								
5. Maintain coherence within sentence.		*						
6. Maintain coherence among sentences, paragraphs, and sections of a piece of writing.		*	*			*		*
7. Develop patterns or organization for their ideas.	*	*	*	*		*		*
8. Use knowledge of potential audience expectations and values to shape a test.	*	*						*
9. Create and use an organizational plan.			*	*				
10. Organize their writing in order to emphasize the most important ideas and information within sentences and larger units such as paragraphs.				*	*			*
11. Cluster similar ideas.								
12. Provide a context for the document in the introduction.								
13. Set up signposts such as table of contents, indexes, and side tabs.								
14. Demonstrate patterns of reasoning in their writing.								
Other dimensions are covered generally.	*	*	*	*	*	*	*	*

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Local Essay Tests					Commercial Essay Tests		
Components	TASP	CLAST	SEEW	IIEP	NJCBSPT	SMSU	College Base	Praxis I
Drafting								
1. Avoid common grammatical errors of standard written English.	*	*	*			*		
2. Quote accurately.	*							
3. Establish and maintain a focus.								
4. Write effective introductions and conclusions.								
5. Write effectively under pressure and meet deadlines.								
6. Make general and specific revisions while they write their drafts.								
7. Move between reading and revising of their drafts to emphasize key points.								
8. Refine the notion of audience(s) as they write.								
Other dimensions are covered generally.								
Collaborating								
1. Collaborate with others during reading and writing in a given situation.								
Other dimensions are covered generally.								
Revising								
1. Correct grammar problems.								
2. Revise to improve word choice.								
3. Select, add, substitute, or delete information for a specified audience.								
4. Reduce awkward phrasing and vague language.								
Other dimensions are covered generally.								
Features of Written Products								
1. Use active or passive voice where appropriate.								
2. Use language their audience understands.				*				
3. Define or explain technical terms.								
4. Use concise language.		*						
5. Use correct grammar, syntax (word order), punctuation, and spelling.	*	*	*	*	*	*	*	*
6. Use correct reference forms.								
7. Use the specific language conventions of their academic discipline or professional area.						*		
Other dimensions are covered generally.		*	*	*	*	*		*

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Local Essay Tests				Commercial Essay Tests			
Components	TASP	CLAST	SEEW	IIEP	NJCBSPT	SMSU	College Base	Praxis I
Written Products 1. Write memoranda. 2. Write letters. 3. Write formal reports. 4. Write summaries of meetings. 5. Write scripts for speeches or presentations. 6. Complete pre-printed forms that require written responses. 7. Write step-by-step instructions. 8. Write journal articles. 9. Write policy statements. Other dimensions are covered generally.								
Other 1. Style. 2. Avoidance of figurative language. 3. Shifts in construction. 4. Analyzing rhetoric. 5. Ambiguity/wordiness. 6. Insightful support for ideas. 7. Point of view exemplified. 8. Maintenance of a consistent tone. 9. Effective opening and closing. 10. Avoidance of generalizations, cliches. 11. Awareness, insight into complexities of prompt. 12. Separating relevant from irrelevant information. 13. Depth, complexity of thought. 14. Sentence variety.	 *	 *	 * 	 * 	 * * * * 	 * 	 * 	 * *

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Commercial Essay Tests							
Components	COMP	A. Profile	CAAP	MCAT	TWE	GMAT	SAT-II	CLEP
Awareness and Knowledge of Audience 1. Consider how an audience will use the document. 2. Choose words that their audience can understand. 3. Understand the relationship between the audience and the subject material. 4. Address audiences whose cultural and communication norms may differ from those of the writer. 5. Clearly understand their audiences' values, attitudes, goals, and needs. 6. Understand the relationship between the audience and themselves. Other dimensions are covered generally.	*							
Purpose of Writing 1. State their purpose(s) to their audience. 2. Use vocabulary appropriate to their subject and purpose(s). 3. Arrange words within sentences to fit the intended purpose(s) and audience. 4. Make appropriate use of creative techniques of humor and eloquence when approaching a writing task. 5. Draw on their individual creativity and imagination to engage their audience. Other dimensions are covered generally.		*	* *	*		*		*
Prewriting Activities 1. Discuss their piece of writing with someone to clarify what they wish to say. 2. Research their subject. 3. Identify problems to be solved that their topic suggests. Other dimensions are covered generally.								

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Commercial Essay Tests							
Components	COMP	A. Profile	CAAP	MCAT	TWE	GMAT	SAT-II	CLEP
Organization								
1. Organize the material for more than one audience.			*					
2. Include clear statements of the main ideas.								
3. Demonstrate their method of organization to their audience(s) by using informative headings.								
4. Write informative headings that match their audiences' questions.								
5. Maintain coherence within sentence.								
6. Maintain coherence among sentences, paragraphs, and sections of a piece of writing.			*	*				
7. Develop patterns or organization for their ideas.			*	*				*
8. Use knowledge of potential audience expectations and values to shape a test.								
9. Create and use an organizational plan.				*				
10. Organize their writing in order to emphasize the most important ideas and information within sentences and larger units such as paragraphs.								
11. Cluster similar ideas.				*				
12. Provide a context for the document in the introduction.								
13. Set up signposts such as table of contents, indexes, and side tabs.		*		*				
14. Demonstrate patterns of reasoning in their writing.	*	*	*	*	*	*	*	*
Other dimensions are covered generally.								

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Commercial Essay Tests							
Components	COMP	A. Profile	CAAP	MCAT	TWE	GMAT	SAT-II	CLEP
Drafting								
1. Avoid common grammatical errors of standard written English.		*						
2. Quote accurately.								
3. Establish and maintain a focus.		*						
4. Write effective introductions and conclusions.								
5. Write effectively under pressure and meet deadlines.								
6. Make general and specific revisions while they write their drafts.								
7. Move between reading and revising of their drafts to emphasize key points.								
8. Refine the notion of audience(s) as they write.								
Other dimensions are covered generally.								
Collaborating								
1. Collaborate with others during reading and writing in a given situation.								
Other dimensions are covered generally								
Revising								
1. Correct grammar problems.								
2. Revise to improve word choice.								
3. Select, add, substitute, or delete information for a specified audience.								
4. Reduce awkward phrasing and vague language.								
Other dimensions are covered generally.								
Features of Written Products								
1. Use active or passive voice where appropriate.	*					*		
2. Use language their audience understands.								
3. Define or explain technical terms.								
4. Use concise language.								
5. Use correct grammar, syntax (word order), punctuation, and spelling.	*	*	*	*	*	*	*	*
6. Use correct reference forms.								
7. Use the specific language conventions of their academic discipline or professional area.								
Other dimensions are covered generally.	*			*	*	*	*	*

Table 10—Dimensions of writing reflected in assessment methods—Continued

	Commercial Essay Tests							
Components	COMP	A. Profile	CAAP	MCAT	TWE	GMAT	SAT-II	CLEP
Written Products								
1. Write memoranda.								
2. Write letters.								
3. Write formal reports.								
4. Write summaries of meetings.								
5. Write scripts for speeches/presentations.								
6. Complete pre-printed forms that require written responses.								
7. Write step-by-step instructions.								
8. Write journal articles.								
9. Write policy statements.								
Other dimensions are covered generally.								
Other								
1. Style.								
2. Avoidance of figurative language.								
3. Shifts in construction.								
4. Analyzing rhetoric.								
5. Ambiguity/wordiness.								
6. Insightful support for ideas.		*	*		*	*	*	*
7. Point of view exemplified.								
8. Maintenance of a consistent tone.								
9. Effective opening and closing.								
10. Avoidance of generalizations, cliches.		*						
11. Awareness, insight into complexities of prompt.		*						
12. Separating relevant from irrelevant information.				*				
13. Depth, complexity of thought.				*				
14. Sentence variety.					*	*	*	

Key to Abbreviations:

CLEP—College-Level Examination Program
SAT-II—Scholastic Aptitude Test
AP—Advanced Placement
CAAP—Collegiate Assessment of Academic Proficiency
COMPASS—Computerized Adaptive Placement Assessment and Support System
TASP—Texas Academic Skills Program
CLAST—College-Level Academic Skills Test
SEEW—Scale for Evaluating Expository Writing
IIEP—Illinois Inventory of Educational Progress
NJCBSPT—New Jersey College Basic Skills Placement Test
COMP—College Outcome Measures Program
MCAT—Medical College Admission Test
TWE—Test of Written English
GMAT—Graduate Management Test

3.3 Issues Relevant to Writing Assessment

The Portfolio Approach

In response to the many concerns regarding essay tests, several writing professionals have advocated portfolio assessment as a viable alternative to the timed essay. In portfolio assessment, already constructed documents are used instead of generating new ones. Advocates of the portfolio approach emphasize the use of "real writing" not produced under artificial conditions, the ability to track the development of student abilities over time, congruence with the process model, and the enhanced opportunities to measure writing defined in terms of higher-order thinking. Murphy (1994) notes that portfolios represent curricula products and, as such, they provide a wealth of information regarding experiences in the classroom (both the course content and the manner in which it is communicated). Murphy further points out that because portfolios indirectly reveal a wealth of information pertaining to the philosophical assumptions and beliefs about teaching and learning that frame educational experiences, reflective analysis of portfolio contents can aid both teachers and policymakers seeking to enhance the quality of instruction.

However, White (1993) noted that portfolio assessment gives rise to a host of several issues that were not previously encountered in writing assessment. For instance, decisions must be made regarding (1) what is to be included in the portfolio, (2) who is responsible for collection and verification of materials, (3) what kind of scoring is practically possible, (4) how upper-level assessment can be made fair to students coming from majors requiring varying amounts of writing, (5) whether the original instructor's grades and comments should remain on the submissions, and (6) what the most appropriate methods are to employ for demonstrating reliability and validity.

Shortcomings associated with the portfolio approach as it is commonly implemented are beginning to be identified as well. For example, Witte et al. (1995) have voiced concern that portfolio assessment is often oriented toward the performance of school tasks that may not correlate with workplace and citizenship tasks, rendering portfolio assessments incongruent with the forms of assessment advocated by the National Education Goals Panel through America 2000. Reliability has also been a particularly problematic issue with portfolio assessment. Although holistic scoring is the most frequently applied scoring approach, this method can be potentially problematic in that readers must examine several samples, often written within many different genres and intended for a number of different audiences and purposes with discrepant levels of success, and then must score the whole set of writing samples on a single scale (Callahan 1995). With several different types of writing included in the portfolio, the rubrics must be general enough to capture the essence of good writing across multiple forms; and with less specificity in the rubric anchor points, interpretation becomes more open to judgment and is likely to compromise inter-rater reliability. Callahan (1995) outlined additional problems with the portfolio approach, including competency of readers for evaluating a wide variety of writing forms and the impact of the order of pieces on the reader. The complexity, expense, and labor-intensive nature of portfolios are discussed by Callahan as well.

Finally, it is vital to remain cognizant of the fact that when direct assessment techniques are applied to the measurement of writing skills, they represent true *direct* measures only to the extent that the skills of interest are actually reflected in the written products (Power, Fowles, and Willard 1994). Moreover, as pointed out by Messick (1992) (cited in Powers, Fowles, and Willard (1994)), any measurement of skills or knowledge cannot in actuality be measured, and there is always an inference from performances and products to underlying abilities even when the methods seem to be the most direct or authentic.

Writing Competency

Adherents of a single factor model of writing ability would argue that attempts to delineate skills characteristic of effective writing result in a limited perspective devoid of an appreciation for the synthesis of capacities that emerge during the act of writing. The multifactor approach, on the other hand, is derived from the premise that writing ability is based on the learning and development of discrete skills that can be identified individually. The manner in which one conceptualizes writing ability has implications regarding assessment that will be discussed below.

Holistic Scoring

Proponents of a global definition of writing ability are typically strong proponents of holistic rating scales that are believed to capture the overall essence or quality of writing products. As noted by Breland et al. (1987), the primary assumption underlying holistic scoring is that the whole composition is more than the sum of its parts. According to Cooper (1977), holistic scoring involves matching a written document with a graded series of writing samples, scoring a document for evidence of features central to a particular type of writing, or assigning a letter or number grade. Moreover, according to Cooper, the assessment should transpire quickly and "impressionistically" following training.

Holistic scoring, which yields one general numerical rating of the overall quality of a writing product, possesses the obvious benefit of speed, rendering it more practical than the analytic scoring approach, which requires ratings on several different factors. Efficiency in scoring is an important consideration when assessments are large; yet a critical limitation of the holistic approach is the lack of diagnostic information produced pertaining to individual students' strengths and weaknesses.

Carlson and Camp (1985) have pointed out that despite rigorous efforts devoted to training scorers, there is always some degree of subjective judgment involved in holistic ratings; and these personal judgments may be particularly problematic when the writer and the scorers possess discrepant sets of cultural conventions and expectations. Research has also shown that ratings are affected by the type of writing scored, by various personality dimensions of the writer, and even by personality attributes of the scorer (Carrell 1995). For example, Carrell found that narrative essays tended to be rated more highly than argumentative pieces, the essays of introverts were often rated higher than those of extraverts, and feeling-oriented raters tended to give higher scores than their "thinking-oriented" counterparts. Interestingly, in Carrell's work, there was a lack of significant differences between the scores of raters who were trained versus those who were untrained, raising questions pertaining to the impact and utility of training.

Elbow and Yancey (1994) have suggested that holistic scoring is based on the potentially erroneous assumption that a complex, multi-dimension performance can be reduced to a single quantitative dimension. Although this scoring methodology was developed to preserve and capture the essence of the entire writing sample, it may ironically turn out to be far more reductionistic than the analytic approach, which at least captures the quality of writing on separate dimensions.

When single holistic scores are used, it is critically important for readers to agree on how to score essays that present skill discrepancies, as when the mechanics and ideas developed are good, but the organization is poor (Carlson and Camp 1985). Carlson and Camp raise another potentially problematic situation that can arise in the context of holistic scoring. Specifically, there must be agreement on issues such as how to rate attempts to compose complex sentences that contain errors versus refraining from the use of complex sentences and presenting correct but simple sentences. Compromised reliability is one of the most frequently cited disadvantages of holistic scoring. Unfortunately, the most commonly employed estimate of reliability with holistically scored essays is inter-rater reliability, which actually tends to be an

inflated estimate, suggesting that reliability may be a problem of greater magnitude than it seems at first glance.

The reliability of holistic scales can be enhanced substantially by designing rubrics with scale points that are clearly defined and differentiated with objective criteria, as opposed to using vague descriptors that are open to subjective interpretation. The inclusion of more than one essay requirement and the use of multiple raters should also increase the reliability of holistically scored tests.

Analytic Scoring

Those who view writing as a set of distinct skills rather than as a global generalized ability tend to prefer analytic scoring methods, based on the notion that individual writers may have strengths in some areas and deficiencies in others. In analytic scoring, the traits of good writing are broken down into categories such as organization, development, awareness of the audience, mechanics, and coherence. Within each category the rater makes a judgment regarding how the paper fares on each of the particular dimensions using a numerical scale typically ranging from a high of "5" or "6" to a low of "1." Each subscale is usually accompanied by a rubric containing detailed descriptors of the characteristics of essays meriting a particular score. Scores on the subscales are then typically added to derive a total score.

Due to the fact that analytic scoring yields more scores than holistic scoring, not only is this methodology more useful for assessing various dimensions of individual students' abilities, but it is also potentially more valuable for prescribing educational interventions for individuals. Further, in cases where several students exhibit similar patterns of deficits, assessment can lead to curriculum reform. In a review of holistic versus analytic scoring, Huot (1990) reported that analytic scales tend to have higher reliability estimates than holistic methods.

In terms of disadvantages of analytic scoring, one of the most frequently cited disadvantages pertains to increased time needed for development of the scales and for the actual scoring of essays. Also, opponents of analytic scoring often voice concerns related to missing an assessment of the writing sample as a unified whole, when the components of successful writing are broken down into smaller units. On a slightly different note, Carlson and Camp (1985) remind us that the reader's general impression often influences ratings on separate dimensions, thereby rendering the advantage of useful separate score information potentially less meaningful.

Computerized Writing Assessment

Computer-administered writing assessments are not extremely widespread at this point in time; however, computer-adapted testing is becoming increasingly prevalent. For example, the COMPAS Writing Skills Placement Test developed by ACT is a multiple-choice, objective test of writing skills that requires the student to find and correct errors in essays, without any prompting pertaining to the regions of the essays containing flawed segments. ACT plans to have an essay segment available in the future. Advances are also being made in the development of computerized writing assessment programs that allow for computerized scoring through counting and analysis of targeted numeric indicators in text files. The Computerized Inventory of Developmental Writing Traits (CIDWT), developed by a research team from the Alaska Writing Program headed by McCurry (see McCurry 1992) provides an efficient, inexpensive means for scoring large numbers of essays with reference to fluency, sentence development, word choice, and paragraph development. Computerized scoring of essays is likely to provide a valid addition to the available measures, particularly in view of the fact that scores on the CIDWT have been found to correlate highly with teacher ratings. However, it is unlikely that computerized scoring will be

able to assess all of the essential components of effective writing. The rating of qualities such as organization, tone of voice, originality of ideas, etc. are not readily conducive to computerized scoring.

Takayosh 1996 pointed out that several scholars have identified changes in the actual processes of writing (invention, drafting, and revision) resulting from the extensive use of computers to compose text. More specifically, she notes how many contend that the fluid and recursive nature of writing is becoming more visible with the generation of electronic text, and the writing process is becoming best conceptualized as a "seamless flow." Moreover, with the stages of the writing process becoming less well defined, Takayosh foresees the need for assessment strategies to reflect this transformation.

Overriding General Issues

Individuals involved in assessment of higher education outcomes, such as writing competency, need to begin the process with a well-formulated definition of writing. Such a definition should not only be formulated within a process framework, but it should also include sensitivity to both the specific skills that are easily defined (e.g., use of appropriate grammar) as well as the more complex or higher order skills (e.g., developing an argument) that may require careful thought and research to delineate precisely. The definition opted for should likewise be consistent with the skills developed in the curriculum to ensure that the selection or design of measures is closely integrated with the objectives and standards of the educational experiences that students encounter. Once an operational definition is developed, assessment personnel should examine the specific purpose of the assessment (how the outcome data will be used, what inferences will be made from the data generated, and what changes are likely to result), in addition to considering the conceptual and methodological criteria outlined above, to select an appropriate existing measure or to help guide the development of a new assessment strategy.

When the advantages and disadvantages of direct vs. indirect measures are carefully analyzed, most professionals arrive at the conclusion that for a complete description of writing ability, a combination of the two forms provides the most thorough, methodologically sound, and reasonable solution (Miller and Crocker 1990; Swanson, Norman, and Linn 1995). To entirely replace selected response measures with essay-type tests or portfolios could be detrimental to writing assessment. As Breland (1996) noted, the decontextualized skills measured with multiple-choice type tests represent skills that are perhaps more readily taught than teaching students how to generate high-quality text. Moreover, skills such as learning to recognize problematic elements in writing are important to many life- and job-related tasks. The combination of selected and constructed response items enables coverage of both the drafting and revision stages of the writing process. Breland has further pointed out that as we increasingly include free-response writing in our assessment efforts, research should be devoted to identifying the effects of assessment changes on the actual development of students' writing abilities. At this point in time data are not available to demonstrate that the new assessment strategies result in the improvement of students' writing abilities.

3.4 Writing Templates

Over the last three decades a number of process-oriented theoretical models have been generated by various writing experts. In 1964, Rohman and Wlecke proposed a model of writing that entailed conceptualization of the writing process as a linear sequence of activities, each of which could be analyzed at a given point in time. Rohman and Wlecke further discussed division of the process into a prewriting stage, which occurs prior to the actual construction of a document, and a writing phase, which also incorporates rewriting activities. Rohman and Wlecke emphasized a distinction between thinking and

writing, yet focused on the importance of stimulating, spontaneous, and original thinking as a prerequisite to high-quality, expressive writing.

Several theorists subsequently adopted a slightly different approach, continuing to adhere to the idea of writing as a process, but preferring a more dynamic, less sequential conceptualization. Research conducted by Emig (1971), Faigley et al. (1985), and Sommers (1980) revealed not only that the composing process did not necessarily follow a linear path as previously believed, but also that revision strategies employed by experienced writers differed qualitatively from those of college freshmen. Zemelman (1977), whose ideas about writing clearly diverge from the earlier, linear approach, defined writing as "a complex process combining many mental activities, each depending on and influencing others: enumerating, categorizing, developing terms, gaining a sense of active participation in a subject, sensing and analyzing one's reactions to a situation, abstracting, seeing new connections and underlying patterns, developing arguments, [and] developing hierarchies of significance" (p. 228).

One of the most prominent models of the writing process to develop out of this second wave of theoretical work was one originally proposed by Flower and Hayes (1981) and updated by Hayes (1996). The emphasis in their framework is on the writer's inner, cognitive processing, with "planning," "translating," and "reviewing" constituting the major classes of mental events that engage the writer. Flower and Hayes also delineated several subprocesses corresponding to each major process, and they contend that the writer monitors his or her movement through different parts of the process based on individualized goals, writing habits, and writing style. By incorporating the work of developmental psychologists such as Piaget and Vygotsky, Britton (1975) arrived at the conclusion that language is not a passive means for transcribing knowledge, but is instead inextricably intertwined with thinking and learning.

A third line of theoretical work was initiated by Bizzell (1982), among others, who felt that although the model offered by Flower and Hayes provided very useful information pertaining to how writers compose, the model neglected the social element of writing. Bizzell described the social context of writing as involving more than just a connection to the audience, incorporating the expectations of the community with which the writer is affiliated as well. Similarly, Faigley et al. (1985) have suggested that an attempt to understand fully the writing process requires that we "look beyond who is writing to whom [and look instead] to the texts and social systems that stand in relation to the act of writing" (p. 539).

TEMPLATES – WRITING COMMERCIALLY DEVELOPED TESTS

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
CLEP General Exam in English Composition persuasive essay and multiple-choice items **Author** Test Development Committee: Paul Tucci (chair), Richard Bellairs, Rosentene Purnell, and Susan Schiller **Publisher** The College Board P.O. Box 6601 Princeton, NJ 08541–6601 **Date** 1993 **Testing Time** 45 minutes **Cost** $43 *For use with all university students*	Award college credit for exemption from gen. ed. requirements in English composition Score of 420–500 on full exam for awarding college credits (American Council on Education)	Total score based equally on essay and multiple-choice items (200–800 points) Centralized scoring by English faculty throughout the U.S.; training involves reading hundreds of essays, finding exemplars of each point on the scale so that scoring standards are set Focus on postwriting, although a polished product is not expected with the time constraint 2 raters per essay, third rater used when scores are discrepant by more than 2 points **Holistic scoring rubrics** 0 Off topic/blank pages 2–4 Falls short of basic requirements 5 Basic command of English grammar, adequate sentence structure, word choice, organization, and logically presented ideas w/ examples 6–8 Surpasses basic requirements, strong dev. of argument Additional standards for each topic are developed No prescribed analytic guidelines	Based on low reliabilities of essays, important decisions should not be made based on the essay component alone	Used for fulfillment of gen. ed. requirements in English comp. at many universities (authors) No differences across adult age groups for total score (earlier version) (Clark 1988) CLEP Eng. comp. passing rate of 41% for GED recipients vs. 52% for all other students (Turner 1993) Minimal instructional utility, information pertaining to specific competencies and deficits not provided	English grades (earlier version) .47 (Kelly 1973) GED writing skills test/CLEP English comp. .70 No better predictability based on age, gender or last grade completed (Turner 1993)

Name	Scoring	Definition	Reliability	Validity	Correlation with other measures
CLEP General Exam in English Composition multiple-choice items: 2 sections **Author** Test Development Committee: Paul Tucci (chair), Richard Bellairs, Rosentene Purnell, and Susan Schiller **Publisher** The College Board P.O. Box 6601 Princeton, NJ 08541–6601 **Date** 1993 **Testing Time** 45 minutes per section **Cost** $43 *For use with all university students*	Total score (200–800)	<u>Skills at sentence level</u> Sentence boundaries Economy/clarity of expression Agreement: subject/verb, verb tense, pronoun reference, shift, number Active/passive voice Diction and idiom Syntax: parallelism, coordination, dangling modifiers Sentence variety <u>Skills in context</u> Main idea, thesis Organization of ideas Relevance of evidence, sufficiency of detail, levels of specificity Audience and purpose (effect on style, tone, language, or argument) Logic of argument (inductive, deductive reasoning) Coherence within/between paragraphs Rhetorical emphasis Sustaining tense or point of view	.91, .92 alternate forms reliability for forms 1 and 2 respectively .92 internal consistency of both forms 30.40, 30.08 scaled SEM for forms 1and 2	Used for fulfillment of gen. ed. requirements in English comp. at many universities (authors)	

Name	Purpose	Scoring	Definition	Reliability	Validity	Correlation with other measures
SAT II—Writing Test (essay component—33%) timed impromptu essay knowledge of specific content required **Author** **Publisher** The College Board P.O. Box 6200 Princeton, NJ 08541–6200 **Date** 1994 **Testing Time** 20 minutes **Cost** $23 *Essays can be used for instructional purposes* *For use with all university students*	College entrance exam, first-year placement, and/or exemption from first-year composition courses Designed to assess ability to express ideas clearly and effectively with sensitivity to language and meaning Assesses knowledge gained both in and outside of the secondary general English curriculum Developed to replace TSWE and ECT tests	Centralized holistic (1–6) (incorporates sensitivity to organization, word choice, sentence structure, and punctuation) Two experienced high school and/or college teachers score each essay on a 6-point scale Discrepancies of 3 or more points are resolved with a third scorer	*see next page*	Coefficients obtained with National Test Population .58 for essay component (.87 internal consistency for total test)	Based on total scores—Essay + M.C. Concurrent—Correlation between SAT II Writing and high school GPA = .4 Construct—Students with relevant coursework (composition, grammar, speak/listen, American lit., British lit., historical lit., and other lit.) achieved higher total scores on the SAT II than students without such experience Predictive—Correlation with college English grades: 4-yr. schools sampled in the southern, southwestern, middle, and western U.S.—coefficients ranged from .23–.50 2-yr. schools sampled in the middle and western U.S.—.32–.47 U.S.—coefficients ranged from .32–.47 (Bridgeman and Bonner 1994)	SAT II writing essay w/ AP lang. and lit. essays .4 (observed) .7 (corrected for attenuation) SAT II writing tools w/ SAT—V .72 (observed); .85 (corrected for attenuation) TSWE .79 (observed); .91 (corrected for attenuation) ECT total .86 (observed); .99 (corrected for attenuation) ECT (essay) =.58 ECT multiple-choice .85 (observed) .96 (corrected for attenuation) AP lang. total = .7 (observed); .8 (corrected for attenuation) AP lang. total = .7 (observed); .8 (corrected for attenuation)

SAT II Writing Test (essay component)

Scale Definition/Rubric/Specificity of Anchor Points

6—Demonstrates clear and consistent competence though it may have occasional errors. Such a paper does the following:

- efficiently and insightfully addresses the writing task;
- is well organized and fully developed, using clearly appropriate examples to support ideas; and
- displays consistent facility in the use of language, demonstrating variety in sentence structure and range of vocabulary.

5—Demonstrates reasonably consistent competence though it will have occasional errors or lapses in quality. Such a paper does the following:

- effectively addresses the writing task;
- is generally well organized and adequately developed, using appropriate examples to support ideas; and
- displays facility in the use of language, demonstrates some syntactic variety and range of vocabulary.

4—Demonstrates adequate competence with occasional errors and lapses in quality. Such a paper does the following:

- addresses the writing task;
- is organized and somewhat developed, using examples to support ideas;
- displays adequate but inconsistent facility in the use of language, presenting some errors in grammar or diction; and
- presents minimal sentence variety.

3—Demonstrates developing competence. Such a paper may contain one or more of the following weaknesses:

- inadequate organization or development;
- inappropriate or insufficient details to support ideas; and
- an accumulation of errors in grammar, diction, or sentence structure.

2—Demonstrates some incompetence. Such a paper is flawed by one or more of the following weaknesses:

- poor organization;
- thin development;
- little or inappropriate detail to support ideas; and
- frequent errors in grammar, diction, and sentence completion.

1—Demonstrates incompetence. Such a paper is seriously flawed by one or more of the following weaknesses:

- very poor organization;
- very thin development; and
- usage and syntactical errors so severe that meaning is obscured.

Note: Many of the descriptors used in this scoring guide are subject to readers' personal interpretations (e.g., "competence," "effectively," and "development"); and distinctions between some components of the different anchor points are not well defined (e.g., is there a difference between "inappropriate or insufficient details to support ideas" associated with a score of "3" vs. "little or inappropriate detail to support ideas" associated with a score of "2"?)

Name	Scoring	Definition	Reliability	Validity	Correlation with other measures
SAT II—Writing Test (multiple-choice component—66%) **Author** **Publisher** The College Board P.O. Box 6200 Princeton, NJ 08541–6200 **Date** 1994 **Testing Time** 40 minutes **Cost** *For use with all university students*	Total score (200–800) Item-type subscores (identifying sentence errors, improving sentences, improving paragraphs)	The test covers a number of writing problems including the following: Being consistent: sequence of tenses shift of pronoun parallelism noun agreement pronoun reference subject/verb agreement Expressing ideas logically: coordination and subordination logical comparison modification and word order Being clear and precise: ambiguous and vague pronouns diction wordiness improper modification Following conventions: pronoun case idiom comparison of modifiers sentence fragment double negative	Internal consistency .89	(Refer to information under essay component for total scores)	SAT II—Writing, multiple-choice with AP lang. multiple-choice .7 (observed); .8 (corrected for attenuation) AP lit. multiple-choice .7 (observed); .8 (corrected for attenuation)

Name	Purpose	Scoring	Definition	Reliability	Validity	Correlation with other measures
Advanced Placement (AP) English language and composition (essay component, 55%) **Author(s)** Development committee—college and high school faculty from around the U.S. **Publisher** The College Board 45 Columbus Avenue New York, NY 10023–6992 **Date** Revised annually **Testing Time** 2 hours (typically 3 questions) **Cost** $74 *For use with all university students*	College placement, credit, and exemption Allows personnel to make decisions regarding students' competencies and placement and may facilitate evaluation of instructional emphases	Centralized Holistic Scorers are encouraged to judge overall quality and avoid dividing the essay into content and style Prior to scoring, faculty consultants receive intensive training using many student samples	*see next page*	Reader reliability coefficients (essay) =.62–.82 Composite-score reliability (essay + multiple-choice) .80–.88 SEM for composite scores 6.1–7.8	Correspondence between AP grades (composite scores) and college course grades: AP exam performance by AP candidates receiving an AP score of 3 was > than that of college students receiving a course grade of B and only slightly below the performance of college students receiving a course grade of A AP candidates with scores of 4 or 5 received AP scores > than those earned by students receiving a course grade of A (Modu and Wimmers 1981) AP students' at U of MD received significantly higher grades in courses beyond the intro level than their non-AP counterparts Content validity—Annual exams are developed over 2 years by a development committee (college and high school faculty) Each question is repeatedly reviewed for accuracy and clarity of language. The full exam is evaluated to ensure breadth of content and skills required in a comparable college course.	Correlation between AP examination grades (composite) and college English instructor readings .46

Advanced Placement (AP) English Language and Composition

(essay)

Scale Definition/Rubric/Specificity of Anchor Points

General instructions: Scores assigned should reflect the quality of the essay as a whole. Reward the writers for what they do well. The score for a particularly well-written essay may be raised by one point from the score otherwise appropriate. In no case may a poorly written essay be scored higher than a 3.

Score of 7–9—Demonstrates an understanding of argumentation by acknowledging both sides of the argument and by making a cohesive, well-supported case for the chosen side. Aptly supports what is said, and demonstrates stylistic maturity by an effective command of sentence structure, diction, and organization. Reveals ability to choose from and control a wide range of the elements of composition to present ideas clearly.

Score of 4–6—Discusses some of the issues raised by the question although with less detail or supporting examples than the best papers. May concentrate on one side of the argument and dismiss the other with little or no attention. Essays that use the question as a starting point for a generalized essay may score no higher than a 4. Arguments are sound, but may be presented with less maturity than the top papers. Some lapses in diction or syntax may be evident, but writing demonstrates sufficient control of the elements of composition to present ideas clearly.

Score of 1–3—Likely to have one or more of these flaws: a restatement or summary of the passage with little argument; an argument that consists almost entirely of asserting without specific or persuasive supporting evidence; excessive attention to the deleted articles or the principal actions; and/or imprecise or incomplete treatment of the constitutional issues. Although sufficient to convey the writer's ideas, writing may suggest weak control over diction, syntax, or organization. May contain consistent spelling errors or some flaws of grammar or other basic elements of composition.

Name	Purpose	Definition	Scoring	Reliability	Validity	Correlation with other measures
Advanced Placement (AP) English language and composition (multiple-choice component—45%) **Author(s)/Publisher** The College Board 45 Columbus Avenue New York, NY 10023–6992 **Date** Revised annually **Testing Time** 1 hour **Cost** $74	For college-level credit by exam	Tests the student's skills in analyzing rhetoric of prose passages	Total scores 1–5	Internal consistency (KR-20) .84	Correlation between multiple-choice and essay components .47	*see total scale information provided under rating scale section*

Name	Purpose	Scale Definition	Scoring	Reliability	Validity	Correlation with other Measures
Collegiate Assessment of Academic Proficiency (CAAP) essay component (there is also a 72-item multiple-choice segment that assesses punctuation, grammar, usage, sentence structure, strategy, organization, and style) **Author/Publisher** American College Testing Program Iowa City, Iowa 52243 **Date** **Testing Time** Two 20-min. essays **Cost** $8.80/student per objective test ($13.90 for more than one) Essay: $2.60 local scoring w/purchase of an objective test. $4.15 for local scoring $8.80 for use of ACT scoring $13.90 for writing package (objective and essay tests) *Used by colleges and universities throughout the U.S.*	To measure writing skills that are considered foundational for performance in upper-level college courses Student required to read a passage, and then given a specific context, to write an essay that argues a particular point Required knowledge is consonant with the training and experience of college sophomores Level of proficiency—curriculum based	The design of the essay test is based on the assumption that the skills most frequently taught in college-level writing courses and required in upper-level courses across the curriculum include: Formulating an assertion about an issue Supporting that assertion with evidence Organizing and connecting major ideas Communicating using good writing skills (mechanics, sentence structure, and command of the language) *rubric on next page*	Centralized (or local if preferred) Holistic	Internal consistency Sophomores Form 88 A .95 Form 88 B .93 Freshmen Form 88 A .93 Form 88 B .93 (for multiple-choice) SEM Sophomores Form 88 A 3.44 Form 88 B 3.47 Freshmen Form 88 A 3.65 Form 88 B 3.47	Content validity established through the use of experts during the development and refinement of the measure Black examinees did not perform as well as white examinees on the essay test Differences between the two groups were of similar magnitude to differences found on the multiple-choice component (Welch 1989) Evidence for the validity of the CAAP as a measure of educational change: entering freshmen pre-tested and then post-tested after their sophomore year at Lehigh County Community College—resulting median difference score of .9	All for multiple-choice Median (across institutions) correlation between writing skills and sophomore English GPA .37, with a range from .26 to .57 Writing skills and sophomore cumulative GPA .36 Writing skills and junior year English grades .25 Enrollment in courses in foreign languages, music, philosophy, sociology, and communications associated with improvement between administrations of the CAAP Essay (Jones and Nugent 1996)

CAAP Scoring Guide

Upper-range papers—Engages the issue identified in the prompt and demonstrates superior skill in organizing, developing, and conveying in standard, written English the author's ideas about the topic.

6 Exceptional—Take a position on the issue defined in the prompt and support that position with extensive elaboration. Organization is unified and coherent. While there may be a few errors in mechanics, usage, or sentence structure, outstanding command of the language is apparent.

5 Superior—Take a position on the issue defined in the prompt and support that position with moderate elaboration. Organization is unified and coherent. While there may be a few errors in mechanics, usage, or sentence structure, command of the language is apparent.

Mid-range papers—Demonstrates engagement with the issue identified in the prompt but does not demonstrate the evidence of writing that would mark it outstanding.

4 Competent—Take a position on the issue defined in the prompt and support that position with some elaboration or explanation. Organization is generally clear. A competency with language is apparent, even though there may be some errors in mechanics, usage, or sentence structure.

3 Adequate—Take a position on the issue defined in the prompt and support that position, but with only a little elaboration or explanation. Organization is clear enough to follow without difficulty. A control of the language is apparent, even though there may be numerous errors in mechanics, usage, or sentence structure.

Lower-range papers—Fails in some way to demonstrate proficiency in language use, clarity of organization, or engagement of the issue identified in the prompt.

2 Weak—While these papers take a position on the issue defined in the prompt, they may show significant problems in one or more of several areas, making the writer's ideas often difficult to follow. Support may be extremely minimal; organization may lack clear movement or connectedness; or there may be a pattern of errors in mechanics, usage, or sentence structure that significantly interferes with understanding the writer's ideas.

1 Inadequate—These papers show a failed attempt to engage the issue defined in the prompt, lack support, or have problems with organization or language so severe as to make the writer's ideas very difficult to follow.

Name	Purpose	Scale Definition	Scoring	Reliability	Validity
The Academic Profile optional, content-related essay (there is also a multiple-choice writing section) **Author/Publisher** ETS Princeton, NJ 08541–0001 **Date** 1989 **Testing Time** 45 minutes **Cost** $300 annual institution fee and per test fees based on the number ordered (e.g., 500 exam booklets $15 and essay = $1.50) *Used by colleges and universities throughout the U.S.*	Designed to assist institutions with their general education outcome assessment Essay requires students to apply concepts to material read or studied in related coursework Focuses on generating an analytic essay integrating appropriate examples from coursework Can help in assessing student growth/change through the use of pre-/postassessments Can be used as performance standard for upper-level courses	Multiple-choice segment assesses students' ability to: Recognize the most grammatically correct revision of a clause, sentence, or sentences Organize units of language for coherence and rhetorical effect Recognize and reword figurative language Organize elements of writing into larger units of meaning *rubric on next page*	Essay total scores 1–4 On multiple-choice total scores range from 100–130 (36 items) Local scoring guide, holistic Proficiency levels achieved on the full exam (essay and multiple-choice) are assigned in addition to numerical reports Level 1—Basic understanding of appropriate writing Level 2—Intermediate level; can recognize and use the elements of good writing Level 3—Can make fine distinctions and solve complicated and subtle writing problems, characteristic of mature writing	Using IRT-based procedures—for multiple-choice segment, reliability .76 SEM 2.54	Content validity established during development with the aid of a committee of college and university faculty members Construct validity—Extensive testing by ETS has shown that as examinees' GPAs increased, the percentage of the core curriculum completed increased. Academic Profile scores also increased (Marr 1995) Writing scores (multiple-choice) and percentage core completed—Spearman rank .19 MANOVA procedure indicated sig. differences between Academic Profile scores among students in different GPA groups Range of GPA 1.0–4.0 Range of writing score means 114.7–120.56

Academic Profile Essay Scoring Guide

The 4 paper:

1. Demonstrates the ability to use the discourse and analysis appropriate to the academic discipline.
2. Displays a clear understanding of the quotation and the task presented in the topic.
3. Sustains a focused discussion.
4. Uses evidence to support a point (e.g., uses consistently well-developed, well-chosen examples).
5. Demonstrates an awareness of or insight into the complexities implied in the quotation.
6. Avoids an awareness of or insight into the complexities implied in the quotation.
7. Avoids sweeping generalizations, cliches, and unsupported assertions.
8. Displays a level of writing skill that supports and enhances the discussion.

The 3 paper:

1. Demonstrates the ability to use the discourse and analysis appropriate to the academic discipline.
2. Displays a clear understanding of the quotation and the task presented in the topic.
3. Sustains a focused discussion.
4. Uses evidence to support a point (e.g., uses a single well-developed example or presents several pertinent, though not thoroughly developed, examples).
5. Displays a level of writing skill that does not interfere with the conveying of information.

The 2 paper:

1. Demonstrates an understanding of the quotation but fails to address the task in one or more of the following ways:
   • depends on poorly selected or inaccurate examples from coursework;
   • fails to develop examples adequately;
   • merely lists (phrases, theories, authors, concepts);
   • provides abstractions and generalizations related to the discipline or topic, but fails to develop, explain, or effectively incorporate them into the essay; or
   • addresses only one part of the task.
2. Provides well-developed examples but does not relate them to the topic.

The 1 paper:

1. Fails to address the task presented in the topic in one or more of the following ways:
   • fails to demonstrate understanding of the quotation and/or the task presented by the topic;
   • is so incoherent that the paper cannot be followed; or
   • depends on feelings, beliefs, or cliches to develop the essay rather than the knowledge of relevant coursework.
2. Displays writing deficiencies so severe that the essay does not convey information

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
College Outcome Measures Program (COMP) writing skills assessment **Authors/Publisher** ACT P.O. Box 168 Iowa City, Iowa 52243 **Date** 1976 **Testing Time** 80 minutes three 20-minute writing assessments based on audiotaped stimulus materials of 3–4 min. duration *For use with all university students*	To measure knowledge and skills acquired as a result of general education programs and that are important to effective adult functioning Assists in program evaluation, not developed for making judgments about individual students The emphasis is on practical application rather than on academic focus; students write a personal letter to a U.S. senator and a letter to a radio station The content areas of social science, technology, and fine arts are covered in the three essays	Total COMP score and 3 subscores: Audience Organization Language Scoring is local or centralized Norm-referenced and criterion-referenced interpretation available A holistic and analytic evaluation is used Postwriting draft is evaluated	Average inter-rater agreement total scores .94 Audience .93 Organization .83 Language .79 Parallel forms total scores .69–.75 Audience .51–.68 Organization .53–.67 Language .62–.81 Cronbach's alpha, freshmen, and seniors respectively Total scores .77, .79 Audience .53, .53 Organization .65, .62 Language .81, .83 Generalizability coefficients total scores (holistic) .76–.84 Audience .48–.79 Organization .74–.86 Language .83–.91 Total analytic .82–.90	COMP writing scores were sensitive to difference expected to occur over 4 years of college Freshmen mean 17.2 Senior mean 19.8 47% of freshmen and 59% of seniors from six institutions passed an arbitrary criterion of middle-level proficiency No meaningful differences in senior COMP writing scores based on age or major Freshmen and senior women scored significantly higher than men on the COMP writing	COMP total score and ACT: Freshmen .50 Senior .42 COMP total score with senior GPA .35

COMP Scoring Guidelines

Audience

Level A—Uses a writing form appropriate to the situation, clearly addresses the intended audience, and consistently attends to the perspective of the audience.
Level B—Uses a writing form appropriate to the situation, addresses the intended audience, and shows some attention to the probable perspective of that audience.
Level C—Uses a writing form appropriate to the situation, yet is so involved in the message that little positive contact is made with the intended audience.
Level D—May not have used an appropriate letter form or generally ignores the audience due to involvement with the content; may lose (talk about rather than address) the specified audience in the body of the letter.
Level E—Does not address the intended audience; may have written an essay to no one in particular.

Organization

Level A—Writes an essay that develops all three points called for in detail in a direct fashion with tight control of language and transition, and more than one level of abstraction (examples and details).
Level B—Writes an essay that treats each of the points called for, developing at least two in detail, with attention to language and transition, and more than one level of abstraction.
Level C—Writes an essay that at least touches upon all three points called for, although development is uneven, with some attention to transition, but few examples and details.
Level D—Writes an essay that elaborates on one point and ignores one or both of the others, and may be somewhat loose or unorganized.
Level E—Writes an essay that has no apparent organization or makes one or more assertions with no elaboration or development of points.

Language

Level A—Writes in a precise or in a lively manner, with originality and sustained effort to use interesting or clever phrases, and few scribal errors.
Level B—Writes in a clear manner that shows some energy and effort at originality with some interesting word choices, and few scribal errors.
Level C—Message is generally clear, although tends to use the wording of the points listed, with some scribal errors that mildly distract from or obscure the message.
Level D—Writes in generalities, tending to repetitious or awkward phrases, with a distracting number of scribal errors.
Level E—Writes in an illiterate manner (incomplete sentences, errors in tense, number or person, etc., with trite or clumsy phrases and many distracting scribal errors).

Name	Purpose	Scoring	Definition	Reliability	Validity	Correlation with other measures
COMPASS writing skills placement test computerized adaptive testing system (an essay segment is planned) **Author/Publishers** ACT 2201 North Dodge P.O. Box 168 Iowa City, Iowa 52243–0168 **Cost** Annual license fee $500 Prices for placement test, diagnostic tests, and creation of student record with background, needs, and goal information vary based on the number of total units purchased and the diagnostic assessment system *For use with all university students*	course placement	Diagnostic scores available in 8 areas Local scoring Writing diagnostic scores: Punctuation Spelling Capitalization Usage Verb formation/ agreement Relationship of clauses Shifts in construction Organization Each domain consists of 42 items that are adaptively selected	Requires students to find and correct errors in essays Global multiple-choice items related to the passages follow revision exercise			

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
College Basic Academic Subjects Examination (College BASE) essay **Author** Steven Osterlind, Director, Center for Educational Assessment, University of Missouri, Columbia, MO **Publisher** The Riverside Publishing Co. **Date** 1989–90 **Testing Time** Available in 3 forms: Long 4 hrs. Short 2 hrs. Institutional matrix 50 min. **Cost** Long $17.10 Short $14.85 Institutional matrix: $6.30 (prices are per student and include scoring) *For use with all university students*	To assess competencies usually achieved through a general education curriculum Typically administered at the end of the sophomore year, but users are encouraged to test at different times to assess change resulting from college experiences Useful for diagnosing strengths and weaknesses of individual students and curricula, not designed for student selection into particular programs	Centralized 40 scores 1 ea. of subjects (English, math, science, social studies) 9 clusters (one is writing) 23 subskills including: expository writing sample (*see rubric*), conventions of written English, and writing as a process 3 competencies including: interpretive reasoning, strategic reasoning, and adaptive reasoning Cluster scores range from 400–560	Internal consistency (KR-20)/reliability estimate based on IRT using average standard error Writing as a process .32/.33 Conventions of written English .56/.56 Writing cluster .59 English .89	Factor analytic studies with over 2,000 examinees showed factor composites were consistent with the intended structure Extensive statistical screening of items for ethnic heritage, cultural, gender, and regional bias	English scores and ACT .61 SAT—V .46 SAT—Q .35 GPA .43 (manual)

TEMPLATES – WRITING LOCALLY DEVELOPED TESTS

Name	Purpose	Definition	Scoring	Validity	Correlation with Other Measures
Praxis I: Academic Skills Assessment Pre-Professional Skills Test (PPST)—Writing Academic Skills Assessment—Writing (CBT) (content is similar, only the form of administration differs between the two tests) essay components—50% (each assessment also has an error recognition multiple-choice component) **Publisher** ETS CN-6057 Princeton, NJ 08541–6057 **Testing Time** PPST 30 min./full test 60 minutes CBT 40 min./full test 66 minutes *Currently used by school districts, colleges, state agencies, and licensing boards*	For use in selection, admissions, evaluation, and certification. Does not require specialized knowledge	General characteristics: State or imply the writer's position or thesis Develop and organize ideas logically and make clear connection between them Support ideas with well-chosen reasons, examples, and/or details Demonstrate effective sentence variety Display facility in the use of language Demonstrate writing generally free from errors in grammar, usage, and mechanics	Total (range: 150–190) Centralized by experienced college professors Holistic, based on the assumption that the elements evaluated are not independent	Content validity for writing test—96% of the items (including the essay) considered relevant by an expert panel of judges at Brigham Young University (Sudweeks 1991) No significant gender differences on the writing component (Daly 1987)	PPST writing and COMP total scores .49 (Sibert 1989)

Pre-Professional Skills Test

6—Demonstrates a high degree of competence in response to the assessment but may have a few minor errors. An essay in this category is well organized and coherently developed; clearly explains or illustrates key ideas; demonstrates syntactic variety; clearly displays facility in the use of language; and is generally free from errors in mechanics, usage, and sentence structure.

5—Demonstrates clear competence in response to the assignment but may have minor errors. An essay in this category is generally well organized and coherently developed; explains or illustrates key ideas; demonstrates some syntactic variety, displays facility in the use of language; and is generally free from errors in mechanics, usage, and sentence structure.

4—Demonstrates competence in response to the assignment. An essay in this category is adequately organized and developed; explains or illustrates some of the key ideas; demonstrates adequate facility in the use of language; and may display some errors in mechanics, usage, or sentence structure, but not a consistent pattern of such errors.

3—Demonstrates some degree of competence in response to the assignment but is obviously flawed. An essay in this category reveals one or more of the following weaknesses: inadequate organization or development; inadequate explanation or illustration of key ideas; a pattern of accumulation of errors in mechanics, usage, or sentence structure; and limited or inappropriate word choice.

2—Demonstrates only limited competence and is seriously flawed. An essay in this category reveals one or more of the following weaknesses: weak organization or very little development, little or no relevant detail, and serious errors in mechanics, usage, sentence structure, or word choice.

1—Demonstrates fundamental deficiencies in writing skills. An essay in this category contains serious and persistent writing errors, or is incoherent, or is underdeveloped.

Name	Purpose	Definition	Scoring	Validity
Graduate Management Admissions Test (GMAT) Analytical Writing **Author** **Publisher** ETS P.O. Box 6106 Princeton, NJ 08541–6106 **Date** **Testing Time** 1 hour (two 30 min. sections) **Cost** $125 *Currently used by graduate management programs throughout the U.S.*	Selection of applicants for graduate study in management and for financial aid based on academic potential Analysis of an issue Analysis of an argument Differentiates applicants based on academic promise (technically not an achievement test)	See next page	Total (200–800) Mathematical (0–60) Verbal (0–60) Analytical writing skills (0–6) Centralized Holistic	Based on data generated from over 35,000 examinees Within white, African-American, and Hispanic/Latino groups, women scored significantly > than men on analytical writing assessment In the Asian American group, men scored > on the analytical (Bridgeman and Frederick 1996)

GMAT- Analysis of an Issue

6 Outstanding—Presents a cogent, well-articulated analysis of the complexities of the issue and demonstrates mastery of the elements of effective writing. A typical paper in this category does the following:

- explores ideas and develops a position on the issue with insightful reasons and/or persuasive examples;
- is clearly well organized;
- demonstrates superior control of language, including diction and syntactic variety; and
- demonstrates superior facility with the conventions (grammar, usage, and mechanics) of standard written English but may have minor flaws.

5 Strong—Presents a well-developed analysis of the complexities of the issue and demonstrates a strong control of the elements of effective writing. A typical paper in this category does the following:

- develops a position on the issue with well-chosen reasons and/or examples;
- is generally well organized;
- demonstrates clear control of the language, including diction and syntactic variety; and
- demonstrates facility with the conventions of standard written English but may have minor flaws.

4 Adequate—Presents a competent analysis of the issue and demonstrates adequate control of the elements of effective writing. A typical paper in this category does the following:

- develops a position on the issue with relevant reason and/or examples;
- is adequately organized;
- demonstrates adequate control of language, including diction and syntax, but may lack syntactic variety; and
- displays control of the conventions of standard written English but may have some flaws.

3 Limited—Some competence in analysis of the issue and in control of the elements of writing, but is clearly flawed. A typical paper in this category has one or more of the following characteristics:

- is vague or limited in developing a position;
- is poorly organized;
- is weak in the use of relevant reasons or examples;
- uses language imprecisely and/or lacks sentence variety; and
- contains occasional major errors or frequent minor errors in grammar, usage, and mechanics.

2 Seriously flawed—Demonstrates serious weaknesses in analytical writing skills. A typical paper in this category has one or more of the following:

- is unclear or seriously limited in presenting or developing a position on the issue;
- is disorganized;
- provides few, if any, relevant reasons or examples;
- has serious and frequent problems in the use of language and sentence structure; and
- contains numerous errors in grammar, usage, or mechanics that interfere with meaning.

1 Fundamentally deficient—Demonstrates fundamental deficiencies in analytical writing skills. A typical paper in this category has one or more of the following characteristics:

- provides little evidence of the ability to organize a coherent response to the topic;
- has severe and persistent errors in language and sentence structure; and
- contains a pervasive pattern of errors in grammar, usage, and mechanics that severely interfere with meaning.

0 Any paper that is totally illegible or obviously not written on the assigned topic.

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
Test of Written English (TWE) narrative, expository, and persuasive writing put in the form of letters, reports, scripts, etc. Administered with the TWE as a Foreign Language (TOEFL) **Author/Publisher** ETS Princeton, NJ **Date** 1986 **Testing Time** .5 hour **Cost** No separate fee beyond $55 cost of the TOEFL *For use with all U.S. and Canadian university students*	Allows examinees whose native language is not English to demonstrate the ability to express ideas in acceptable written English TWE aids in the evaluation of the academic proficiency of ESL and EFL students TWE is not designed to predict academic performance or to assess scholastic aptitude, motivation, language-learning aptitude, specific knowledge, or cultural adaptability	A total TWE score is obtained by averaging two ratings of a first draft; if the ratings differ by two or more points, a third rater is requested TWE score appears separate from the TOEFL score on the report Readers are primarily English and English-as- a-second-language (ESL) writing specialists affiliated with accredited colleges, universities, and secondary schools in the U.S. and Canada Readers use a holistic approach by considering the organization, examples, and conventions of standard written English used Only scores are provided to the institution, which makes assessing individual strengths and weaknesses difficult	Internal consistency with coefficient alpha: first six administrations .85–.88 Score discrepancy rates: first six administrations .02–.05	Content—Employs writing tasks that are comparable to those required of North American colleges and universities (Bridgeman and Carlson 1983) Construct—Of examinees whose TOEFL scores were above 600, 92.25% scored 4.0 or above on the TWE Of those with scores below 400, 97.44% obtained TWE scores below 4.0	Compare/contrast topic type scores (requires examinee to describe pros and cons of each side of an argument and take a position) and TOEFL total scores .65 Chart/graph topic type scores (requires description and interpretation) and TOEFL total scores .65

Test of Written English (TWE) Scoring Guide

Score of 6—Demonstrates clear competence in writing at both the rhetorical and syntactic levels, though it may have occasional errors. A paper in this category:
• effectively addresses the writing task;

• is well organized and well developed;

• uses clearly appropriate details to support a thesis or illustrate ideas;

• displays consistent facility in the use of language; and

• demonstrates syntactic variety and appropriate word choice.

Score of 5—Demonstrates clear competence in writing at both the rhetorical and syntactic levels, though it will probably have occasional errors. A paper in this category:
• may address some parts of the task more effectively than others;

• is generally well organized and well developed;

• uses details to support a thesis or illustrate an idea;

• displays facility in the use of language; and

• demonstrates some syntactic variety and range of vocabulary.

Score of 4—Demonstrates minimal competence in writing at both the rhetorical and syntactic levels. A paper in this category:
• addresses the writing topic adequately but may slight parts of the task;

• is adequately organized and developed;

• uses some details to support a thesis or illustrate an idea;

• displays adequate but possibly inconsistent facility with syntax and usage; and

• may contain some errors that occasionally obscure meaning.

Score of 3—Demonstrates some developing competence, but it remains flawed at either the rhetorical and syntactic levels, or both. A paper in this category may reveal one or more of the following weaknesses:
• inadequate organization or development;

• inappropriate or insufficient details to support or illustrate generalizations;

• a noticeably inappropriate choice of words or word forms; and

• an accumulation of errors in sentence structure or usage.

Score of 2—Suggests incompetence in writing. A paper in this category is seriously flawed by one or more of the following weaknesses:
• serious disorganization or underdevelopment;

• little or no detail, or irrelevant specifics;

• serious and frequent errors in sentence structure or usage; and

• serious problems with focus.

Score of 1—Demonstrates incompetence in writing. A paper in this category does the following:
• may be incoherent;

• may be underdeveloped; and

• may contain severe and persistent writing errors.

Name	Purpose	Scale Definition	Scoring	Reliability	Validity
MCAT Essay **Author/Publisher** Association of American Medical Colleges Medical College Admission Test 2450 N. Street, NW Washington, DC 20037 **Date** 1985 **Testing Time** Two 30-min. essays *Used by colleges and universities throughout the U.S.*	Medical school entrance exam Each writing prompt provides a context for writing a response to a statement expressing an opinion, discussing a philosophy, or describing a policy related to a field of general interest such as business, politics, history, art, or ethics	Designed to assess skills in: Developing a central idea Synthesizing concepts and ideas Separating relevant from irrelevant information Developing alternative hypotheses Presenting ideas cohesively and logically Writing clearly with grammar, syntax, punctuation, and spelling consistent with timed, first draft composition (*see rubric on next page*)	Centralized Holistic—Based on general impression of overall quality	If the two readers' scores are discrepant by > 1 point, the paper is read by a more experienced resolution reader who determines the total score for the essay (fewer than 5%) Fall 1985 administration Inter-rater reliability .84 SEM .90 (Mitchell and Anderson 1986) Inter-rater reliability estimates for first three administrations using generalizability theory ranged from .70 to .73 (Mitchell and Anderson 1987) Test-retest (corrected for restriction in range) with a piloted 45-min. version ranged from .38 to .58	No average score differences between examinees grouped by gender, rural/urban status, age, or number of years of postsecondary education (Mitchell and Anderson 1987)

MCAT Holistic Scoring Guide

6—These papers show clarity, depth, and complexity of thought. The treatment of the writing assignment is focused and coherent. Major ideas are substantially developed. A facility with language is evident.

5—These essays show clarity of thought, with some depth or complexity. The treatment of the writing assignment is generally focused and coherent. Major ideas are well developed. A strong control of language is evident.

4—These essays show clarity of thought and may show evidence of depth or complexity. The treatment of the writing assignment is coherent, with some focus. Major ideas are adequately developed. An adequate control of language is evident.

3—These essays show some clarity of thought but may lack complexity. The treatment of the writing assignment is coherent but may not be focused. Major ideas are somewhat developed. While there may be some mechanical errors, control of language is evident.

2—These essays may show some problems with clarity or complexity of thought. The treatment of the writing assignment may show problems with integration or coherence. Major ideas may be underdeveloped. There may be numerous errors in mechanics, usage, or sentence structure.

1—These essays may demonstrate a lack of understanding of the writing assignment. There may be serious problems with organization. Ideas may not be developed. There may be so many errors in mechanics, usage, or sentence structure that the writer's ideas are difficult to follow.

X—These responses avoid the assigned topic altogether. They may be blank, illegible, or written in a language other than English; consist entirely of an obvious effort to ignore the purpose of the writing sample, such as a drawing; or address a topic other than the one assigned.

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
Texas Academic Skills Program (TASP) writing test essay component and a 40-item multiple-choice segment (used only with a failing grade on the essay by one or both raters) **Author/Publisher** Texas Academic Skills Program P.O. Box 140347 Austin, TX 78714–0347 **Date** 1989 **Testing Time** 5 hours to complete the writing component (basically untimed) **Cost** $24 *For use with all university students*	The TASP test is a power test designed to insure that all students attending public higher education institutions have the basic skills necessary to perform effectively	Centralized by NES in Texas Holistic (National Evaluation Systems) Final draft with revisions made Individual diagnostic utility leading to informed placement decisions and remediation as needed	Discrepancies between raters are resolved by a third rater	Significantly fewer black and Hispanic students passed the writing test compared to Caucasian students Females exhibited a significantly higher passing rate than males Students with high school GPAs below 2.5 had a significantly lower rating compared to their counterparts with higher GPAs The percentage of transfer student passing was significantly lower than for nontransfers (Bell and Olney 1990) Trend analysis showed that passing rates for writing test have increased over the past several years 1989–94 for all minorities except Asians (Texas Higher Education Coordinating Board 1995)	

The following characteristics are incorporated into scoring essays:
Appropriateness—Extent to which the student addresses the topic and uses language and style appropriate to the given audience, purpose, and occasion.
Unity and focus—The clarity with which the student states and maintains a main idea or point of view.
Development—The amount, depth, and specification of supporting detail the student provides.
Organization—The clarity of the student's writing and logical sequence of the student's ideas.
Sentence structure—The effectiveness of the student's sentence structure and the extent to which the student's writing is free of errors in sentence structure.
Usage—The extent to which the student's writing is free of errors in usage and shows care and precision in word choice.
Mechanical conventions—The student's ability to spell common words and use the conventions of capitalization and punctuation.
The multiple-choice segment assesses the following:
Elements of composition, including recognition of purpose, audience, and appropriate organization.
Sentence structure, usage, and mechanics, including recognition of effective sentences and edited American English usage.

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
College-Level Academic Skills Test Essay (CLAST) narrative/persuasive essay (multiple-choice available) **Author/Publisher** Florida State Dept. Of Education **Date** 1984 **Testing Time** 1 hour *All information from author (1994) unless otherwise stated*	Advance-ment to upper division courses	Holistic scoring; Range of scores on essay 2–8 (sum of 2 raters); total score for each writing subtest (essay and multiple-choice); passing score 5 Essays read in 1–2 minutes; given score from 1–6 based on the following elements: Definite purpose Clear thesis Organized plan Well-developed supporting paragraphs Specific, relevant details A variety of effective sentence patterns Logical transitions Effective word choice Correct standard English usage	For multiple-choice KR-20 .71–.73 SEM 1.89–2.06 Percent rater agreement 47–53%	Students who failed ACT freshman placement test failed the CLAST at a rate of 38.5%, compared to 10.7% who passed the placement test With a GPA of 2.0, the passing rate was 72.7% w/ increasing passing rates corresponding to higher GPAs (Nickens 1992)	

CLAST Scoring Rubric

Score of 6—Implied or stated thesis that is developed with noticeable coherence. Ideas are substantive, sophisticated, and carefully elaborated. Choice of language and structure is precise and purposeful. Control of sentence structure, usage, and mechanics, despite an occasional flaw, contributes to the writer's ability to communicate the purpose.

Score of 5—Presents an implied thesis and provides convincing, specific support. Ideas are usually fresh, mature, and extensively developed. Command of language and use of a variety of structures are demonstrated. Control of sentence structure, usage, and mechanics, despite an occasional flaw, contributes to the writer's ability to communicate the purpose.

Score of 4—Presents a thesis and often suggests a plan of development that is usually carried out. Enough supporting detail to accomplish the purpose of the paper is provided. Makes competent use of language and sometimes varies sentence structure. Occasional errors in sentence structure, usage, and mechanics do not interfere with the writer's ability to communicate the purpose.

Score of 3—Presents a thesis and often suggests a plan of development that is usually carried out. Support that tends toward generalized statements or a listing. In general, support is neither sufficient nor clear enough to be convincing. Sentence structure tends to be pedestrian and often repetitious. Errors in sentence structure, usage, and mechanics sometimes interfere with the writer's ability to communicate the purpose.

Score of 2—Paper usually presents a thesis. The writer provides support that tends to be sketchy and/or illogical. Sentence structure may be simplistic and disjointed. Errors in sentence structure, usage, and mechanics interfere with the writer's ability to communicate the purpose.

Score of 1—Paper generally presents a thesis that is vaguely worded or weakly asserted. Support, if any, tends to be rambling and/or superficial. The writer uses language that often becomes tangled, incoherent, and thus confusing. Errors in sentence structure, usage, and mechanics frequently occur.

Name	Purpose	Definition	Scoring	Reliability	Validity	Correlation with Other Measures
New Jersey College Basic Skills Placement Test (NJCBSPT) **Author/Publisher** **Date** 1978 **Testing Time** Essay 20 min. Rest of test 2 hrs., 45 min. **Cost** *Currently used by publicly supported colleges in NJ and a number of private schools*	To determine which students admitted to college need remedial instruction in basic skill areas in order to successfully complete college programs (proficiency)	Writing unified paragraphs, organization of ideas, development of a logical argument, provision of specific examples, use of complete sentences and correct spelling, maintains a consistent tone, and can express ideas precisely	Holistic Essay Composition (a composite based on sentence sense and essay sections) English (a composite based on reading comprehension, sentence sense, and essay sections) High level of refinement not expected due to time limit	If scores differ by > than one point on the 4-point scale, a third reader scores	Median predictive validity coefficients: Sentence structure .34 Essay .21 Reading comprehension .26 Median concurrent validity coefficients: Sentence structure .33 Essay not available Reading comprehension .27 Results of two content validity questionnaires revealed NJ college instructors were in general agreement that the test content was appropriate and important to assess (Hecht 1980) 68–98% of students believed by instructors to be appropriately placed 60–98% of students who thought they were placed correctly (Hecht 1980)	NJCBSPT and GPA of college students attending South Central Comm. College in CT .11 $p > .05$ (Hasit and DiObilda 1996) Grades in writing courses in college and NJCBSPT: Sentence structure .16 to .47 Essay -.04 to .40 Reading comprehension .16 to .52 (Hecht 1980) NJCBSPT reading comprehension and scores on comparative guidance and placement (CGP) reading test .75 (Hecht 1980) NJCBSPT sentence structure and CGP sentences .73 (Hecht 1980) Reading comprehension and SAT—V .74 TSWE .68 Sentence structure and SAT—V .66 TSWE .75 Essay and SAT—V .50 TSWE .55 (Hecht 1978)

Organization/Content

1—May not have an opening and/or a closing. These papers are on topic and demonstrate at least a minimal attempt to respond to the topic by stating a subject or giving a list of subjects. Some of the lengthier papers are disorganized, making them consistently difficult to follow. Others will relate to the topic but will have an uncertain focus. In these papers the reader has to infer what the focus is. The overriding characteristic of many of these papers is a lack of control with no sense of planning. Details may be random, inappropriate, or barely apparent.

2—May not have an opening and/or a closing. These responses will exhibit an attempt at organization. In other words, there will be some evidence the writer attempted to control the details. The responses relate to the topic, but in some papers, the writer drifts away from the primary focus or abruptly shifts focus. In other papers, there is a single focus but there are few, if any, transitions, making it difficult to move from idea to idea. Details are presented with little, if any, elaboration.

3—May not have an opening and/or a closing. The responses relate to the topic and usually have a single focus. Some of these papers may drift from the focus or abruptly shift focus; however, in these papers, at least one of the subjects focused upon clearly meeting the criteria for a three. For example, some "3" papers are sparse—they have several details with a little elaboration, but they are organized and controlled; some "3" papers will ramble somewhat, repeating ideas and resulting in a lengthy response that otherwise would be sparse; and other "3" papers have elaborate ideas and details, but the writing sample is interrupted by organizational flaws/lapses or by a lack of transition between ideas or between clusters of ideas.

4—Generally will have an opening and closing. The responses relate to the topic. They have a single focus and are organized. There is little, if any, difficulty moving from idea to idea. Ideas may ramble somewhat and clusters of ideas may be loosely connected; however, an overall progression is apparent. In some papers, development is uneven, consisting of elaborated ideas interspersed with bare, unelaborated details.

5—Have an opening and a closing. These responses relate to the topic and have a single focus. They are organized and progress logically from beginning to end. The key ideas are developed with appropriate and varied details. Clusters of ideas are strongly connected. Some writers take compositional risks and are, for the most part, successful. Although these papers are flawed, they have a sense of completeness and unity.

6—Have an opening and closing. The responses relate to the topic and have a single focus. They are well developed, complete compositions that are organized and progress logically from beginning to end. A variety of cohesive devices are present, resulting in a fluent response. Many of these writers take compositional risks resulting in highly effective, vivid responses.

Usage

1—May display numerous errors in usage. This includes problems in tense formation, subject-verb agreement, pronoun usage and agreement, and word choice.

2—May have severe problems with usage, but they are not totally out of control.

3—May display a pattern of errors in usage.

4—May display some errors in usage, but no consistent pattern is apparent.

5—Have few errors in usage.

6—Have very few, if any, errors in usage.

NJCBSPT Rubrics—Continued

Sentence Construction

1—May demonstrate an assortment of grammatically incorrect sentences and/or incorrect rhetorical modes. Statements may be either incoherent or unintelligible.

2—May demonstrate excessive monotony in syntax and/or rhetorical modes. There may be numerous errors in sentence construction.

3—May demonstrate an excessive monotony in syntax structure and/or rhetorical modes. There may be errors in sentence construction.

4—May demonstrate a generally correct sense of syntax. They avoid excessive monotony in syntax and/or rhetorical modes. There may be a few errors in sentence construction.

5—Demonstrate syntactic and verbal sophistication through an effective variety of sentences and/or rhetorical modes. There are few, if any, errors in sentence construction.

6—Demonstrate syntactic and verbal sophistication through an effective variety of sentence and/or rhetorical modes. There will be very few, if any, errors in sentence construction.

Mechanics

1—May display errors in mechanics so severe as to detract from the meaning of the response.

2—May display numerous serious errors in mechanics.

3—May display a pattern of errors in mechanics.

4—May display some errors in mechanics, but these errors will not constitute a consistent pattern.

5—Have few errors in mechanics.

6—Have very few, if any, errors in mechanics.

Name	Purpose	Definition	Scoring	Reliability	Validity
Illinois Inventory of Educational Progress-Writing Assessment **Author/ Publisher** Illinois State Board of Education **Date** 1983 **Testing Time** 25 min. **Cost** *Currently used by public institutions in Illinois*	To describe the current status of Illinois students' writing abilities and to monitor skill development over time High instructional utility—provides detailed info about individual strengths and weaknesses and helps to identify areas of instructional need Emphasizes stages of development and avoids pejorative classifications	Functional Writing—students write essays in which they explain their points of view on certain issues or convey ideas or events to inform or convince the reader	6-point analytic ratings for 4 elements of clear writing: Focus Organization Support Elaboration Mechanics Also info. pertaining to whether or not mechanical skills (sentence construction, usage, spelling, punctuation, capitalization, and paragraph format) are at or below mastery The holistic rating is conceptualized as a global judgment of how effectively the composition generally incorporates the 4 elements and addresses the assignment	Inter-rater at least .80 for all subscales except for focus (.74) Total (.92) Generalizability coefficients .81 to .98	Aggregate writing ability scores and inferential reading/grammar multiple-choice .50 (Chapman, Fyans, and Kerins 1984)

Illinois Inventory of Educational Progress Rubric

Focus
1—The subject may be unclear. There is no discernible main point.
2—The subject is still clear. There may be more than one main idea developed. The reader must work very hard to infer a main idea.
3—The subject is clear. Opening or closing statements may specify more or fewer points or subtopics than are actually developed in the paper. The reader must, but can, infer the main idea.
4—The subject is clear. The main idea or view is stated. There is no attempt to specify points that are developed. The beginning and end may relate, but do not contradict each other.
5—The subject is clear. The main idea or view is stated. The general number or type of key points or subtopics are mentioned. Opening and closing statements may relate to or follow from each other.
6—The essay can stand alone. The subject is clear. The main idea or view is stated. The key points or subtopics that are developed are specifically named. Opening and closing statements match or logically relate to the text and to each other.

Support
1—There is little or no support. Support is very confusing or at the same level of generality as the point it is intended to develop.
2—Support is attempted, but few major points are elaborated. Little of the elaboration is precise or clear. The support may be redundant.
3—Only some major points are elaborated. Only some elaboration is specific. It may be a list.
4—Many major points are further elaborated. Much of the elaboration is specific. Much of the elaboration is second order.
5—Most major points are elaborated. Most elaboration is specific and second order.
6—The essay's main idea or view and all major subtopics are elaborated and explained by specific detail.

Organization
1—There is no evidence of a plan. Almost no points are logically related.
2—A plan is attempted, but the reader must work very hard to infer it. There are few or no transitions signaling major points. There are few logically developed points.
3—The plan is noticeable, but the reader must infer it. Only some major points are signaled by transition. There are some logically connected points. There may be some major digressions.
4—The plan is clear. Many major points are signaled by transitions and in paragraphs. Most points are logical. There may be a few minor digressions, but no major ones.
5—The plan is clear. Most major points are separated into paragraphs and signaled by transitions. All points are logically developed to each other. There may be a few minor digressions but no major ones.
6—The essay plan is very evident. The plan is signaled by the division of major points into paragraphs. The plan is also signaled by use of transitions.

Grammar/Mechanics
1—Errors are so numerous and serious that they interfere with communication.
2—There are many gross errors, causing some confusion.
3—There are numerous minor errors and some gross errors. Sentence construction is below mastery.
4—There are a few common errors. A few may be gross.
5—There may be a few minor errors, but no more than one gross error.
6—There are few or no minor errors. There are no gross errors.

Name	Purpose	Scoring	Reliability	Validity	Correlation with other measures
Writing Proficiency Exam **Southeast Missouri State University** 2-part essay: first part based on personal experience, second part based on readings about content of first essay **Author/Publisher** Correspondence: Nancy Blattner Director of Writing Assessment Southeast Missouri State Cape Girardeau, MO 63701 **Date** 1997 **Testing Time** 75 minutes *All information from author*	Exit exam Graduation requirement Monitor changes in writing skills Pre-/post-essay test: following course in written expression and after completion of 75 hrs.	Local, holistic approach See attached rubrics			

Southeast Missouri State University Writing Proficiency Exam—Scoring Rubric

Score 6

A. Focus—Main idea is very clearly stated, and the topic is effectively limited.

B. Organization—A logical plan is signaled by highly effective transitions; the essay's beginning and end are effectively related to the whole.

C. Development—All major ideas are set off by paragraphs that have clearly stated or implied topics; the main ideas and all major topics are supported by concrete, specific evidence.

D. Style—Sentences relate to each other and to the paragraph topic and are subordinate to the topic; word and phrase choice is felicitous; tone is consistent and appropriate.

E. Correctness—There are no major mechanical errors (e.g., agreement) and only a few minor errors.

F. References—Source material is incorporated logically, insightfully, and elegantly; sources are documented accurately, elegantly, and emphatically.

Score 5

A. Focus—Main idea is clear, and the topic is limited.

B. Organization—A logical plan is signaled by some transitions; the essay's beginning and end are clearly and effectively related to the whole.

C. Development—Almost all major ideas are set off by paragraphs that, for the most part, have clearly stated or implied topics; the main idea and all major topics are supported by concrete, specific detail.

D. Style—Paragraphs are built on logically related sentences; word and phrase choice is consistent and accurate; tone is nearly consistent and appropriate.

E. Correctness—There is only one major mechanical error or a few minor errors.

F. References—Source material is incorporated logically and proficiently; sources are documented accurately.

Score 4

A. Focus—Main idea is clear or clearly implicit, and the topic is partially limited.

B. Organization—A logical plan is signaled by transitions; the essay's beginning and end are somewhat effective.

C. Development—Most major ideas are set off by paragraphs that mainly have stated or implied topics; the main idea and almost all major points are supported by concrete, specific detail.

D. Style—Sentences in paragraphs are subordinate to topics; word choice is almost accurate; tone is sometimes appropriate.

E. Correctness—There may be a few major mechanical errors or a few minor errors.

F. References—Source material is incorporated logically and adequately; sources are documented accurately for the most part.

Score 3

A. Focus—Main idea is unclear, and the topic is only partially limited.

B. Organization—There is an attempted plan that the reader must infer; the essay's beginning and end may be ineffective.

C. Development—Some major ideas are set off by paragraphs that may have stated or implied topics; some major points in paragraphs are supported by concrete, specific detail.

D. Style—Sentences may not be subordinate to topics; word choice is generally accurate; tone is often inappropriate.

E. Correctness—Some major and minor mechanical errors are present.

F. References—Source material is incorporated but sometimes inappropriately or unclearly; documentation is accurate only occasionally.

Score 2

A. Focus—Main idea is unclear, and the topic is unlimited.

B. Organization—There is no clear plan; the essay's beginning and end are not effective.

C. Development: few major ideas are set off by paragraphs; few paragraphs have stated or implied topics; supportive detail is imprecise, unclear, or redundant.

D. Style—Sentence relationships at times are confusing; word choice is frequently inaccurate; tone is inappropriate.

E. Correctness—Many major and minor mechanical errors cause confusion.

F. References—Source material is inappropriately or unclearly incorporated; documentation is infrequent.

Score 1

A. Focus—The subject and the main idea are unclear; no apparent attempt has been made to limit the topic.

B. Organization—There is no discernible plan; no attempt is made to compose an effective beginning and end.

C. Development—Major ideas are not set off by paragraphs; only one, if any, paragraph has a stated or implied topic; little or no supporting detail is used.

D. Style—Sentence relationships must be inferred; word choice is often confusing; tone is inappropriate or distracting.

E. Correctness—Many varied major and minor errors occur, making the paper difficult to read.

F. References—Source material is never incorporated or incorporated appropriately or clearly; documentation is inaccurate.

Score 0

Designates an essay that is clearly not written on the assigned topic or makes no attempt to answer the given question.

Name	Purpose	Definition	Scoring	Reliability	Validity	Correlation with other measures
Miami University's Portfolio **Authors** Laurel Black, Donald Daiker Jeffrey Sommers, Gail Stygall **Publisher** Department of English Miami University Oxford, OH **Date** 1996	To award entering students college credit and advanced placement in composition based on their best high school writing	See content descriptions of 4 pieces (reflective letter, story or description, persuasive essay, response to text)	A total holistic score (1–6) is derived from 4 equally important pieces of prose writing *See attached rubric below*			

Miami University's Portfolio Content Descriptions

1 A reflective letter—This letter, addressed to Miami University writing teachers, introduces you and your portfolio by thoughtfully reflecting upon and analyzing your writing or yourself as a writer. Your reflections should give readers a better understanding of who you are as the writer of this portfolio. Your letter may discuss important choices in creating the portfolio, describe your development as a writer, evaluate the strengths and weaknesses of your writing, or combine these topics.

2 A story or a description—This narrative or descriptive piece should be based upon your own experience. Its aim is to communicate a significant experience rather than explain it. Your writing will most likely be personal and informal. A short story is acceptable.

3 An explanatory, exploratory, or persuasive essay—It may be formal or informal in style, but it should have a strong focus and a clear central idea or direction. The aim of both an explanatory or exploratory essay is to be informative and enlightening, but an explanatory essay answers questions whereas an exploratory essay raises them. The aim of a persuasive paper is to be convincing, to change the reader's mind or heart or both. A paper that explains a physical process—a "how-to" paper—is not appropriate. Neither is a research paper that merely assembles information from other sources and is not based on your own ideas.

4 A response to a written text—This essay should respond to a short story, novel, poem, play, or piece of nonfiction prose written by a professional, a classmate, or yourself. It may interpret all or part of the test, evaluate it, show how it works, explain its significance, compare it to other texts, relate it to personal experience and values, or combine these approaches. Even if some secondary sources are used, readers should come away with a strong sense of your own response to the text. (If the text is not commonly known, a copy of it should be included in the portfolio.)

Miami University Portfolio Scoring Scale

6 range—An excellent portfolio; its numerous and significant strengths far outweigh its few weaknesses. Writer demonstrates an ability to handle varied prose tasks successfully. Substantial and original in content (both length and development) and/or in style.

5 range—A very good portfolio; its many strengths clearly outweigh its weaknesses. Writings suggest an ability to handle varied prose tasks successfully. Engages the material and explores issues, but not to the same extent as in a "6" portfolio.

4 range—A good portfolio; its strengths outweigh its weaknesses, but the reader may want to be more fully convinced of the writer's ability to handle varied prose tasks successfully. Portfolio shows genuine intellectual efforts and moments of sharp focus that compensate for its possible predictability.

3 range—A competent portfolio; its strengths and weaknesses are about evenly balanced. There is some evidence of the writer's ability to handle varied prose tasks successfully. Some pieces may be too brief, underdeveloped, general, or predictable, but the writing is competent.

2 range—A fair portfolio; its weaknesses outweigh its strengths. There is little evidence of the writer's ability to handle varied prose tasks successfully. Usually thin in substance and undistinguished in style but perhaps clear and error free.

1 range—A poor portfolio; its many weaknesses clearly outweigh its strengths. It appears to have been put together with not enough time or thought.

Name	Purpose	Definition	Utility/Applicability	Reliability/Validity	Correlation with other measures
Missouri Western State College Portfolio includes resume, reflective essay, and several writing pieces from major courses **Author/Publisher** MWSC English Dept. Faculty Correspondence: Jane Frick Missouri Western State College St. Joseph, MO 64507 **Date** 1992 **Testing Time** N/A **Scores** N/A	Exit survey for 3 English major concentrations (technical communications, public relations, and writing) Portfolio assessment using a "course approach" for designating pieces of writing Assessment was developed in response to state law requiring public higher education institutions to establish majors exit exams Faculty devised this assessment approach as an alternative to commercially available exams due to a discrepancy between course content in three of their English emphases and the GRE, NTE, or ETS exams, which emphasize literature	Three faculty members judge each portfolio to be complete or incomplete, adding evaluative comments if they wish; if two of the three readers view the portfolios to be incomplete, students are required to meet with their academic advisors, rework, and then resubmit the portfolio for reevaluation	Provides information for faculty regarding student perceptions of the curriculum, the value of internship experiences (through review of student materials produced in the work of world), and types of assignments given by colleagues Greater variety and depth of assignments Innovative teaching methods have resulted Has insured continuation of programs and adequate funding		

Name	Purpose	Scoring	Validity	Correlation with other measures
The Computerized Inventory of Developmental Writing Traits (CIDWT) **Authors** Niki McCurry, Writing Theory James Nivette, Statistical Design William Wresch, Programming Alan McCurry, Instructional Plan **Publisher** Developed by a research team from the Alaska Writing Program Box 80210 Fairbanks, Alaska *Enables comparisons across colleges and states*	Direct assessment of student writing to measure curriculum improvements in the context of program evaluation Assess process of writing with normed scores provided in exchange for contributing to the national database CIDWT is an MS-DOS program with 35 counts and analyzes targeted numeric indicators in text files CIDWT counts several variables and calculates weighted scores, t-scores, and norms	Score counts on variables and a total weighted score Centralized (scored at CIDWT, the database center in CA) Computerized (runs on IBM compatible computers) CIDWT can score 40–44 essays per minute; word processing files need only be saved as a basic text file to be transferred to CIDWT for analysis	Four factors emerged across numerous studies: fluency, sentence development, word choice, and paragraph development CCNY college freshmen (82 cases) El Paso Community College (243 samples) San Jose State sophomores (75 samples) Including Caucasian, Hispanic, Black, and Asian students	Scores correlate very well and consistently with teacher ratings (as high as .85, with San Jose samples)

Essay Scoring

Numeric indicators

Total words, standard sentence length, average word length, standard word length, percentage of unique words, average sentence length, percentage of most common words, percentage of uncommon words, percentage of common words, number of semi-common words, number of uncommon words, number of semi-uncommon words, number of common words, number of very common words, number of most common words	number of prepositions, number of articles, number of pronouns, number opinion words, number of transitions, number of slang words, number of THEs, number of punctuation marks, number of subordinates, number of -ion words, number of vague words, number of conditionals, number of coordinates, number of TO BE verbs, total paragraphs

Name	Purpose	Scoring	Validity	Reliability	Correlation with Other Measures
University of Southern California Freshman Writing Center Program Portfolio Assessment **Authors** USC English Department faculty **Date** 1991 **Currently used by USC English Department**	Evaluation of the freshman writing program and affiliated tutoring center Specifically, to address questions such as, how do writing center visits affect student grades? What aspects of the writing process should be emphasized during writing center visits?	End of semester portfolios are graded by one instructor familiar with the student's work and one who is not Midterm portfolio submission—A course paper is selected by the student for diagnosis of strengths and weaknesses, and is revised; no grades assigned Required documents in the final portfolio include two previously submitted papers that can be extensively revised and an impromptu essay written in class as a guard against cheating			

Name	Purpose	Scoring	Reliability
Scale for Evaluating Expository Writing (SEEW) (revised form: Expository Scale V) **Author** Edys Quellmalz **Publisher** Center for the Study of Evaluation UCLA Los Angeles, CA **Date** 1978–82 (construction)	Designed as a criterion-referenced scale to describe levels of writing skill development for basic essay elements at intermediate, postsecondary, and adult levels Program assessment High level of instructional utility given the inclusion of 5 analytic subscales; the inclusion of analytic scales enables the provision of diagnostic feedback to students, parents, teachers, and program personnel	Local, holistic/analytic The holistic judgment of the General Impression Scale requires the rater to assess the overall quality with which the writer engages the topic to achieve the intended output for the intended audience Raters may include subjective reactions to freshness of idea, originality, and style Analytic scales call for quality ratings based on a specified set of basic elements Rubrics for General Impression, General Competence, and Essay-Coherence scales are provided (*see bottom of page*)	At the end of a structured training session, generalizability coefficients indicating rater agreement on the subscales ranged from .93 to .97 Percentages of rater agreement after rating ranged from .89 to .91 on the subscales

Expository Scale V Rubrics

General Impression

6—An excellent example of exposition.
5—A good, adequate example of exposition.
4—An adequate example of exposition.
3—A marginal example of exposition.
2—A poor example of exposition.
1—A very poor example or barely readable paper, completely off the topic.

General Competence

Based on their first or second readings of the essay, raters decide how competently the writer formed the essay, with reference to the following elements: main idea, essay organization, paragraph organization, support, and mechanics.

Expository Scale V Rubrics—Continued

Master
6—Very competent. The paper executes all the elements competently. There are no serious errors. The paper has a clear main idea; logical organization; relevant, detailed support; and a command of basic mechanics. There are no major flaws.

5—Definitely competent. The paper is competent in all of the basic elements, but there may be a few minor flaws.

4—Adequately competent. The paper is adequately competent in all of elements. There may be a few flaws, some of which may be serious.

Nonmaster
3—Almost competent. The paper lacks competence in one or two elements, and there are several flaws.

2—Not very competent. The paper has two or more of the elements. There are many serious flaws

1—Not at all competent. Paper has none or only one of the elements competently executed.

Essay Coherence

This subscale focuses on the flow of ideas throughout the entire paper and between paragraphs. The emphasis is on vertical relationships of ideas throughout the essay.

Master
6—The subject is identified. The main idea is stated or implied in opening and/or closing statement. Opening and closing statements must match or logically relate to the text and to each other. The topic is limited through reference to key points or lines of reasoning. The essay plan is logical. The essay plan is clearly signaled by transitions. The essay plan is consistently maintained (no digression or extraneous material).

5—The subject is identified. The main idea is stated or implied in opening and/or closing statement. Opening and closing statements relate to or follow from the text and from each other. The topic is partly limited by indicating number and type of key points. The plan is logical. The plan is signaled by appropriate transitions. There may be digression or an elaboration.

4—The subject is identified. The main idea is identified or implied. There may or may not be an attempt to limit the topic or give directions to subsequent reasoning. There may be a few minor digressions from the plan, but no major digressions. Subtopics can be reshuffled.

Nonmaster
3—Subject is clear. Main point may not be very clear. There may be a major digression or several minor digressions. A plan is attempted that may need to be inferred.

2—Subject is clear. Main idea not very clear and/or there may be more than one. There are many digressions. The plan is attempted but not consistently or not completely carried out.

1—Subject is unclear. Main idea is absent or very unclear. No plan is attempted or followed.

REFERENCES

American Association for Higher Education Assessment Forum. (1997). In L.F. Gardiner, C. Anderson, and B.L. Cambridge (eds.), *Learning Through Assessment: A Resource Guide for Higher Education*. Washington, DC: American Association for Higher Education.

American College Testing Program. (1989). *Report on the Technical Characteristics of CAAP, Pilot Year 1: 1988–1989*. Iowa City, IA: American College Testing Program.

American College Testing Program. (1991). *CAAP Technical Handbook*. Iowa City, IA: American College Testing Program.

American College Testing Program. (1992). *Clarifying and Assessing General Education Outcomes of College, College Outcome Measures Program: Technical Report 1982–1991*. Iowa City, IA: American College Testing Program.

Banta, T.W., and Pike, G.R. (1989). Methods of Comparing Outcomes Assessment Instruments. *Research in Higher Education 30*(5): 455–69.

Bell, J.D., and Olney, R.J. (1990). An Analysis of Factors Which Predict Failure Rates on Writing Competency Tests. Paper presented at the Annual Meeting of the Association for Business Communication. (ERIC Document Reproduction Service No. ED324694.) San Antonio, TX.

Berger, A. (1985). Review of Watson-Glaser Critical Thinking Appraisal. In J.V. Mitchell (ed.), *Ninth Mental Measurement Yearbook* (pp. 1692–93). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Bizzell, P. (1982). Cognition, Convention, and Certainty: What We Need to Know About Writing. *Pre/Text 3*: 213–43.

Blankentstein, K.R., Flett, G.L., and Batten, I. (1989). Test Anxiety and Problem-Solving Self-Appraisals of College Students. *Journal of Social Behavior and Personality 4*(5): 531–40.

Brabeck, M.M. (1983). Critical Thinking Skills and Reflective Judgment Development: Redefining the Aims of Higher Education. *Journal of Applied Developmental Psychology 4*: 23–24.

Breland, H.M. (1983). *The Direct Assessment of Writing Skills: A Measurement Review*. College Board Report No. 83–6. (ETS RR No. 83–32.) NY: College Entrance Examination Board.

Breland, H.M. (1996). *Writing Skill Assessment: Problems and Prospects*. Princeton, NJ: Educational Testing Service.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., and Rock, D.A. (1987). *Assessing Writing Skill*. NY: College Entrance Examination Board.

Brems, C., and Johnson, M.E. (1989). Problem-Solving Appraisal and Coping Style: The Influence of Sex-Role Orientation and Gender. *The Journal of Psychology 123*(2): 187–94.

Bridgeman, B., and Bonner, M. (1994). *New SAT Special Report No. 5*. New York, NY: College Entrance Examination Board.

Bridgeman, B., and Carlson, S. (1983). *Survey of Academic Tasks Required of Graduate and Undergraduate Foreign Students*. Technical report. (ERIC Document Reproduction Service No. ED241543.) Princeton, NJ: Educational Testing Service.

Bridgeman, B., and Frederick, M. (1996). *Gender and Ethnic Group Differences on the GMAT Analytical Writing Assessment*.Technical report. (ERIC Document Reproduction Service No. ED400330.) Princeton, NJ: Educational Testing Service.

Britton, J. (1975). *The Development of Writing Abilities*. London: Macmillan Education Press.

Callahan, C.M. (1995). Review of the California Critical Thinking Disposition Inventory. In J.C. Conoley and J.C. Impara (eds.), *Twelfth Mental Measurements Yearbook* (p. 57). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Callahan, S. (1995). Portfolio Expectations: Possibilities and Limits. *Assessing Writing 2*: 117–52.

Carlson, S.B., and Camp, R. (1985). *Relationships Between Direct and Indirect Measures of Writing Ability*. Paper presented at the annual meeting of the National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 255 543.) Chicago, IL.

Carrell, P.L. (1995). The Effects of Writers' Personalities and Raters' Personalities on the Holistic Evaluation of Writing. *Assessing Writing 2*: 153–90.

Carscaddon, D.M., Poston, J.M., and Sachs, E. (1988). Problem-Solving Appraisal as It Relates to State-Trait Personality Factors. *Psychology, A Journal of Human Behavior 25*(3/4): 73–76.

Chaffee, J. (1991). *Thinking Critically (3rd ed.)*. Boston: Houghton Mifflin Company.

Chapman, C.W., Fyans, L.J., and Kerins, C.T. (1984). Writing Assessment in Illinois. *Educational Measurement: Issues and Practice 3*: 24–26.

Chynoweth, G.H. (1987). Perceived Problem-Solving Processes: A Multidimensional Analysis. *Dissertation Abstracts International 47*(7-A): 2507–08.

Chynoweth, G.H., Blankinship, D.A., and Parker, M.W. (1986). The Binomial Expansion: Simplifying the Evaluation Process. *Journal of Counseling Development 64*: 645–47.

Clark, I.L. (1993). Portfolio Evaluation, Collaboration, and Writing Centers. *College Composition and Communication 44*: 515–24.

Clark, R.S. (1988). Age Cohort Performance on the College-Level Examination Program (CLEP) General Education in English Composition. Requirement for Directed Study. (ERIC Document Reproduction Service No. ED300390.) University of Pittsburgh.

Cole, N.S., and Moss, P.A. (1989). Bias in Test Use. In R. Linn (ed.), *Educational Measurement (3rd ed.)*. New York: Macmillan Publishing Company.

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

Cooperman, S., Bloom, J., and Bloom, D.S. (1983). *The Registered Holistic Scoring Method for Scoring Student Essays: Scoring Guide for Training*. Trenton: New Jersey State Department of Education.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Fort Worth, Florida: Harcourt Brace Jovanovich College Publishers.

Daly, R.F. (1987). A Comparative Study of the Pre-Professional Skills Tests (PPST) Test Scores of Pre-Service Teachers at Mankato State University. Technical report. (ERIC Document Reproduction Service No. ED291630.) Manko State University, MN.

Davey, T., Godwin, J., and Mittelholz, D. (1987). Developing and Scoring an Innovative Computerized Writing Assessment. *Journal of Educational Measurement 34*: 21–41.

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

Dressel, P.L., and Mayhew, L.B. (1954). *General Education*. Washington, DC: American Council on Education.

Dunbar, S.B. (1993). Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective. *JGE: The Journal of General Education 42*(2): 83–104.

Educational Testing Service. (1990). *The Academic Profile: A User's Guide, Part 1*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1994). *Tasks in Critical Thinking: Core Scoring Manual*. Princeton, NJ: Educational Testing Service.

Educational Testing Service and The College Board. (1989). *New Jersey College Outcomes Evaluation Program: A Report on the Development of the General Intellectual Skills Assessment*. Princeton, NJ: Educational Testing Service.

Elbow, P., and Yancey, K.B. (1994). On the Nature of Holistic Scoring: An Inquiry Composed on E-mail. *Assessing Writing 1*: 91–108.

Elliott, T.R., Sherwin, E., Harkins, S.W., and Marmarosh, C. (1995). Self-Appraised Problem-Solving Ability, Affective States, and Psychological Distress. *Journal of Counseling Psychology 42*(1): 105–15.

Emig, J. (1971). *The Composing Processes of Twelfth Graders*. Urbana, IL: National Council of Teachers of English.

Ennis, R.H., Millman, J., and Tomko, T.N. (1985). *Cornell Critical Thinking Tests Level X and Level Z—Manual*. Pacific Grove, CA: Midwest Publications.

Ewell, P.T., and Jones, D.P. (1993). Actions Matter: The Case for Indirect Measures in Assessing Higher Education's Progress on the National Education Goals. *The Journal of General Education 42*: 123–48.

Facione, N.C., and Facione, P.A. (1992). *The California Critical Thinking Skills Test: Forms A and B—Test Manual*. Millbrae, CA: The California Academic Press.

Facione, P.A. (1990a). *The California Critical Thinking Skills Test—College Level Technical Report 1: Experimental Validation and Content Validity*. Millbrae, CA: California Academic Press.

Facione, P.A. (1990b). *The California Critical Thinking Skills Test—College Level Technical Report 2: Factors Predictive of Critical Thinking Skills*. Millbrae, CA: California Academic Press.

Facione, P.A. (1990c). *The California Critical Thinking Skills Test—College Level Technical Report 3: Gender, Ethnicity, Major, Critical Thinking, Self-Esteem, and the CCTST*. Millbrae, CA: California Academic Press.

Facione, P.A. (1990d). *The California Critical Thinking Skills Test—College Level Technical Report 4: Interpreting the CCTST*. Millbrae, CA: California Academic Press.

Facione, P.A., Facione, N.C., and Giancarlo, C.A. (1992). *The California Critical Thinking Disposition Inventory: Test Manual*. Millbrae, CA: The California Academic Press.

Facione, P.A., Sanchez, C.A., Facione, N.C., and Gainen, J. (1995). The Disposition to Critically Think. *The Journal of General Education 44*(1): 1–25.

Faigley, L., Cherry, R.D., Jolliffe, D.A., and Skinner, A.M. (1985). *Assessing Writers' Knowledge and Processes of Composing*. Norwood, NJ: Ablex Publishing Corporation.

Feldt, L.S., and Brennan, R.L. (1989). Reliability. In R. Linn (ed.), *Educational Measurement (3rd ed.)*. New York: Macmillan Publishing Company.

Florida State Dept. of Education. (1994). *The College-Level Academic Skills Test, Technical Report 1993–1994*. (ERIC Document Reproduction Service No. ED381570.) Tallahassee, FL.

Flower, L., and Hayes, J.R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication 32*: 365–87.

Follman, J.C., and Anderson, J.A. (1967). An Investigation of the Reliability of Five Procedures for Grading English Themes. *Research Teaching of Education I*: 190–200.

Fontana, D., Lotwick, G., Simon, A., and Ward, L.O. (1983). A Factor Analysis of Critical, Convergent and Divergent Thinking Tests in a Group of Male Polytechnic Students. *Personality and Individual Differences 4*(6): 687–88.

Forrest, A., and Steele, J.M. (1982). *Defining and Measuring General Education Knowledge and Skills—College Outcome Measures Project: Technical Report 1976–81*. Iowa City, IA: American College Testing Program.

Frisby, C.L. (1992). Construct Validity and Psychometric Properties of the Cornell Critical Thinking Test (Level Z): A Contrasted Groups Analysis. *Psychological Reports 71*: 291–303.

Gadzella, B.M., Ginther, D.W., Bryant, G.W., and Caldwell, G. (1996). *Differences Among Course Grade Groups on Critical Thinking*. Paper presented at annual meeting of the Southwest

Educational Association Convention. (ERIC Document Reproduction Services No. ED 393 394.) New Orleans, LA.

Garrett, K., and Wulf, K. (1978). The Relationship of a Measure of Critical Thinking Ability to Personality Variables and to Indicators of Academic Achievement. *Educational and Psychological Measurement 38*: 1181–87.

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

Gray, M.J., and Grace, J.D. (1997). *Enhancing the Quality and Use of Student Outcomes Data. Final Report of the National Postsecondary Education Cooperative Working Group on Student Outcomes from a Data Perspective* (NCES 97–992). Washington, DC: National Postsecondary Education Cooperative.

Hasit, C., and DiObilda, N. (1996). Portfolio Assessment in a College Developmental Reading Class. *Journal of Developmental Education 19*: 26–31.

Hayes, J.R. (1996). A New Model of Cognition and Affect in Writing. In C.M. Levy and S. Ransdell (eds.), *The Science of Writing*. Mahwah, NJ: Erlbaum.

Hecht, L.W. (1978). Measuring Student Behavior During Group Instruction. *Journal of Educational Research 71*(5): 283–90.

Hecht, L.W. (1980). *Validation of the New Jersey College Basic Skills Placement Test*. Technical report. (ERIC Document Reproduction Service No. ED214945.) Princeton, NJ: Educational Testing Service.

Helmstadter, G.C. (1985). Review of Watson-Glaser Critical Thinking. In J.V. Mitchell (ed.), *Ninth Mental Measurements Yearbook* (pp. 1693–94). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Heppner, P.P. (1988). *The Problem-Solving Inventory Manual*. Palo Alto, CA: Consulting Psychologists Press.

Heppner, P.P., and Anderson, W.P. (1985). The Relationship Between Problem-Solving Self-Appraisal and Psychological Adjustment. *Cognitive Therapy and Research 9*: 415–27.

Heppner, P.P., and Peterson, C.H. (1982a). The Development and Implications of a Personal Problem-Solving Inventory. *Journal of Counseling Psychology 29*: 66–75.

Heppner, P.P., and Peterson, C.H. (1982b). Personal Problem-Solving: A Descriptive Study of Individual Differences. *Journal of Counseling Psychology 29*(6): 580–90.

Heppner, P.P., Reeder, B.L., and Larson, L.M. (1983). Cognitive Variables Associated with Personal Problem-Solving Appraisal: Implications for Counseling. *Journal of Counseling Psychology 30*: 537–45.

Hughes, J.N. (1992). Review of Cornell Critical Thinking Tests. In J.J. Kramer and J.C. Conoley (eds.), *Eleventh Mental Measurements Yearbook* (pp. 241–43). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research 60*: 237–63.

Illinois State Board of Education. (1984). *Write on Illinois! A User's Guide to Scoring Student Essays*. (ERIC Document Reproduction Service No. ED264298.) Springfield, IL.

Jacobs, S.S. (1995). Technical Characteristics and Some Correlates of the California Critical Thinking Skills Test Forms A and B. *Higher Education Research 36*: 89–108.

Jones, E.A., Dougherty, B.C., Fantaske, P., and Hoffman, S. (1997). *Identifying College Graduates' Essential Skills in Reading and Problem-Solving: Perspectives of Faculty, Employers and Policymakers*. (Contract No. R117G10037/CDFA84.117G.) University Park, PA: U.S. Department of Education/OERI.

Jones, E.A., Hoffman, S., Moore, L.M., Ratcliff, G., Tibbetts, S., and Click, B.A. (1995). *National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking* (NCES 95–001). Washington, DC: U.S. Government Printing Office.

Jones, E.A., and Nugent, M. (1996). *Enhancing the Teaching and Learning Process of Undergraduate Education at Large Universities*. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED394426.) New York, NY.

Kelley, P.H. (1973). Relation of Scores on the CLEP Subject Examination in Analysis and Interpretation of Literature to Performance in English 314K: Introduction to Literature I. *Research Bulletin* 73-11. (ERIC Document Reproduction Service No. ED173420.)

King, P.M., and Kitchener, K. (1994). *Developing Reflective Judgment*. San Francisco CA: Jossey-Bass Publishers.

King, P.M., Wood, P.K., and Mines, R.A. (1990). Critical Thinking Among College and Graduate Students. *The Review of Higher Education 13*(2): 167–86.

Koehler, C., and Neer, M. (1996). *An Investigation of Variables That Define Collaborative and Critical Thinking: The Effects of Argumentative Style on Critical Thinking Competence and Social Tolerance*. Paper presented at the Annual International Conference on Critical Thinking and Educational Reform. (ERIC Document Reproduction Service No. ED 399575.) Rohnert Park, CA.

Koenig, J., and Mitchell, K. (1988). An Interim Report on the MCAT Essay Pilot Project. *Journal of Medical Education 63*: 21–29.

Knefelkamp, L.L., and Slepitza, R. (1976). A Cognitive-Developmental Model of Career Development: An Adaptation of the Perry Scheme. *Counseling Psychologist 6*(3): 53–58.

Knefelkamp, L.L., Widick, C.C., and Stroad, B. (1976). Cognitive-Developmental Theory: A Guide to Counseling Women. *Counseling Psychologist 6*(2): 15–19.

Kurfiss, J.G. (1988). *Critical Thinking: Theory, Research, Practice, and Possibilities*. (AAHE-ERIC Higher Education Report No. 2.) (ERIC Document Reproduction Service.)

Langer, P., and Chiszar, D. (1993). Assessment of Critical Thinking Courses. *Perceptual and Motor Skills 77*: 970.

Lehigh Community College. (1992). Critical Test Review. In *Report to the President on Student Learning Competencies: Lehigh County Community College*. (ERIC Document Reproduction Service No. ED 345764.)

Malcoln, K.K. (1992). Review of Cornell Critical Thinking Tests. In J.J. Kramer and J.C. Conoley (eds.), *Eleventh Mental Measurements Yearbook* (pp. 243–44). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Marr, D. (1995). *An Investigation of the Construct Validity of the Long Form of the Academic Profile*. Princeton, NJ: Educational Testing Service.

McCammon, S., Golden, J., and Wuensch, K.L. (1988). Predicting Course Performance in Freshman and Sophomores Physics Courses: Women Are More Predictable Than Men. *Journal of Research in Science Teaching 25*(6): 501–10.

McCurry, N. (1992). *The Computerized Inventory of Developmental Traits*. Program description. (ERIC Document Reproduction Service No. ED357375.)

McMillan, J. (1987). Enhancing College Students' Critical Thinking: A Review of Studies. *Research in Higher Education 26*(1): 3–29.

Messick, S. (1981). Evidence and Ethics in the Evaluation of Tests. *American Psychologist 30*: 955–66.

Messick, S. (1989). In R.L. Linn (ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing Company.

Messick, S. (1992). The Interplay of Evidence and Consequencs in the Validation of Performance Assessments. Paper presented at the annual meeting of the National Conference on Measurement in Education. San Francisco, CA.

Miami University of Ohio. (1996). *The Best of Miami University's Portfolios: 1996 Scoring Guide for Portfolios*. Unpublished paper, appendix D. Oxford, OH: Miami University.

Miller, M.D., and Crocker, L. (1990). Validation Methods for Direct Writing Assessment. *Applied Measurement in Education 3*: 285–96.

Mines, R.A., King, P.M., Hood, A.B., and Wood, P.K. (1990). Stages of Intellectual Development and Associated Critical Thinking Skills in College Students. *Journal of College Student Development 31*: 538–47.

Mitchell, K.J., and Anderson, J.A. (1986). *Reliability of Holistic Scoring for the 1985 MCAT Essay*. Paper presented at the annual meeting of the American Psychological Association. (ERIC Document Reproduction Service No. ED285913.) Washington, DC.

Mitchell, K.J., and Anderson, J.A. (1987). *Estimation of Interrater and Parallel Forms Reliability for the MCAT Essay*. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED283837.) Washington, DC.

Modjeski, R.B., and Michael, W.B. (1983). An Evaluation by a Panel of Psychologists of the Reliability and Validity of Two Tests of CT. *Educational and Psychological Measurement 43*: 1187–93.

Modu, C., and Wimmers, E. (1981). The Validity of the Advanced Placement English Language and Composition Examination. *College English 43:* 609–20.

Moore, W.S. (1990). *The Measure of Intellectual Development: An Instrument Manual*. Olympia, WA: Center for the Study of Intellectual Development.

Moss, P.A. (1994). Validity in High Stakes Writing Assessment. *Assessing Writing 1*: 109–28.

Murphy, S. (1994). Portfolios and Curriculum Reform: Patterns in Practice. *Assessing Writing 1*: 175–206.

Neal, G.W., and Heppner, P.P. (1986). Problem-Solving Self-Appraisal, Awareness, and Utilization of Campus Helping Resources. *Journal of Counseling Psychology 33*(1): 39–44.

Nezu, A.M. (1986). Efficacy of a Social Problem-Solving Therapy Approach for Unipolar Depression. *Journal of Consulting and Clinical Psychology 54*(2): 196–202.

Nickens, J. (1992). *Research Studies Related to the College-Level Academic Skills Test*. Technical report. (ERIC Document Reproduction Service No. ED390883.) Talahassee, FL.

Ochoa, S.H. (1995). Review of the California Critical Thinking Disposition Inventory. In J.C. Conoley and J.C. Impara (eds.), *Twelfth Mental Measurements Yearbook* (pp. 57–58). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Pascarella, E.T. (1989). The Development of Critical Thinking: Does College Make a Difference? *Journal of College Student Development 30*: 19–26.

Pascarella, E.T., Bohr, L., Nora, A., and Terenzini, P.T. (1996). Is Differential Exposure to College Linked to the Development of Critical Thinking? *Research in Higher Education 37*(2): 159–74.

Pike, G.R. (1989a). *A Comparison of the College Outcomes Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) Exams*. (ERIC Document Reproduction Service No. ED 320911.) Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G.R. (1989b). *The Performance of Black and White Students on the ACT-COMP Exam: An Analysis of Differential Item Functioning Using Samejima's Graded Model*. Research Report 89–11. (ERIC Document Reproduction Service No. ED 314461.) Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G.R. (1997). Assessment Methods: The California Critical Thinking Skills Test. *Assessment Update 9*(2): 10–11.

Powers, D.E., Fowles, M.E., and Willard, A.E. (1994). Direct Assessment, Direct Validation? An Example from the Assessment of Writing. *Educational Assessment 2*: 89–100.

Quellmalz, E. (1982). *Scale for Evaluating Expository Writing (SEEW)*. (ERIC Document Reproduction Service No. ED 236670.) Los Angeles: University of California, Center for the Study of Evaluation.

Ritchey, K.M., Carscaddon, D.M., and Morgan, C.H. (1984). Problem-Solving Appraisal Versus Hypothetical Problem-Solving. *Psychological Reports 55*: 815–18.

Rohman, D.G., and Wlecke, A.O. (1964). *Pre-Writing: The Construction and Application of Models for Concept Formation in Writing*. (Cooperative Research Project No. 2174.) East Lansing, MI: Michigan State University.

Saucier, B.L. (1995). Critical Thinking Skills of Baccalaureate Nursing Students. *Journal of Professional Nursing 11*(6): 351–57.

Scott, J.N., and Markert, R.J. (1994). Relationship Between Critical Thinking Skills and Success in Preclinical Courses. *Academic Medicine 11*: 920–24.

Scriven, M. (1991). *Review of Report on General Intellectual Skills Assessment in New Jersey*. (Contract no. 75.105.) Washington, DC: National Center for Education Statistics.

Sell, G.R. (1989). *Making Assessment Work: A Synthesis and Future Directions*. In P.J. Gray (ed.), *New Directions for Higher Education*. San Francisco: Jossey-Bass.

Sibert, P.C. (1989). *Relationships Among the ACT, PPST, NTE, ACT COMP, and the GRE*. (ERIC Document Reproduction Services No. ED 305374.) Cookeville, TN: Center for Teacher Education Evaluation, Tennessee Technological University.

Sommers, N.I. (1980). Revision Strategies of Student Writers and Experienced Adult Writers. *College Composition and Communication 31*: 378–88.

Southeast Missouri State University. (1997). *Writing Proficiency Exam: Student Essays Evaluated and Annotated*. (ERIC Document Reproduction Services.) Cape Girardeau, MO.

Stephenson, B.W., and Hunt, C. (1977). Intellectual and Ethical Development: A Dualistic Curriculum Intervention for College Students. *Counseling Psychologist 6*(4): 39–42.

Sudweeks, R. (1991). Validation and Standard Setting Studies for the Pre-Professional Skills Test for the Teacher Education Program at Brigham Young University. *Contemporary Issues in Reading 6*: 95–107.

Swanson, D.B., Norman, R.R., and Linn, R.L. (1995). Performance-Based Assessment: Lessons from the Health Professions. *Educational Researcher 24*: 5–11.

Takayoshi, P. (1996). The Shape of Electronic Writing: Evaluating and Assessing Computer-Assisted Writing Processes and Products. *Computers and Composition 13*: 245–57.

Taube, K.T. (1995). *Critical Thinking Ability and Disposition as Factors of Performance on a Written Critical Thinking Test*. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Services No. ED 387510.) San Francisco, CA.

Teacher Programs and Services. (1996). *Praxis I: Academic Skills Assessments*. Princeton, NJ: Educational Testing Service.

Terenzini, P. (1997). *Student Outcomes Information for Policymaking. Final Report of the National Postsecondary Education Cooperative Working Group on Student Outcomes from a Policy Perspective* (NCES 97–991). Washington, DC: National Postsecondary Education Cooperative.

Texas Higher Education Coordinating Board. (1995). *TASP and the Effectiveness of Remediation*. Annual report. (ERIC Document Reproduction Services No. ED 394478.) Austin, TX.

Turner, A.P. (1993). *Predictability Research Study Between General Educational Development Writing Skills Test and College Level Examination Program General English Composition*. Technical report. (ERIC Document Reproduction Service No. ED366371.) Danvers, MA: North Shore Community College.

U.S. Department of Education. (1996, November). *The National Assessment of College Student Learning: An Inventory of State-Level Assessment Activities* (NCES 98–862). Research and development report. Washington, DC: National Center for Education Statistics.

Watson, G., and Glaser, E.M. (1980). *Watson-Glaser Critical Thinking Appraisal Manual: Forms A and B*. San Antonio, TX: The Psychological Corporation.

Welch, C. (1989). *Differential Performance on a Direct Measure of Writing Skills for Black and White College Freshmen*. ACT Research Report Series. (ERIC Document Reproduction Service No. ED320955.) Iowa City, IA.

Wertheimer, L.C. (1980). Relations Among Developmental Dimenstions in Jane Loevinger's Model of Ego Development. *Dissertation Abstracts International 41*(6-B): 2362.

White, E.M. (1993). Assessing Higher-Order Thinking and Communication Skills in College Graduates Through Writing. *JGE: The Journal of General Education 42*: 2.

White, E.M. (1994). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance*. San Francisco, CA: Jossey-Bass Publishers.

White, E.M., and Thomas, L. (1981). Racial Minorities and Writing Skills Assessment in the California State University and Colleges. *College English 43*: 276–83.

Wilson, D.G., and Wagner, E.E. (1981). The Watson-Glaser Critical Thinking Appraisal as a Predictor of Performance in a Critical Thinking Course. *Educational and Psychological Measurement 41*: 1319–22.

Witte, S.P., Flach, J., Greenwood, C., and Wilson, K.E. (1995). More Notes Toward an Advanced Ability to Communicate. *Assessing Writing 2*: 21–66.

Yarbough, D.B. (1991). *Assessing Cognitive General Education Outcomes: Conclusions from a Decade of Research on the ACT COMP Measures*. Iowa City, IA: American College Testing Program.

Zemelman, S. (1977). How College Teachers Encourage Student Writing. *Research in the Teaching of Student English 11*: 227–34.