Testing Guidelines

Guidelines for Computer-Based Testing

Acknowledgements

I would like to thank the tireless efforts of a number of individuals in putting together this first edition of the Guidelines for Computer Based Testing. This was an effort that was started in 1999 and involved numerous meetings, the backing of a number of organizations, many drafts and the selfless dedication of a group of professionals who had tasked themselves with the goal of spearheading this important effort. I would also like to thank the members of the Association of Test Publishers and others in the measurement community who provided feedback on the Guidelines.

So, without further ado, I would like to first thank James Olsen, Vice Chair of the Computer Based Testing ("CBT") Guidelines Committee along with the following individuals and organizations who contributed so significantly to this publication: Contributors: James Adair (deceased), David Arnold, Betty Bergstrom, Michael Callans, Jerry Christensen, Alan Epstein, David Foster, John Fremer, Richard Gershon, Kathy Gialluca, Wade Gibson, Arana Greenberg, Steve Greenfield, Chris Gruber, William G. Harris, Lee Hayes, Lauren Hebert, Michelle Issing, J. Patrick Jones, G. Gage Kingsbury, Jamie Mulkey, Larry Newman (deceased), Bror Saxberg, Michael Segovia, Lauren Scheib, Janice Scheuneman, Eric Shepherd, Alan Thiemann, Eric Wenck, Kathleen Williams and Anthony Zara.

Companies: Alpine Media Corporation, American Guidance Service, Assessment Systems Inc., Association of Test Publishers, Consulting Psychologists Press, Designing Events, Educational Testing Service, Excelsior College, Hewlett Packard, IBM Lotus, Knowledge Testing Enterprises, Knowledge Universe, Microsoft Corporation, NASD Regulation, Northwest Evaluation Association, PES, Performance Testing Summit, Promissor, Psychological Services Inc (PSI), Pearson Reid London House,Quality Assessment Services, Questionmark, Richard Gershon Associates, Thiemann, Aitken & Vohra, L.L.C., Thomson Prometric, VUE/NCS Pearson, Western Psychological Services, Wonderlic, Inc.

Again, my heartfelt thanks for your time, expertise and support. (And my sincere apologies if I have inadvertantly left any individual or entity off this list!)

Cyndy Fitzgerald, Ph.D.
Microsoft Corporation
Chair, CBT Guidelines Committee
ATP Chairman of the Board, 2002-2003

This document contains recommended guidelines for computer-based testing. Nothing contained herein should be interpreted as legal advice or legal opinion. These guidelines are to be used in conjunction with other professional and legal guidelines for the proper implementation of computer-based tests. If you have legal questions, please consult an attorney.

Table of Contents

PART 1: BACKGROUND & EXPLANATIONS
Chapter 1: Introduction

Computers are now standard and pervasive tools that significantly impact our daily lives. In testing too, computers have changed the ways in which tests are developed and administered. In the face of the rapid growth of computer-based testing, the Association of Test Publishers (ATP) has sponsored the development of guidelines to help assure high measurement quality of computer-based tests and to provide direction for the principles, procedures and best practices used for developing and administering these tests. These guidelines are intended to supplement, extend, and elaborate on the *Standards for Educational and Psychological Testing* (1999, "Standards") as they apply to computer-based testing (CBT). The Guidelines are written to reflect guidelines for best practices in computerized testing.

These guidelines are written to inform and improve professional practice and judgment, but are not to be viewed as a set of explicit checklists or standards that must be followed in minute detail.

AUDIENCES FOR THE GUIDELINES

The discussion and guidelines were prepared to be appropriate for a wide range of audiences. The following table suggests how various audiences can use the content and guidelines presented in this document.

Audience	**How the Guidelines Can Be Useful**
Test development organizations	for specifying procedures for designing, developing, field-testing, and validating computer-based tests
Test publishers and administrators and test delivery organizations	for establishing common industry guidelines for communication of test items, exam scores, and item response information to and from computer-base test center locations |
Test delivery organizations	for providing information about how to achieve consistent high quality delivery of computer-based tests
Test-takers	for providing information about and descriptions of the types of test items, tests, test score interpretations, and test orientation they might encounter when they take a computer-based test
Research and evaluation specialists	for providing information on current and expected future uses of computer-based tests
Administrators of computer-based tests	for providing information about interpreting test scores from these exams and using these scores appropriately, and about helping students to prepare appropriately for computer-based tests
Advanced technology companies based	as aids in determining how to create products and services that might be beneficial in solving current or future problems and issues in computer-testing
General readers	for providing interesting and useful information
Licensure and certification boards	for developing and delivering computer-based tests.

COMPUTER-BASED TESTING

Computer-based tests are tests administered by computer in either stand alone or networked configurations or by other technology devices linked to the Internet or World Wide Web. Like traditional tests, computer-based tests come in a wide variety of forms. They may be loosely categorized according to the rules or models used for administration of the test items, the type of response required from the test-taker, and the testing environment, which may be for high or low stakes decisions about test-takers.

Test Administration Models

Linear or Sequential Administration
The most common method for administering computer-based tests is linear or sequential administration. In its simplest form, all test-takers receive the same items though possibly not in the same order. Alternately, some number of fixed test forms can be developed. Each form is built to the same test specifications and a test-taker is assigned one of these comparable forms for administration. Using computer-assisted test assembly procedures, the number of forms can become very large, depending on the size of the item pool and the degree of overlap (i.e., the number of items in common between two forms) that can be tolerated. Forms can also be developed instantly with the computer selecting items for administration during the testing session. Linear test administration differs from adaptive or tailored administration in that the selection of the test items is based only on the test specifications and is not influenced by the test-taker's responses to previous items.

Module or Testlet Administration
In one model of administration, items are administered using intact sets of items or modules, called testlets. In the simplest form of modular administration, the computer selects modules according to test specifications and every test-taker sees the same number of modules, not unlike linear/sequential computer-based testing. Modules may also be administered adaptively so that the next module is selected to be easier or more difficult according to the test-taker's performance (score) on the previous module.

Mastery Testing using Modules or Testlets
In another model, which is appropriate when only a pass/fail decision is needed, modules that are parallel in content and difficulty are administered successively until an adequate degree of precision or certainty has been achieved to determine the pass/fail status (mastery model). In this model, the number of modules administered will differ with more modules administered to test-takers whose knowledge or ability is close to the passing point (cutting score) and fewer modules to test-takers who clearly pass or fail.

Computerized Adaptive Testing
In computer adaptive testing, the computer selects the next item based on the test-taker's responses to previous items. This procedure eliminates items from the administration that the test-takers have a high probability of getting either correct or wrong. The adaptive test may be either a fixed length test with the same number of items administered to each test-taker or a variable length test. Variable length tests are administered with some stopping rule so that administration ceases when a degree of precision or certainty is reached concerning the test-taker's score. Adaptive mastery models are also used when the decision is to assign the test-taker either a passing or failing score. For many test-takers, adaptive testing will result in the same or greater precision of measurement with fewer items than linear or sequential testing.

Summary
Additional test administration models will be developed and distinguished as the computer-based testing profession matures. The foregoing models illustrate that computer-based testing can apply to multiple models of test administration.

Test-Taker Response Types

Computers create the possibility of presenting and scoring a wider variety of item types and test-taker response modes. Item and response types might be classified along a continuum with one end defined by objective items where responses are limited to a selection among the options that are provided. This would include multiple-choice items, true false and matching items. Also included may be items where the test-taker must locate an object using a "point and click" procedure using a computer mouse.

Further along this continuum are constructed response items where the test-taker generates brief text, numeric, or graphic answers to the questions posed. Examples are fill-in-the blank items, numeric answers to math or science problems, or short sentence or paragraph responses to open-ended questions. A test-taker might also be asked to assemble something, arrange steps into a proper sequence, or place elements into their proper locations in an overall structure or system using computer

methods such as drag and drop. In all these instances, responses are produced by the test-taker, but are still somewhat limited in their form. The responses are relatively short and computer scoring is fairly straight forward. The correct answer, a list of possible correct responses, or fairly simple rules for determining the correct answer can be readily identified by the test developer.

At the far end of the continuum are complex performance assessments such as essays, multiple-step exercises, or simulations. A complex performance item may require additional problem solving and reasoning from the test-taker than responding to an objective or open-ended question. The time to complete a response to these assessments is long and the test-taker will have leeway in how to formulate a response. Usually, the test-taker's responses are not clearly right or wrong but are more or less acceptable given the performance environment or task that was set. Scoring is based on complex rubrics or algorithms and may require human judgment to select the appropriate score for a given response. Computer scoring procedures may be used for some performance assessments, but these procedures require a considerable investment in research and development.

Low- and High-Stakes Tests

Computer-based tests are administered in both low- and high-stakes testing environments. The testing purpose determines how important the decisions and interpretations from the test scores are for test takers and test users.

Low-stakes tests are situations where the decisions and interpretations from the test scores have minor or indirect consequences for the test-taker, the program, or the institution involved. In many low-stakes testing environments, the test-taker has an opportunity to repeat the test and to improve his or her score without restrictions or outside intervention. Examples of low-stakes testing environments include practice tests, diagnostic tests, instructional unit tests, and classroom quizzes.

High-stakes tests are situations where decisions and interpretations from test scores have important, critical, and direct consequences for the test-taker, the program, and the institutions involved. Examples of high-stakes testing environments include final examinations for a college course, college and graduate school admissions examinations, professional certification and licensure tests, and job selection tests.

Current computer-based test delivery and administration services are delivering high-stakes computer-based tests. These tests include information technology certification tests, college and graduate school entrance tests, and professional licensure and certification tests. The increased use of practice tests and simulation environments, and the increased ratio of students to computers within schools and work environments are expected to expand the volume of low-stakes computer-based testing environments. The Internet and World Wide Web can be used for administration of both high-stakes computer-based tests with appropriate proctoring or for administration of low-stakes computer-based tests.

SCOPE OF THE GUIDELINES

Although computer-based testing is used in a wide spectrum of testing environments, the Guidelines are aimed at high stakes computerized education, certification and licensure tests, high stakes computerized aptitude and industrial/organizational psychological tests, and personality, interest testing and biographical data surveys used for employment testing.. To the extent, therefore, that the issues surrounding low stakes applications are different from those of high stakes testing, they will not be discussed. In addition to the low stakes applications mentioned above, areas not addressed by the Guidelines include computer-based administration, scoring, or interpretation of personality measures or other measures of psychological constructs, placement examinations, interest inventories, or attitude surveys. Those who are interested in the use of those types of measures are referred to the Joint Standards.

OVERVIEW

This book is divided into two parts. The first part has four chapters, the introductory chapter and chapters providing an overview of validity and test design, test development and analysis, and test administration. Although much of this material is essentially the same as for paper-and-pencil testing, the overview provides a context in which to discuss those issues and parts of the process that differ or require particular

attention for computer-based testing. The following chapters focus primarily on those tests that make use of the computers for modular or adaptive administration formats and those tests that use item formats requiring computer interaction. These chapters also provide the terminology and foundation for the later guidelines.

Part 1 provides some background and rationale for the different parts of test design and development for those involved in computer-based testing who have little training in measurement. Those readers are referred to the *Joint Standards* or a measurement textbook for elaboration of the topics presented here. Recall that these Guidelines are intended to supplement, extend, and elaborate on the *Standards for Educational and Psychological Testing (*1999, "Standards") as they apply to computer-based testing ("CBT").

Part 2 describes the Guidelines for development and use of computer-based tests. The Guidelines cover areas of test development and uses that are the responsibilities of the test sponsors, test development organization, test delivery organization, and score users. (Note that these are not necessarily different organizations and some organizations may consist of only one or two individuals). This section is divided into six chapters, which cover planning and test design, test development, test administration, scoring and score reporting, psychometric analysis, and stakeholder communications.

PART 1: BACKGROUND & EXPLANATIONS
Chapter 2: Validity and Test Design

For any test, proper planning and design enhances the validity of the measurement instrument by assuring that the development process focuses on the purpose. Validity is the extent to which a test measures what it claims to measure. Carefully defining the test purpose, designing specifications, and constructing the test items to meet those specifications leads to a test that possesses validity evidence for the inferences and interpretations of test scores.

Computer-based tests differ from paper-and-pencil administered tests in that the planning phase takes into consideration a number of additional factors such as hardware and software limitations. Test development may involve many item types that differ from the standard multiple-choice item found in most paper-and-pencil examinations. If performance assessments are computer-delivered, sophisticated software is used to administer the assessments and to capture and possibly to score complex test-taker generated responses.

PLANNING THE TEST

Planning for computer-based testing is similar in many aspects to planning for other types of testing.

Testing Purpose

The purpose of an examination may be described in general terms. For example, a computerized test may determine that candidates for certification have the necessary level of the knowledge, skills, and abilities for practice in the occupation or profession. A computerized test may be used for employee selection or career advancement. A computerized test may be used to determine the degree to which applicants for admission to post secondary colleges or professional schools have the basic academic knowledge, skills, and abilities required.

For test design, the actual knowledge, skills, and abilities to be measured should be described, broadly at first and in more detail as planning proceeds. The documentation should include both the primary and secondary purposes. For example, a certification test may be used for licensing in some jurisdictions or used by schools or colleges to evaluate the effectiveness of their programs.

The purpose of the test dictates other decisions. For example, what types of score or scores are needed? Will a simple pass/fail designation suffice or will scaled scores be necessary? Will information such as normative data be required for score interpretation? Will information be available to identify subgroups of test-takers?

The purpose defines the construct measured and determines what validity evidence is required. For example, a certification or licensing examination will often be linked to job analysis information, which is collected before the test specifications are completed and before the test is given. For admissions testing, where future performance of the test-takers is desired, validity evidence may relate to how successfully students perform once admitted, evidence that is collected after the test is administered.

Needs Analysis

A needs analysis is one approach to gathering the information needed to complete a detailed statement of test purpose and a first step in test planning and design that leads to the test specifications. A needs analysis identifies design and environment constraints, specifies program objectives, and determines the program requirements from multiple perspectives.

At least five groups of participants in the testing process should be considered in the needs analysis. These are the test-takers, the test developers, the test delivery service, the score users, and the test sponsors.

A first step in the needs analysis is to carefully describe the target population of test-takers. Some relevant characteristics to consider include:

- expected age range
- reading levels
- basic educational requirements
- curriculum or training courses completed
- expected performance standards
- relevant knowledge and skill levels
- geographic locations (domestic vs. foreign)
- native and secondary language familiarity
- computer literacy

Computer literacy, familiarity with computers, or experience with using a mouse or other response devices are important for computer-based testing. For some tests, simple keystrokes or point-and-click operations that a test-taker can easily learn in a tutorial or practice session are sufficient to respond to the test questions. Items that require drag-and-drop operations are likely to require more familiarity. Simulations or other complex performance assessments may require extended practice to learn the computer interface, even for experienced computer users.

Another perspective for investigation with the needs analysis is that of the organization that develops the test. This organization develops objectives for the test; creates test specifications (blueprints), identifies subject matter experts, if needed; writes and reviews test items; pretests (field tests) items; and conducts statistical item and test analysis to improve the computer-based test. The needs analysis should identify the limitations, current practices and capabilities, and future growth potential of the test development organization.

The needs of the test publishing and/or test delivery organization should also be considered. The test publishing organization sets up the final computer-based forms for the exams or item pools supplied by the test developer, ensuring that the computer-based tests are published or made available in standardized format to each test administration location. Test publishers should also consider the visual legibility of the test items for examinees. Responsibility for a standardized computerized test administration process is shared by the test delivery organization and the test developer or sponsoring organization.

The needs analysis should specify the capabilities of the test publishing and/or test delivery organizations. The procedures used for handling multiple forms of the same test, the ability of the organization to support modular or adaptive testing, and the availability of audio or other multimedia presentations will be important for many testing programs. The number and location of testing sites, including the availability of international centers, may also be important.

Another target of the needs analysis is the user organization that interprets and uses the test results. The user organization may be a college admissions office, a licensing jurisdiction, or a certification board. The needs analysis should identify the information required by the test users for them to interpret the scores accurately and to use them appropriately. Other requirements that should be addressed in the planning stages for a computer-based test are whether the scores are to be reported at the time of testing or at a later date and if some types of rosters or summary reports are required. If summary data are required for a score user, the plans should include when and how these data are to be compiled. Additional considerations for test user organizations are the accessibility to and security of tests, answer keys, and test examinee data.

The final perspective is that of the test sponsor or test developer or user. An example might be a professional organization that sponsors the development of a licensing examination or an organization like the Graduate Management Admissions Council that sponsors an admission test for schools of business. The perspective of the test sponsor is likely to concern costs and public relations. The funding available for development of a computer-based test will, of course, have considerable impact on the options that can be considered in planning the examination. The public relations aspect can be particularly important

for computer-based testing. For some examinations' selling the concept of computer-based testing to the various stakeholders may be a crucial step in the success of a program. The plans should include educational efforts to acquaint the test-takers and sometimes the score users with the nature of the computer-based testing and its advantages.

Validation Plan

The validation process is the accumulation of evidence to support the inferences and interpretations made from test scores. The interpretations of test scores for specific uses of a test are validated, not the test itself. The validation plan begins with an explicit statement of the proposed interpretation of test scores and a rationale for the proposed use. The validity of test score interpretations relates to the construct that the test measures. This construct could be, for example, competence in a profession, mathematics achievement, or reading ability.

The next part of the validation plan is to generate a list of inferences that are required to move from the construct to the scores and from the scores to the test use. For example, for a test that is intended to determine if a person has learned the material presented in a course of study, one inference would be that the content measured by the test matches the content of the curriculum. Once such inferences have been specified, the type of evidence needed to support each inference can be established.

One source of validity error that needs to be considered in the plan is construct irrelevant variance. Construct irrelevant variance refers to factors in the test or testing environment that affect a test-taker's performance but are not associated with the construct being measured. Designers of computer-based tests should be mindful that the construct being measured is not an examinee's computer skills or familiarity with computer applications. The testing interface used should be easy to understand and use. With computer-based testing, this might include difficulties with using the computer, anxiety about the computer-testing mode, or lack of understanding of the computer interface. For novel item types or performance tasks, it might also include difficulties in understanding what response is required or how to formulate a response in the proper form for the computer. Designers of computer-based tests need to pay special attention to these issues.

Construct under representation is another source of validity error that may arise with simulations and performance assessments, unless they are given in combination with substantial numbers of multiple-choice items or other objective item types. Construct under representation means that the sample of test-taker performance is too narrowly defined to measure the construct well. A related concern is task specificity, i.e., how well test-takers do on one task may be weakly associated with how well they will do on another task from the same performance domain. These issues should be addressed in the validation plan.

Finally, simulations and performance assessments have the issue of how scores are assigned. Different scoring procedures will not be equally effective in capturing the underlying construct that is measured. The validity plan should include how test scores are established and validated. Differences in score assignment with simulations and performance assessments may be considered analogous to differences in rater reliability with multiple judges.

TEST SPECIFICATIONS

The test specifications or blueprint provides the details, which connect the test purpose to the test content. The specifications are shaped during the planning process and provide the foundation that will guide item writing, pilot tryouts, statistical item and test analysis, selection of test items, and test construction. The test specification describes:

- the purpose of the test
- characteristics of the test-takers, if available
- the test content or performance domain (e.g., knowledge, cognitive abilities, skills, processes, diagnostic elements)
- the item formats (e.g., standard multiple-choice, singly or in sets; point-and-click; fill in the blank)

- the computer display formats used for presenting the test and its various items or tasks
- the response methods to be used by the test-taker
- the psychometric characteristics of the items and tests

The test specifications also document the time requirements for items and tests. The test specifications should describe the procedures used to create items; pilot, field test, or pretest items; select final items and compute item statistics. For linear administrations, it should also include the number of forms to be constructed as well as any item weighting rules or algorithms. Procedures to be used in administering adaptive or modular tests should be specified in some detail including factors such as the method for choosing the next item or module, the maximum number of items or modules to be administered, and the stopping rules for variable length tests.

Interactive simulation and software performance environments require extensions of the test specification of procedures and methodology. Interactive simulations and software performance environments, for example, may provide a simulation shell that functions as a template or exemplar from which different items and performance exercises can be developed, administered, and scored.

The Job/Task Analysis

For many tests, the test content will come from the results of a job or task analysis (also called a role delineation study or practice analysis). The job/task analysis will establish the knowledge, skills, and abilities required for work in an occupation or profession or the types of tasks to be performed by people in a given job or area of competence. The specification for the job/task analysis presents meaningful groupings of knowledge areas, job elements, tasks, and subtasks that are performed by qualified individuals within an occupation, profession, or job specialty. The job/task analysis specification often uses structured interviews, focus groups, or surveys to identify key elements of practice or job performance.

After the initial job/task analysis specification is prepared, the document may undergo multiple revision cycles of analysis (task breakdowns) and synthesis (task regroupings) to determine whether the identified groupings are meaningful and useful. The job/task analysis specification may also indicate the tasks and subtasks, which can be used to classify qualified or unqualified job performers. The statements prepared through this process can then be evaluated by rating them for difficulty, importance, and frequency by panels of subject matter experts or by large numbers of practitioners in the field.

Hardware and Software Specifications

Because of the variety in computer hardware and software configurations, test specifications should specify the minimum hardware and software test delivery configuration required for the computer-based test. Where possible, the test delivery software should be designed so that it will not initiate the computerized test until the minimum hardware and software configuration is available. The minimum hardware delivery specification should include the following components:

- computer type
- central processor type and speed
- random access memory capacity
- hard or removable disk storage capacity
- monitor display size, resolution and pixel pitch
- test response input device (keyboard, mouse, touch screen, keypad, etc.)
- test data exchange device (diskette, CD-ROM, modem)
- network type and speed
- single user or laptop configuration
- printer type, resolution, and speed (pages/minute)

The *minimum* software delivery configuration should include the following components:

- computer operating system and version
- network operating system software
- test driver, test administration and management system

EVALUATION PLAN

The value of the computer-based-testing should be carefully evaluated throughout the entire test development, field testing and administration process. The nature and extensiveness of the evaluation plan should be incorporated in the test planning and design stage. A test evaluation plan should include the key evaluation questions to be addressed, the evaluation procedures, evaluation instruments, data collection procedures, data analysis procedures, and report and dissemination procedures. The purpose for the evaluation is to judge the quality or added value of the computer-based testing method. The evaluation plan should also check the integrity of the complete test display and test scoring systems.

PART 1: BACKGROUND & EXPLANATIONS
Chapter 3: Test Development and Analysis

At present the development of many computer-based tests does not differ markedly from the development of paper-and- pencil tests. Different tools or procedures may be needed, however, for tests administered adaptively or in modules or for tests that use formats that depend on computer-administration. Even tests with linear/sequential administration of multiple-choice items may find that more forms and items are needed than for paper-and-pencil testing, making good item management tools important. This chapter discusses item banking, test assembly, test fairness issues, statistical analysis of item and test data, and monitoring and tracking of the computer-based test.

ITEM BANKING

Developers of computer-based tests should carefully consider the capabilities of item banking and test development software systems. A number of item banking systems are commercially available, but may not have all the features that might be desired for a given computer-based test. In some instances, particularly with simulations or performance assessments, test developers may need to create their own item banking systems to include the features they require.

Because computer technology allows innovative item types, the item bank for a test using these items should have the capability of effectively handling item types other than standard multiple-choice items. These item types would include multiple-selection items (more than one answer correct), fill-in-the-blank, point-and-click, drag-and-drop items and other item types that may not be readily usable in a paper-and-pencil format. Line drawing and photographs, along with video, audio, and animation, can be embedded as appropriate into test items. Item banking systems should be able to store items in the appropriate format and display the items as they would appear on the test for review by subject matter or test development experts.

An item-banking database should include any text for the item and other accompanying material (or references to that material that may be stored elsewhere) such as graphics, video displays, or software applications. The item database should also contain codes linking the item to the test specifications, keep a record of the use of items in forms or modules, and record any statistical information that has been obtained from pretesting or field testing as well as from operational use. Status of the item with regard to review, pretesting, or availability for use in test administration should also be included. For linear forms or modules, the system should have the capability of providing reports on the properties of the items selected for inclusion, including average difficulty of the module or form and its overall match to the test specifications. For computer adaptive testing, the item bank should be formulated for easy transfer of the active item pool to the test delivery organization.

The item banking system should also provide summary reports of the items in the pool by content category, item format, difficulty, or other dimensions important for a particular test. It should be able to identify items that are available for use or are not available because of poor statistics, over-exposure in adaptive testing, or other reasons. One of the important functions of the item bank for computer-based testing is to provide information needed to describe and monitor the status of the item pool to inform later test development efforts.

Item banking systems can also accommodate the use of test-taker surveys to be delivered before or after the test. Keeping the data from these surveys as part of the item banking system permits the results to be linked easily to the tests, thus providing information that may be useful for making test improvements and for marketing. These data can also be used to correlate test-takers' responses to the survey with their test scores. Analysis of the survey results may be particularly important for novel item types where the comfort of the test-takers in responding within the computer environment is of concern.

TEST ASSEMBLY

The assembly of forms or modules for a computer-based test can be a computer-assisted process. Although many procedures can be used, the more sophisticated versions of automated test assembly will use mathematical models or heuristic devices to select items that meet a number of rules and design constraints on the test form or module. These rules will relate to factors such as the test content, item format, and statistical properties, as well as possible limitations on other factors such as the number of items related to a particular application or work setting that may not be explicitly indicated in the specifications. If the number of forms or modules to be developed is quite large, such a system may be a necessity. If desired, the output of the process can be immediately publishable on a test delivery vendor's registration, scheduling, and distribution network.

For computer adaptive testing, the items taken by one test-taker are not selected until the test is administered. For this application, test assembly is replaced by the rules used by the computer for the selection of items. These will include the statistical model on which the algorithms are based (e.g., Bayesian models or sequential probability models), the procedures for estimating item parameters (usually one-parameter, two-parameter, or three-parameter item response theory methods) and ongoing computations such as test proficiency, item and test information functions, and standard error computations. Other rules will concern how the test administration is to be terminated and the procedures for deriving scores. These rules may be implemented in software developed by the test developer or by the test delivery organization. Software for adaptive testing is also available commercially and may be purchased for use if it meets the needs of the testing program.

TEST FAIRNESS

As with other types of testing, developers of computer-based tests should strive toward fairness and equity for all test-takers. A fair and unbiased examination provides a context that permits all test-takers to demonstrate their knowledge and abilities. Fairness and equity issues concern the appropriateness of test content, item and response formats, test administration procedures, and test scoring, reporting, and interpretation. Test fairness should be considered throughout the test planning, item writing, and evaluation phases of test development.

Efforts should be made to ensure that language is gender fair, with both men and women represented as appropriate to the test content. Women and people of color should not be portrayed in stereotypic roles or situations. Power and status should not be portrayed as the exclusive province of a single group nor should poverty or disadvantage. If people with disabilities are portrayed, then the material should emphasize their abilities and positive accomplishments rather than emphasize their disabilities.

When testing internationally, item writers should avoid vocabulary and references to institutions or entities from a particular country that might not be known to test-takers from other international countries. Even well known acronyms may become a source of irrelevant difficulty to test-takers from another nation or culture where those acronyms may not be used in some testing applications, common terminology may differ across national boundaries.

When tests are administered worldwide, many individuals will be tested whose primary or native language is not the same as the language in which the computer-based test was developed. When computer-based tests are administered to test-takers who lack reading and listening fluency in the target language of the test, the test can be measuring language skills in addition to the intended knowledge, skill, or performance. Language proficiency tests are an exception to this statement because language proficiency tests are designed explicitly to test for test-taker competence in using a particular target language.

Some tests administered worldwide may be adapted into other languages. Since adaptation is rarely perfect, it could result in tests that are easier in one language than in another language. When tests are given in different languages, they should be analyzed to assure that they are of equivalent difficulty in the different language versions. In performance measures, scoring procedures and rubrics also require special attention to assure that the same scores have equivalent meaning in the two language versions.

For tests that are administered in multiple languages or in multiple countries, a test fairness issue is the familiarity with computers across languages and cultures. Every attempt should be made to ensure that concepts, constructs, and test items are relevant and have the same meaning for examinees taking the exam from different countries.

One final test fairness issue for consideration is differences in item display time with different computer hardware, network and communication configurations or with different browsers for Internet or web-based testing.

ITEM AND TEST ANALYSIS

In traditional paper-and-pencil testing, the tests are often given on a single occasion and data analyses are performed afterwards. For many computer-based testing applications, it is desirable to report test scores at the time of testing. If more than one test form or set of performance assessments is administered, some of the data analyses are typically performed in advance of the test implementation. This almost always requires field-testing or pretesting of items to obtain the data for analysis.

Pretesting provides information on the individual items. Classical item difficulty and discrimination values may be obtained as well as item response theory ("IRT") parameter estimates. If the number of test-takers for the items is adequate for parameter estimation, item response theory permits the test developer to project the characteristics of a test made up of pretested items before the test is administered. This includes the ability to determine comparable scores for people who took different sets of items. IRT is essential for adaptive testing and modular testing models.

TRACKING AND MONITORING

The long-term performance of the test can be regularly monitored. When problems are discovered with the test as a whole or with individual items, test developers may be informed by an automated test tracking and monitoring system. Item exposure rates can be tracked and item and test statistics calculated at regular intervals. Plans should be made to update item files on a routine basis.

Test-taker's performance can also be monitored. Measures of their performance can include the evaluation of test taking strategies and patterns, use of item skipping and review, appropriateness of timing for an individual testing session, and reactions to the testing through survey items. While tracking and evaluating scores and score variances based on gender, age, education level, nationality, etc., may seem to be relatively academic in nature, this information is important for ensuring the integrity and fairness of the test, procedures, delivery system, and testing specifications.

If problems are identified, rapid test revision is possible. Items that function poorly on a computer-based test may be removed as soon as problems are found and new tests can replace the old ones, sometimes within a few days. Items in adaptive testing pools that exceed exposure control criteria can be retired automatically. Test developers may want to develop procedures to review and update items and forms as appropriate for their testing volumes, administrative schedule (continual testing versus administration windows), and rapidity of change in the content and skill areas tested.

For testing programs that include both computer and paper-administered versions of the exam there is a need for accurate record keeping and audit trails to track changes in items and tests.

PART 1: BACKGROUND & EXPLANATIONS
Chapter 4: Guidelines for Test Administration

Standardized test administration is an important aspect of validity. In computer-based testing, the responsibility for a standardized administration process is shared by the test delivery vendor and the test developer or test sponsor, depending on how the testing program is organized. The goal of the test administration should be to provide a common, standardized experience for all test-takers who are taking the same test, regardless of where the test is delivered.

The test developer/sponsor is responsible for stipulating and documenting delivery-related requirements. These requirements include all characteristics of the testing session, such as time allowed for tutorials or practice tests, time for the actual examination, and the materials that test-takers are permitted to access during the testing sessions.

Documentation for testing center personnel should be consistent with the test delivery organization having primary responsibility for final review, approval, and implementation of the directions for test administrators and test-takers. The test developer/ sponsor is responsible for providing test-takers information about the administration including the appropriate means of test registration and scheduling, and general procedures to be followed at the testing center or other authorized test delivery location.

Test delivery organizations should adhere to the standardized procedures for administration and scoring specified by the test developer. Specifications regarding instructions to test-takers, item display and test time limits, the form of test presentation or response, and test materials or equipment should be strictly observed. Exceptions should be made only on the basis of carefully considered professional judgment and should be documented to permit review and evaluation of the soundness of the decision.

Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test-takers to attain elevated scores by fraudulent means. Security of the test material and establishing identity of the test-taker are responsibilities of the test delivery organization. Procedures to guard against unauthorized access to examinations should be in place and physical layout of the testing centers and other authorized test delivery locations should discourage test-takers from attempts to see the responses of other test-takers. In general, test sponsors will specify the requirements for forms of identification needed from test-takers.

A major advantage of computer-based tests is the possibility of on-demand, global availability of tests. Such availability offers greater opportunity for the unauthorized release of test content by previous test-takers to test-takers currently registered for the test. Test sponsors must balance the availability of the test against the security considerations required to support the validity of the assessment.

Test delivery systems are also responsible to provide special accommodations for test-takers with disabilities. The test delivery organization should also implement an administrative process so that each test administrator knows what to do if he or she received an accommodation request. Special test accommodations may include an untimed test administration or a private location to limit distraction for a test-taker with attention deficit disorder. A second example of test accommodations may involve the use of an exam reader for a visually impaired test-taker. Alternately, accommodations may use special equipment such as larger monitors, a larger font for text, a zoom feature for graphics, or audio interpretation of visual material. The test delivery organization should plan to make available reasonable accommodations at each of its centers or other authorized test delivery locations.

Performance assessments, such as live application testing or simulations, may also be anticipated in the test delivery environments. When delivering performance assessments, the computer must provide a consistent tool for delivering, assessing, and scoring the computer-based test. Where possible, the computer platform should be consistent at every testing facility or other authorized test delivery location to ensure test standardization and to promote test-taker equity. Factors such as screen display time, plotting speed and version control of computer operating system and testing delivery software will affect test standardization and validity.

For computer-based tests which are delivered via Internet or World Wide Web there may be a single version of the computer-based software that resides on a central server. The client workstations that access the computer-based testing software may be running different browsers and different operating systems. The test developers should specify the minimum browser and operating systems versions that are acceptable for accessing and delivering the computer-based test.

PART 2: COMPUTER-BASED TESTING GUIDELINES

Chapter 1: Planning and Design

This chapter provides Guidelines to plan the computer-based test and to document the information that should be provided in the test specifications. When documentation is desirable, the specifications suggested below may include proprietary or sensitive scoring information, such information would not necessarily be available to all parties described in the introduction of the guidelines.

1.1 The purpose of a computer-based test should be investigated through a needs analysis to determine characteristics and requirements relevant to the test purpose for the test-takers, the test users, the test developers, the test publishers/test deliverers, and the test sponsors. If a needs analysis is conducted, the results should be documented.

1.2 Computer-based tests should be designed and developed with an underlying sound systematic and scientific basis. The test developers should compile and document evidence to support the validity of the inferences of the test scores. if the test-takers receive different fixed test forms or if items are drawn either randomly from item pools or by adaptive testing approaches, documentation should include a plan for collecting evidence concerning validity and reliability. This evidence will aid test-takers and users in interpreting test scores and using test scores in line with the test purpose.

1.3 Computer-based tests can be designed and developed to meet different purposes. Design documentation should include:

- test purpose
- construct or content domain definitions
- content structure for the test items
- required response formats for the test items
- sample test items illustrating the response formats
- number of items to be developed and administered
- scoring and reporting formats and procedures
- test administration procedures and the mechanics of test delivery, e.g. item marking, skipping, and review

1.4 Design documentation should include the procedures concerning the required content for the test including how it will be obtained and reviewed. The review mode for the computer-based test should replicate the expected test delivery conditions to the greatest extent possible.

1.5 Information should be provided about the test delivery environment, minimum computer hardware for the computer-based-test development and delivery environments, and minimum software configurations and requirements for computer-based-test development and delivery environments.

1.6 Specifications for computerized adaptive tests should describe the method for administration including the algorithms used for test item selection, the measurement model employed, test proficiency estimation algorithms, item and test information functions, test standard error computations, test termination procedures, and test scoring and scaling.

1.7 When a computer-based test is to be administered with multiple test delivery methods, any necessary comparative studies should be planned and documented.

1.8 Qualifications and experience of individuals who write and/or review test items should be specified.

1.9 When a test is delivered in multiple languages the rationale for linguistic modifications and accommodations should be specified.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 2: Test Development

These Guidelines relate to development of the computer-based test and provide direction on using computers in the process of test development. They apply to test administration only indirectly.

2.1 Characteristics of the potential testing delivery environment(s) should be specified to match the needs of the test. It is important to make sure that items can be properly displayed in the test delivery environment and that test-taker input and results can be collected, aggregated, and reported. For example, graphics constraints need to be identified so that item writers do not create items that have too many colors or require a screen resolution that is too high for the current test delivery environment.

2.2 The test development process should include a quality assurance review of the computer-based test event. Internal review versions of the test should be created before release so that quality assurance reviews can be performed. The quality assurance reviews should use the same hardware and software conditions as the actual test delivery environments.

2.3 Item types for computer based tests should be selected to match the test purpose and target assessment objectives.

2.4 Where necessary, item writers should be trained in the technology needed to author computer based testing items that match target assessment constructs or objectives. Item writers may also collaborate with graphics designers and programmers in developing items and item display sequences.

2.5 When the presentation modality may affect test taker performance, item writers and reviewers, should view the item under conditions as close as possible to the actual testing delivery environment.

2.6 The test development process should include the development of test introduction and instruction screens. These screens welcome the test-taker and provide information about the test and test administration conditions. (e.g. time limits, number of questions, whether skipping and returning to items are allowed). Similar screens might be needed to separate subsections of the test and to prepare the test-taker for tasks to follow.

2.7 Appropriate stakeholders should minimize the impact of construct error by ensuring that test takers are given the opportunity to understand all navigation and response features required to complete the test.

2.8 Instructions, tutorials or practice tests should be created for each computer-based test to familiarize test-takers with the features of the test and items to minimize construct irrelevant variance due to computer modality. Characteristics to consider include use of specific item types (e.g., drag-and-drop, point-and-click) and test navigation features (e.g., item skipping).

2.9 Test developers should consider and design appropriate computerized test timing features. Several factors should be considered such as minimizing construct irrelevant variance; the use of introduction screens, surveys and tutorials; and item display time, item response latency, and speededness.

2.10 Developers of computer-based tests should consider how aspects of computer delivery might impact fairness and equity and take appropriate action to minimize their effect. These factors may include aspects of test design, content, specific items, or format elements.

2.11 Developers of computer-based tests should plan for, specify and document reasonable accommodations that can be useful for test-takers who have specific conditions that may make the standard form of a computer-based test an inaccurate measure of the test-taker's knowledge skill, or performance. Test developers will need to take into account the extent to which an accommodation may affect the validity of the inferences made from test scores.

2.12 Developers of computer based test should consider, specify and document item selection algorithms, scoring rules, delivery features and reporting options consistent with measurement objectives.

2.13 The test developer should consider processes and technologies to aid in item authoring and review, item classification, tracking of item changes, test version control, management of results data, etc.

2.14 Test developers should consider the use of standard item and data interchange formats to facilitate access and retrieval of information.

2.15 Computer-based tests should record the information necessary to recreate the sequence of items and test-

taker responses on completed tests, and on interrupted tests as appropriate.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 3: Test Administration

This chapter provides Guidelines about the administration of computer-based tests. It includes information for the test-taker, and information on the testing interface, the testing environment, hardware and software requirements, special accommodations for test-takers with disabilities, test security, and disaster recovery.

3.1 Test sponsors, developers and deliverers should establish acceptable hardware and software requirements (e.g. configuration standards) for the administration of computer-based tests.

3.2 Variability across testing environments should have no meaningful impact on test scores. In addition to factors such as test-taker comfort, noise level, amount of workspace, and lighting, appropriate steps should be taken to ensure that the test environments meet the specified hardware and software requirements.

3.3 The test-taker should be provided with an opportunity to become familiar with, and to demonstrate facility with, the computer testing interface and functionality. For example, a test-taker should have the opportunity to take a set of practice questions or a tutorial prior to or during the actual testing administration. instructions to test-takers should clearly indicate how to select, enter, or construct responses and use any special equipment.

3.4 When delivering in a proctored test environment, in addition to standard test administration procedures, directions to test administration personnel for computer-based testing should include information on hardware and software requirements, means of obtaining technical support, data transfer procedures, test security, how to handle problems and whom to contact with questions.

3.5 When delivering in a non-proctored test environment, the test-taker should be made aware of hardware, software and security requirements; and procedures on how to obtain technical support, how to handle problems and whom to contact with questions.

3.6 Reasonable accommodations in the test event should allow for a test experience that is fair and equitable for all test-takers. Modifications to the test event should be made in accordance with the procedures outlined by the appropriate stakeholders on the basis of carefully considered professional judgment. Examples of reasonable accommodations include a larger monitor for a test-taker with a visual impairment, the availability of an alternate means for responding to a test item for an individual with a physical or psychomotor impairment, and extended test time.

3.7 Appropriate stakeholders should establish procedures to address test event irregularities (e.g. interrupted sessions), disaster recovery and anomaly reporting.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 4: Scoring and Score Reporting

This chapter contains *Guidelines* concerning factors that influence scores, score interpretation, or reporting of scores on computer-based tests.

4.1 Any aspect of the hardware or software that might affect the interpretation of test scores should be thoroughly described in the test user's manual or other test documentation.

4.2 If test scores from different modes of administration are considered interchangeable, equivalency across modes of administration should be documented in a test users or technical test manual. In some testing applications, this documentation might not involve empirical studies of the specific referenced test, but instead rely upon well established research conducted on similar types of tests.

4.3 In explaining the nature of the test administration process, the test publisher should provide a clear description of appropriate test taking strategies that might affect the test-taker's overall score. For example, the test-taker might be unable to skip items or go back and change answers.

4.4 If computer-based tests are normed or standardized with paper-and-pencil-test data, comparability or equivalence of the paper-and-pencil scores to the computer-based scores should be established before norms are applied to scored test data. This may be especially important if the computer-based test is time constrained, or includes extensive reading passages, scrolling text, or graphical images.

4.5 The accuracy of computer scoring algorithms should be established before implementation of the computer-based test. The test score should be reproducible.

4.6 When test score information is released to stakeholders, appropriate interpretations of individual or group summary test scores should be available. The interpretations should clearly describe the meaning of the scores, the precision of the scores, common misinterpretations of the scores, and the use of the scores.

4.7 Test developers should provide information about the types of score reports that are available and the level of authorization to access each report.

4.8 When computer-generated interpretations of test responses and test scores are reported, the sources, rationale, and empirical basis for these interpretations and their limitations described should be available to the appropriate stakeholders.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 5: Psychometric Analysis

This chapter covers Guidelines on test reliability, test validity, item characteristics, and other psychometric properties. Many of the guidelines provided in this section are not unique to computer-based exams and are addressed in the 1999 *Standards for Educational and Psychological Testing*. Computer- based tests present a greater variety of testing paradigms and a unique set of measurement challenges.

5.1 When test-takers are given different items, appropriate indices of measurement accuracy should be used.

5.2 Where appropriate, the effects of test speededness on test-taker responses should be investigated.

5.3 Where appropriate, item, subtest, and test statistics should be reported on an ongoing basis (e.g. when item exposure issues may be important; to monitor parameter drift).

5.4 In cases where the content domain changes rapidly, statistical analysis of model fit of the items to the proposed test model should be evaluated regularly.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 6: Stakeholder Communications

There may be a need to educate test-takers and the public to the benefits, limitations, and capabilities of computer-based tests. Since test-takers may not understand the differences between computer-based and paper-based tests, these differences need to be communicated to all appropriate stake holders. These differences could include but are not limited to test scheduling and delivery, tutorials, practice tests, marketing, web design, scoring and results, item types and presentation models, test design, development and delivery, and copyright ownership.

6.1 To ensure a successful transition from a paper-and-pencil based to computer-based format, test organization sponsors should develop and execute a well-conceived test communications and education program. Elements of such a program might include:

- identification of key stakeholders
- explanation of benefits of computer-based testing
- explanation of reasons for the proposed changes in test design
- explanation of specific benefits of the proposed changes in test design
- impact of proposed changes on program cost
- opportunity for questions and answers
- opportunity to access practice materials

The transition plan should include ample time to ensure appropriate stakeholder understanding.

6.2 Where appropriate, test developers should provide information concerning the test purpose and test content specification to test users prior to the availability of the test.

6.3 Test-takers who are unfamiliar with a computer keyboard or a mouse should have access to practice questions, practice exams or tutorials. Any practice questions could be given prior to or during the computerized test administration process. When test-takers have had experience using previous versions of the testing system, it may be worthwhile to emphasize administration instructions specific to any changes.

6.4 Appropriate stakeholders should inform test takers about test administration procedures. For example, picture identification might be required, a calculator might not be allowed, and the testing session might be recorded.

6.5 Copyright and ownership of the test and/or test items should be protected.

PART 2: COMPUTER-BASED TESTING GUIDELINES
Chapter 7: Security

The electronic development, delivery, scoring and reporting of tests creates new security threats that test publishers and test sponsors should anticipate and work to prevent. Protecting the security of the testing process is fundamentally important to the test developer, the sponsoring organization, and the test takers. These Guidelines address some of the key security concerns for technology-based testing.

7.1 Test publishers and test sponsors should develop and implement a written test security plan that describes who is authorized to access the test items, the scores of test takers, the scoring algorithm and other sensitive test-related information. When feasible, the authorization plan should be task specific, time limited, and periodically updated. The plan should allow access to secure areas of the test development, delivery, scoring and reporting processes through authorized access and authentication with users rights, pass words, personal identification numbers (PlNs), or other secured authentication methods.

7.2 Appropriate procedures should be established to promote and maintain the security of the computer-based test through all phases of test development, distribution, administration, scoring, reporting, and review. Factors to consider include, but are not limited to, computer networks, test administration centers, and test receipt and score and report upload communication software and hardware. In many educational testing applications the security of the computer-based test may be the responsibility of the school testing personnel.

7.3 Access to items should be controlled through user rights, passwords, and/or personal identification numbers (PlNs). The security system should not allow access to all items and item banks, but only to the content areas, item banks and items the person is authoring or modifying.

7.4 When high-stakes tests are electronically distributed, test publishers and test sponsors should use industry-accepted cryptographic or related techniques to secure the tests, item displays, scoring algorithms and test-taker scores throughout the test distribution and reporting channels.

7.5 The testing environment should provide for the secure delivery of computer-based testing. Factors affecting test security include, but are not limited to, test-taker workspace, access to personal materials, level of test-taker monitoring, and the ability to encrypt and decode test material and test-taker scores and reports.

7.6 Test publishers and test sponsors should establish formal procedures to correctly identify authorized test takers.

7.7 The exam candidate should sign a non-disclosure agreement.

7.8 Appropriate procedures should be established to ensure the confidentiality of the test content, test responses, electronic images, and test data collected at a test center or other authorized test delivery location. Access to confidential test and test-taker data should be controlled and limited to authorized individuals.

7.9 The electronic distribution and administration of copyrighted tests, items, and related materials should be protected. Test publishers and sponsors should take reasonable steps to prevent the unauthorized disclosure or copying of sensitive test-related information.

7.10 Appropriate procedures should be established to limit access to authorized users and to ensure the confidentiality of test content, its electronic storage and display representation, and the responses and scores of the test-taker. Test publishers should also establish procedures to minimize the intentional or accidental alteration of test-taker data, scores, or reports.

7.11 Means to protect the copyright of test materials should be addressed in the development and delivery of the test. For high stakes tests, an acknowledgment that binds the test-taker to preserve the confidentiality of test material should occur prior to the administration of the test.