1991

10. Guidelines for Computer Testing

Bert F. Green
Johns Hopkins University

Follow this and additional works at: http://digitalcommons.unl.edu/buroscomputerdecision

Part of the Educational Assessment, Evaluation, and Research Commons, and the Educational Methods Commons

http://digitalcommons.unl.edu/buroscomputerdecision/12

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in The Computer and the Decision-Making Process by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Testing by computer is big business. Many companies are offering software enabling a psychologist to test a client by seating him or her at a computer terminal and pressing Return. The software presents the instructions on the screen, guides the test taker through some sample items to see if the instructions are understood, and then presents the test, automatically recording the responses. After one or more tests have been completed, the equipment scores the responses, and delivers test scores. But it doesn’t stop there. It then continues by printing out a complete test interpretation in fairly well-constructed narrative prose. The prose often shows a few signs of having been pasted together out of standard phrases, sentences, and paragraphs, but then so do many reports written by real psychologists.

The proliferation of testing systems and automated test interpreters has generated consternation among some clinical psychologists. Matarazzo (1983) cried “Wolf” in an editorial in Science, and went a little far, seeming to condemn all computerized testing. I replied (Green, 1983b) that there is much less concern about the computer giving the test than about the computer interpreting the test. In fact, a group at the Navy Personnel Research and Development Center in San Diego (McBride & Martin, 1983; Moreno, Wetzel, McBride, & Weiss, 1984) had just successfully transferred the Armed Services Vocational Aptitude Battery to the computer, with no major difficulties.

The Navy group used Computerized Adaptive Testing (CAT), the most important advance in cognitive testing (Green, 1983a; Weiss, 1985). In a CAT, the computer chooses the next item to be administered on the basis of the responses to the previous items. This procedure requires a new kind of test theory—classical test theory is not adequate. The new theory is called item response
theory (IRT), and is now quite well developed, although it is still new and cumbersome. Using IRT, a computer can readily tailor the test to each test taker. The Navy group has successfully used the technique to administer the Armed Services Vocational Aptitude Battery (ASVAB). It has been found that a conventional test can be replaced by an adaptive test with about half the items, at no loss of reliability or validity. For many test takers, a conventional test has a lot of wasted items—items that are too easy for the good students, items that are too hard for the poor students. If the items are chosen to be most informative about the individual test taker, a lot of time can be saved. Of course, this means developing an estimate of the test taker’s ability as the test progresses, and it implies many intermediate calculations, but the computer is good at that. An interesting by-product of CAT is that nearly everybody who takes it likes it. Such a test provides more success experiences than the lower half of the ability spectrum is used to, and does not seem to disconcert the high scorers. Also, the computer is responsive. As soon as an answer is input, another item appears on the screen; The computer is attending to the test taker in an active way that an answer sheet cannot emulate. Hardwicke and Yoes (1984) report that one recruit said, of the CAT version of the ASVAB, “It’s faster, it’s funner, and it’s more easier.”

Although computerized administration seemed to be working well in the cognitive area, there was more concern about personality tests. The American Psychological Association began getting several calls each week from its members asking about, or complaining about computerized testing. Apparently, some guidelines were needed for the users and the developers of computer-based tests and assessments. We hoped to stimulate orderly, controlled growth in an important and volatile field. The Guidelines (APA, 1986; see Appendix) address the development, use, and technical evaluation of computerized tests and test interpretations. They emphasize personality tests and personality assessments, but are relevant to all computer testing.

Why develop guidelines when we have just finished congratulating ourselves about the new joint Testing Standards (APA, AERA, & NCME, 1985)? Because the Testing Standards cover this situation only in a generic sort of way, and deserve amplification in particular details; especially computer-based assessments, that is, narrative interpretations. The new Guidelines are viewed as a special application of the new Testing Standards and as subordinate to them in case of any perceived conflict.

Some credits are in order here. Although the Guidelines can be viewed as a derivative of the Testing Standards they didn’t really grow out of the Standards, except generically. Another precursor was a set of guidelines for computerized adaptive cognitive tests, prepared for the Navy by Green, Bock, Humphreys, Linn, and Reckase (1984). However, the document that eventually evolved into the Guidelines was first prepared by Paul Hofer and Don Bersoff for a computer-
testing company (Bersoff & Hofer, 1986; Hofer, 1985). These authors drew on the Standards, the adaptive tests guidelines, and many earlier guidelines adopted by state psychological associations. Much work was also done by Barbara Wand, a member of the APA’s Committee on Professional Standards. The final revision, taking into account hundreds of useful comments from many interested APA members was a task assigned to Lyle Schoenfeldt and myself, with the able assistance of Debra Boltas of the APA staff.

The general purpose of these Guidelines is to interpret the Testing Standards as they relate to computer-based testing and test interpretation. When the circumstances of computer testing are essentially equivalent to those of conventional tests, it was presumed that the issue was covered in the Testing Standards. For example, test security is essential to the integrity and meaning of scores on any test, whether the test is administered conventionally or by computer. Users should guard computer software for a test as diligently as they would guard booklets of a conventional test, so no special mention was deemed necessary.

As a matter of fact, guarding software probably does deserve special mention, because of the peculiar standards of morality that have arisen in copying software. Many people who own personal computers have pirated some software, and don’t even feel very badly about it. We only start worrying when piracy threatens us. We are in the awkward position of saying that copying someone’s word processor is naughty but copying someone’s test is profoundly unethical. The concern is not so much the copying, but the chance that the copy won’t be guarded as diligently as the original.

An aspect of security that the Guidelines do mention is privacy and confidentiality (Guideline 15). The scores must be kept in a way that only people with a legitimate need to know may have access to them. That is one of the problems in academic record automation at universities. Once the student’s transcript is in a computer, there is the lurking fear that it can be altered by students, coaches, or others. Severe competition for grades has caused many colleges and universities to abandon the honor system, and we must beware of the possibility that an unscrupulous person might get access to the grade files, or in our case today, files of test scores, and cause real trouble.

If the Guidelines are tacit on test security, they do treat many other issues. This chapter discusses four main areas of concern: equivalence, administration, interpretation, and review.

**ESTABLISHING THE EQUIVALENCE OF SCORES**

When a conventional test is transferred to a computer, the computer scores can be interpreted using norms from the conventional test only if the conventional and
computer forms are equivalent, that is to say, essentially parallel. The Guidelines say

Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode.

Roughly speaking, the two aspects of equivalence are first, correlations and second, score distributions (see Hofer & Green, 1985, for more detail). If the cross-mode correlation is low, there is no point in going further, because the test is measuring different things in the two modes. If the cross-mode correlation is high, there is still the matter of test score distribution. If the means, standard deviations, and shapes of the score distributions are different, the computer scores will have to be rescaled, or calibrated to the conventional scale before using the conventional scale norms.

An excellent example of establishing correlational equivalence was reported by Vicino and Hardwicke (1984). They described the Navy’s validity study comparing computer and conventional versions of the Armed Services Vocational Aptitude Battery (ASVAB). With 10 tests in each mode, a $20 \times 20$ correlation matrix was obtained. Four correlated factors emerged, as they usually do with the ASVAB. The factor patterns were remarkably similar for the two modes. There are a few minor subtleties, but plainly the modes are giving essentially the same information.

Not many differences should be expected in cognitive tests due to mode of administration, but there are some. Two different studies (Lee, Moreno, & Symson, 1984; Sachar & Fletcher, 1978), done several years apart at NPRDC, show a mean shift in a test of math knowledge, but no mean difference in verbal comprehension; the correlations were very high in both cases. The mean shift was slight, amounting to about one raw score point, or about 0.25 standard deviations, in favor of the conventional test. Careful work showed that the results were attributable to not permitting review of earlier items on the test. If the math test is given in a paper version of the computer, one item per sheet, with no looking at earlier sheets, the difference disappears.

Although software could be modified to permit review, it would be awkward, and psychometrically it is better to keep items independent. However, if the computer is not to permit review, the score scale may need adjustment before using conventional norms, because the conventional format permits review.

Paragraph comprehension tests of reading can also be a problem. Some paragraphs won’t fit on the screen along with several associated questions. One can think of shifting back and forth between the paragraph screen and the query screen, which could be awkward, or the paragraph could be shortened, with only
one question about it. However, with short paragraphs there is a chance that the test might become more of a vocabulary test. Of course it would be fun to prohibit rereading the paragraph once the query is encountered, but that is clearly a different task.

Time limits are critical to equivalence. The computer is a one-on-one test administration, and in that mode, much more liberal time constraints would be possible. Tests are timed mainly as a matter of administrative convenience. But changing the timing will change the score distributions.

Highly speeded tests pose an especially severe problem. Two tests on the ASVAB, numerical operations and coding speed, are simple clerical tests. Very few errors are made. The issue is how fast the examinee can do the task. Since pressing a computer key takes much less time than marking an answer sheet, scores on computer versions of clerical tests tend to hit the ceiling. Greud and Green (1986) compared several different computer scoring schemes, and got best results by recording the time taken by an examinee to finish a fixed number of items, and then calculating a rate measure, the mean number of correct items per minute. With this score the conventional and computer forms of the test could be made equivalent by rescaling.

In the personality domain, Allred (1986) found a big difference in conventional and computer administrations of the Adjective Check List (ACL) (Gough & Heilbrun, 1980). This instrument asks respondents to examine 300 adjectives and check those that apply to them. People tend to select many more adjectives in the computer mode. The conventional ACL is a checklist; on a checklist, non-response can mean either. "This doesn't apply to me," or "I didn't read the item." The computer forces the respondent to step through all the adjectives, so pressing the key marked NO can only mean, "No, this doesn't apply to me." The effect can be reduced simply by changing the labels on the keys from YES-NO to CHECK-NEXT ITEM, but there is still a tendency to check more adjectives on the computer. Worse, most of the additional adjectives are favorable. When forced to say more about themselves, people tend to say more nice things. Partly for this reason, the cross-mode correlations are not as high as one would like, and, of course, the score distributions are very different. Forcing a response is not likely to be a problem on tests of skills and knowledge, but it could well be a problem in interest inventories, or attitude surveys.

The MMPI has a similar problem. The conventional form asks for a response of yes or no, but instructs test takers that if in a few cases they cannot say, they may leave the item blank. A test protocol with more than a few blanks is considered suspect. Again the computer cannot permit an item to be left blank passively. If a category called "cannot say" is added as a third possible response, it creates a response demand. That is, people use the cannot-say response too much. White, Clements, and Fowler (1985) claim that the effect can be minimized by not using the "cannot say" option on the computer. Very little difference was observed in their studies. The mean differences are nonzero but
slight. Individual correlations are not given, but are reported as a group to be between .5 and .7. The high scale index, greatly admired by MMPI interpreters of both electronic and human types, are not very stable in either medium. This casts uncertainty on all MMPI interpretations.

TEST ADMINISTRATION

In test administration, the computer can often be much more helpful than a test supervisor in an ordinary group testing situation, or a clinician in an office. Many people feel uncomfortable about asking for help in taking the test. For one thing, the computer can monitor the test taker’s readiness for the test (Guideline 6; see also Guidelines 3–5). By demanding active responses to the instructions and the practice problems, the computer can determine whether the test taker understands the task. The computer can refuse to move on to the main test until the demonstration and practice items have been successfully negotiated. This is a great advantage over the conventional test, where one can only hope that the test takers have read and worked through the preparatory material.

Many people are concerned that some students will be unfamiliar with computers and will therefore be at an unfair disadvantage. Guideline 7 says,

Test takers should be trained on proper use of the computer equipment, and procedures should be established to eliminate any possible effect on test scores due to the test taker’s lack of familiarity with the equipment.

This concern seems to be exaggerated. Remember that the test taker is not being asked to program the computer or even to use some special software. He or she has only to press one of a few buttons—indeed it may be wise to replace the full keyboard by a special response box. Remember also that computers are no more novel to young people today than are VCRs and phonographs. The computer is part of their world and they accept it—indeed they welcome it.

It is not the young we must worry about, it is their elders. The computer is not a part of their world, especially the older ones. The elderly need careful training—with detailed explanation of the equipment, and demonstration of what to do if some trouble occurs.

The computer is a boon when testing the handicapped (Guideline 8). Pressing keys can be made easy. The computer is especially good for the deaf. Whether it is as good as large print for the near-blind remains to be determined. Letters can be made any size, but at the expense of reduced screen capacity. Creativity is still needed here.

Adaptive testing is a major contribution of computation. In a CAT, the system’s facility in matching the item difficulty to the examinee’s ability leads to important efficiency. Whether content should also be balanced in these custom-
ized tests is still a matter of technical debate (Yen, Green, & Burket, 1986). Many have argued that tests should do more than provide a score, and do more than adapt to the overall ability level of the candidate. Tests should diagnose specific difficulties. If Johnny can’t read, where is his trouble? If Suzy can’t subtract, what is she doing wrong? In arithmetic, that can be done today. In other areas it will not be as easy but it can be done. Diagnosis is easier when assessment is built into computer-based instruction, or computer drill.

However, when new tests are to be devised, the Guidelines bow to the Testing Standards. Apart from some special opportunities in test administration, a computer-administered test is still a test, and ordinary methods apply. The sooner we start devising new tests that take advantage of the computer’s power, rather than transporting our tired old paper-and-pencil tests to the computer, the sooner some of these Guidelines can fade away.

TEST INTERPRETATIONS

Equivalence of test scores, and computer administration of tests are psychometric challenges, which are not particularly exciting to clinical psychologists. What gets the clinicians so exercised is not automated test scores, but the subsequent step of automated interpretation. If the clinician merely signs the printout and hands it over to the patient or to some third party, professional care has not been maintained. Matarazzo tells of a man who indicated, in response to some test questions, that he stayed home most of the time, and didn’t get out much. The computer diagnosed him as reclusive and withdrawn, when in fact the fellow was bedridden with a broken hip. Guideline 9 points out that any automated report should be adjusted by the clinician to take into account the context of the particular examinee.

On the other hand, the Guidelines also comment,

A long history of research on statistical and clinical prediction has established that a well-designed statistical treatment of test results and ancillary information will yield more valid assessments than will an individual professional using the same information. Only when the professional uses more information than the statistical system will the professional be in a position to improve the system’s results. Therefore, if the system has a statistical, actuarial base, the professional should be wary of altering the system’s interpretation. Likewise, if the system represents the judgments and conclusions of one or more skilled clinicians, the professional must recognize that changing the computerized interpretation means substituting his or her judgment for that of the expert.

The Guidelines then come down firmly on both sides of the issue. “The final decision must be that of a qualified provider with sensitivity for nuances of test
administration and interpretation. Altering the interpretation should not be done routinely, but only for good and compelling reasons."

When judging the appropriateness of an individual test interpretation, users need general information about the validity of the interpretive system. If the system has an actuarial base, the user needs to know the empirical facts. If validity is based on clinical judgment, as in an expert system, then the qualifications of the experts should be reported. The most useful information would come from empirical studies of the validity of interpretations produced by the system.

Test interpretation is branching out to other areas than personality assessment. Vale and Keller (1984) report developing an interpretive system for executive personnel evaluation that combines personality and ability measures. The Psychological Corporation is now marketing a system to prepare automatic assessments of a child's need for special education, the McDermott Multidimensional Assessment of Children (McDermott & Watkins, 1985). The system is well designed, and provides a lot of diagnostic information. Career guidance is also highly computerized, and the evaluation of ordinary educational progress is likely to follow.

We must consider the field as evolving its methods and standards. For the user to evaluate a test interpretation system, the user must have some idea of the basis for the various statements. A good, extensive manual is essential. In fact, there should be both a standard users manual and also a technical manual describing the technical basis for the interpretation (Guidelines 25–29).

Interpretations are often triggered by score profiles, and even response patterns. The reliability with which persons can be classed into categories becomes an issue. Consequently, discussing the reliability and validity of the narratives requires new methods. This area cries out for more technical work.

Review

The Guidelines do not suggest that all aspects of the algorithms and statement files of computer-based test interpretation systems be available to reviewers. Instead, Guideline 31 says,

Adequate information about the system and reasonable access to the system for evaluating responses should be provided to qualified professionals engaged in a scholarly review of the interpretive service.

An early version of this guideline did suggest that reviewers be permitted access to the entire system, but it quickly became clear that system publishers would not accept such guidance. Their counterproposals led to the present language.

Actually, reviewers probably could not make good use of the source code and file listings. Deciphering programs is usually difficult, and examining the code to determine what the system will do in a variety of circumstances is virtually
impossible. There are too many interactive contingencies. Moreover, most programs are not adequately annotated.

A much better reviewing strategy, it would seem, would be to use the system. The reviewer could enter sets of responses and examine the resulting interpretations. Some shortcuts could be provided. The reviewer may want to enter one response pattern, and then to alter a few of the responses to see what difference it makes. Also, for comparative purposes, it would be useful to see how each of several systems react to the same response patterns. Systems should probably be reviewed together in batches, as is now commonly done with introductory texts.

Another relevant question is the vulnerability of the system to inadvertent or malevolent responding, which can best be determined by exercising the system.

With review, as with many other areas of the Guidelines, the profession will learn as it proceeds. The Guidelines should be viewed as a living document, which will require regular attention and frequent revision. Today, they provide an important start.

REFERENCES


APPENDIX

Guidelines for Computer-based Tests and Interpretations

Committee on Professional Standards and Committee on Psychological Tests and Assessment

Guidelines. The use of computers in psychological testing and assessment is not a recent development. With the introduction of user-friendly microcomputers and software within the economic grasp of the individual practitioner, however, the variety of such uses has increased at a hitherto unequalled rate. These uses include computer administration of psychological tests, computerized test scoring, and computer-generated interpretations of test results and related information. The rapid increase in the availability and use of these applications of computer technology has served as the impetus for the writing of this document.

In addition, the market is swiftly expanding for automated test scoring services, computerized test interpretations, computer-administered tests, and software to perform these functions. It is essential that the users, developers, and distributors of computer-based tests, scoring services, and interpretation services
apply to these technological innovations the same ethical, professional, and technical standards that govern the development and use of traditional means of performing these functions.

The American Psychological Association (APA) first adopted interim standards on “Automated Test Scoring and Interpretation Practices” many years ago (Newman, 1966, p. 1141). The 1974 Standards for Educational and Psychological Tests (APA) included several references to computerized assessment. The 1985 Standards for Educational and Psychological Testing (APA) contains even more. The guidelines that follow are a special application of the revised Testing Standards and relate specifically to the use of computer administration, scoring, and interpretation of psychological tests.

Purpose

In January 1984 the APA Board of Directors instructed the Committee on Professional Standards (a committee of the Board of Professional Affairs) and the Committee on Psychological Tests and Assessment (a committee of the Board of Scientific Affairs) to develop guidelines for computer-based test administration, scoring, and interpretation. During the development of these Guidelines the Committee on Professional Standards has consisted of Susan R. Berger, William Chestnut, LaMaurice H. Gardner, Jo–Ida Hansen, Carrie Miller, Marlene Muse, Lyle F. Schoenfeldt, William Schofield (chair), and Barbara Wand. The Committee on Psychological Tests and Assessment has consisted of Wayne F. Cascio, Fritz Drasgow, Richard Duran, Bert F. Green (chair, 1984), Lenore Harmon, Asa Hilliard, Douglas N. Jackson (chair, 1985), Trevor Sewell, and Hilda Wing. Central Office staff assistance was provided by Debra Boltas and Rizalina Mendiola.

These Guidelines were written to assist professionals in applying computer-based assessments competently and in the best interests of their clients. The Guidelines were designed also to guide test developers in establishing and maintaining the quality of new products.

Specific reference is made to existing APA standards of particular relevance to computerized testing, which are abbreviated as follows: the Ethical Principles of Psychologists (Ethical Principles; APA, 1981); the Standards for Educational and Psychological Testing (Testing Standards; APA, 1985); and the Standards for Providers of Psychological Services (Provider Standards; APA, 1977). In addition, use has been made of selected sections of Standards for the Administration and Interpretation of Computerized Psychological Testing (Hofer & Bersoff, 1983).

The general purpose of these Guidelines is to interpret the Testing Standards as they relate to computer-based testing and test interpretation. They are intended to indicate the nature of the professional’s responsibilities rather than to provide extensive technical advice, although some technical material of particular rele-
vance to computer-based assessment has been included. The Testing Standards provide complete technical standards for testing. Technical guidance in computerized adaptive cognitive testing can be found in Green, Bock, Humphreys, Linn, and Reckase (1982, 1984).

When the circumstances of computer testing are essentially equivalent to those of conventional tests, it is presumed here that the issue is covered in the Testing Standards. For example, test security is essential to the integrity and meaning of scores on any test, whether the test is administered conventionally or by computer. Users should guard computer software for a test as diligently as they would booklets of a conventional test, so no special mention was deemed necessary.

The Guidelines are deliberately slanted toward personality assessment and the migration of conventional tests to the computer form of presentation. Many new tests are now being developed specifically for computer presentation, including many tests requiring novel responses. In general, the Testing Standards provides pertinent guidance for the development of such tests and should be considered to take precedence over these Guidelines.

In preparing these Guidelines, the Committee on Professional Standards (COPS) and the Committee on Psychological Tests and Assessment (CPTA) were aware that the sale and use of computerized test scoring and interpretive services extends beyond the membership of APA and that the guidelines may be of some relevance to others. Nevertheless, as an APA document, it has been appropriate to refer to APA documents throughout, even though they are binding only on APA members.

The Committees were further aware that APA standards refer to the obligations of individual members, whereas computerized testing services are usually the products of incorporated companies. The purpose of these Guidelines is to alert APA members to their personal obligations as professional psychologists when they use, develop, or participate in the promotion or sale of computerized test scoring or interpretive services, either alone or as an agent or director of a company. Furthermore, the Guidelines apply to the administration and use of tests for individual decision making. When the test results are to be used only in research or in general group evaluation, the Guidelines should be treated as advisory and in no way restrictive.

Participants in the Testing Process

Test Developer. The Testing Standards identifies the test developer as an individual or agency who develops, publishes, and markets a test. For purposes of this document it is useful to distinguish among the following: (a) the test author, who originally develops a test; (b) the software author, who develops the algorithm that administers the test, scores the test and, in some cases, provides
interpretive statements; and (c) the *test or software publisher*, who markets the computer software and accompanying documentation for the test.

**Test User.** The professional who requires the test results for some decision-making purpose. In some cases the test user provides the scores or an interpretation of the results to some separate decision maker, such as a probation officer or a director of college admissions. In that case, both parties bear responsibility for proper test use.

**Test Taker.** The individual who takes the test. In some cases, such as in a self-directed guidance system, the test taker may be the ultimate consumer and is in this sense both test taker and test user. When the test taker is the ultimate consumer, special care is needed in providing an appropriate context for understanding the test results.

**Test Administrator.** The individual who actually supervises and has professional responsibility for administering the test. In cases where the test administrator delegates the proctoring of test administration to another person, the administrator retains responsibility for adherence to sound professional practice.

Responsible actions of these various parties all contribute to the effective delivery of services to clients. Many of these responsibilities have been set forth in the *Ethical Principles and Provider Standards*. Reference is made here to these documents even though it is recognized that the parties might not be psychologists in all cases. Although binding only on psychologists, these documents provide sound advice for any person responsible for developing and offering computer-based administration, scoring, and interpretation of psychological tests.

**THE USER’S RESPONSIBILITIES**

Some aspects of testing can be carried out advantageously by a computer. Conditions of administration of some tests can be better standardized and more accurately timed and controlled when the test is administered by a computer. Test scoring can be done more efficiently and accurately by a computer than it can by hand. Test score interpretation based on complex decision rules can be generated quickly and accurately by a computer. However, none of these applications of computer technology is any better than the decision rules or algorithm upon which they are based. The judgment required to make appropriate decisions based on information provided by a computer is the responsibility of the user.

The test user should be a qualified professional with (a) knowledge of psychological measurement; (b) background in the history of the tests or inventories
being used; (c) experience in the use and familiarity with the research on the tests or inventories, including cultural differences if applicable; and (d) knowledge of the area of intended application. For example, in the case of personality inventories, the user should have knowledge of psychopathology or personality theory.

The responsibilities of users are expressed by the following clauses from the *Ethical Principles*.

**Principle 1: Responsibility**

In providing services, psychologists maintain the highest standards of their profession. They accept responsibility for the consequences of their acts and make every effort to ensure that their services are used appropriately.

*Interpretation:* Professionals accept personal responsibility for any use they make of a computer-administered test or a computer-generated test interpretation. It follows that they should be aware of the method used in generating the scores and interpretation and be sufficiently familiar with the test in order to be able to evaluate its applicability to the purpose for which it will be used.

**Principle 2: Competence**

Psychologists recognize the boundaries of their competence and the limitations of their techniques. They only provide services and only use techniques for which they are qualified by training and experience. They maintain knowledge of current scientific and professional information related to the services they render.

2e. Psychologists responsible for decisions involving individuals or policies based on test results have an understanding of psychological or educational measurement, validation problems, and test research. *Provider Standards* 1.5 and 1.6 further underscore the nature of the professional's responsibility:

1.5 Psychologists shall maintain current knowledge of scientific and professional developments that are directly related to the services they render.
1.6 Psychologists shall limit their practice to their demonstrated areas of professional competence.

*Interpretation:* Professionals will limit their use of computerized testing to techniques with which they are familiar and competent to use.

**Principle 6: Welfare of the Consumer**

Psychologists fully inform consumers as to the purpose and nature of an evaluative . . . procedure.
Principle 8: Assessment Techniques

8a. In using assessment techniques, psychologists respect the right of clients to have full explanations of the nature and purpose of the techniques in language the clients can understand, unless an explicit exception to this right has been agreed upon in advance. When the explanations are to be provided by others, psychologists establish procedures for ensuring the adequacy of these explanations.

8c. In reporting assessment results, psychologists indicate any reservations that exist regarding validity or reliability because of the circumstances of the assessment or the inappropriateness of the norms for the person tested. Psychologists strive to ensure that the results of assessments and their interpretations are not misused by others.

Interpretation: The direct implication of Principles 8a and 8c for the user of computer-based tests and interpretations is that the user is responsible for communicating the test findings in a fashion understandable to the test taker. The user must outline to the test taker any shortcoming or lack of relevance the report may have in the given context.

GUIDELINES FOR USERS OF COMPUTER-BASED TESTS AND INTERPRETATIONS

The previous references to the Ethical Principles, Provider Standards, and Testing Standards provide the foundation for the following specific guidelines for computer-based tests and interpretations.

Administration

Standardized conditions are basic to psychological testing. Administrative procedures for tests are discussed in Chapters 15 and 16 of the 1985 Testing Standards. The main technical concern is standardization of procedures so that everyone takes the test under essentially similar conditions. Test administrators bear the responsibility for providing conditions equivalent to those in which normative, reliability, and validity data were obtained. The following guidelines are of particular relevance to the computerized environment.

1. Influences on test scores due to computer administration that are irrelevant to the purposes of assessment should be eliminated or taken into account in the interpretation of scores.

2. Any departure from the standard equipment, conditions, or procedures, as described in the test manual or administrative instructions, should be
demonstrated not to affect test scores appreciably. Otherwise, appropriate calibration should be undertaken and documented (see Guideline 16)

COMMENT: A special problem with computerized administration may arise with the use of different equipment by different professionals or use of equipment different from that for which the system originally was intended. Where equipment differences are minor, it may be determined on the basis of professional judgment that test scores are unlikely to be affected. In other cases, users . . . should demonstrate empirically that the use of different equipment has no appreciable effects on test scores.

3. The environment in which the testing terminal is located should be quiet, comfortable, and free from distractions

COMMENT: The overall aim is to make the environment conducive to optimal test performance for all test takers. Ideally, a separate cubicle for each terminal is recommended. If this is not possible, at a minimum, terminals should be located in a comfortable, quiet room that minimizes distractions. Users should be prepared to show that differences in testing environments have no appreciable effect on performance.

The test administrator should be careful to ensure that the test taker is free from distraction while taking the test and has adequate privacy, especially for tests or inventories involving personal or confidential issues. The environment should be quiet, free of extraneous conversation, and only the test administrator and test taker should be in a position to see either the test items or the responses. In addition to maintaining consistency in the testing environment, this helps to prevent inadvertent item disclosure.

4. Test items presented on the display screen should be legible and free from noticeable glare.

COMMENT: (See Testing Standards, 1985, 15.2) The placement of the equipment can introduce irrelevant factors that may influence test performance. Proper design and position of the display screen will avoid reduction in the legibility of the test materials by reflections from windows, ceiling lights, or table lamps.

5. Equipment should be checked routinely and should be maintained in proper working condition. No test should be administered on faulty equipment. All or part of the test may have to be readministered if the equipment fails while the test is being administered.
COMMENT: Proper equipment design and optimum conditions do not ensure against malfunctioning equipment. To prevent disruptions such as sticky keys or dirty screens that may adversely affect test performance, there should be a schedule of regular and frequent maintenance, and the equipment should be checked for each test taker prior to its use.

6. Test performance should be monitored, and assistance to the test taker should be provided, as is needed and appropriate. If technically feasible, the proctor should be signalled automatically when irregularities occur.

COMMENT: Monitoring test performance is essential so that the user can remedy any problem that might affect the psychometric soundness of the eventual score or interpretation. For users who test a few individuals, this can be done by simply looking in on the test taker; users who regularly test large numbers of people may wish to monitor automatically. This can be done by using computer programs that notify the test proctor if a test taker is responding too quickly or slowly or outside the range of response options. Peculiar responses might generate a warning to the proctor that the test taker does not understand the test directions, is not cooperating, or that the terminal is malfunctioning. In most cases, help should be immediately available to the test taker. In the case of self-administered tests for guidance and instruction, help may not be urgently needed, but some provision should always be made for assisting the test taker.

7. Test takers should be trained on proper use of the computer equipment, and procedures should be established to eliminate any possible effect on test scores due to the test taker’s lack of familiarity with the equipment.

COMMENT: It is important to ensure that test takers are so familiar with the equipment and procedures that they can devote their full attention to the substance of the test items. Adequate training should be given to those who need it. This may require an ample store of sample items. It is very likely that such practice will reduce anxiety, increase confidence, and improve the reliability and validity of test results.

8. Reasonable accommodations must be made for individuals who may be at an unfair disadvantage in a computer testing situation. In cases where a disadvantage cannot be fully accommodated, scores obtained must be interpreted with appropriate caution.

COMMENT: Computerized testing may facilitate testing persons with some
physical disabilities by providing especially large type or especially simple response mechanisms. In other cases, the computer may place persons who have certain handicapping conditions at a disadvantage. Chapter 14 of the 1985 *Testing Standards* addresses the testing of persons who have handicapping conditions.

Although tests have been successfully administered by computer to large numbers of both younger and older adults, some older people may need special reassurance and extended practice with the equipment and can be expected to respond more slowly than younger test takers. Of course, no accommodation is appropriate when the disadvantage is what is being tested. A person with poor eyesight is at a disadvantage in a test of visual acuity; it is precisely that disadvantage that is being assessed.

**Interpretation**

9. Computer-generated interpretive reports should be used only in conjunction with professional judgment. The user should judge for each test taker the validity of the computerized test report based on the user's professional knowledge of the total context of testing and the test taker's performance and characteristics.

COMMENT: A major concern about computer-generated reports is that they may not be as individualized as those generated in the conventional manner. Some information, such as demographic characteristics of the test taker, can be included in interpretation programs so that the computer will use more appropriate norms or base rates, if they exist, and qualify interpretations to take into account the particular test taker's characteristics. But no assessment system, whether computer based or conventional, can, at this time, consider all the unique relevant attributes of each individual.

A test user should consider the total context of testing in interpreting an obtained score before making any decision (including the decision to accept the score). Furthermore, a test user should examine the differences between characteristics of the person tested and those of the population for whom the test was developed and normed. This responsibility includes deciding whether the differences are so great that the test should not be used for the person (*Testing Standards*, 1985, 7.6). These, as well as other judgments (e.g., whether conditions are present that could invalidate test results), may be ones that only a professional observing the testing situation can make. Thus, it is imperative that the final decision be made by a qualified professional who takes responsibility for overseeing both the process of testing and judging the applicability of the interpretive report for individual test takers, consistent with legal, ethical, and professional requirements. In some circumstances, professional providers may need to edit or amend the computer report to take into account their own observa-
tions and judgments and to ensure that the report is comprehensible, free of jargon, and true to the person evaluated.

A long history of research on statistical and clinical prediction has established that a well-designed statistical treatment of test results and ancillary information will yield more valid assessments than will an individual professional using the same information. Only when the professional uses more information than the statistical system will the professional be in a position to improve the systems results. Therefore, if the system has a statistical, actuarial base, the professional should be wary of altering the system’s interpretation. Likewise, if the system represents the judgments and conclusions of one or more skilled clinicians, the professional must recognize that changing the computerized interpretation means substituting his or her judgment for that of the expert. The final decision must be that of a qualified provider with sensitivity for nuances of test administration and interpretation. Altering the interpretation should not be done routinely, but only for good and compelling reasons.

THE DEVELOPER’S RESPONSIBILITIES

Developers of computerized test administration, scoring, and interpretation services are referred to the Testing Standards (1985), which provides standards for test development. The following general principles from the Ethical Principles and the Provider Standards also are relevant.

From Ethical Principles:

8b. Psychologists responsible for the development and standardization of psychological tests and other assessment techniques utilize established scientific procedures and observe the relevant APA standards.

8d. Psychologists recognize that assessment results may become obsolete. They make every effort to avoid and prevent the misuse of obsolete measures.

8e. Psychologists offering scoring and interpretation services are able to produce appropriate evidence for the validity of the programs and procedures used in arriving at interpretations. The public offering of an automated interpretation service is considered a professional-to-professional consultation. Psychologists make every effort to avoid misuse of assessment reports.

8f. Psychologists do not encourage or promote the use of psychological assessment techniques by inappropriately trained or otherwise unqualified persons.

From the Provider Standards:

1.5 Psychologists shall maintain current knowledge of scientific and professional development that are directly related to the services they render.

3.4 Psychologists are accountable for all aspects of the services they provide and shall be responsible to those concerned with these services.
When advertising and selling computer-based testing services, the following from the *Ethical Principles* are relevant.

**Principle 4: Public Statements**

Public statements, announcements of services, advertising, and promotional activities of psychologists serve the purpose of helping the public make informed judgments and choices. Psychologists represent accurately and objectively their professional qualifications, affiliations, and functions, as well as those of the institutions or organizations with which they or the statements may be associated. In public statements providing psychological information or professional opinions or providing information about the availability of psychological products, publications, and services, psychologists base their statements on scientifically acceptable psychological findings and techniques with full recognition of the limits and uncertainties of such evidence.

4b. Public statements include, but are not limited to, communication by means of periodical, book list, directory, television, radio, or motion picture. They do not contain (i) a false, fraudulent, misleading, deceptive, or unfair statement; (ii) a misinterpretation of fact or a statement likely to mislead or deceive because in context it makes only a partial disclosure of relevant facts; (iii) a testimonial from a patient regarding the quality of a psychologist’s services or products; (iv) a statement intended or likely to create false or unjustified expectations of favorable results; (v) a statement implying unusual, unique, or one-of-a-kind abilities; (vi) a statement intended or likely to appeal to a client’s fears, anxieties, or emotions concerning the possible results of failure to obtain the offered services; (vii) a statement concerning the comparative desirability of offered services; (viii) a statement of direct solicitation of individual clients.

4e. Psychologists associated with the development or promotion of psychological devices, books, or other products offered for commercial sale make reasonable efforts to ensure that announcements and advertisements are presented in a professional, scientifically acceptable, and factually informative manner.

4g. Psychologists present the science of psychology and offer their services, products, and publications fairly and accurately, avoiding misrepresentation through sensationalism, exaggeration, or superficiality. Psychologists are guided by the primary obligation to aid the public in developing informed judgments, opinions, and choices.

4j. A psychologist accepts the obligation to correct others who represent the psychologist’s professional qualifications, or associations with products or services, in a manner incompatible with these guidelines.

4k. Individual diagnostic and therapeutic services are provided only in the context of a professional psychological relationship. When personal advice is given by means of public lectures or demonstrations, newspaper or magazine articles, radio or television programs, mail, or similar media, the psychologist utilizes the most current relevant data and exercises the highest level of professional judgment.
And from the *Provider Standards*:

2.3.1 Where appropriate, each psychological service unit shall be guided by a set of procedural guidelines for the delivery of psychological services. If appropriate to the setting, these guidelines shall be in written form.

**GUIDELINES FOR THE DEVELOPERS OF COMPUTER-BASED TEST SERVICES**

The *Testing Standards* (1985) and the previous cited sections of the *Ethical Principles* and *Provider Standards* provide the foundation for the following specific guidelines for the developers of computer-based test services.

**Human Factors**

10. Computerized administration normally should provide test takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats.

**COMMENT:** For tests that involve a discrete set of response alternatives, test takers should be able to verify the answer they have selected and should normally be given the opportunity to change it if they wish. Tests that require constructed responses (e.g., sentence completion tasks) typically require more extensive editing facilities to permit test takers to enter and modify their answers comfortably. Tests that involve continuous recording of responses (e.g., tracking tasks) can make use of a variety of visual, auditory, or tactile feedback sources to maximize performance and minimize examinee frustration.

11. Test takers should be clearly informed of all performance factors that are relevant to the test result.

**COMMENT:** Instructions should provide clear guidance regarding how the test taker is to respond and the relative importance of such factors as speed and accuracy. If changes are permitted, directions should explain how and when this is to be done. Before the actual test begins, the testing system itself or the proctor should check that these instructions are understood and that the examinee is comfortable with the response device.

The availability of screen prompts, an on-line help facility, or a clock display (in the case of timed performances) may be used advantageously to guide the examinee through the test instructions, test practice, and possibly the test itself. If used during the test, such devices become a part of the test itself, and cannot be changed without recalibrating the test.
12. The computer testing system should present the test and record responses without causing unnecessary frustration or handicapping the performance of test takers.

COMMENT: Advances in hardware and software design have provided a wide range of ways to transmit information to the computer. Computer test design should explore ways that are most comfortable for test takers and allow them to perform at their best. For example, a touch-sensitive screen, light pen, and mouse may all be perceived as being significantly less confusing than a standard computer keyboard. When a standard keyboard is used, it may be appropriate to mask (physically or through software control) all irrelevant keys to reduce the potential for error.

The type of test and test item may create special design problems. Speed tests must have especially quick and uniform time delays between items to minimize frustration. Tests that require reading of long passages or that have complicated directions to which test takers may want to refer occasionally require procedures that allow display changes and recall. Diagrams with fine detail require displays with greater resolution capacity than normal. If such modifications are not possible, the test takers should be provided with the diagrams or instructions in booklet form.

13. The computer testing system should be designed for easy maintenance and system verification.

COMMENT: When teleprocessing is involved, reasonable efforts should be made to eliminate transmission errors that could affect test scores. Software design should permit ways of checking that scoring and interpretive parameters recorded on a disk, for example, remain intact and accurate.

14. The equipment, procedure, and conditions under which the normative, reliability, and validity data were obtained for the computer test should be described clearly enough to permit replication of these conditions.

15. Appropriate procedures must be established by computerized testing services to ensure the confidentiality of the information and the privacy of the test taker.

COMMENT: Several services that provide computerized administration of clinical instruments maintain confidentiality by avoiding any use of test takers' names. (See Chapter 16 of the 1985 Testing Standards.)

Psychometric Properties

16. When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should
be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode.

COMMENT: If individuals obtain equivalent scores from both conventional and computer administration, computer-specific factors will have been shown to have no appreciable effect, and the computer version may legitimately be used in place of the conventional test. If condition (a) is not met, the tests cannot be claimed to be measuring the same construct and should not be used interchangeably. If (a) is met but (b) is not, then one set of scores can be rescaled to be comparable with scores from the other test. If conventional norms are being used, then the computer test scores must be rescaled. If condition (b) is met but (a) is not, then scaling will produce similar distributions, but test equivalence has not been demonstrated. If the tests are not equivalent, new norms must be established. Chapter 4 of the Testing Standards (1985) concerns norming and score comparability. Testing Standard 4.6 states that data on form equivalence should be made available, together with detailed information on the method of achieving equivalence (see also the comment on Standard 2.11, pp. 22–23).

A number of research designs can be used to study equivalence. Differences in the means, dispersions, or shapes of computer and conventionally obtained test score distributions all indicate a lack of strict equivalence when equivalent groups are tested. Although perfect equivalence may be unattainable (and unnecessary), the following condition should be satisfied if one wishes to use norms from a conventionally developed test to interpret scores from a computerized test. Computer-obtained test scores should preserve, within the acceptable limits of reliability, the ranking of test takers. If ranking is maintained, then scale values can be transformed through such procedures as linear or equipercentile equating so that test takers receive the same score as they would have obtained through conventional administration. In this way, cutting scores, validity estimates, norms, and other data generated from the conventional scale can be applied to the computer-obtained scores. The same considerations would apply (with the obvious changes) to a test developed entirely in the computer medium that was later printed in paper-and-pencil format. The equivalence of the forms should be established before norms developed for the computer version are used in interpreting the derivative paper-and-pencil format.

The present Guidelines are conservative in suggesting empirical information about equivalence for each test that is rendered in a different presentation mode. At present some tests in some situations show differences; others do not. As the literature expands, generalizations presumably will permit accurate expectations of the effect of presentation mode.
17. The validity of the computer version of a test should be established by those developing the test.

COMMENT: Procedures for determining validity are the same for tests administered conventionally and by computer (see Chapter 1 of the 1985 Testing Standards). A new computer test should be validated in the same way as any other test. If equivalence has been established between the conventional and computer-administered forms of a test, then the validity of the computer version can be generalized from the validity of the conventional version. If equivalence has not been established, the validity and meaning of the computer version should be established afresh. At present, there is no extensive evidence about the validities of computerized versions of conventional tests. Until such evidence accumulates, it will be better to obtain new evidence of predictive and construct validity.

18. Test services must alert test users to the potential problems of nonequivalence when scores on one version of a test are not equivalent to the scores on the version for which norms are provided.

COMMENT: This will most often be a problem when comparing a computer version of a test with a conventional paper-and-pencil version, but it can also be a problem when comparing tests presented on two different computer systems. Screens of very different size, or special responding devices such as a light pen, could in some circumstances affect test norms. This is especially an issue with timed responses, which are known to vary in speed for different types of required responses. Until enough information accumulates to permit generalization about the relevance of equipment variation, caution is prudent. When a test is offered on different equipment the offerer should provide assurance of comparability of results, and the accompanying manual should reflect the different equipment.

19. The test developer should report comparison studies of computerized and conventional testing to establish the relative reliability of computerized administration.

20. The accuracy of computerized scoring and interpretation cannot be assumed. Providers of computerized test services should actively check and control the quality of the hardware and software, including the scoring, algorithms, and other procedures described in the manual.

21. Computer testing services should provide a manual reporting the rationale and evidence in support of computer-based interpretation of test scores.

COMMENT: The developer is responsible for providing sufficient information in the manual so that users may judge whether the interpretive or classifica-
tion systems are suited to their needs. Chapter 5 of the 1985 *Testing Standards* summarizes the information that should be presented in the manual.

### Classification

Certain classification systems depend on the determination of optimal cutting scores. The determination of the cutting score is, in turn, dependent on a number of statistical and practical variables including (a) the base rate of the characteristic to be inferred, (b) the error of measurement at various points along the test score scales, (c) the validity of the tests for the inference to be made, and (d) the costs of errors of classification. Balancing all these considerations is as difficult in making computerized test interpretations as it is in making clinical interpretations.

22. The classification system used to develop interpretive reports must be sufficiently consistent for its intended purpose (see Chapter 2 of the 1985 *Testing Standards*). For example, in some cases it is important that most test takers would be placed in the same groups if retested (assuming the behavior in question did not change).

**COMMENT:** There is a tradeoff between consistency and precision. The more classification decisions the test is asked to make, the less consistent will such assignments be. Making too few classifications may lead test users to ignore meaningful differences among test takers; too many may lead test users to overestimate the precision of the test.

Classification systems should be sufficiently consistent so that most test takers would be placed in the same groups and given the same interpretations if retested, and sufficiently precise to identify relevant differences among test takers. Consistency depends both upon the reliability of the test and the size of the score intervals in each class. Precision requires that the test be capable of discriminating meaningfully among test takers. Cutting scores and decision rules should take into account the discriminability of the test at different points of the measurement scale and the purposes for which the interpretations will be used. At a minimum, classification categories must represent rational decisions made in the light of the goals users have in mind. The more important the consequences for the test taker, the more assurance there should be that the interpretation and ultimate decisions are fair and accurate. Developers of interpretive systems must exercise discretion in deciding how many and what kinds of classifications will be useful.

23. Information should be provided to the users of computerized interpretation services concerning the consistency of classifications, including, for example, the number of classifications and the interpretive significance of changes from one classification to adjacent ones.
Validity of Computer Interpretations

24. The original scores used in developing interpretive statements should be given to test users. The matrix of original responses should be provided or should be available to test users on request, with appropriate consideration for test security and the privacy of test takers.

25. The manual or, in some cases, interpretive report, should describe how the interpretive statements are derived from the original scores.

COMMENT: Professionals who provide assessment services bear the ultimate responsibility for providing accurate judgments about the clients they evaluate. It should be possible to fulfill these ethical demands without infringing on the testing service’s proprietary rights. To evaluate a computer-based interpretation, the test user must know at least two facts: (a) the nature of the relationship of the interpretations to the test responses and related data, and (b) the test taker’s score or scores on the relevant measures. (In addition, raw data or item responses often will be very useful.) For example, the test developer could describe the organization of interpretive statements according to the scale on which they are based, otherwise provide references for statements in the report, or provide in the manual all the interpretive statements in the program library and the scales and research on which they are based. Each test taker’s test and scale profile can be printed along with the narrative interpretations, together with the original set of responses where appropriate.

26. Interpretive reports should include information about the consistency of interpretations and warnings related to common errors of interpretation.

COMMENT: Test developers must provide information that users need to make correct judgments. Interpretive reports should contain warning statements to preclude overreliance on computerized interpretations. Unusual patterns of item responses can lead to seemingly inconsistent statements within a single report (“the respondent shows normal affect,” “the respondent may have suicidal tendencies”). Either the manual or the introductory comments on the interpretation might indicate that inconsistent statements result from inconsistent test responses, which may indicate that the result is not valid.

27. The extent to which statements in an interpretive report are based on quantitative research versus expert clinical opinion should be delineated.

28. When statements in an interpretive report are based on expert clinical opinion, users should be provided with information that will allow them to weigh the credibility of such opinion.

COMMENT: Some interpretations describe or predict objective behavior,
whereas others describe states of mind or internal conflicts. Some interpretations are quite specific. Others are very general. Some make statements about the test taker's present condition; others make predictions about the future. Some make use of well-established, consensually understood constructs, others use terms drawn from ordinary language. The type of interpretation determines the nature of the evidence that should be provided to the user.

29. When predictions of particular outcomes or specific recommendations are based on quantitative research, information should be provided showing the empirical relationship between the classification and the probability of criterion behavior in the validation group.

COMMENT: Computerized interpretation systems usually divide test takers into classes. It is desirable to present the relationship among classes and the probability of a particular outcome (e.g., through an expectancy table) as well as validity coefficients between test scores and criteria.

30. Computer testing services should ensure that reports for either users or test takers are comprehensible and properly delimit the bounds within which accurate conclusions can be drawn by considering variables such as age or sex that moderate interpretations.

COMMENT: Some reports, especially in the area of school and vocational counseling, are meant to be given to the test taker. In many cases, this may be done with limited professional review of the appropriateness of the report. In such cases, developers bear a special burden to ensure that the report is comprehensible. The reports should contain sufficient information to aid the test taker to understand properly the results and sufficient warnings about possible misinterpretations. Supplemental material may be necessary.

Review

31. Adequate information about the system and reasonable access to the system for evaluating responses should be provided to qualified professionals engaged in a scholarly review of the interpretive service. When it is deemed necessary to provide copyrighted information or trade secrets, a written agreement of nondisclosure should be made.

COMMENT: Arrangements must be made for the professional review of computer-based test interpretation systems by persons designated as reviewers by scholarly journals and by other test review organizations, including the Buros–Nebraska Institute of Mental Measurement. Such reviewers need more information than a regular consumer could absorb, but generally will not need access to
the computer code or the entire array of statements from which interpretations are fashioned. At present, there is no established style for reviewing a CBTI system, and different reviewers may want different information. At a minimum, a reviewer should be able to communicate freely with technically qualified, knowledgeable persons associated with the test developer, who can answer questions about the system. Access to the system should be provided for trying actual or simulated test responses and for exercising the offered components of the system.

In some cases it may be necessary to impart trade secrets to the reviewer, in which case a written agreement should state the nature of the secret information and the procedures to be used to protect the proprietary interests of the test author, the software author, and the test publisher. As a rule, however, it is advisable to make readily available enough information for a reviewer to evaluate the system. This would certainly include the general structure of the algorithms and the basis for transforming test responses into interpretive reports, but it might not extend to the entire library of interpretive statements or to the specific numerical values of the cutting point and other configural definitions. The general size of the statement library or equivalent process of generating interpretations should be provided, along with information about its source. The algorithms can usually be explained in reasonable detail without disclosing trade secrets.

ACKNOWLEDGEMENT

This material is reprinted from the APA Guidelines for Computer-Based Tests and Interpretations (1986). Copyright 1986 by the American Psychological Association. Permissions fee paid. Further reproduction of this material without the expressed written permission of the American Psychological Association is prohibited. Copies of the Guidelines may be ordered by writing the APA Order Department, P.O. Box 2710, Hyattsville, MD 20784.

REFERENCES


