College Board Report No. 99-3
GRE Board Research Report No. 96-12R

Writing Assessment in Admission to Higher Education: Review and Framework

HUNTER M. BRELAND, BRENT BRIDGEMAN,
and MARY E. FOWLES

GRE® ETS Educational Testing Service The College Board

College Board Report No. 99-3
GRE Board Research Report No. 96-12R
ETS RR No. 99-3

Writing Assessment in Admission to Higher Education: Review and Framework

HUNTER M. BRELAND, BRENT BRIDGEMAN,
and MARY E. FOWLES

This report presents the findings of a research project funded by and carried out under the auspices of the College Board and the Graduate Record Examinations Board.

College Entrance Examination Board, New York, 1999

Hunter M. Breland is senior research scientist at ETS.
Brent Bridgeman is principal research scientist at ETS.
Mary E. Fowles is assessment specialist II at ETS.

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

Founded in 1900, the College Board is a not-for-profit educational association that supports academic preparation and transition to higher education for students around the world through the ongoing collaboration of its member schools, colleges, universities, educational systems and organizations.

In all of its activities, the Board promotes equity through universal access to high standards of teaching and learning and sufficient financial resources so that every student has the opportunity to succeed in college and work.

The College Board champions—by means of superior research; curricular development; assessment; guidance, placement, and admission information; professional development; forums; policy analysis; and public outreach—educational excellence for all students.

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Additional copies of this report (item #217798) may be obtained from College Board Publications, Box 886, New York, New York 10101-0886, (800) 323-7155. The price is $15. Please include $4 for postage and handling.

Acknowledgments

The authors thank Peter Cooper, Claudia Gentile, Donald Powers, and Lawrence Stricker for helpful comments on an earlier draft of this report.

Figure 1 on page 2: From *Gender and Fair Assessment* (p. 230; fig. 7.1), by W. W. Willingham and N. S. Cole, 1997, Mahwah, NJ: Erlbaum. Copyright © 1997 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.

Printed in the United States of America.

Contents

Tables

Figures

Abstract

A comprehensive review was conducted of writing research literature and writing test program activities in a number of testing programs. The review was limited to writing assessments used for admission in higher education. Programs reviewed included ACT, Inc.'s ACT™ program, the California State Universities and Colleges (CSUC) testing program, the College Board's SAT® program, the Graduate Management Admissions Test® (GMAT®) program, the Graduate Record Examinations® (GRE®) test program, the Law School Admission Test® (LSAT®) program, the Medical College Admission Test® (MCAT®) program, and the Test of English as a Foreign Language™ (TOEFL®) testing program. Particular attention was given in the review to writing constructs, fairness and group differences, test reliability, and predictive validity. A number of recommendations are made for research on writing assessment.

Key words: writing, research, literature, review, testing, higher education

Introduction

This review summarizes the results of writing assessment research and describes programmatic activities in large-scale writing assessment programs. The focus on assessments used for admission to higher education reflects the interests of the Graduate Record Examinations Board and the College Board, our sponsors. Accordingly, many important writing assessments—such as those not used for admissions (e.g., National Assessment for Educational Progress, or NAEP) and those intended primarily for writers of English as a second language (e.g., the Test of English as a Foreign Language, or TOEFL)—are not reviewed in detail, although data from such programs may be cited for reference. We also decided to exclude assessments that involve writing but do not assess writing proficiency. For example, the College Board's Advanced Placement Program (AP®) and the PRAXIS Series: Professional Assessments for Beginning Teachers®, Subject Matter Examinations require much writing but focus on content such as history, literature, or biology, not on writing ability.

The objective of this review is to survey broadly the entire field of writing assessment, as it is used in admission to higher education, and to identify researchable problems. It is anticipated that this review will provide a useful framework for research committees of the College Board and the Graduate Record Examinations Board as they discuss various research issues related to writing assessment.

Test Design: The Construct to Be Assessed

Definition

Construct is a convenient term to use when referring to a theoretical domain of knowledge and skills. Constructs vary greatly in complexity and scope. The construct of typing ability, for example, can be defined quite narrowly compared to such complex constructs as mathematical ability or musical aptitude. The construct of writing—or writing ability—is so complex that, as one researcher noted,

> *Logically and empirically, there seems little merit in references to a unitary construct [of] writing.* (Purves, 1992)

Reporting a news event, narrating a story, describing a scene, critiquing an argument, proposing a solution, revising a memo, deciding which grammatical construction to use—all of these belong to the theoretical domain of writing ability, as do a range of cognitive skills (e.g., interpreting, analyzing, synthesizing, organizing) and various kinds of knowledge (e.g., knowing effective ways of introducing a topic or understanding linguistic structures).

Even if a unitary construct of writing *could* be defined, no single test could possibly assess the full domain. Therefore, testing programs need to be very clear about the specific skills and knowledge a test is designed to assess so that test users can determine whether the construct that is actually assessed is appropriate for a particular purpose and population.

To understand the difference between the theoretical construct (writing ability) and the construct actually tested, it may be useful to refer to the schema (Figure 1) used by Willingham and Cole (1997) to represent three facets related to test construct: knowledge, skills, and task format. The three facets overlap, and the construct tested lies somewhere in the center of this overlap. Task formats are represented with a dashed circle to indicate that they help determine the construct but are not the trait being assessed. Willingham and Cole explain the contribution of format as follows:

> *Format is one of the facets because the assessment method may affect the nature of the knowledge*

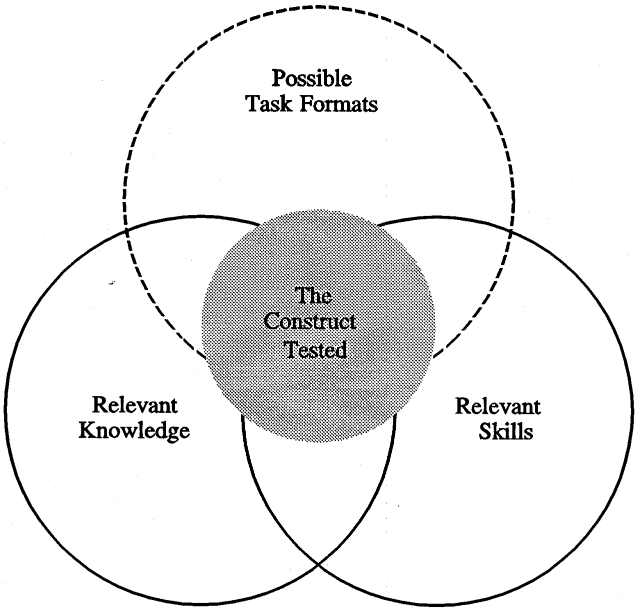

Figure 1. Three-facet representation of a test construct. (From Willingham and Cole, 1997.)

> *and skills that are measured. Format is subsidiary because, strictly speaking, we are interested only in the knowledge and skills. Nonetheless, assessment format may be critical either in the limitations it imposes or the opportunities it creates in measuring different constructs (p. 230).*

In writing assessment, task formats typically include open-ended essay prompts, multiple-choice questions, or paragraphs needing revision. The format may have multiple parts, and its stimulus may include any kind of text (e.g., a brief statement, an extended passage, dialogues, e-mail messages) or graphics. There are many options, and test designers need to base their decisions on meaningful data as well as on clear consensus from all relevant perspectives.

How Constructs Are Defined— The Test Development Process

The process of defining a valid construct for a writing test is multiphased: It may include surveys, focus-group sessions, committee deliberations, and reviews of research literature as well as writing curricula and standards. The process also involves extensive pilot testing designed to evaluate the effectiveness of various formats, directions, topics, and evaluation criteria and to find out what the students themselves think about the tasks they are asked to perform. The goal of these activities is the articulation of the specific aspects of writing ability to be assessed. This articulation usually takes the form of test specifications, which include a verbal description of the process employed, the construct definitions, and justification for these definitions. Specifications also include a description of task formats, directions, timing, response mode and scoring criteria/methods as well as justification of how these relate to the construct being assessed.

At the center of nearly every writing assessment program is a committee of teachers who understand the writing abilities and interests of the examinees as well as the writing expectations of the schools to which the students are applying. Members of the GRE Writing Test Advisory Committee, for example, are undergraduate and graduate teachers of nursing, English, history, economics, physiology, psychology, and political science. Although they represent different academic perspectives, all require writing as an integral part of the curriculum. Through a process of discussing the writing traits that are important for advanced study in all disciplines and then exploring different ways of assessing those traits, the Committee obtained the information it needed to define a construct for the GRE Writing Test. Its recommendations were then further validated through a series of research studies, many of which are summarized in this paper.

After the construct has been defined and the task formats, directions, timing, response mode, scoring method, and scoring criteria have been specified, the Committee and/or test developers begin writing questions designed to meet the test specifications. After reviews for fairness, clarity, and appropriateness of content, the questions are administered to a representative group of students. At this point, the Committee members (or, for some programs, a group of experienced readers) evaluate students' responses before recommending that a question be used in an operational form of the test. Among the factors they consider is the extent to which each question is worded clearly; seems accessible to all examinees; and elicits appropriately complex, varied, and engaging responses from the writers.

Influence of Writing Theory and Practice

Despite the considerable body of research in writing instruction and writing assessment, the question of how best to identify which writing skills should be assessed has rarely been the subject of formal research. However, the process of making these decisions is influenced to varying degrees by current theory and practice.

In the previous discussion of the construct of writing, it was observed that, for practical reasons, the construct tested is always less comprehensive than what might be idealized as a theoretical construct. The question then arises: What important aspects of the theoretical construct are missing from the construct that is assessed?

One approach to answering this question is to examine the results of protocol analysis of the writing process (e.g., Collins and Gentner, 1980; Flower and Hayes, 1981; Hayes, 1996). Figure 2 presents a recursive model of the writing process consisting of three basic elements: the task environment, cognitive writing processes, and the writer's long-term memory. The cognitive writing processes consist of planning, text generation, and revision. Comparing this model of writing to the usual writing assessment situation, with its imposed time constraints, reveals that planning and revision are two aspects of the model that assessments rarely emphasize. A 40- to 60-minute task may allow some time for planning, and, if the examinee uses the time wisely, some time for revision as well. Nevertheless, these two aspects of writing must necessarily receive less attention than text generation.

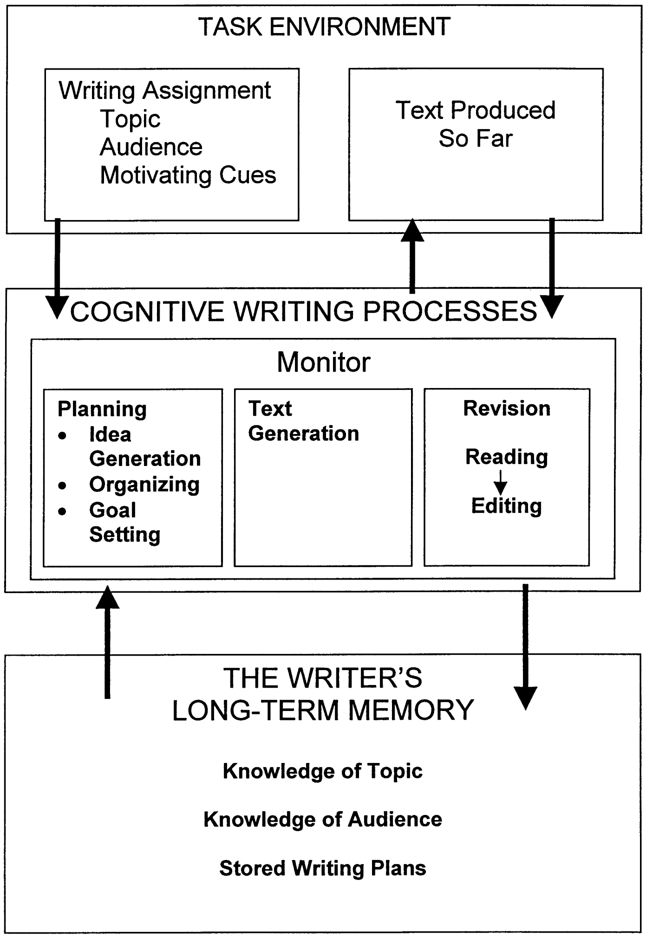

Figure 2. The Flower and Hayes Model of Writing.

The National Assessment of Educational Progress (NAEP) also recognizes that planning and revision are difficult to incorporate into its writing assessment. Beginning in 1992, a blank planning page was added, and students were encouraged to use this page to plan their responses. In addition to rating the quality of the students' writing, the NAEP raters coded whether students had used the planning page. The results for 1992 showed that only 18 percent of the twelfth graders made use of the planning page provided. Those who did make use of the planning page performed significantly higher on the writing task than did those who left it blank (Mullis et al., 1994).

Questions in the 1992 NAEP writing assessments also asked students about the revising and editing strategies they used, including their attention to writing conventions and to the structure and organization of the text as a whole. Eleventh graders who reported that they revised frequently (e.g., more than half the time corrected their grammar, changed words, added ideas, took out parts they did not like, moved sentences or paragraphs) outperformed students who reported that they never or hardly ever did so (Mullis et al., 1994).

If planning and revision are important parts of the writing process, as both the protocol analyses and the NAEP results suggest, then it would seem important to include them in assessments of writing. As Resnick and Resnick (1990, 1992) argue, if it is not tested it will not be taught. Just how to integrate planning and revision into writing assessments is not clear, however. Additional time could be added in test administration procedures to allow for revision, but it may not be possible to isolate an examinee's revision skills from other writing skills and to score them separately. Thus, allowing additional time for revision would probably not improve reliability, although writing performance might improve. Including planning in writing assessments would appear to be an even more difficult challenge. The GRE Writing Committee developed an essay writing/reflection task and administered it experimentally to study, among other things, the way students might have planned and composed a different essay had they additional time and resources or a slightly different assignment. However, after reading and discussing the students' reflective pieces, the Committee identified so many scoring problems that it rejected the task format.

Computer scoring may be well suited for assessing revision. Davey, Godwin, and Mittleholtz (1997) report on an assessment developed by ACT Inc.'s ACT program that presents an examinee with text that needs correcting. The entire assessment is administered and scored by computer. The GRE Board has also sponsored some experimental revision assessments (Breland, 1998) that could be scored on the computer.

Construct Validity

The construct of a writing test cannot include all relevant knowledge and skills; conversely, a poorly designed test may assess knowledge and skills that are irrelevant to the purpose of the test, and, as a result, weaken the validity of a test. According to Messick (1989, 1994), the two greatest threats to construct validity are *construct irrelevant difficulty* and *construct underrepresentation.*

Construct-irrelevant difficulty occurs when the tasks include features that to some degree prevent an examinee from demonstrating the relevant traits. Irrelevant difficulty may occur in a writing assessment when any part of it (e.g., an essay topic about living in the city or a requirement that the essay be word processed) is so unfamiliar to some examinees that they cannot adequately display their ability to write an essay. Construct-irrelevant difficulties can arise from problems in coherence (examinees may not understand the directions), vocabulary (examinees may misinterpret idiomatic phrases in a test question), or content (a topic could mistakenly assume that all examinees live in the United States or have had certain kinds of educational experiences). Construct irrelevancy can also be caused by decisions to require all examinees—even those with little computer experience—to word process their essays. To avoid construct irrelevance caused by problematic topics, most testing programs try out their essay questions with a representative sample of examinees and revise or reject those questions that create unintended difficulties. To avoid construct irrelevance caused by other design issues (position in test, choice of topic), programs should research or debate these issues and provide rationales for each decision.

Construct underrepresentation occurs when it is not feasible to assess all important aspects of a construct. Examples of construct underrepresentation include the use of only multiple-choice questions testing grammar and sentence structure when the assessment is intended to assess critical writing skills; the use of extremely constrained time limits, allowing no time for planning or revision; or the use of only narrative writing tasks when analytical writing skills are an objective. To avoid construct underrepresentation, testing programs conduct surveys and appoint faculty committees to help develop test specifications and scoring guidelines to ensure that the most important writing skills are assessed in the most appropriate context.

Number and Type of Tasks

The challenge for writing committees is to design an assessment that measures an appropriate range of writing skills and yet is also cost-effective and efficient. Rarely does a testing program administer a single writing task without a multiple-choice section or additional writing tasks. Exceptions are some “low-stakes” tests such as placement exams or outcome measures that do not report individual scores. A single essay test is not likely to yield individual scores that are reliable enough for admission purposes.

Different programs use different task formats, depending on the particular skills and knowledge they wish to assess. If descriptive-writing skills are important, the task might present pictures for the writers to describe. If the goal is to assess summary-writing skills, the task might include texts for writers to read and then summarize. To assess analytical writing, a task will probably present data or text for the writer to analyze. Task formats need to be well suited to the construct being assessed.

Multiple-choice tests tend to assess students’ ability to choose revisions that would improve text by making its language more precise, correcting its grammar or punctuation, improving its sentence structure or organization of ideas, adding relevant reasons or details, and so on. Essay tests typically require students to use these same writing skills in an integrated way, and at the same time display their ability to effectively conceive, synthesize, and articulate their own thoughts about an issue, an argument, a proposal, or some other stimulus. If time permits, test takers may revise their essays, but rarely do standardized tests evaluate the test takers’ ability to revise their own writing.

Effects of Technology on Test Design

Until recently, writing assessments required that all examinees write on the same topics on the same day, using only paper and pencil to compose their responses. Hundreds of readers would gather at a single site to evaluate the papers, and scores were reported a few weeks later. Although that model still exists, technological developments have allowed other models to emerge, not only in the way tests are delivered and scored but also in the way they are designed.

Multimedia Writing Assessment

Multimedia writing assessment currently under development uses CD-ROM technology to incorporate audio and video stimuli into writing assessment tasks. In the Core Skills Series: Communication Skills Module, for example, students perform reading, writing, and listening tasks in a simulated workplace environment.

They listen to voice mail messages, attend virtual meetings, select documents from files, write e-mail messages, and draft complex position papers on the computer. It is expected that schools will have the option of all-automated scoring or scoring by both the computer and faculty-readers.

Programmatic Considerations

Each testing program responds to the challenges of writing assessment in slightly different ways. ACT, GRE, GMAT, LSAT, MCAT, SAT II: Writing Subject Test, PRAXIS, and TOEFL all assess writing proficiency, yet—because their purposes and populations differ—they assess somewhat different writing skills and knowledge. Table 1 summarizes and compares the writing assessments used in a number of different programs.

Task Design

Timing of Tasks

Essay-writing tasks vary greatly in their timing. Differences in timing may be explained to a large degree by differences in the test's goals, construct, population, and use of test scores. For example, the SAT II: Writing Subject Test includes a multiple-choice section and a 20-minute essay that is based primarily on personal experience; analysis is not specified in the essay directions or scoring guide. TOEFL's Test of Written English allows 30 minutes for a similar kind of personal-experience essay, but here the testing population speaks English as a second language. When the test construct includes analytical writing, several major programs (GRE and California State University's English Placement Test, for example) allow 45 minutes for a well-reasoned response. The University of California allows even more time for its Subject A essay—120 minutes—because students must read and think about a long stimulus—two pages of text—before they begin writing. In this case, the longer time is necessary because the construct includes both college-level reading and analytical writing skills.

Decisions about how much time to allow examinees to write should be based primarily on pretest evidence. Practical considerations such as cost and scheduling also influence timing decisions, but writing committees need first to examine the quality of writing that representative groups of examinees produce under actual testing conditions before they can recommend the timing for a particular kind of task.

Allowing more time to write essays appears to result in a moderate increase in scores for TOEFL (TWE®) essays (Hale, 1992), prototype SAT II: Writing Subject Test essays (Bridgeman and Schmitt, 1997), and prototype GRE essays (Powers and Fowles, 1996). In each study, the essays were intermingled and evaluated by readers who did not know which timing was used for a given essay. Hale compared 30- and 45-minute timings and found that the extra 15 minutes yielded score gains of about .3 on the TWE 6-point holistic scale. Bridgeman and Schmitt reported that, on an 11-point SAT II: Writing Subject Test holistic scale, with a standard deviation of about 2.0, scores were about 0.8 points higher for the longer (30-minute) essays than for the essays written in 20 minutes. Powers and Fowles compared 40- to 60-minute time limits on two different pairs of essay topics and found very similar results (differences of 0.7 to 1.0 on an 11-point holistic scale with standard deviations of 1.9 to 2.2).

Merely finding that scores are higher with more time does not indicate whether the different timings have an impact on the construct assessed or on the relative standing of groups or individuals. Hale found that, despite the mean-score gain, students were rank-ordered the same way with both 30- and 45-minute timings, and that gains were comparable for students in intensive English programs versus those in regular academic courses. Bridgeman and Schmitt, on the other hand, found that changes in the time limit realigned the relative positions of men and women. Specifically, they observed that gender differences (favoring females) increased as the time limit became shorter. Note that this result cannot be explained by a reliability decline with shorter essays because differences *increased* with the stricter time limits. However, in a study comparing 55-minute to 90-minute essay scores for the California Bar Examination, Klein (1981) failed to find a differential gender impact of increased time. Furthermore, Powers and Fowles found that the stricter time limit did not especially disadvantage students who identified themselves as slower-than-average writers who usually feel pressured on timed tests. Powers and Fowles also examined correlations of essay scores with independent estimates of writing skills in a number of areas and concluded that "there was no detectable effect of different time limits on the meaning of essay scores, as suggested by their relationship to several non-test indicators of writing ability" (p. 433).

Topic Choice

An important goal of standardized assessment is to make certain that all examinees face the same task. On a multiple-choice examination, this is typically

TABLE 1

Programmatic Comparisons

Program	*Test Purpose*	*Tasks*	*Mode of Response*	*Presentation of Topics*	*Choice of Topic*	*Unique Writing Skills Assessed*	*Special Knowledge Assumed*
GMAT	Admission to graduate schools of business management	30-minute critique: Analysis of an Argument 30-minute essay: Analysis of an Issue	Examinees must word process their responses	Computer selects topic from a prepublished pool Prompts often focus on business related topics	No	Identify important features of the argument Present a well-reasoned analysis of the complexities of the issue	An interest in business related issues Word processing skills
GRE	Admission to graduate schools, all disciplines	30-minute critique: Analysis of an Argument 45-minute essay: Present Your Perspective on an Issue	Examinees choose whether to word process or handwrite	Computer selects topics from prepublished pool	Yes (choose 1 out of 2)	Identify important features of the argument Present a well-reasoned analysis of the complexities of the issue	
LSAT	Admission to law school	30-minute Argument	Handwrite	Impromptu	No	Write an argument to a specified audience	
MCAT	Admission to medical colleges	Two 30-minute writing samples	Handwrite	Impromptu	No	In a unified response, interpret a statement, describe a situation that supports (or contradicts) the statement, and discuss relevant factors or principles	
SAT II: Writing Subject Test	Undergraduate college admission	20-minute essay 40-minute 60 M-C items	Handwrite	Impromptu	No	Identifying sentence errors, improving sentences, and improving paragraphs	
TWE	Admission of foreign students to U.S. colleges	30-minute essay	Handwrite	Impromptu	No		

accomplished by giving the same questions to all examinees or using different questions that have been equated by rigorous statistical procedures. On an essay examination, standardization may be interpreted to mean that all examinees respond to the same question, but allowing examinees some topic choice may increase the validity of a writing assessment. For example, suppose an essay topic were to "Explain what you admire most about Thomas Edison." This assignment is highly standardized in that all examinees are focused on the same person. However, examinees may differ widely in how much they know about Thomas Edison. Thus, if the question was intended to assess what the student knew about this particular person, it could be valid and fair. But if the question was intended to elicit expository writing skills, then the highly standardized version would be both unfair and invalid. Allowing some degree of choice within a single question (e.g., "Identify a person you admire and explain what you most admire about her or him") would improve standardization and validity for a test of writing skills.

One way of allowing choice is to write broad topics (such as the "Identify a person . . ." above) that allow the writer considerable freedom within the specified topic; another way of giving writers more control over their test-taking experience is to allow choice among several different topics. When the PRAXIS Writing Committee recommended topic choice (one out of two), Powers and Fowles (1998) studied the effects of offering choice. A sample of undergraduate students read 20 topics and indicated their preferences. Then each student wrote two 50-minute essays. For each essay, the student selected one of two topics. Comparisons were made between performance on their high- and low-preferred topics and between the relationship of certain variables (e.g., undergraduate grades and admission test scores) to performance on each kind of topic. Although student preferences varied considerably (e.g., topics most preferred by some students were often least preferred by others), these preferences exhibited little if any relationship to essay scores. The results of this study, then, suggest that offering choice has little or no impact on scores and yet supports the argument from writing instructors that students should write about topics that draw on their individual interests and background knowledge.

Topic choice has also been examined in tests of content knowledge where, despite efforts to keep all topics at comparable levels of difficulty, it is presumed that some topics may be harder than others. Using essays from the Advanced Placement Program (AP), Pomplun, Morgan, and Nellikunnel (1992) showed that some topics seemed to yield higher scores than others, even for examinees who had comparable scores on the mandatory parts of the examination. Although examinees can potentially be hurt by making bad choices, one study suggests that students taking AP history examinations are more likely to make the right choice than the wrong choice (Bridgeman, Morgan, and Wang, 1997). That is, most examinees can identify the topic on which they can write a better essay and receive a higher score. Furthermore, the essays that these students wrote on their preferred topics seemed to be better indicators of their ability in history than the essays they wrote on their less-preferred topics. Specifically, correlations with an independent history test were higher for the preferred topic than for the less-preferred topic.

Powers and Fowles (1998), on the other hand, found that preference for proposed GRE "Present Your Perspective on an Issue" topics was only weakly related to performance on those topics. This relatively weak relationship, compared to the stronger relationship for AP history essays, may reflect the general nature of the GRE topics as opposed to the AP topics that require knowledge of specific historical events.

Although there is evidence to support offering topic choice, other factors could pose problems. If, for example, examinees need to read and think about the stimulus for a considerable time before making a decision, the act of choosing could seriously interfere with the time left for composing a response. The GRE "Analysis of an Argument" task, for example, does not present a choice of topics because of the testing time and intellectual energy that test-takers would need to spend making informed decisions about which argument to analyze. Another factor to consider is the difficulty in maintaining comparable grading standards across a variety of different topics.

Topic Disclosure

Computer-based programs such as PRAXIS, GMAT, and GRE disclose (prepublish) all essay topics for reasons of equity. In order to meet the demands of computer-based testing, a large number of essay topics must be developed and reused. Because essay topics tend to be brief and easily remembered, programs would find it difficult to keep the topics secure. Some test-takers would undoubtedly gain access to the topics and be unfairly advantaged.

Prepublishing, it is thought, is fairer since everyone has the same access to the entire pool of questions. The released pool needs to be large enough (usually 70 to 150 topics per task) so that examinees cannot memorize prepared responses. PRAXIS, GMAT, and GRE encourage examinees to think about the topics and even to sketch out a few ideas before taking the test. Little is known, however, about how disclosed prompts are used for test preparation purposes once a writing assessment has become operational.

Test Administration

Test Delivery

Computer-administered tests

Several large-scale writing assessments (e.g., GMAT, PRAXIS, and TOEFL) are currently being administered on the computer, and others will soon follow (e.g., GRE in 1999). Test administration schedules are usually flexible: Examinees have many more opportunities to take the test at their own convenience. Before taking the test, examinees can read over the entire pool of essay topics, and, in some cases, practice using the word-processing software they will use at the testing center. GMAT requires everyone to use the word processor, whereas PRAXIS allows examinees to choose whether to word process or handwrite their responses. (Although a sizeable percentage (around 30 percent) of PRAXIS examinees handwrote their essays when the program became operational in 1993, more and more examinees are now choosing to word process their essays.) The computer selects a topic (or, if the program allows choice, the computer selects two prompts) from a large pool—anywhere from 50 to 200 operational topics. When the test is over, the essay is sent electronically to a certified reader who evaluates it on a computer screen. Scores are reported fairly quickly (within a few business days), and GMAT is investigating the possibility of "real-time" scoring, which would allow examinees to receive their essays scores as soon as they complete the other parts of the test.

Response Mode

Word-processed and handwritten essays

Programs that offer choice between word processing or handwriting must be certain that scores in the two formats are comparable—a difficult challenge, since several studies have revealed a tendency for readers to give higher scores to handwritten essays. In a study conducted by Powers et al. (1992), each participant handwrote one essay and word processed another. Then the original essays were transcribed into the other mode, and all four variations were evaluated. The handwritten essays (both original and transcribed) tended to get higher scores. In follow-up interviews, the authors found that readers were more forgiving of certain kinds of writing features (e.g., spelling errors) when they occurred in handwritten form. In a second phase of the study, readers were given special training that addressed these differences, and the essays were rescored. Although the score difference was reduced, there was still a tendency to give higher scores to the handwritten essays.

Accommodations for Examinees with Documented Disabilities

Reasonable testing accommodations are provided to allow candidates with documented disabilities (recognized under the Americans with Disabilities Act) an opportunity to demonstrate their skills and knowledge. The ADA mandates that test accommodations be individualized. This means that no single type of test accommodation may be adequate or appropriate for all individuals with any given type of disability.

Examples of accommodations that may be approved include extended testing time, additional breaks, a writer to record answers, a reader to dictate test questions, a sign language interpreter, Braille test formats, enlarged print, and audiocassettes.

Scoring and Reporting

Scoring Issues

Essay scoring for high-stakes testing programs has traditionally involved gathering a large number of readers together in one place for several days of intensive reading. Training takes place during the first day of a scoring session, and various procedures are used throughout the session to ensure that scoring criteria are applied fairly and consistently. A table leader, who is one of the most experienced readers, monitors the scoring process at his or her table, provides additional training if necessary, and adjudicates essays that received discrepant scores from two different readers (e.g., scores of "4" and "2" on a 6-point scale). Readers are monitored to determine how accurately they evaluate prescored sample essays and how closely their operational scores match scores from other readers. Some readers tend to consistently score too high or too low, whereas others may be erratic in their application of the scoring criteria. Although statistical procedures have been developed to adjust for consistent differences in severity and leniency of ratings across readers (e.g., Braun, 1988; Longford, 1994), these procedures are not being used operationally, in part because scoring differences are rarely systematic. Careful reader training and monitoring appear to be the only effective ways to minimize the problems of discrepant scoring. As noted in the section on reliability, these efforts have been largely successful in maintaining high levels of reader reliability. If a program shifts from a group scoring model to remote scoring or other decentralized models, however, the program will need to reevaluate training and monitoring procedures since it cannot be assumed that the effects of group training will still apply.

Remote scoring

One alternative to bringing readers together for scoring in a conference setting is to allow readers to score essays independently in their own homes or offices. The same sample papers and scoring instructions used in the conference setting can be provided to readers in an independent setting, but asking questions and consulting with others on problem papers may be more difficult. Moreover, these readers lack the group interaction that leads to a shared understanding of the construct and interpretation of the scoring criteria. A study comparing the reliability and validity of ratings obtained in the different settings suggests a small advantage for the conference approach (Breland and Jones, 1988). This study relied on mailed instructions; there was no opportunity for interaction among the remote readers and no ongoing monitoring of the scoring. However, with the advent of computerized testing and the Internet, the possibilities of remote scoring are being revisited. GMAT, in fact, is making the transition to Web scoring in 1999. Reader training will be provided at a secure Web site, and monitoring will be conducted by scoring leaders who will be available via telephone and e-mail for one-on-one consultation when essays are being scored. Multiple copies of essays can easily be distributed for scoring, allowing scoring leaders in a central location to compare readers' scores and provide help to any reader who is scoring inaccurately.

Diagnostic scoring over the Web

Other computer-delivered services are changing the way examinees prepare for essay tests. In 1997, the College Board introduced a diagnostic scoring service which operates as follows: Students can prepare to take the SAT II: Writing Subject Test or an AP essay test by logging on to their computers, selecting the appropriate essay questions (actual released forms of the test), and writing an essay response. Their writing is then sent electronically to qualified AP or SAT II: Writing Subject Test readers who evaluate the essays in their homes or offices. Within a few days, students receive their essays with diagnostic comments linked to the scoring guide of the particular test. In fall 1998, the University of California and California State University pilot tested a Web-based diagnostic service that California teachers can use with their students to help them meet the writing standards established by the two university systems.

Computer scoring

Several testing programs are currently exploring the feasibility of using automated scoring either in addition to or in place of human readers in an effort to find more timely and cost-effective methods of scoring essays and other constructed responses. Using Natural Language Processing (NLP) techniques, researchers follow a procedure in which they collect a large number of responses (about 500 per prompt) that have been evaluated by expert readers and that represent the full range of scores. Then the computer analyzes the prescored responses at each score point, looking for certain variables, such as the particular structures, length, and words that distinguish the writing at each score level. The computer assigns weights to those variables to define a model for scoring the responses to a particular prompt. Although research is still preliminary, results have been promising: computer scores range from 89 to 97 percent exact or adjacent agreement with expert human scores.

As research continues with the NLP tools, attention is being paid to the likely consequences of computer scoring and its effects on test validity. Some of the emerging questions include: What are the instructional implications of moving from human readers to computer scoring, particularly the notion of writing to communicate to a particular audience? Are certain kinds of constructed responses more or less suitable for computer scoring? Why or why not? Even if computer-with-human correlations on the first 500 essays are high, will the computer "model" of a particular set of weighted variables generalize over time—or will it be necessary to monitor the computer scoring with routine checks by human readers? Under what conditions is computer scoring actually more efficient and cost-effective than human scoring? Could automated scoring be used to help monitor the accuracy of human scorers, or should computer-generated scores be reported? What features of the construct can computers not assess? And, finally, what are the most compelling reasons for moving or not moving to computer scoring, and how does the public understand/endorse those arguments? Currently we know little about the broader implications of automated scoring.

Reporting Issues

Reporting practices vary across testing programs. At one time, College Board writing tests were not scored at all; instead, copies of the student's essay were sent to designated colleges, as is currently the practice for the writing sample of the Law School Admission Test (LSAT). For many years, College Board writing tests have consisted of a holistically scored essay and a multiple-choice component and have, until relatively recently, reported only a single score based on both components. Since 1994, however, with the introduction of the SAT II: Writing Subject Test, separate scores are reported for the essay and for the multiple-choice component. The Graduate Management Admissions Test (GMAT) administers two essays but reports only a single score: the average of the raw scores. The Medical College Admission Test (MCAT) writing assessment also consists of two essays, but scores are reported only on an alphabetic scale because of concerns that the reliability of the writing part of the MCAT does not meet traditional standards. GMAT, upon request, sends out score reports that include not only the average of the raw essay scores but also copies of the examinees' essays.

In deciding on reporting policies, testing programs must weigh concerns of reliability and test consequences as well as the interests of test users. If two different kinds of topics are administered, test users may want to receive separate scores as well as a combined score. However, scores for individual tasks are less reliable than a combined essay score, and, although a testing program may state that the two writing tasks have equal value, test users might nevertheless rely on the single score to make unwarranted decisions about the examinees.

Consequences of Writing Test Formats

When choosing among available writing test formats, at least two kinds of consequences need to be considered. The first relates to bias. If it can be shown that one writing test format favors some groups of examinees over others for reasons unrelated to the construct, then that is evidence of bias. The second relates to educational consequences. If the writing test format influences curriculum because teachers "teach to the test" or because they or their students place undue emphasis on question types found on national tests, then the test can be viewed as having an influence on the educational system. If people agree that this influence is positive, that is all to the good. The problems arise when at least some groups claim that the format has a negative influence on instruction and learning.

Bias

The results of the present review confirm that gender differences in writing skills do exist: On average, women perform better than men do on both essay and multiple-choice tests of writing (Willingham and Cole, 1997, p. 286). Women who take the SAT II: Writing Subject Test average a little more than a tenth of a standard deviation better than men on both essay and multiple-choice parts of the test. The situation is somewhat different for ethnic and language groups. Minority ethnic groups and groups for whom English is not their best language perform less well on all kinds of English writing tests than do white groups whose first language is English, and the difference is larger for multiple-choice tests. Group comparisons for multiple-choice and essay writing test formats show, however, that minority ethnic groups and groups for whom English is not their best language perform relatively less well on multiple-choice tests of writing than they do on essay tests of writing (Pomplun, Wright, Oleka, and Sudlow, 1992). Compared to white students, African-American students average about two thirds of a standard deviation less on multiple-choice tests of writing but less than half a standard deviation on essay tests of writing. Similarly, Hispanic students average about a half of a standard deviation lower on multiple-choice writing tests but less than a third of a standard deviation lower on essay tests of writing. It has been hypothesized that the differences between multiple-choice and essay-writing test formats for these groups are due to reliability differences in the assessments—that is, essay tests are less accurate measures than multiple-choice tests, and are thus less likely to show differences. Although statistical adjustments are at times made to correct for reliability differences (e.g., Bridgeman and McHale, 1996), it does not usually erase differences.

Educational consequences

Since the mid-1980s, there has been considerable discussion in the literature about the educational consequences of testing. Resnick and Resnick (1990, 1992) have argued that traditional standardized tests distort the school curriculum because teachers tend to teach to whatever tests are being used. Similar arguments were made by Frederiksen and Collins (1989) in questioning the validity of educational tests when the educational system adapts itself to the characteristics of tests. Wiggins (1989, 1993), among others, referred to the kinds of assessments needed as "authentic," and Linn, Baker, and Dunbar (1991) characterized these assessments as open-ended problems, essays, hands-on science problems, computer simulations of real-world problems, and portfolios of student work. Thus, in writing assessment, arguments have been made for more essay testing because traditional multiple-choice measures of writing skills are viewed as having the unintended side effect, or negative consequence, of causing teachers to teach and students to focus on sentence-level problems while ignoring the more global aspects of writing.

Evidence that teachers are influenced by tests used in school settings has been presented by Shepard (1988), Archbald and Porter (1990), and others. In a carefully designed experiment in Ireland, however, where no standardized testing of any kind had been previously used, Kellaghan, Madaus, and Airasian (1982) found few effects on teachers' practices of introducing standardized testing over a four-year period. Whether national tests used for admission to higher education influence school curricula has been debated, but no empirical studies have been conducted.

Several research investigations have concluded that the best assessments use a combination of assessment formats (e.g., Ackerman and Smith, 1988; Miller and Crocker, 1990; Swanson, Norman, and Linn, 1995; Willingham and Cole, 1997). The negative consequences of one test format may be offset by the positive consequences of another test format. Unfortunately, timing constraints do not always allow for the use of multiple test formats.

Differential Item Functioning in Writing Tests

Differential item functioning (DIF) occurs when two groups possessing the same ability levels have significantly different scores on an item. Consequently, DIF refers to the differential subgroup performance on a test item beyond what would be expected by differences in the mean abilities of the subgroups.

Most available methods for detecting DIF have been developed for multiple-choice test items. Problems have been encountered when attempting to apply these same methods to non-multiple-choice test items, such as those used in direct writing assessments. Three problems, in particular, have been encountered:

1. *Polytomous Scoring.* DIF procedures for multiple-choice (M-C) items are designed for dichotomous scores (correct or incorrect), while scores for direct writing assessment items are polytomous (having more than two scores, as in a six-point holistic scale). Polytomous items present a more complex set of problems in scoring and interpretation. DIF can occur in all the polytomous score categories, or it can occur in only a few of the score categories (French and Miller, 1996).
2. *Reliability of Matching Variable.* The matching variable used for M-C DIF procedures should be as reliable as possible, and it usually is for a long M-C test. In direct writing assessment, however, no more than two or three items are usually used; thus, the reliability of the total scores is often quite low. As a result, any matching variable based only on direct writing assessment items is likely to be unstable. Attempts to remedy this problem by using a combination of direct and indirect assessments of writing have so far not been successful (Welch and Miller, 1995).
3. *Unidimensionality Assumptions.* Most available DIF procedures assume that the items studied for DIF measure the same dimension as the matching variable (Dorans and Potenza, 1994). This assumption can be especially troublesome for writing assessments that combine M-C assessments of writing and direct assessments of writing, since these two formats are usually viewed as measuring different constructs of writing skill.

Several attempts have been made to adapt available DIF procedures to polytomously scored items (Chang and Mazzeo, 1994; Chang, Mazzeo, and Roussos, 1996; French and Miller, 1996; Muraki, 1993; Miller and Spray, 1993; Welch and Hoover, 1993; Zwick, Donoghue, and Grima, 1993; Zwick and Thayer, 1996). None of these attempts has been totally successful for a number of reasons. The choice of a variable to use for matching on ability represents a special challenge. Questions that are scored polytomously tend to take longer to complete. Consequently, these tend to have fewer questions and a lower reliability if the matching variable is limited to these kinds of items. In tests consisting of both dichotomous and polytomous items, as for some writing tests, the use of both types in the matching variable increases its reliability but risks combining items which may not measure the same skills (Allen and Donoghue, 1994).

The disappointing results of studies that have attempted to apply existing DIF methodologies to polytomous items suggest that the greatest promise may lie in entirely new methods. One new approach is to use background variables for matching. For estimating DIF on a writing test, for example, self-reported writing ability and type of English class enrolled in might be used for matching. There are conceptual problems with this approach, however, because background variables may relate to group membership. IRT-based methods represent another possibly promising approach (Muraki, 1998), especially multi-dimensional IRT-based measures (Oshima, Raju, and Flowers, 1997). Until better statistical methods are developed for detecting DIF in direct writing assessments, judgmental evaluations represent the best available means for the examination of DIF in these kinds of assessments.

Equating Different Forms of Writing Tests

A final fairness issue is whether examinees taking a test at different administrations have equivalent opportunities to score well. Essay prompts are rarely equated unless they are part of a larger writing test that includes multiple-choice writing questions. The SAT II: Writing Subject Test uses multiple-choice writing questions (usage and sentence completion) to adjust the raw (readers') scores. This process is generally considered valid when the two sections (multiple-choice and essay) assess at least part of the same construct, and the correlation between the two is high. It is not feasible to equate different forms of essay-only tests, such as the GMAT Analytical Writing Assessment and the GRE Writing Assessment, because two essays do not provide a statistical basis for equating. Since it is not feasible to equate different forms of essay-only tests, special attention needs to be given to the comparability of the prompts used for different administrations.

Reliability and Generalizability

The reliability of a test refers to the consistency of scores obtained over different occasions. Generalizability is similar to reliability but is a more precise term for statistical purposes. In this paper, the two terms are used more or less interchangeably. Whatever the terminology, the principal objective is to estimate the likelihood that a given assessment will yield the same score for a given examinee when administered over two or more occasions.

Essay Tests

For essay tests of writing, reliability refers to the generalizability of scores both across raters and across tasks. According to Anastasi (1982), a reliable assessment of an individual's writing skill is based on several essays on different topics, written on different days, and judged by different readers. Practical constraints, however, limit the number of essays that can be administered in most large-scale testing situations.

The importance of the number of topics or tasks and the number of raters of each is indicated in experimental research on writing. Godshalk, Swineford, and Coffman (1966) administered five different topics to over 600 high school students, and the responses of each student were read and scored by five different raters working independently. The results, summarized in Table 2, suggest that a reasonably high reliability (.84) can be obtained using five different topics, each read and scored by five different raters, but that relatively low reliabilities result when fewer topics and fewer raters are used. For example, for a single topic rated by two readers, the estimate of reliability was only .38, and for two topics—each read by two raters—the estimate of reliability was .55. Careful examination of the procedures used for this study show that although 25 ratings of each student's writing were obtained, the ratings were made on only a three-point holistic scale—"superior," "average," and "inferior"—and that the readers had no experience in evaluating brief, extemporaneous essays. Moreover, although examinees were allowed 40 minutes (per essay) to write two of the essays, they were allowed only 20 minutes for each of the three other essays. Consequently, questions can be raised about the accuracy of the reliability estimates obtained, especially given the major advancements made in reader training since 1966.

Because of the possible limitations of the Godshalk et al. study and because of questions about the meaningfulness of its results over time, an improved replication of the study was conducted 20 years later. In this second study, Breland, Camp, Jones, Morris, and Rock (1987) used six different essay topics requiring three different modes of discourse. The topics were administered both as in-class and as take-home assignments. Two narrative essays and two expository essays were written in class, with 45 minutes allowed per essay. Two persuasive essays were begun in class but finished as take-home assignments. Each essay was read and scored by three highly experienced readers using a six-point holistic rating scale. As shown in Table 3, a reliability of .88 was obtained using all six essays and three readers per essay. Table 3 also shows that if one extrapolates the data to nine topics, with four readings of each, a reliability of .93 is estimated. As in the Godshalk study, lower reliabilities were estimated for fewer topics and fewer readers. One topic with two readers yielded a reliability estimate of .53 (compared to Godshalk's estimate of .38). Two topics with two readers of each yielded a reliability estimate of .70 (compared to Godshalk's estimate of .55).

TABLE 2

Essay Test Reliability Estimates From a 1966 Study*

Readings per Topic	*Number of Topics*				
	1	*2*	*3*	*4*	*5*
1	.26	.42	.52	.59	.64
2	.38	.55	.65	.71	.75
3	.44	.62	.71	.76	.80
4	.49	.66	.74	.79	.83
5	.52	.68	.76	.81	.84**

*From Coffman (1966)
**Based on empirical data

TABLE 3

Essay Test Reliability Estimates From a 1987 Study*

Number of Modes	*Topics per Mode*	*Total Essays*	*Readers per Essay*			
			1	*2*	*3*	*4*
1	1	1	.42	.53	.58	.63
	2	2	.59	.70	.74	.77
	3	3	.69	.77	.81	.84
2	1	2	.57	.68	.72	.76
	2	4	.73	.81	.84	.86
	3	6	.80	.86	.88	.90
3	1	3	.66	.75	.79	.82
	2	6	.80	.86	.88**	.90
	3	9	.85	.90	.92	.93

*From Breland et al. (1987)
**Based on empirical data

In 1991, Dunbar, Koretz, and Hoover reviewed reliability estimates from a number of studies, including those of Godshalk et al. (1966) and Breland et al. (1987). A principal focus in this review was the relative importance of the number of tasks as compared to the number of readers of each task. In the six studies reviewed, the number of tasks used was shown to markedly increase reliability, although the number of readers had little effect beyond about three readers. The average reliability for a single writing task was about .50, and that for 10 tasks was about .88. However, the average reliability using a single reader was about .60, but average reliability for 10 readers increased to only about .70. This review, as well as individual studies, indicates that what is most important in maximizing the reliability of essay tests is the number of tasks used.

In a hypothetical analysis of writing portfolios, Reckase (1995) has shown that the reliability of the portfolio depends not only on the number of entries in the portfolio but also on the correlations between the scores for each entry. In other words, it is better if the entries are similar rather than disparate. A writing portfolio consisting of five entries, each having a reliability of .55, would have an overall reliability of .79 if the inter-entry correlation were .28, but the overall reliability would decrease to .73 if the inter-entry correlation were only .16. If the reliability of each entry were only .43 and the inter-entry correlation were only .16, a five-entry writing portfolio would have an overall reliability of only .65. In this analysis, the reliability of portfolio components and the intercorrelations among these components were estimated from actual portfolio data.

Writing assessments using two essays are of particular interest because the cost and practicability of administering more than two essays is usually problematic for large-scale assessments. In addition to the Godshalk et al. (1966) and Breland et al. (1987) estimates, reliabilities have been reported for two-essay assessments by Clemson (1978), Finlayson (1951), Moss, Cole, and Khampalikit (1982), and Steele (1979). The mean and median reliabilities reported for two-essay assessments for all samples in all of these studies are .71 and .73, respectively. Therefore, it seems reasonable to expect other well-designed and well-conducted two-essay assessments of writing to have similar reliabilities.

With the above experimental and hypothetical analyses as background, the reviewers examined reliability estimates obtained in various testing programs that administer two essays. Since 1992, the Medical College Admission Test (MCAT) has included a writing assessment. Examinees write two 30-minute essays, and each essay is evaluated by two trained readers working independently. Koenig (1995) reports that, between 1992 and 1994, split-half reliability estimates ranged from .77 to .80 for six test forms, with an average of .78. For the same period, generalizability estimates ranged from .76 to .86, with an average of .79. Test-retest reliability estimates, based on examinees repeating the assessment, ranged from .61 to .68, with an average of .64. The split-half and generalizability estimates for the MCAT are somewhat higher than Godshalk's in 1966 (.55 reliability estimate for a two-essays with two readers of each essay) and slightly higher than Breland's in 1987 (.70 estimate for similar assessment). Of all the studies reviewed that estimate reliability for two-essay assessments, only one (Finlayson, 1951) reported a reliability higher (.88) than those reported for the MCAT. The lower test-retest estimates for the MCAT were undoubtedly related to the time interval between assessments, which ranged from four to seven months.

The Graduate Management Admissions Test (GMAT) has included since 1994 an essay writing assessment consisting of two 30-minute tasks. The two tasks are different: one presents an issue to analyze (writers develop their own ideas about the issue), and the other presents an argument to analyze (writers critique the logical soundness of the argument). Each essay is scored on a 0 to 6 scale by two trained readers working independently. Wightman, (1996, personal communication) reported that, for three test administrations in 1995 (January, March, and June), the alpha reliability estimates for Western Hemisphere examinees ranged from .66 to .72, with an average of .69. These results are very close to the Breland et al. (1987) generalizability estimate of .70 for a two-essay assessment, two readers per essay. The GMAT estimates are also close to the .71 mean and the .73 median for two-essay assessments. Bridgeman and McHale (1996) estimated reliability of the GMAT writing assessment by the split-half method (correlating scores for the two tasks) using an October 1994 sample and obtained estimates ranging from .51 to .77 for different gender and ethnic groups. The lowest estimate was for white females and the highest for Asian Americans of both genders. The authors suggest that the higher reliabilities for nonwhite ethnic groups could reflect a greater range of scores within those groups. However, the standard error of measurement (which takes into account within-group score variability) indicates that measurement error was fairly comparable across all subgroups. Gender differences within ethnic groups were quite small. Wightman (1997, personal communication) made similar estimates for January, March, and June 1995 gender, ethnic, and language samples. Generalizability estimates for females averaged .67 (ranging from .65 to .69), and for males estimates averaged .66 (ranging from .66 to .68). Estimates for examinees reporting that English was their best language averaged .55 (ranging from .54 to .56), whereas estimates for examinees reporting that English was not their best language averaged .76 (ranging from .74 to .76).

Multiple-Choice Tests of Writing

Multiple-choice tests of writing tend to have higher reliabilities than essay tests of writing. Donlon (1984) reported KR-20 reliability estimates for eight different administrations of the Test of Standard Written English (TSWE), a 30-minute test with 50 sentence-level items on grammar and sentence structure, and obtained estimates ranging from .86 to .90, with an average of .88. Also reported were test-retest reliabilities for six different administrations based on repeaters; these test-retest reliabilities ranged from .80 to .83, with an average of .82. The longer (90-item) achievement test in English Composition was reported to have a KR-20 reliability of .92. Breland and Gaynor (1979) also obtained test-retest reliabilities for the TSWE in a study of four colleges in which the TSWE was administered three times using different test forms each time: first when the students were applying for college, next at the end of their first semester of college, and then at the end of second semester. Correlations among the scores obtained on these three occasions represent test-retest estimates of reliability. The correlations were .83 between the first and second administrations, .84 between the first and third administrations, and .84 between the second and third administrations. A total of 836 students were represented in the first correlation, 493 in the second correlation, and 333 in the third correlation. The Advanced Placement English Language and Composition Examination multiple-choice test, a 60-item test, had a KR-20 reliability of .84 reported for 1986 (College Board, 1988).

Other Types of Writing Tests

In contrast to essay and multiple-choice tests, a format known as the "interlinear exercise" presents an essay already written, with errors. The examinee's task is to revise the essay by deleting, inserting, or rearranging words or phrases. Completed exercises are scored by readers in a way similar to that used for holistic essay scoring except that a more detailed scoring guide specifies acceptable changes to the original text. Reliabilities for interlinear exercises once used in College Board testing programs ranged between .77 and .86 with an average reliability of .83 over 18 different exercises.

Summary of Reliability and Generalizability

The general picture that emerges from the studies reviewed is that multiple-choice tests of writing and composites of multiple-choice and essay tests of writing have the highest reliabilities, with reliability estimates usually in the range from .80 to .90. Two-essay assessments, in which both essays are read and scored independently two different times, tend to have slightly lower reliabilities, most often in the range from .70 to .80—although some types of reliability estimates, notably test-retest estimates with a considerable time interval in between test administrations, are somewhat lower. Writing assessments based on single essays, even those read and scored twice, have extremely low reliability—usually less than .60. Thus, it is the sampling of tasks, rather than the number of raters, that is the major source of unreliability.

Predictive Validity of Writing Assessments

Several studies have reported correlations between scores on various types of writing tests and criteria such as grade-point average (GPA), grades in English courses, essay writing performance, instructor's ratings of writing ability, student self-assessments of writing ability, and student-reported accomplishments in writing. Not all of the correlations reported are predictive in the temporal sense in which the predictor data were obtained at an earlier time than the criterion data, but they can still be interpreted as predictive correlations in a statistical sense.

Predictive correlations have been reported for multiple-choice tests, free-response essay tests, combinations of multiple-choice and essay tests, and other types of tests (such as editing exercises for which examinees improve the text presented). Essay tests used as predictors have consisted of either one or two essays, whereas essay tests used as criteria always consist of multiple essays.

The following sections summarize three kinds of studies: (1) those that used grades, grade-point average (GPA), or instructors' ratings as criteria; (2) those that used writing performance as the criterion; and (3) those that used a variety of examinee background indices as criteria.

Studies Using Grades as Criteria

Table A-10 (Appendix) summarizes studies that have used grades, GPA, or instructors' ratings as criteria. In one of the earliest studies (Huddleston, 1954), 763 high school students provided essays and multiple-choice test scores of composition skill. Scores from the essays and multiple-choice tests were then correlated with the students' high school English grades and their high school teachers' ratings of their writing ability. Essay scores were found to correlate .43 with English grades and .41 with teachers' ratings. The multiple-choice writing test scores correlated .60 with English grades and .58 with teachers' ratings.

In the Breland (1977) study, scores on the multiple-choice Test of Standard Written English (TSWE) and an essay pre-test administered in freshman English courses were used to predict first-semester English grades. The mean correlations over four institutions were equal for the TSWE and for the essay pre-test (r = .34). As a reference, the SAT verbal was also correlated with grades, and the mean correlation obtained was .29.

Michael and Shaffer (1978) correlated essay scores from the English Placement Test (EPT), used in California State Universities, with fall GPA and fall English grades and obtained correlations of .21 and .31, respectively. An in-class essay was also used, which correlated .25 and .32 with fall GPA and fall English grades. Michael et al. (1980) studied 30-minute essays as predictors of cumulative college GPA and obtained a correlation of .40. All essays were scored holistically by two different readers.

In a study with six different institutions, Breland et al. (1987) used the TSWE and a single 45-minute essay scored twice holistically to predict fall English grades and obtained correlations of .41 and .43 for the TSWE and the essay, respectively. The SAT verbal appeared to correlate about as well as the writing tests (r = .44). The summation of holistic scores from two essays were also used, and a correlation of .46 was obtained. However, the highest correlation (.50) was obtained when a single 45-minute essay was combined with the TSWE.

Bridgeman (1991) reported on two studies of college students. In one study involving students at four New Jersey colleges, scores on the New Jersey Basic Skills Test (a multiple-choice test of "sentence sense" and a 20-minute essay scored twice holistically) were correlated with freshman GPA. An average correlation of .32 was obtained for the multiple-choice test, compared to .21 for the essay test. When the two scores were combined into a composite, the average correlation was .31. The SAT verbal score was also correlated with freshman GPA, and an average correlation of .31 was obtained. In the second study, essay and multiple-choice test scores from the English Composition Test (ECT) for more than 6,000 students in 21 colleges were correlated with freshman GPA. The mean correlations over the 21 colleges were .30 for the multiple-choice test and .17 for the essay test. The mean correlation for the SAT verbal, used as a reference, was .29.

In a study involving students at 14 colleges, Bridgeman and Bonner (1994) examined the predictive validity of the new SAT II: Writing Subject Test, which consists of a 40-minute multiple-choice test combined with a 20-minute essay. The weighted average correlation of the SAT II: Writing Subject Test with grades in regular English courses was .32, compared to a figure of .27 for the SAT verbal score and .31 for the TSWE.

A more recent study (Breland et al., 1999) examined each SAT II: Writing Subject Test component, as well as the composite score, as a predictor of various criteria, including fall English grades. Weighted average correlations for predicting fall English grades at eight colleges were .28 for the essay and .39 for the multiple-choice component, while the composite score weighted average correlation was .49.

Several studies have used writing skill measures to predict performance in law school. Olsen (1956) studied five different multiple-choice tests of writing: combining sentences, editing, organization of ideas, error recognition, and expression situations (a measure of sensitivity to overtones in writing, awareness of audience, and the purposes of writing). Each was a 30-minute test except for combining sentences, which required 40 minutes. Scores on these measures were correlated with average law school grades in four different law schools. The weighted average correlations with average law-school grades were as follows: combining sentences (.12), editing (.20), organization of ideas (.14), error recognition (.26), and expression situations (.13). Although these correlations seem low, the Law School Admission Test (LSAT), which was at that time 2¾ hours in length, correlated only .29 on average over the four law schools.

Pitcher (1962) predicted first-year law-school grades from two experimental tests—an interlinear writing task (editing) and a combining-sentences test. Scores on the experimental tests were correlated both with first-year average law grades and with first-year writing-course grades. The average correlation with first-year average grades was .19 (for both tests). The average correlation with first-year writing course grades was .26 for the interlinear exercise and .23 for combining sentences. LSAT scores correlated .31 on average with average law grades but only .14 with first-year writing course grades.

Pitcher (1962) also correlated writing test scores with first-year law school grades and obtained an average correlation of .21 over 10 different law schools, while the LSAT correlated on average .29 for these same schools. Other analyses were conducted to determine the incremental validity of the writing test, and it was concluded that the writing test did not add to the predictive effectiveness of the LSAT.

Breland, Carlton, and Taylor (1998) studied a new prototype Diagnostic Writing Skills Tests (DWST) for entering law school students. The DWST is a 45-minute multiple-choice test of writing in which examinees are questioned about two hypothetical student responses to a writing task. Scores on the DWST were correlated with first-year writing course grades and with law school grade average. The DWST was a better predictor of performance in legal writing courses than was either the LSAT or undergraduate grade-point average, and it was second only to the LSAT in predicting law-school grade average. Correlations with first-year writing course grades (across five institutions) were .29 for the DWST, .23 for undergraduate GPA, and .21 for the LSAT. Weighted average correlations with law school GPA were .37 for the LSAT, .31 for the DWST, and .19 for undergraduate GPA.

In summary, studies of writing tests as predictors of GPA, English grades, and instructors' ratings suggest that a multiple-choice test may be a slightly better predictor of GPA and that the essay test may be a slightly better predictor of English grades. An unresolved issue is whether a composite test (multiple-choice and single essay) or a two-essay test is the better predictor of GPA and/or English grades. Although only one study was found that examined an essay score based on two essays, the correlation obtained with fall English grades (.46) would appear to make two essays a promising alternative.

Studies Using Writing Performance as the Criterion

Because many variables can influence grades, some studies have used writing performance as a criterion. These studies used a comprehensive criterion consisting of several essays, each scored multiple times by different readers. While it would not usually be practical to administer multiple essay assessments for admissions purposes, such assessments can be highly reliable and serve as a direct measure of the skill that is of interest. A summary of studies using writing performance as a criterion is shown in Table 4.

The first study to use a comprehensive writing performance criterion (Godshalk, Swineford, and Coffman, 1966) examined several types of writing assessment consisting of brief impromptu essays and several types of multiple-choice tests. The criterion used was an essay performance criterion based on a set of five essays, each written on a different topic and each scored by five different readers working independently. The summation of scores on each set of essays (the essay performance criterion) was then correlated with scores on several types of writing assessments. Two multiple-choice tests, usage and sentence correction, yielded the highest correlations with the essay-performance criterion (.71 and .70). Two interlinear (editing) exercises also correlated well with the essay-performance criterion (.67 and .64), as did a construction-shift exercise (.64). Three other multiple-choice tests correlated less well with the essay performance criterion: paragraph organization

TABLE 4

A Summary of Studies Examining Validity of Writing Test Scores: Correlations Predicting Writing Performance

Study/Criterion	*Program*	*Year*	*Sample*	*Verbal Test*	*M-C Test*	*One Essay*	*Two Essays*	*M-C + Essay*
Breland et al. (1987)/	CB/TSWE	1985	267	.54	.62	.61	.72*	.72
Five essays, three readings							.66*	
							.58*	
	CB/TSWE		141	.53	.68	.63	.73*	.76
	CB/ECT				.62		.70*	.73
Breland et al. (1999)/	SAT II	1996	173		.53	.38		.56
Eight essays, two readings			112	.58	.44	.21		.48
Coffman (1966)/		1961	300+			.56	.67	
Four essays, five readings								
Godshalk et al. (1966)/	CB/PSAT/	1961	296	.56	.66	.48		.73
Four essays, five readings	NMSQT™		279	.63	.67	.47		.72
			237	.69	.67	.55		.75
			254	.63	.68	.49		.74
Osterlund & Cheney (1978)/	CB/TSWE	1976	42		.61			
Two essays								
Means				*.59*	*.62*	*.49*	*.68*	*.69*

*Based on three readings. All other predictor essay scores based on two readings.

(.45), prose groups (.57), and error recognition (.59). Each of the five essays in the criterion was evaluated as a predictor by correlating its scores with the summation of the remaining four essays. The essay scores did not correlate as well with the performance criterion as did the best multiple-choice tests (usage, sentence correction, interlinear, and construction shift). When only two readings were used for the essay predictors, the usual situation in large-scale testing, the correlations with the four-essay performance criterion ranged from .48 to .55.

Coffman (1966), reporting on the same study, noted that to obtain validity coefficients for essay assessments that were comparable to a one-hour multiple-choice test of English composition would require either two essays, each scored by five different readers, or three essays, each scored by three different readers. Although the multiple-choice tests of English composition correlated better with the writing performance criterion than did single essay tests, the best correlations were obtained when multiple-choice and essay tests were combined. For example, the combination of an essay scored twice and two multiple-choice tests of English composition correlated .75 with a four-essay performance criterion.

One limitation of this study by Godshalk, Swineford, and Coffman was that the essay performance criterion consisted of all brief, impromptu essays with time limits of 20 to 40 minutes. Accordingly, the criterion and the predictor essays were very similar. It would be expected, therefore, that if longer or take-home essay examinations were used for the criterion, the validity coefficients would decrease.

An improvement over the Godshalk design appeared in the Breland et al. study (1987), which used colleges instead of high schools. The six colleges represented a wide range of abilities and were widely dispersed geographically. One was a historically black college and one was an all-woman's college. The criterion was a set of six essays, of which four were written in class with a 45-minute time limit and two were started in class, discussed with peers, and completed as take-home assignments. Finally, two of the essays were to be written in a narrative mode of discourse, two in an expository mode, and two in a persuasive mode. As in the previous study, essay scores for single essays were correlated with a performance criterion based on the remaining five essays. Scores from the TSWE multiple-choice test of writing, obtained when students applied to college, were also correlated with the five-essay criterion. TSWE correlations with the five-essay criterion ranged from .62 to .70, depending on the sample used. Single essays with two readings correlated between .61 and .65 with the five-essay performance criterion. As in the Godshalk study, the highest correlations were obtained when essay and multiple-choice tests were combined to predict the essay performance criterion. For example, when the TSWE was combined with a single essay read twice, the predictive correlations with the criterion ranged from .72 to .77.

The more recent Breland et al. (1999) study also used a writing performance criterion. Unlike the criteria used by both Godshalk et al. (1966) and Breland et al. (1987), in which all subjects wrote on the same topics, the criteria used for this study consisted of various samples of writing produced in English courses of participating institutions. Accordingly, scoring these writing samples was more difficult than scoring essays on the same topics, and it cannot be expected that the scoring reliability will be equivalent. Table 4 shows that the predictive correlations obtained are generally less than those obtained by Godshalk et al. (1966) and Breland et al. (1987). Nevertheless, one advantage of this study is that the criterion data, collected during the fall semester of college, were obtained at a much later time than the predictor data, which were collected when students applied for admission to college. Thus, the correlations shown are predictive in the temporal sense. These correlations suggest that the multiple-choice portion of the SAT II: Writing Subject Test is a better predictor of writing performance in college than the essay portion of the same test. The SAT II: Writing Subject Test composite score, however, was the best predictor of writing performance.

In general, these studies using writing performance as a criterion yield much higher correlations than do studies using grades, GPA, or instructors' ratings as criteria. They indicate that a multiple-choice test of writing is a better predictor of writing performance than is a single essay scored twice, and that a composite score based on a combination of essay and multiple-choice scores is an even better predictor. Some evidence suggests, however, that an assessment consisting of two essays, each scored twice, may be equivalent (as a predictor of writing performance) to an assessment combining one essay and one multiple-choice test.

Studies Using Student Self-Reports as Criteria

The validity studies described above, whether using grades as a criterion or writing performance as a criterion, are often difficult to carry out. Students have to be followed over some length of time to collect the necessary data, or, in the case of writing performance criteria, several samples of writing from each student must be collected and scored. A much simpler approach is to ask students to complete a questionnaire asking about their writing experiences: their grades in English or writing courses, their writing accomplishments, and their own assessments of how well they write in comparison to their peers. Table 5 summarizes two studies which have used this approach.

Powers, Fowles, and Boyles (1996) studied several types of writing tests being considered for the Graduate Record Examination (GRE), including one-essay and two-essay assessments. Table 5 indicates that the one-essay assessment was limited as a predictor of students' self-assessments of writing ability, self-reported writing grades, and self-reported writing accomplishments. A composite of these assessments correlated only .19 with a single 20-minute essay assessment scored twice. A measure consisting of two essays, timed at 40 minutes and 60 minutes, yielded a much higher correlation (.59) with the self-reported composite score.

Breland et al. (1999) also examined student self-reports as criteria. Table 5 shows that a self-report composite score correlated .46 with the SAT II: Writing Subject Test multiple-choice score, .43 with the SAT II: Writing Subject Test essay score, and .51 with the SAT II: Writing Subject Test composite score. The SAT verbal score correlated .38 with the self-reported composite scores.

Like the studies that used grades and writing performance as criteria, these studies using student self-reports suggest that, while a multiple-choice test of writing may be a better predictor than a single-essay test, a two-essay assessment may predict as well as a combined essay and multiple-choice test.

Summary of Predictive Validity Studies

The primary validity evidence available for writing assessments comes from studies of predictive validity. Table 6 summarizes the evidence available for the four kinds of writing assessment. Three types of criteria were used in these studies: English course grades, writing performance, and writing self-assessments. English course grades are usually for college freshman composition classes. Writing performance criteria consist of multiple essays scored by multiple readers. Writing self-assessments are obtained from questionnaires that ask students to rate themselves in relation to their peers on writing skill, to provide grades in writing courses, and to indicate their writing accomplishments (e.g., won essay contest, edited newspaper, published written pieces, etc.).

Table 6 is limited because validity data for certain assessment types and criteria are available from only a single study. In fact, writing performance is the only criterion for which multiple observations of predictive validity are available for all four types of assessment. The median predictive validity coefficients for this criterion show the single essay assessment to have the lowest median predictive validity (.49) and the combined essay and multiple-choice assessment to have the highest median predictive validity (.73). Two-essay assessments and all multiple-choice assessments had similar predictive validities (medians of .67 and .66, respectively) in studies using a writing performance criterion.

TABLE 5

A Summary of Studies Examining Validity of Writing Test Scores: Correlations Predicting Student Self-Reports

Study/Criterion	*Program*	*Year*	*Sample*	*Verbal Test*	*M-C Test*	*One Essay*	*Two Essays*	*M-C + Essay*
Breland et al. (1999)	SAT II	1996						
Self-assessment of writing			206/246	.25	.30	.37		.36
Self-reported GPA			206/245	.30	.42	.30		.43
GPA in writing courses			206/245	.37	.32	.24		.33
Writing accomplishments			206/245	.23	.27	.27		.30
Self-report composite			156/242	.38	.46	.43		.51
Powers, Fowles, & Boyles (1995)		1994						
Self-assessment of writing			299			.22	.44	
Self-reported writing grades			290			.17	.42	
Writing accomplishments			141			.04	.34	
Self-report composite			136			.19	.59	
Means				*.31*	*.35*	*.25*	*.45*	*.39*

TABLE 6

Predictive Validity Evidence for Different Writing Assessments Summarized Across Studies (Correlation Coefficients)

Assessment Description	*Predicting English Course Grades*		*Predicting Writing Performance*		*Predicting Writing Self-Assessment*	
	Range	*Median*	*Range*	*Median*	*Range*	*Median*
1 essay	.28–.43	.36	.27–.63	.49	.04–.43	.24
2 essays	.42–.46	.44	.31–.73	.67	.34–.59*	.41*
M-C	.04–.60	.33	.61–.68	.66	.46*	.46*
1 essay + M-C	.13–.56	.38	.67–.76	.73	.51*	.51*

*Results from a single study only.

Another method for comparing the four writing assessment types is to examine their relationship *within* a single study. Such a comparison is possible within the Breland et al. (1987) study, with some cautions. Note that Table 7 compares predictive validity evidence for a sample of 267 students. The three criteria used were freshman English course grades, writing performance, and instructors' ratings of student writing skill. Two cautions for interpreting these results are: (1) the two-essay assessment is based on *three* readings of each essay (rather than the two readings for all other essays), and (2) both the grade and instructor rating criteria are slightly contaminated in that the essays used as predictors could also have influenced the grades and instructors' ratings. Thus, the predictive validity correlations involving essays could be slightly inflated. The criterion contamination problem does not occur for the writing performance criterion, however.

The predictive correlations for the writing performance criterion would suggest that the predictive validity of a single essay and a multiple-choice test of writing is roughly equivalent (.61 vs. .62). Similarly, two-essay assessments and single-essay plus multiple-choice assessments would appear to be roughly equivalent with respect to predictive validity (both .72, although the two-essay assessment was based on three readings). The two other criteria, grades and instructors' ratings, suggest a similar conclusion for the single-essay versus multiple-choice comparison: They have roughly equivalent predictive validities. Grades and instructors' ratings, however, yield predictive validities slightly less for two-essay assessments than for single-essay plus multiple-choice assessments.

Table 8 shows a comparison of three of the four types of writing assessments using the results from the Breland et al. (1999) study. Unfortunately, there were no data for two-essay assessments in this study, since the data were for the SAT II: Writing Subject Test, which used only a single 20-minute essay. Still, the results are roughly parallel to those of other studies. Single-essay assessments yield the lowest predictive validities; combined essay and multiple-choice assessments yield the highest predictive validities. These results are of special interest because the predictor data were obtained when students applied for admission to college, while the criterion data were obtained during the first semester of college. For most students, this time interval ranged from several months to over a year.

TABLE 7

Predictive Validity Evidence for Different Writing Assessments in a 1987 Study* (Correlation Coefficients)

Assessment Description	*Predicting English Course Grades*	*Predicting Writing Performance*	*Predicting Instructors' Ratings of Student Writing Skill*
1 essay	.43	.61	.47
2 essays	.46**	.72**	.48**
M-C	.41	.62	.48
1 essay + M-C	.50	.72	.56

*From Breland et al. (1987).
**Based on three readings of each of two essays. All other essay scores based on two readings.

TABLE 8

Predictive Validity Evidence for Different Writing Assessments in a 1999 Study* (Correlation Coefficients)

Assessment Description	*Predicting College English Course Grades*	*Predicting College Writing Performance*	*Predicting Student Self-Assessment of their Writing Skill and Experience*
1 essay	.35	.52	.43
2 essays			
M-C	.37	.61	.46
1 essay + M-C	.49	.67	.51

*From Breland et al., 1999.

Fairness and Group Differences

Gender, Ethnic, and Language Effects

This section summarizes differences among gender, ethnic, and language groups based on their performance on essay writing assessments in comparison with multiple-choice tests. Several facts must be considered when interpreting these results. First, group differences will almost always be smaller on a less reliable measure than on a more reliable measure of the same construct. Unreliability adds random noise to scores, and this random component is, by definition, distributed equally among all groups, so the higher the proportion of noise in the scores the smaller the group differences. To the extent that essay assessments are less reliable than multiple-choice tests, they would be expected to show smaller group differences. If an essay assessment is made more reliable, perhaps by using two or three essays rather than a single piece of writing, group differences typically increase. Second, to the extent that essay and multiple-choice tests were, by design, measuring somewhat different constructs, the size and even direction of group differences could shift from one format to the other even if both assessment methods were perfectly reliable. The multiple-choice part of a writing test may require students to recognize specific grammar errors, and the essay portion of the test may require students to organize arguments into a convincing presentation. Though both are measures of writing-related skills, there is no reason to believe that individuals or groups that excel in one of these skills would necessarily excel in the other. As the skills assessed in the two formats diverge even more, group differences may become even more distinct. Thus, group differences on a measure of verbal reasoning and reading comprehension could be substantially different from group differences on a direct writing assessment.

Gender differences across a number of writing assessments are summarized in Table 9 and in Table A-1 (Appendix). Testing programs represented include medical college admission (MCAT), college admission (ECT, TSWE, and SAT II: Writing Subject Test), graduate management school admission (GMAT), and undergraduate and graduate admission for foreign students (TOEFL/TWE). Because the various tests have different standard deviations, direct comparisons of gender differences in the original scale units of the tests are not meaningful. Therefore, differences are expressed in standard deviation units (*d*); the mean for males was subtracted from the mean for females and this difference was divided by the average standard deviation (sum standard deviation for males and standard deviation for females and divide sum by 2). In the table, the "Reference

TABLE 9

Gender Differences on Verbal Tests, Multiple-Choice Writing Tests, and Essays

		Mean *d*			
Test	*Number of Samples*	*Ref. Test**	*M-C Writing*	*Essay*	*Essay & M-C*
ECT	7	–.13	.13	.14	.14
TOEFL/TWE	8	–.14		.08	
GMAT	5	–.19		.07	
SAT II: Writing	6		.07	.13	.10
MCAT	1	–.06		.13	

*Reference test for ECT is SAT-V; for TOEFL it is TOEFL Total (includes reading, listening, structure and written expression); for GMAT it is the verbal score.

Test" refers to a multiple-choice test, which is primarily verbal reasoning, but it may include other constructs in some cases (e.g., in the case of TOEFL, it includes English language proficiency in reading and listening). The M-C test is a multiple-choice test of writing-related skills such as recognizing grammatical and structural problems in sentences. The Essay Test is typically a single expository essay with a time limit of 20 to 40 minutes, although the GMAT was the sum of scores on two essays. Essay scores for all tests were based on the sum of ratings from two independent raters.

Scores on the reference tests consistently favored males by a small margin (average *d* of –.13) while essay tests favored females by about the same margin. Across several samples, the multiple-choice Test of Standard Written English and the multiple-choice portions of the English Composition Test (the forerunner of SAT II: Writing Subject Test) had gender differences, favoring females, that were just about as large as the essay test differences. Because the size and even the direction of gender differences can be quite sensitive to such factors as the selectivity of the sample (Willingham and Cole, 1997), these results should not be generalized beyond the admission tests studied. Indeed, the data presented by Willingham and Cole suggest that in less selective samples the female advantage on essay tests is much larger as indexed by a *d* of over .5 (p. 87).

Ethnic differences on the same three types of tests (verbal test, multiple-choice test of writing-related skills, and essay tests) are presented in Table 10 and in Tables A-2, A-3, and A-4 for African-American, Hispanic, and Asian-American groups, respectively. The African-American/white differences were nearly identical for the verbal reasoning and multiple-choice writing skills tests, but were somewhat reduced for the essay tests. Smaller differences for the essay tests may be at least partly accounted for by reliability differences between essays and multiple-choice assessments. Note, for example, that differences for each of the GMAT essays considered separately are smaller than the differences for the more reliable composite that combines scores from both essays. Bridgeman and McHale (1996) used a statistical adjustment to correct for reliability differences and found that the means for African-American males and for white males differed by the same amount on the essay as on the verbal reasoning test. However, even with the reliability adjustment, the gender difference remained: The mean for African-American females was closer to the white male mean on the essay test than on the verbal reasoning test. This is consistent with the overall pattern of gender differences mentioned above. Low reliability can shrink differences between groups, but it cannot reverse the direction of those differences.

Hispanic/white and Asian-American/white patterns also suggest that differences are smaller for the essay test, though after reliability adjustment this difference disappears. Tables A-5 and A-6 show differences for Hispanic and Asian-American examinees whose first language (FL) is English, and Table A-7 shows differences Asian-American and Hispanic examinees for whom English is a second language (ESL). The general patterns observed previously remain, with differences smallest for the essay examinations. Although it might seem that essay tests would be especially problematic for ESL examinees, the multiple-choice format may be even more difficult for at least two reasons: 1) a student can avoid unfamiliar structures or vocabulary when writing an essay but must recognize errors in these areas on a multiple-choice test, and 2) essays are evaluated for many features, including organization and development of ideas, whereas multiple-choice tests tend to focus on problems in grammar and syntax.

TABLE 10

Ethnic Differences on Verbal Tests, Multiple-Choice Writing Tests, and Essays

		Mean *d*			
Test	*Group*	*Ref. Test**	*M-C Writing*	*Essay*	*Essay & M-C*
ECT	Asian Am.	–.63	–.62	–.38	–.63
	Hispanic	–.43	–.46	–.25	–.45
GMAT	African Am.	–.83		–.67	
	Hispanic	–.68		–.53	
	Asian Am.	–.73		–.75	
MCAT	African Am.	–1.13		–.65	
	Hispanic	–.65		–.32	
	Asian Am.	–.32		–.00	

*Reference test for ECT is SAT-V; for TOEFL it is TOEFL Total (includes reading, listening, structure and written expression); for GMAT it is the verbal score.

Summary and Conclusions

Test Design: The Construct

- *Establishing a valid construct is central to all other considerations about a test.*

  Writing assessment programs should therefore avoid misleading or vague descriptions such as "general writing ability" and instead provide clear contextual descriptions of both what will be assessed and how it will be assessed.

- *Various types of writing assessment are used for admissions to higher education.*

  The constructs (knowledge and skills) assessed by writing tests vary depending on such factors as the educational level and primary language of the examinees, the purpose of the test, the model of writing theory valued by the test designers, assumptions made about writers' skills and interests, expected use of test scores, costs, and the role of technology.

- *The format used to assess writing ability affects the construct that is measured.*

  Multiple-choice tests tend to assess the ability to choose the revision or revision strategy that could improve the coherence or correctness of text. Essay tests, on the other hand, tend to assess the ability of examinees to reflect on a topic and then engage in the process of conceiving, synthesizing, and articulating their own thoughts about the topic. If time permits, test takers may revise their essays, but rarely do standardized tests evaluate the test takers' ability to revise their own writing.

- *Time constraints in essay testing limit the kinds of skills that can be assessed.*

  Because of the time required to compose a thoughtful response, essay-writing tests are necessarily restricted in the kinds of skills and knowledge they can assess. Most large-scale writing assessments present a brief stimulus (usually just a sentence or paragraph) for test takers to read and think about before they begin writing. For reasons of reliability and construct validity, it is often more important to collect a second writing sample than to extend the time for a single task.

- *Test designs should avoid construct irrelevance.*

  Construct irrelevance occurs when the assessment is poorly designed—that is, when unintended dimensions or facets are included that make the task unnecessarily difficult.

- *Test designs should, to the extent possible, avoid construct underrepresentation.*

  Construct underrepresentation occurs when the assessment is too narrow and fails to include important dimensions or facets of the construct. No single test can fully represent a universally accepted construct of writing ability, but a well-designed assessment represents as much as is reasonably possible.

- *Test designs that use different writing tasks and/or formats maximize construct validity by assessing a broader range of writing skills.*

  It is rare for a testing program to administer only one writing task. Most writing assessments—especially if their results are used for admission or selection purposes—include either a writing task and a multiple-choice test (SAT II: Writing Subject Test, PRAXIS) or two writing tasks (GMAT, MCAT). By administering two *different* tasks, GMAT assesses a broader construct than does MCAT, which administers two versions of the same writing task.

Task Design

Timing of Tasks

- *Time limits vary according to the complexity of the task, the purpose of the test, the constraints of the testing environment, and the results of field test trials or other studies.*

  For all essay tests, the timing specifications should be based on the results of well-designed field tests in which the same or similar populations demonstrate that they can write effective responses in the time allowed.

- *Essay scores are usually higher with longer time limits, but additional time may have little impact on the rank ordering of students or on the meaning of essay scores as reflected in other measures of writing ability.*

  One study, comparing 40-minute and 60-minute performances on GRE Issue topics, concluded that "there was no detectable effect of different time limits on the meaning of essay scores, as suggested by their relationship to several non-test indicators of writing ability."

- *Limited data suggest that women do relatively better with shorter time limits than men do.*

  Data collected for the field trials of the SAT II: Writing Subject Test indicated that women scored about .49 standard deviation units higher than men on a 15-minute essay, about .34 units higher on a 20-minute essay, and .26 higher on a 30-minute essay.

Choice of Topics

- *Even though students' topic preferences vary considerably, studies show that allowing students to choose between alternative topics has little if any relation to their essay scores.*

  For some testing programs, allowing topic choice may increase the validity of the assessment and make it seem fairer to the examinee, especially when the choice is easy to make. GRE and PRAXIS offer choice primarily to help test takers feel more confident about the issue they discuss and thus more comfortable with the essay-writing experience. Although other research has addressed the subject of offering a choice of questions that assess content knowledge, we found no other studies that explored the advantages or disadvantages of offering a choice of essay topics in writing assessment.

Comparability of Topics

- *All topics undergo extensive reviews to ensure that examinees receive questions of comparable difficulty across test administrations.*

  As the first step in establishing topic comparability, testing programs submit potential questions to multiple reviews to confirm that they meet test specifications for appropriate content, cognitive complexity, clarity of phrasing, and fairness.

- *All topics selected for operational forms of a writing test must meet the pretest standards established by the testing program.*

  Essay topics are pretested on a representative sample of examinees for a particular program. Then teams of expert faculty-readers meet with test developers to read the responses, discuss how well each prompt performed according to a predetermined set of criteria (e.g., fairness, complexity, comparability, and range of scores), and recommend whether or not a prompt should be used operationally.

Disclosure of Topics

- *Prepublishing essay topics appears to have little effect on writing performance.*

  Although students perceive that their writing skills are better displayed by writing on disclosed topics than on previously unseen topics, prepublishing topics appears to have little effect on writing performance. Little is known, however, about how disclosed prompts are used for test preparation purposes once a writing assessment has become operational.

Test Administration

Test Delivery

- *The number of large-scale computer-administered essay tests is increasing.*

  In 1994, PRAXIS introduced a computer-administered writing test. In 1997, GMAT moved its paper-based Analytical Writing Assessment to computer delivery. TOEFL made the same transition in 1998, and GRE plans to introduce its writing assessment on the computer in 1999. The main advantages appear to be nearly continuous testing (instead of designated testing dates), efficiency (reduced paper handling), quick turnaround time for score reporting, and the opportunity for writers to word process their responses if they wish.

- *New forms of writing assessment may use multimedia.*

  New technology poses new possibilities for the content and contexts of assessment. ETS is currently exploring a writing assessment concept in which business students are placed in a simulated workplace environment where they attend virtual meetings, listen to voice-mail messages, send e-mail messages, and create overheads for presentations. This kind of interactive test administration may be especially well suited for instructional programs with embedded assessment.

Response Mode

- *Some computer-based tests require word-processed essays; others allow examinees to choose the response mode (either handwriting or word processing) they prefer.*

  The issue is primarily one of fairness to the examinees and responsiveness to the needs of the programs to which they are applying. GMAT now requires all examinees to word process their essays (except in special cases) on the rationale that graduate schools of management expect students to be skilled in word processing. PRAXIS offers a choice because much of its population prefers to handwrite and because word-processing skills are not necessarily relevant to all teaching positions.

- *When choice is allowed, the trend is toward greater numbers of test takers choosing to use the word processor.*

  In 1994, about 30 percent of the GRE examinees participating in the essay pretests handwrote their essays; by 1997, the number had dwindled to about 6 percent. Whereas the PRAXIS figures are also declining, a substantial number of PRAXIS examinees still choose to handwrite their CBT essays—about 30 percent in 1997, down from a high of nearly 50 percent in 1993.

- *Readers tend to give higher scores to handwritten than to word-processed essays.*

  In a 1991 PRAXIS study, handwritten essays were typed, and word-processed essays were transcribed by hand. All versions were scored. Average scores were higher for essays scored in the handwritten mode, regardless of the mode in which essays were originally composed. Changes in reader training helped mitigate the differences, but scoring comparability needs to be routinely monitored in operational scoring.

Scoring and Reporting

Scoring Method

- *Holistic scoring with published criteria is the most widely used method of evaluating essays for purposes of admission, placement, and outcomes assessment.*

  Most ETS tests of writing ability use some form of holistic scoring, as do most university placement exams and writing assessments developed by ACT and other companies. Holistic scoring is often endorsed by writing faculty because it supports the view of writing as a constructive mental process that involves the complex integration of many different skills and choices about content, writing strategies, and use of language and conventions. Holistic scoring also allows programs to develop and publish task-specific scoring criteria that can be applied consistently across topics. A review of writing assessment programs reveals that their holistic scoring criteria vary to emphasize unique features of the writing task or the writing evidenced in the essay responses. The scoring guides for GMAT and GRE "Argument" tasks, for example, include references to the writer's ability to analyze the argument's line of reasoning, whereas the scoring guide for TOEFL's Test of Written English focuses on descriptions of writing produced by writers of English as a second language.

- *Holistic scores correlate moderately with other indicators of writing quality, including faculty judgments, self-reported writing grades, writing achievements, etc.*

  When graduate educators judged the quality of GRE essays without knowing their scores, their judgments, on average, corresponded closely to holistic scores assigned by trained readers. However, faculty judgments were not nearly as consistent or as reliable as those of trained readers.

- *Computer-generated essay scores using Natural Language Processing (NLP) tend to correlate well with readers' holistic scores and have a similar pattern of correlations with V, Q, and A.*

  While correlations look promising, little is known about the consequences of replacing human scores with automated scores. Careful research is needed before this approach can be recommended as valid and cost-effective.

- *New scoring methods and uses of technology may be used to provide diagnostic feedback or other kinds of instructional information.*

  Because holistic scoring does not provide diagnostic information, other, more analytic, methods are better suited for instructional purposes. Thus, through the years, the designers of large-scale writing assessments have experimented with various methods of reporting scores for particular writing features or components. Information about why particular approaches succeeded or failed has not been captured systematically across programs. Also, recent developments in Web-based scoring procedures and refinements in Natural Language Processing offer new options for programs to provide individual students with diagnostic feedback about their writing. This is a new and important area to research if large-scale writing assessment is to have a more direct and positive effect on student learning and on the ways teachers integrate assessment into the curriculum.

Reader Training and the Scoring Environment

- *Without specialized training, scorers tend to give higher scores to handwritten rather than word-processed essays.*

  If both handwritten and word-processed essays are to be allowed, additional training and monitoring of readers is needed to counter the possibility of this effect.

- *Traditional scoring with groups of readers brought together at a central location can produce high reader agreement.*

  Discrepancies among readers can range from 0 percent discrepancies to 10 percent or higher, depending on the quality of the training and the experience of the readers. The definition of "discrepant" scores varies. Some programs send papers with two-point differences (on a six-point scale) to adjudication; in other programs, essay scores must be at least three points apart before they are considered discrepant.

- *Remote scoring options (such as scoring over the Internet) may allow more flexibility, but limited* *research suggests that remote scoring might be slightly less reliable as well as less discriminating at the extremes of the score scale.*

Any move to remote scoring should be accompanied by careful evaluation of changes in reader reliability or readers' accurate use of the entire score scale.

Reporting

- *The method of reporting essay scores varies across programs.*

  SAT II: Writing Subject Test reports a composite writing score (multiple-choice and essay) as well as separate scores. GMAT reports only a single averaged score for its two essays. MCAT reports writing scores only on an alphabetic scale because of concerns that the reliability of the assessment does not meet traditional standards.

Test Reliability

- *Test reliability depends primarily on the number of test items used.*

  Multiple-choice tests of writing typically have 50 to 60 items administered in 30 to 40 minutes. Essay tests of writing typically have one or two essays, and each essay represents a single item. The more multiple-choice items or the more essays administered, the higher the test reliability.

- *Essay test reliability is influenced by at least two major sources of measurement error: (1) the degree to which the assessment domain is sampled, and (2) rater disagreement.*

  In general, the assessment domain is not well sampled by a single essay because only one topic and only one mode of writing can be used. This is the principal reason for the relatively low reliability of single-essay assessments of writing. Two essays represent a much better sampling of the domain. Rater disagreement represents an additional source of error.

- *Writing tests with one or two essays usually have reliabilities in the .50 to .80 range.*

  Reliabilities of single-essay tests (with two readers) are usually in the .50 to .60 range, although interrater reliabilities (i.e., recognizing only error due to raters) are somewhat higher. Reliabilities of two-essay tests (each essay read by two readers) are usually in the .70 to .80 range. The GMAT Analytical Writing Ability Test, which is similar to the planned GRE writing measure, has had reliability estimated in the .66 to .79 range over several test administrations.

- *Reliabilities of multiple-choice tests of writing, or writing tests using combined multiple-choice and essay formats, are usually in the .80 to .90 range.*

  For example, over several test administrations, the SAT II: Writing Subject Test has had reliabilities estimated in the .88 to .91 range for its 40-minute (60-item) multiple-choice component and .85 to .91 for its composite score.

Predictive Validity

- *Most validity evidence for writing tests is based on correlations between test scores and criteria such as English course grades, students' self-assessments, faculty ratings, and student writing performance.*

  English course grades are usually limited as a criterion because factors other than writing ability are often involved in the assignment of grades. Student self-assessments of their own writing ability are relatively easy to obtain, but their accuracy may be questionable. Faculty ratings of students' writing abilities are sometimes better than English grades as a criterion, but usually they yield about the same level of predictive correlations as grades. Probably the most compelling criterion to use for predictive validity studies is "writing performance"—a collection of student writing samples obtained under controlled conditions and scored centrally by experienced readers.

- *Predictive validity studies comparing essay tests and multiple-choice tests of writing ability usually show that multiple-choice assessments have higher predictive correlations, but the highest predictions are made using combinations of multiple-choice and essay questions.*

  In a study of the SAT II: Writing Subject Test, for example, correlations with writing performance (i.e., a collection of writing samples) were .61 for the multiple-choice component, .52 for the essay component, and .67 for the composite score.

- *Two-essay assessments of writing, however, yield predictive correlations roughly equivalent to multiple-choice plus single-essay assessments of writing.*

  Although predictive validity studies using two-essay assessments are rare, one study obtained correlations with writing performance ranging from .66 to .73. In contrast, a multiple-choice plus single-essay writing assessment, in the same study, yielded correlations with writing performance in the .72 to .76 range.

Fairness and Group Differences

- *Gender differences on multiple-choice writing tests are comparable in size to differences on essay writing tests.*

  Differences are generally small, but when present, most often show women scoring relatively higher on the essay than on the multiple-choice portion of the examination.

- *Writing tests, whether essay or multiple-choice, yield smaller gender differences than do verbal ability tests.*

  Because of this difference, adding a writing assessment to most admissions tests would tend to result in the selection of more women. If the addition of a writing assessment also results in a quantitative score's receiving less weight, even more women would tend to be selected.

- *Essay tests yield smaller differences among ethnic groups than do multiple-choice tests of writing ability.*

  Asian-American, African-American, and Hispanic groups all score relatively higher on essay assessments than on multiple-choice writing assessments. Further research is needed to determine how much of this relative advantage may be attributed to the lower reliability of the essay assessments. In general, group differences will be smaller on a less reliable test.

- *Relative to their performance on multiple-choice verbal questions, ESL students perform better on essays.*

  Although ESL students would seem to be at a disadvantage writing essays in English, they may be at an even greater disadvantage attempting to deal with difficult vocabulary or sentence-level features that appear in multiple-choice questions.

Research Recommendations

- *Although numerous validity studies of writing assessment at the high school and college level have been conducted, no studies that address the validity of writing assessments for predicting performance in graduate school were identified in this review.*

  One of the difficulties in conducting studies of this type is that of obtaining adequate criterion data. Most graduate schools do not offer writing courses, although many business and law schools do. As a result, there exists no convenient and known setting in which to collect criterion data on writing performance or grades for other graduate students. One approach might be to collect faculty ratings of graduate student writing performance, especially in courses where students write extensively.

- *Studies that address the public or professional acceptance of computer-scored essay assessments are needed.*

  Relatively simple surveys or focus group sessions with parents, students, faculty, or other professional groups could be conducted to determine the general level of acceptance as well as the acceptance of specific kinds of essay scoring practices (e.g., having only one of two scores assigned by computer).

- *Studies that explain differences in writing test scores for different ethnic groups on different writing test formats are needed.*

  In studies of tests of writing, differences between white examinees and African-American or Hispanic examinees were almost always larger for multiple-choice tests than for essay tests. These observed differences are at times attributed to reliability differences between the two types of test format, but no demonstration of that hypothesis has been conducted. Research studies could be designed to test this and other possible reasons for these format differences.

- *Studies that address the construct validity of writing assessments need to be conducted.*

  Demonstrations of construct validity are especially important when consistent differences between groups are observed, as is true for most academic tests. Little is known about how test content or format affect different groups. This kind of research could be useful in the development of instructionally relevant test preparation materials.

- *Little is known about the effects on essay scores of special test preparation materials or strategies.*

  Most writing assessment programs publish information about the test (sample questions and essay responses, scoring criteria, etc.), and some (e.g., SAT II: Writing Subject Test and PRAXIS) have developed diagnostic services and/or instructional programs. We found no studies about the role of test preparation tools or strategies in helping students improve their writing performance.

- *No adequate statistical methods currently exist for evaluating the differential item performance of essay tests of writing.*

If statistical methods cannot be developed, nonstatistical techniques will be needed to examine the comparability of essay prompt types and topics for different gender and ethnic groups.

- *No adequate statistical methods currently exist for equating essay scores.*

  If statistical methods cannot be developed, nonstatistical techniques will be needed to ensure that all examinees have an equal chance for success.

- *Some writing tests allow written responses only in a paper-and-pencil mode, some allow responses only in a computer mode, and others allow a choice between handwritten and word-processed responses, but there has been little research examining the differences in these modes of administration.*

  It would seem to be important that research be conducted to examine any differences in scores obtained from these different modes of administration.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*(2), 117–128.

Allen, N., & Donoghue, J. (1994). Detecting differential item functioning: Current methods and continuing problems. 1994 Proceedings of the Social Statistics Section of the American Statistical Association, pp. 17–26.

American Association of Medical Colleges (1997). *Characteristics of 1996 MCAT examinees.* Washington, DC: AAMC.

American College Testing (ACT) Program (1997). The high school profile report, Normative data: A description of the academic abilities and nonacademic characteristics of your ACT tested 1997 graduates. Iowa City, IA: ACT.

Anastasi, A. (1982). *Psychological testing.* New York: Macmillan.

Archbald, D. A., & Porter, A. C. (1990). A retrospective and an analysis of roles of mandated testing in education reform. Report prepared for the Office of Technology Assessment. PB92-127596. Springfield, VA: National Technical Information Service.

Braun, H. I. (1988). Understanding score reliability: Experience calibrating essay readers. *Journal of Educational Statistics, 13,* 1–18.

Breland, H. M. (1977). *Group comparisons for the Test of Standard Written English.* College Board Report RDR No. 77-78, No. 1 (ETS RB No. 77–15). Princeton, NJ: Educational Testing Service.

Breland, H. M. (1996). *Writing skill assessment: Problems and prospects.* A Policy Issue Perspective. Princeton, NJ: Policy Information Center, Educational Testing Service.

Breland, H. M. (1998). Writing assessment through automated editing. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Breland, H. M., Bonner, M. W., & Kubota, M. (1995). *Factors in performance on brief, impromptu essay examinations.* College Board Report No. 95-4 (ETS RR No. 95-41). New York: College Entrance Examination Board.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* (Research Monograph No. 11). New York: College Entrance Examination Board.

Breland, H. M., Carlton, S., and Taylor, S. (1998). *Program of research on legal writing, Phase II: Research on a writing exercise.* Law School Admission Council Research Report 96-01. Newtown, PA: Law School Admission Council.

Breland, H. M., & Gaynor, J. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement, 16*(2), 119–128.

Breland, H. M., & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology, 74*(5), 713–721.

Breland, H. M., & Jones, R. J. (1982). *Perceptions of writing skill.* College Board Report No. 82-4 (ETS RR No. 82-47). New York: College Entrance Examination Board.

Breland, H. M., & Jones, R. J. (1988). *Remote scoring of essays.* College Board Report No. 88-3 (ETS RR No. 88-4). New York: College Entrance Examination Board.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). *The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test.* College Board Report No. 99-4. New York: College Entrance Examination Board.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education, 32*(3), 319–331.

Bridgeman, B., & Bonner, M. (1994). SAT-W as a predictor of grades in freshman English composition courses. Unpublished report. Princeton, NJ: Educational Testing Service.

Bridgeman, B. & McHale, F. (1996). Potential impact of the addition of a writing assessment on admissions decisions. *Research in Higher Education, 39,* 663–677.

Bridgeman, B., Morgan, R., & Wang, M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement, 34,* 273–286.

Bridgeman, B., & Schmitt, A. (1997). Fairness issues in test development and administration. In W. Willingham & N. Cole, *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Broch, E. (1997). Summary of ethnic, gender, and language proficiency group performance on the GRE Writing Measure and GMAT Writing Assessment. Unpublished draft report. Princeton, NJ: Educational Testing Service.

Buck, G., Van Essen, T., Tatsuoka, K., & Kostin, I. (1997). Identifying the cognitive attributes underlying performance on the PSAT Writing Test. Unpublished proposal. Princeton, NJ: Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English.* GRE Board Research Report GREB No. 83-2R (ETS RR No. 85-21). Princeton, NJ: Educational Testing Service.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391–404.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–353.

Checketts, K. T., & Christensen, M. G. (1974). The validity of awarding credit by examination in English composition. *Educational and Psychological Measurement, 34,* 357–361.

Clemson, E. (1978). A study of the Basic Skills Assessment direct and indirect measures of writing ability. Princeton, NJ: Basic Skills Assessment Program.

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

College Board (1988). *Advanced placement technical manual.* New York: College Entrance Examination Board.

Collins, A., & Gentner, D. (1980). A framework for a cognitive theory of writing. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive processes in writing.* Mahwah, NJ: Erlbaum.

Davey, T., Godwin, J., & Mittleholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement, 34*(1), 21–41.

Donlon, T. F. (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Dorans, N. J., & Potenza, M. T. (1994). Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning (RR No. 94-49). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Psychological Measurement, 4*(4), 289–303.

Engelhard, G., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiental demand, and gender on the quality of student writing. *Research in the Teaching of English, 26,* (3), 315–336.

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology, 21,* 126–134.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332.

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

Hale, G. (1992). *Effects of amount of time allowed on the Test of Written English.* TOEFL Research Report 39 (ETS RR No. 92-27). Princeton, NJ: Educational Testing Service.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Michael Levy and Sarah Ransdell (Eds.), *The Science of Writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.

Huddleston, E. (1954). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. *Journal of Experimental Psychology, 22,* 165–213.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). *The effects of standardized testing.* Boston: Kluwer-Nijhoff Publishing.

Klein, S. P. (1981). *The effects of time limits, item sequence, and question format on applicant performance on the California Bar Examination.* San Francisco: Committee of Bar Examiners of the State Bar of California and the National Conference of Bar Examiners.

Koenig, J. A. (1995). *Examination of the comparability of MCAT writing sample test forms.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Linn, R. L., Baker, E. V., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Longford, N. T. (1994). *A case for adjusting subjectively rated scores in the Advanced Placement tests* (ETS RR No. 94-58). Princeton, NJ: Educational Testing Service.

Lukhele, R., & Sirici, S. G. (1995). Using IRT to combine multiple-choice and free-response sections of a test onto a common scale using a priori weights. Paper presented at the annual meeting of the National Council on Measurement in Education, April, San Francisco.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement,* New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Michael, W. B., Cooper, T., Shaffer, P., & Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and by professors in other disciplines. *Educational and Psychological Measurement, 40,* 83–95.

Michael, W. B., & Shaffer, P. (1978). The comparative validity of the California State University and Colleges English Placement Test (CSUC-EPT) in the prediction of fall semester grade-point average and English course grades of first semester entering freshmen. *Educational and Psychological Measurement, 38,* 985–1001.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3*(3), 285–296.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminate function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107–122.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement, 19*(1), 37–47.

Mullis, I. V. S., Dossey, J. A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., & Latham, A. S. (1994). *NAEP 1992 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Muraki, E. (1993, April). Implementing item parameter drift and bias in polytomous item response models. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Muraki, E. (1998). Stepwise analysis of differential item functioning based on multigroup partial credit model. Unpublished paper. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress (1994). *NAEP 1992 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Olsen, M. A. (1956). The effectiveness of tests of writing ability and of directed memory for predicting law school grades and ratings of writing ability. Law School Admission Council Report LSAC-56-1. Law School Admission Council.

Oshima, T. C., Raju, N. S., & Flowers, D. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential item functioning of items and tests. *Journal of Educational Measurement, 34*(3), 253–272.

Osterlund, B. L., & Cheney, K. (1978). A holistic essay-reading composite as criterion for the validity of the Test of Standard Written English. *Measurement and Evaluation in Guidance, 11*(3), 155–158.

Pitcher, B. (1962). The prediction of first-year law school grades from experimental tests of writing and reasoning ability and from the Law School Admission Test, 1959–1960. Law School Admission Council Report LSAC-62-2.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). *Choice in Advanced ETS Placement tests* (SR No. 92-51). Princeton, NJ: Educational Testing Service.

Pomplun, M., Wright, D., Oleka, N., & Sudlow, M. (1992). An analysis of English composition essay prompts for differential performance. College Board Report No. 92-4. New York: College Entrance Examination Board.

Powers, D. E., & Fowles, M. E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement,* 33(4), 433–452.

Powers, D. E., & Fowles, M. E. (1998). Test takers judgments about GRE writing test prompts (GRE No. 94-13). Princeton, NJ: Educational Testing Service.

Powers, D. E., Fowles, M. E., & Boyles, K. (1996). Validating a GRE writing test (GRE No. 93-26B; ETS RR No. 96-27). Princeton, NJ: Educational Testing Service.

Powers, D. E., Fowles, M. E., & Farnum, M. (1993). Prepublishing the topics for a test of writing skills: A small-scale simulation. *Applied Measurement in Education, 6,* 119–135.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1992). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31,* 220–233.

Purves, A. (1992). Reflection on research and assessment in written composition. *Research in the Teaching of English, 26*(1), 108–122.

Reckase, M. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*(1), 12–14, 31.

Resnick, L. B., & Resnick, D. P. (1990). Tests as standards of achievement in schools. In: The Uses of standardized tests in American education: Proceedings of the 1989 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B. R. & O'Connor, M. C. (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston: Kluwer-Nijhoff Publishing.

Shepard, L. A. (1988). *Should instruction be measurement driven: A debate.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

Steele, J. (1979). *The assessment of writing proficiency via qualitative ratings of writing samples.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher,* 24(5), 5–11, 35.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1–20.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32*(2), 163–178.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 79,* 703–713.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(3), 200–214.

Wightman, L. (1997). Analyses of GMAT Analytical Writing Assessment data. (Personal Communication).

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Zwick, R., Donoghue, J., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187–201.

Appendix: Tables Summarizing Previous Studies of Mean Differences, Reliability, and Validity

TABLE A-1

A Summary of Studies Examining Gender Differences in Writing Test Scores

Study	*Program*	*Date*	*Sample*	*Ref. Test*[a]	*M-C Test*	*Essay Test*	*M-C + Essay*
AAMC (1997)	MCAT	1996	28,018F, 32,593M	–.08[b]		.13[b]	
ACT	ACT			.00[b]	.16[b]		
ACT (1997)	ACT	1997	540,252F, 419,049M	.05	.15		
Breland (1977)	CB/TSWE	1975	342F, 462M		.41	.32	
Breland & Griswold (1981)	CSUC	1977	5,908F, 4,766M	–.01	.05	.36	.10
Breland & Jones (1982)	CB/ECT	1979	37,977F, 35,367M	–.06	.14	.16	.20
	CB/TSWE				.15		
Breland et al. (1995)	CB/ECT	1990	213F, 187M	.12	.27	.34	.34
	CB/TSWE				.37		
Breland et al. (1999)	CB/SAT	1996	177F, 126M	.20	–.02	–.16	–.05
Bridgeman & Bonner (1994)	SAT-W	1993	494F, 463M	–.17	.03	.32	.15
			150F, 184M	–.12	–.08	.11	–.03
Bridgeman & McHale (1996)	GMAT	10/94	9,210F, 14,041M	–.14		.12	
Broch (1997)	GMAT	95–97	121,509F, 171,832M	–.17		.09	
	GRE	94–97					
	Issue		3,989F, 2,479M	–.25		–.05	
	Argument		1,656F, 950M	–.31		–.06	
Engelhard et al. (1991)	Georgia	89–90	61,620F, 64,136M Grade 8			.49	
Golub-Smith et al. (1993)	TOEFL/	1989	4,317F, 6,197M	–.14		.06	
	TWE		4,042F, 5,840M	–.14		.10	
			4,049F, 5,882M	–.16		.03	
			3,894F, 5,882M	–.12		.15	
			3,975F, 5,908M	–.12		.10	
			3,999F, 5,785M	–.16		.06	
			4,004F, 6,001M	–.14		.05	
			3,850F, 5,651M	–.15		.10	
NAEP (1990, 1994)		1988	Grade 11			.51	
			Grade 8			.53	
			Grade 11			.48	
		1992	Grade 8			.55	
			Grade 12			.54	
Pomplun et al. (1992)	CB/ECT	1983	30,988F, 27,198M	–.15	.12	.25	.19
	CB/ECT	1985	4,412F, 3,801M	–.08	.14	.10	.15
	CB/TSWE				.15		
	CB/ECT	1986	34,517F, 28,916M	–.15	.12	.11	.11
	CB/TSWE				.08		
	CB/ECT	1987	37,785F, 32,098M	–.18	.12	.10	.11
	CB/TSWE				.10		
	CB/ECT	1988	31,860F, 26,587M	–.20	.11	.19	.15
	CB/TSWE				.10		
	CB/ECT	1989	30,108F, 24,676M	–.15	.14	.08	.10
	CB/TSWE				.16		
	CB/ECT	1990	28,171F, 23,559M	–.10	.15	.15	.14
	CB/TSWE				.11		
PSAT (1996)[c]	PSAT	95–96	100,000+F, 100,000+M		.06	.13	.09

(continues)

Table A-1 *(continued)*

Study	*Program*	*Date*	*Sample*	*Ref. Test*[a]	*M-C Test*	*Essay Test*	*M-C + Essay*
SAT (1997)[d]	SAT II	10/95	14,823F, 12,793M		.10	.10	.04
		11/95	25,908F, 22,029M		.00	.00	.04
		12/95	31,633F, 27,713M		.00	.10	.10
		1/96	10,853F, 11,182M		.10	.22	.16
		5/96	10,706F, 8,221M		.10	.28	.19
		6/96	40,339F, 32,188M		.10	.10	.09
Wightman (1996)	GMAT	6/95	16,644F, 23,973M	–.24		–.01	
		3/95	12,963F, 19,069M	–.12		.07	
		1/95	13,890F, 21,437M	–.24		.12	
		10/94	15,723F, 22,526M	–.12		.05	
Mean	NAEP	84–92	5 samples			.52	
Mean	SAT	79–90	9 samples	–.11	.13	.14	.14
Mean	SAT	95–96	6 samples		.07	.13	.10
Mean	TOEFL	89	8 samples	–.14		.08	
Mean	GMAT	94–97	6 samples	–.17		.07	

Differences indicated as Female minus Male in SD units.

[a] The reference test is usually a verbal test. It is the SAT Verbal score for College Board samples, which means a test made up of reading and vocabulary questions. It is also a verbal score for the GMAT samples, but what constitutes that verbal score is more than reading and vocabulary. For TOEFL, the reference score is the total TOEFL Converted score, which includes a number of types of tests including listening comprehension. For ACT, the reference test is ACT Reading.

[b] Reported in Willingham and Cole (1997), Table 3.2, page 84.

[c] This study is based on students who took the PSAT in the Fall of 1993 and the SAT II: Writing Subject Test in the spring or fall of 1994.

[d] The estimates reported here for the SAT II: Writing Subject Test are based on rounded data and thus are not precise.

Table A-2

A Summary of Studies Examining African-American/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
AAMC (1997)	MCAT	1996	5,238B, 36,825W	–1.13		–.65	
ACT (1997)	ACT/English	1997	90,617B, 663,878W	–.85	–.89		
Breland & Griswold (1981)	CSUC/EPT	1977	583B, 5,236W	–1.47	–1.34	–.91	
	CSUC/TSWE				–1.35		
Breland & Jones (1982)	CB/ECT	1979	2,000+B, 50,000+W	–.63	–.68		
	CB/TSWE				–.69	–.48	–.72
Pomplun et al. (1992)	CB/ECT	1983	2,157B, 49,903W	–.67	–.68	–.40	–.67
	CB/ECT	1985	326B	–.60	–.59	–.39	–.61
	CB/TSWE		7,339W		–.56		
	CB/ECT	1986	2,575B	–.59	–.62	–.34	-.61
	CB/TSWE		56,568W		–.63		
	CB/ECT	1987	3,566B	–.64	–.62	–.37	–.62
	CB/TSWE		62,022W		–.62		
	CB/ECT	1988	2,953B	–.62	–.58	–.34	–.59
	CB/TSWE		50,563W		–.62		
	CB/ECT	1989	3,029B, 46,543W	–.64	–.59	–.37	–.60
	CB/TSWE				–.62		
	CB/ECT	1990	2,769B	–.62	–.64	–.37	–.64
	CB/TSWE		43,509W		–.62		
Breland et al. (1995)	CB/ECT	1990	100B	–.65	–.58	–.46	–.65
	CB/TSWE		100W		–.51		
Bridgeman & McHale (1996)	GMAT	Oct 94	2,496B, 23,251W	–.91		–.71	
Broch (1997)	GMAT	95–97	22,761B, 202,850W	–1.04		–.65	
	GRE						
	Issue	94–97	535B, 4,907W	–1.07		–.43	
	Argument		172B, 1,977W	–.87		–.49	
Wightman (1996)	GMAT	Jun 95	2,847B, 24,493W	–.86		–.67	
		Mar 95	2,102B, 18,291W	–.98		–.76	
		Jan 95	2,388B, 20,790W	–.86		–.54	
		Oct 94	2,486B, 23,550W	–.86		–.69	
Mean	CB/SAT	82–90		–.63	–.62	–.38	–.63
Mean	GMAT/MCAT	94–97		–.95		–.67	

Differences indicated as African American minus White in SD units.

Table A-3

A Summary of Studies Examining Hispanic/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
AAMC (1997)	MCAT	1996	1,257H, 36,825W	–.65		–.32	
ACT (1997)	ACT/English	1997	21,511MA, 663,878W	–.55	–.63		
ACT (1997)	ACT/English	1997	26,841PR, 663,878W	–.52	–.57		
Breland & Griswold (1981)	CSUC/EPT	1977	445H, 5,236W	–.91	–.85	–.57	
	CSUC/TSWE				–.89		
Breland & Jones (1982)	CB/ECT	1979	1,200+H, 56,000+W	–.60	–.68	–.48	–.69
	CB/TSWE				–.68		
Breland et al. (1995)	CB/ECT	1990	100H, 100W	–.65	–.76	–.34	–.70
	CB/TSWE				–.74		
Bridgeman & McHale (1996)	GMAT	10/94	1,262H, 23,242W	–.56		–.46	
Broch (1997)	GMAT	95–97	8,929H, 202,850W	–.44		–.25	
	Issue					–.21	
	Argument					–.24	
	GRE	94–97					
	Issue		323H, 4,907W	–.47		–.20	
	Argument		135H, 1,977W	–.61		–.49	
NAEP (1990, 1994)	NAEP	1984	Grade 8			–.53	
			Grade 11			–.64	
		1988	Grade 8			–.41	
			Grade 11			–.59	
		1992	Grade 8			–.55	
			Grade 12			–.35	
Pomplun et al. (1992)	CB/ECT	1983	1,130H, 49,903W	–.60	–.62	–.42	–.62
	CB/ECT	1985	117H, 7,339W	–.45	–.43	–.29	–.44
	CB/TSWE				–.48		
	CB/ECT	1986	987H, 56,568W	–.35	–.38	–.20	–.36
	CB/TSWE				–.43		
	CB/ECT	1987	1,218H, 62,022W	–.33	–.38	–.17	–.36
	CB/TSWE				–.37		
	CB/ECT	1988	1,225H, 50,563W	–.39	–.43	–.22	–.42
	CB/TSWE				–.45		
	CB/ECT	1989	1,339H, 46,543W	–.44	–.48	–.23	–.46
	CB/TSWE				–.45		
	CB/ECT	1990	1,436H, 43,509W	–.44	–.47	–.23	–.46
	CB/TSWE				–.48		
Wightman (1996)	GMAT	6/95	1,683H, 24,493W	–.76		–.55	
		3/95	1,371H, 18,291W	–.69		–.56	
		1/95	1,399H, 20,780W	–.76		–.56	
		10/94	1,560H, 23,550W	–.64		–.51	
Mean	NAEP					–.51	
Mean	CB			–.62	–.72	–.41	–.70
Mean	GMAT			–.68		–.53	

Differences indicated as Hispanic minus White in SD units.

Table A-4

A Summary of Studies Examining Asian-American/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
AAMC (1997)	MCAT	1996	14,032A, 36,825W	–.32		.00	
ACT (1997)	ACT/ English	1997	28,542A, 663,878W	–.17	–.15		
Breland & Griswold (1981)	CSUC/EPT	1977	606A, 5,236W	–.63	–.73	–.45	
	CSUC/TSWE				–.72		
Breland et al. (1995)	CB/ECT	1990	100A, 100W	–.33	–.58	–.36	–.58
	CB/TSWE				–.62		
Bridgeman & McHale (1996)	GMAT	10/94	3,198A, 23,242W	–.74		–.72	
Wightman (1996)	GMAT	6/95	3,267A, 24,493W	–.61		–.73	
		3/95	2,739A, 18,291W	–.86		–.84	
		1/95	2,908A, 20,780W	–.73		–.73	
		10/94	3,079A, 23,550W	–.73		–.71	
Mean	CB			–.48	–.66	–.40	–.58
Mean	GMAT			–.73		–.75	

Differences indicated as Asian-American minus White in SD units.

Table A-5

A Summary of Studies Examining Hispanic/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
Broch (1997)	GMAT	95–97	8,929H, 202,850W	–.44		–.25	
	GRE	94–97					
	Issue		323H, 4,907W	–.47		–.20	
	Argument		135H, 4,907W	–.61		–.49	
Pomplun et al. (1992)	CB/ECT	1983	1,130H, 49,903W	–.60	–.62	–.42	–.62
	CB/ECT	1985	117H, 7,339W	–.45	–.43	–.29	–.44
	CB/TSWE				–.48		
	CB/ECT	1986	987H, 56,568W	–.35	–.38	–.20	–.36
	CB/TSWE				–.43		
	CB/ECT	1987	1,218H, 62,022W	–.33	–.38	–.17	–.36
	CB/TSWE				–.37		
	CB/ECT	1988	1,225H, 50,563W	–.39	–.43	–.22	–.42
	CB/TSWE				–.45		
	CB/ECT	1989	1,339H, 46,543W	–.44	–.48	–.23	–.46
	CB/TSWE				–.45		
	CB/ECT	1990	1,436H, 43,509W	–.44	–.47	–.23	–.46
	CB/TSWE				–.48		
Mean	CB			–.47	–.51	–.29	–.50
Mean	GMAT			–.51		–.31	

For Hispanic examinees who indicated that English was their first language, that they spoke as well in English as in any other language, or that they were most fluent in English—differences indicated as Hispanic minus White in SD units.

Table A-6

A Summary of Studies Examining Asian-American (FL)/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
Broch (1997)	GMAT	95–97	14,526A, 202,850W	–.20		–.12	
	GRE	94–97					
	Issue			–.25		–.21	
	Argument			–.16		–.22	
Pomplun et al. (1992)	CB/ECT	1983	3,405A, 49,903W	–.35	–.33	–.19	–.32
	CB/ECT	1985	284A, 7,339W	–.01	–.13	+.06	–.08
	CB/TSWE				–.15		
	CB/ECT	1986	2,070A, 56,568W	–.07	–.23	–.04	–.19
	CB/TSWE				–.28		
	CB/ECT	1987	2,645A, 62,022W	–.06	–.21	–.03	–.17
	CB/TSWE				–.26		
	CB/ECT	1988	2,278A, 50,563W	–.08	–.23	.00	–.18
	CB/TSWE				–.32		
	CB/ECT	1989	2,521A, 46,543W	–.13	–.14	–.08	–.26
	CB/TSWE				–.34		
	CB/ECT	1990	2,637A, 43,509W	–.04	–.22	–.02	–.19
	CB/TSWE				–.24		
Mean	GMAT/GRE			–.20		–.18	
Median	CB			–.10	–.24	–.04	–.20

For Asian-American examinees indicating that English was their first language, that they spoke as well in English as in any other language, or that they were most fluent in English—differences indicated as Asian-American minus White in SD units.

Table A-7

A Summary of Studies Examining Hispanic (ESL)/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
AAMC (1997)[a]	MCAT	1996	1,018H, 36,825W	–2.00		–2.00	
Pomplun et al. (1992)	CB/ECT	1983	240H, 49,903W	–1.28	–1.38	–1.12	–1.47
	CB/ECT	1985	117H, 7,339W	–1.28	–1.12	–.68	–1.11
	CB/TSWE				–1.07		
	CB/ECT	1986	786H, 56,568W	–1.00	–.88	–.63	–.91
	CB/TSWE				–1.01		
	CB/ECT	1987	1,313H, 62,022W	–1.10	–1.02	–.69	–1.03
	CB/TSWE				–1.04		
	CB/ECT	1988	1,072H, 50,563W	–1.08	–.98	–.66	–1.00
	CB/TSWE				–1.09		
	CB/ECT	1989	1,402H, 46,543W	–1.08	–.91	–.58	–.92
	CB/TSWE				–.98		
	CB/ECT	1990	1,460H, 43,509W	–1.07	–1.03	–.70	–1.06
	CB/TSWE				–1.04		
Mean	CB			–1.26	–1.04	–.72	–1.07

Differences indicated as Hispanic (ESL) minus White in SD units.

[a] Commonwealth Puerto Rican examinees (essay test difference only approximate).

Table A-8

A Summary of Studies Examining Asian-American (ESL)/White Differences in Writing Test Scores

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*Essay Test*	*M-C + Essay*
Pomplun et al. (1992)	CB/ECT	1983	841A, 49,903W	–1.65	–1.30	–1.26	–1.47
	CB/ECT	1985	353A, 7,339W	–.90	–.84	–.69	–.89
	CB/TSWE				–.92		
	CB/ECT	1986	2,755A, 56,568W	–.86	–.79	–.75	–.88
	CB/TSWE				–1.15		
	CB/ECT	1987	3,702A, 62,022W	–.88	–.87	–.71	–.91
	CB/TSWE				–1.14		
	CB/ECT	1988	3,312A, 50,563W	–.84	–.84	–.59	–.86
	CB/TSWE				–1.12		
	CB/ECT	1989	4,133A, 46,543W	–.82	–.88	–.68	–.91
	CB/TSWE				–1.13		
	CB/ECT	1990	4,253A, 43,509W	–.72	–.76	–.62	–.81
	CB/TSWE				–.96		
Mean	CB			–.96	–.98	–.76	–.96

Differences indicated as Asian-American (ESL) minus White in SD units.

Table A-9

Reliability Estimates for Writing Tests

Study	*Program*	*Type*	*Admin. Date*	*M-C Test*	*One Essay*	*Two Essays*	*M-C + Essay*
AAMC (1995)	MCAT	Split–Half	4/92			.78	
			9/92			.77	
			4/93			.78	
			9/93			.80	
			4/94			.80	
			8/94			.78	
		Generaliz.	4/92			.76	
			9/92			.77	
			4/93			.78	
			9/93			.80	
			4/94			.80	
			8/94			.78	
		Test/Retest	9/92			.61	
			4/93			.65	
			9/93			.65	
			4/94			.68	
			8/94			.63	
Breland & Gaynor (1979)	CB/ECT	Test/Retest	1975	.84	.51	.68	
Breland et al. (1987)	None	Generaliz.	1984		.53	.70	
Clemson (1978)					.55	.55	
Coffman (1966)	None	Generaliz.			.38	.55	
Donlon (1984)	TSWE	KR–20	1981	.88			
			1/82	.88			
			3/82	.88			
			6/82	.86			
			11/82	.90			
			3/83	.89			
			5/83	.88			
			6/83	.87			
		Test/Retest	1977	.82			
			1978	.80			
			1979	.83			
			1980	.83			
			1981	.82			
			1982	.82			
Finlayson (1951)					.78	.88	
Lukhele & Sireci (1995)	GED						.87
Moss et al. (1982)						.46	
Steele (1979)					.43	.58	
					.58	.73	
Wightman (1996)	GMAT	Alpha					
Eastern Hemisphere			1/95			.66	
			3/95			.79	
			6/95			.74	
Western Hemisphere			1/95			.69	
			3/95			.72	
			6/95			.66	
Willingham & Cole (1997)	SAT II						.84
Mean				.86	.54	.71	.86

Table A-10

A Summary of Studies Examining Validity of Writing Test Scores: Correlations Predicting Grades, GPA, and Teachers' Ratings

Study	*Program*	*Year*	*Samples*	*Ref. Test*	*M-C Test*	*One Essay*	*Two Essays*	*M-C + Essay*
Breland (1977)/	CB/TSWE	1975	128+	.29	.34	.36		
Fall English Grade			243+	.25	.32	.28		
			332+	.33	.35	.32		
			189+	.28	.34	.40		
Breland et al. (1987)/	CB/TSWE		267	.44	.41	.43	.46[a]	.50
Fall English Grade								
Breland et al. (1999)/	SAT II	1996	200+	.52	.37	.35		.49
Fall English Grade								
Bridgeman (1991)/	NJBSPT		1,499	.32	.35	.23		.34
Freshman GPA			884	.27	.28	.21		.29
			1,635	.26	.26	.08		.20
			1,413	.39	.39	.31		.40
	ECT		6,088	.29	.30	.16		
			2,809	.26	.27	.18		
Bridgeman & Bonner (1994)/	SAT II		957	.32	.33			.36
Fall English Grade			334	.20	.30			.32
			593	.17	.21			.21
			525	.35	.39			.32
			171	.09	.21			.21
			75	.01	.04			.13
			330	.25	.27			.32
			255	.44	.43			.50
			117	.30	.39			.45
			45		.09			.38
			112					.40
			42					.47
			255					.35
Checketts & Christensen (1974)/	CLEP		123					.53
Fall English GPA								
Donlon (1984)	CB/TSWE		25 Colleges		.37[b]			
Huddleston (1954)/			763					.56
Avg. English Grade					.60	.43		
Teachers' Ratings					.58	.41		
Michael & Shaffer (1978)/	CSUC/EPT							
Fall GPA			1,536		.21			
Fall English Grade			637		.31			
Michael et al. (1980)/						.40		
Cumulative College GPA								
Osterlund & Cheney (1978)/	CB/TSWE		42		.42			
English Grades								
Mean (GPA Criterion)				.30	.29	.22		.35
Mean (Eng. Grade Criterion)				.28	.32	.37		.37

[a] Based on 3 readings. All other essay scores based on 2 readings.

[b] Mean for 25 colleges.