3. Computer Technology In Testing

Gale H. Roid

Western Psychological Services

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

http://digitalcommons.unl.edu/burosfuturetesting/4

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 Future of Testing by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Although computers have had an important role in educational and psychological testing for decades, the widespread availability of personal computers has focused interest on the appropriate role of computerization in the development, administration, scoring, and interpretation of tests. Although the early decades of computer usage found hardware and services concentrated in large computer installations, future decades will find hardware and services distributed more widely among individual users. With this rapid diffusion of technology and the lightning speed with which technology is changing, it will be increasingly difficult to predict the directions that computerized testing will take. Therefore, the purpose of this review is to discuss some broad themes in the future of computer technology as applied to testing, but, at the same time, restricting the discussion to methodology and usages that appear feasible for application in the near future.

Although it is always tempting in the discussion of any innovation, this review will resist the urge to view computerized testing or computerized interpretation of tests as a panacea for all of the limitations of nonautomated procedures. In fact, computerized test interpretation raises several new ethical issues and complications that magnify the latent problems of inexperienced test users (Zachary & Pope, 1983). It remains clear that skillful and imaginative clinical use of tests and assessment will always require the reasoned guidance of the experienced professional. Computers remain a tool to the professional, admittedly a more complex tool than previously available to the individual user.

This chapter has two time perspectives: (1) current status, and (2) future directions. Within each of these perspectives, four areas of computerization are...
discussed: (a) aides to test development, (b) test administration, (c) scoring, and
(d) the interpretation of test results.

CURRENT STATUS OF COMPUTER TECHNOLOGY IN
TESTING

Test Development

Perhaps it may seem odd for the reader interested in computerized administration
or scoring of tests to begin with a discussion of the seemingly obvious role of
computers in the development of tests. However, there are a number of steps in
test development that are often invisible to the consumer of tests, and many of
these steps have involved the use of extensive computer analyses, particularly
during the last 2 decades. One need only read the descriptions by Terman and
Merrill (1937) of the use of Hollerith machines in the processing of standardiza-
tion data for the Stanford-Binet Tests of Intelligence to appreciate how far test
development has progressed and how much computing power now lies in the
hands of the individual owner of even the most basic home computer.

Although the home computer shows promise in contributing to the efforts of
test developers, there remain some difficult technical problems and a time lag in
the adaption of packaged statistical and data analysis programs to the microcom-
puter. Perhaps by stimulating further interest in the role of computers in test
development, this chapter can contribute to an awareness of the need to bridge
the existing gap between the software available to test developers at large com-
puter installations and that available to the owner of a personal computer. Build-
ing such an awareness is seen as important because of the beneficial role that
objective procedures and data-based design can have on the generation of test
items (Roid & Haladyna, 1982), the calibration and banking of items and scales
(Bock & Mislevy, 1982; Choppin, 1968; Gorth, Allen, & Grayson, 1971), and
documentation of the psychometric properties of tests, to name only three ele-
ments of test development.

Item Writing. Although extensive reviews of item writing methodology are
provided elsewhere (Roid, 1984; Roid & Haladyna, 1980), a brief overview of
the current status of computerized item writing is given. The majority of comput-
er applications in test item generation have been in the areas of achievement
testing and instructionally based testing systems. The distinctive feature of such
applications is that item programs that direct the computer to assemble a related
set of unique items are stored, not the items themselves. Examples include the
early work by Suppes, Jerman and Groen (1966) and Atkinson and Wilson
(1969) in computer-assisted instruction, the implementation of test generators in
university science courses (Johnson, 1973; Millman, 1980; Olympia, 1975),
military-training applications (e.g., Braby, Parrish, Guitard, & Aagard, 1978), and the assessment of specific skills such as spelling (Fremer & Anastasio, 1969) and computer programming (Vickers, 1973).

The work of Suppes, Atkinson, and their colleagues at Stanford (Atkinson & Wilson, 1969) was of historical importance to the field of computerized instruction and testing because it demonstrated three concepts: (a) that individual student-computer interaction was feasible and cost-effective, (b) that sophisticated hardware and software could be designed for the specialized functions of instruction and testing, and (c) that psychological theories of learning and cognition could be integrated into daily lessons and tests in complex and experimentally meaningful ways. Their heavily funded projects were a stimulus for the development of the IBM 1500 computer-assisted instruction system, the COURSEWRITER II author language, and a lower-cost PDP-1 system that delivered drill-and-practice instruction on teletypes to as many as 3000 students per day. They developed a series of COURSEWRITER macros (computer commands that call up more detailed segments of computer programming) that allowed the individual test-like events in instruction to be varied at will. For example, sentences like “Dan saw the (tan, fat, man, run) hat,” would appear on the computer screen with the taped-recorded message “Touch and say the word that belongs in the sentence.” The segment had been programmed by a series of COURSEWRITER commands that could be varied by a macro command listing the sentence text, alternatives, response time-limit, and other parameters. The selection of sentences and alternatives was guided by a psycholinguistic theory based on *vocalic center groups* which are words containing a vowel nucleus with zero to four preceding or following consonants (e.g., at, cat, scat). Experiments could then be designed to verify rules such as “Vocalic center groups with zero preceeding consonants should be introduced to the student before those having initial consonant clusters (e.g., “at” before “gnat”).

The contribution of the work of Vickers (1973), Olympia (1975), Millman (1980), and others is the development of computer software for computerized item generation without the need for extensive item banks containing prewritten items. Vickers, for example, used a large university computer system to generate test items for a course in FORTRAN programming with an enrollment of 400 students. Random number generators were used to select item types, distractors, and the letters or numbers used to compose the names of variables in FORTRAN statements (e.g., $XY2 = JCEQ5 + N3$). Fremer and Anastasio (1969) contributed methods for programming computers to implement the erroneous rules students use in misspelling words, as part of the computerized generation of spelling test items. Hively, Patterson, and Page (1968) and Osburn (1968) advanced the theory underlying computerized item generation by describing the formal properties of “item forms.” Item forms are sets of specifications that provide a fixed syntactical structure for items and variable elements that are systematically replaced to create unique items (e.g., “What is the standard error
of measurement of a scale with a standard deviation of X and reliability equal to Y?"]. Millman (1980) and Millman and Outlaw (1978) described an extension of the BASIC computer programming language that made possible the programming of item forms with greater ease that would be required if every test were generated by its own unique computer program. All of these computerized testing projects were important in demonstrating the feasibility and methodology of systems that did not require item banks containing prewritten items.

The work of Braby, Parrish, Gitard, and Aagard (1978) moved the technology of computerized achievement testing a step further by developing systems that generated both instructional sequences and the mastery tests used to assess learning from each sequence. They discovered that most of the training programs designed to teach symbols or codes (e.g., Morse code, weather-report codes) had a generic structure that could be computerized. Sequences of teaching materials, followed by practice examples, followed by unit mastery tests were all programmable on a computer. The computer could generate and print not only the tests but the entire training manual as well.

A new and growing area for the application of computers to the generation of items (or, more precisely, test-like events) is in the area of perceptual, cognitive, and memory assessment (e.g., Barrett, Alexander, Doverspike, Cellar, & Thomas, 1982). Posner and his colleagues (e.g., Posner & Osgood, 1980) developed many sophisticated computer-control programs for laboratory computers used in assessing perceptual and memory functions. Mapou (1982) recently implemented tachistoscopic functions on a microcomputer. Kornbrot (1981) developed a specialized computer language and system for creating and running psychological experiments. Recently, increased attention to the assessment of memory functions in the aged has included the development of sophisticated microcomputer programs that involve the dynamic generation of graphic patterns on the computer screen (e.g., Gilmore, Royer, Tobias, & Ruffing, 1983; Hertzog, 1983).

For reasons that are unclear, certain areas of testing such as personality assessment have not pursued any of the technologies of test-item writing (Roid & Haladyna, 1980) that lead to automated item generation as is done in achievement or aptitude measurement. Although computerized versions of standard psychological tests have proliferated, little is currently available that can be truly described as generative of items in the same sense as is done in the composition of a mathematics problem using random number generators. The early work of Colby and associates (e.g., Colby, Watt, & Gilbert, 1966) on the generation of counseling or psychotherapeutical conversations has not been followed by a widespread application to psychological assessment. Perhaps this is because of the extremely complex nature of human dialogue and natural language.

Although it may be difficult to adjust the natural language sequences that occur in personality inventory items, certain key words can be rather easily adjusted to match the characteristics of examinees who are completing psycho-
logical inventories. Johnson, Giannetti, and Williams (1979) have developed some response-contingent systems that adapt to the demographic characteristics of the examinee. For example, if the examinee has only an older sibling, certain items would be reworded to refer to "your older brother" or "your older sister."

**Item Calibration and Item Banking.** Whether test items are generated by the computer, or written offline and simply stored in computer files, it is possible to collect and store them in extensive "item banks" or "item pools." Again, educational measurement has led the way in the development of numerous and extensive item banks at regional or university centers (e.g., Gorth, Allen, & Grayson, 1971), school districts (e.g., Forster & Doherty, 1978), or at the national level (e.g., Popham, 1980; Wood & Skurnik, 1969). Also, recent years have seen a proliferation of published, criterion-referenced achievement tests that feature the possibility for school districts to adapt test content to their particular curricular emphases. These developments are part of the movement toward a closer linking of testing and instruction in educational program evaluation (e.g., Airasian & Madaus, 1983).

Another arena in which educational measurement has broken new ground is in the development of theories and statistical models that are an alternative to classical test theory. The development of these *item response theory* (IRT) models such as the Rasch model (Rasch, 1980; Wright & Stone, 1979) and the 2- and 3-parameter models (Lord, 1980; Hambleton & Swaminathan, 1985) have been the cornerstone of new methods of computerized adaptive testing (Urry, 1977; Weiss, 1979). Implementation of these methodologies to microcomputers (e.g., Bock & Mislevy, 1982) are just now appearing on the horizon. The promise of such methods is that they will save up to 50% on the number of items required for equivalent precision in comparison to longer paper-and-pencil instruments, and will provide more precise measurement by matching the difficulty of items to the functioning level of the examinee (Haladyna & Roid, 1983; Hansen, 1969).

A number of important applications of the Rasch and 3-parameter IRT models to educational achievement tests and aptitude or intelligence batteries have appeared in recent years. Woodcock (1973, 1978) has shown, through the application of Rasch scaling to widely used tests in reading achievement and psychoeducational assessment, that IRT models can help to provide accurate diagnostic information. By showing the positioning of test items (and the skill-content of the items) along the underlying Rasch scale of a test, Woodcock has advanced the interpretability of IRT applications in the practical world of assessment in school psychology and special education. Of similar impact is the recent work by Elliott (1983a) in applying the Rasch model to the development and interpretability of a complex and comprehensive intelligence and achievement battery that includes 23 subtests that span the ages of 2 1/2 to 17. Because each item in each subtest has a calibrated Rasch difficulty, it is possible to assemble
short forms of each subtest, and to link new items to an existing subtest (Elliot, 1983a, pp. 25–29). Also, Elliott (1983a) has argued that Rasch scaling allows subjects tested on different subsets of items to be compared on a common scale, so that younger and older subjects can be compared even though they are given different item sets, and change over time can be measured on a common scale. Numerous studies have proliferated in recent years claiming the inaccuracy of Rasch scaling for vertical equating (e.g., Slinde & Linn, 1979) and other purposes. In a recent rejoinder to such studies, however, Elliott (1983b) expressed concern that comparative research on IRT models has tended to apply the 1-parameter model to preexisting item and test data that were not specifically developed to fit that model. For example, Rasch himself expressed concern about multiple-choice items in describing the military intelligence tests he used to develop his most widely applied model (Rasch, 1980, p. 62)—“V is a test of verbal analogies, formally a multiple choice test, but with so many answers offered that the deficiencies of a multiple choice test are practically eliminated.”

The contributions of Woodcock and Elliott to computer-based item calibration for widely used tests rests on their development of items and tests targeted to the Rasch model, with rigorous attention to model-data fit during the development and field testing of items. In contrast, the indiscriminant application of IRT models to any tests, especially those not specifically designed to fit the model, would seem to be “technology gone wild.” The challenge of computerized item calibration in the future will be to ensure its appropriate application, not as an appendage, but as a central part of an item and test development effort.

Another major example of the application of IRT models to widely used achievement tests is the recent development of the Comprehensive Tests of Basic Skills (CTBS), Forms U and V, (CTB/McGraw-Hill, 1982; Yen, 1982) using the 3-parameter model developed by Lord (1980) and others. By calibrating difficulty, discrimination, and guessing parameters of individual items, it is claimed that the response patterns of each student are treated differently (because each item is weighted in a formula score on the basis of difficulty, discrimination, and guessing parameter weights). Clearly, without computerized item calibration and computerized scoring, such complex weighted scores would not be feasible for users who wish to survey the achievement of large groups of students.

A few isolated examples of the calibration of personality inventory items via item-response theory models have been reported such as the Rasch analysis of the Tennessee Self Concept Scale (Stanwyck & Garrison, 1982), and the Rasch analysis of the Central Life Interest measure (Schmitt, 1981) used to assess an employee’s degree of job orientation. A good deal of conceptual work has been done on the application of latent-trait or item-response theories to personality and attitude scales, but there have been few actual implementations of the item banking concept or computerized-adaptive testing outside of achievement or ability testing.
Damarin (1970) constructed a rather elaborate theory including specification of the probability that a subject responds "true" to an item as a function of the individual's position on one or more latent dimensions. However, Damarin concentrated his applications on the problem of acquiescence and response bias on the MMPI, without extension to the general problem of calibrating items on content dimensions such as depression, anxiety, etc. A more recent theoretical contribution has been made by Thissen, Steinberg, Pyszczynski, and Greenberg (1983) using LISREL analyses (Joreskog & Sorbom, 1981) in attitude scale construction. Perhaps the most extensive discussion of item-response theory analysis of rating scales or questionnaire items is that of Wright and Masters (1982) using the Rasch model. All of these are important theoretical contributions, but there has been a lag of several years in the distribution of computer programs to analyze test items so that item banking and computerized adaptive testing can be implemented. Examples of actual item calibration, followed by computerized-adaptive administration of personality inventories or scales are rare. In a study of the California Psychological Inventory, Sapinkopf (1978) used the adaptive methodologies of Weiss (1979) to tailor items to subjects. A considerable savings in the number of items administered was achieved (67% fewer items), but with some reduction in scale reliability. Although lower reliability may be expected if fewer items are used, adaptive methods draw items from a large pool of potential items, presumably selecting the most precise items for each subject. Thus, Hansen (1969), Weiss (1979), and Hambleton and Swaminathan (1985), have shown empirically that sequential adaptive tests can provide greater reliability and precision with fewer items than conventional tests in the area of achievement assessment.

Developing the Psychometric Properties of Tests. Perhaps more than any other contribution of computer technology to testing, the use of large scale computers in the sophisticated, multivariate analysis of test data has contributed importantly to the overall improvement in the precision and accuracy of educational and psychological tests during this century. Although somewhat invisible to the average user, the monumental contributions of computers to the development of college entrance exams, professional licensure tests, standardized achievement batteries, and objectively scored clinical instruments is staggering. In fact, by the very nature of the Joint Technical Standards for Educational and Psychological Testing (American Psychological Association, 1985) and their requirement for quantitative indexes of reliability and validity, it is difficult to imagine a test that could be developed and published without some form of computerized analysis.

A new emphasis in the psychometric development of tests is the establishment of the validity of computerized interpretations and computer-generated reports. The development of specifications for a fully interpretive computer report forces
one to pose detailed questions relating the research base of a test to specific numerical rules of score interpretation. For example, research on MMPI profile elevations (e.g., identifying the 2 or 3 highest scores on the standard profile) provides a basis for generating descriptive statements when such elevations are found in an individual’s profile. On the Barclay Classroom Assessment System (Barclay, 1983), a specific range of low scores on peer and teacher support in the classroom signals a student at risk for psychosocial stress that may effect learning and achievement. On other interpretive reports (see following section on Test Interpretation) narrative material has been generated based on correlations between clinicians personal observations and profile patterns. Such computerized interpretations require particular kinds of research targeted for eventual use in documenting computerized decision rules. A rigorously valid interpretive report forces the developer to plan verification studies (or the adaption of existing research into the framework of computerized decision rules) that might otherwise not have been so obviously needed. Thus, the computer has an important role in both facilitating the completion of decision-rule verification studies, and in encouraging the validation of specific interpretations of tests. As the new Joint Technical Standards (APA, 1985) have emphasized, there are as many “validities” or kinds of validity evidence for a test as there are interpretations or uses of the test. Computerized reporting may have inadvertently heightened the attention to this important principle. Further discussion of the controversies surrounding computerized test interpretation is discussed under the heading “Test Interpretation.”

Despite the importance of computer analysis to the development of tests, there remains a surprising void in the availability of computer programs specifically designed in an integrated package for the development and analysis of tests and test items. Perhaps because so few centers of test development exist and because the developers of questionnaires and other informal scales are often not simultaneously skilled in their content area, computer programming, and measurement statistics, there has previously been no integrated and widely used set of computer programs for test analysis similar to the statistical packages currently in worldwide use. The computer-assisted data analysis (CADA) package of Novick and colleagues (Novick et al., 1983) has recently been expanded and released in a new package that includes a component on psychometric methods containing score equating and item analysis (norm-referenced and criterion-referenced) modules. The CADA system is probably the most comprehensive test-analysis package available, but it also emphasizes Bayesian statistical method with which some users may be unfamiliar. Consequently, the test developer must use considerable ingenuity in putting together a sequence of computer programs to develop a high-quality test or scale. If measurement is to further develop as a science, it would seem to be crucial to have objective methods of field-test refinement, reliability and validity estimation, item- and test-bias determination, and derivation of methods of test interpretation that exist as an integrated whole. Such an
integration would be possible for the test developer if easily usable sets of computer aides were readily available and tailored to the specific needs of test development.

Test Administration

One of the assumptions of this chapter, which is widely shared by many measurement professionals, is that computer applications are ill-fated unless they provide a new dimension to testing not possible with nonautomated techniques. For example, use of the computer as a "page turning" device to present the same questions appearing on a printed form is clearly degenerative, unless another function such as automatic storage and interpretation of data is coexistent with it. Even the use of a microcomputer time clock may be a frill unless each response is timed and interpreted or some other dimension of timing is implemented that would be too costly or difficult with a hand-held stopwatch.

The best of the current offerings in computerized test administration do add considerable benefits over and above those afforded conventional assessment techniques. Klingler, Johnson, and Williams (1976) have shown that savings in staff time, and increased acceptance by patients, have justified the use of a comprehensive system for the computerized intake assessment of mental health patients in large facilities. Urry (1977), Schmidt, Urry, and Gugel (1978), Weiss (1979), Croll (1981), Weitzman (1982) and others have shown the advantages of computerized adaptive testing in achievement and aptitude assessment in government, military, and university settings.

Adaptive Testing. What are the purported advantages and the possible limitations of adaptive, computerized testing? A review of the literature in adaptive, computerized test design reveals at least five major advantages claimed for the technique:

(a) increased precision of measurement,
(b) improved efficiency and time savings for the examinee,
(c) increased breadth of trait or achievement levels assessable, resulting in more accurate decision making (e.g., identification of students who have mastered a unit of instruction),
(d) improved examinee motivation due to the perceived objectivity of computers or fairness in the choice of items selected for a given examinee, and
(e) possible technical improvements in the selection of test items having desirable statistical properties for certain subjects.

In terms of increased measurement precision for adaptive testing, Hambleton and Swaminathan (1985), Lord (1980), and Weiss (1979) have shown that information curves of adaptive tests are superior to conventional tests, particularly for
examinees who are of lower or higher ability than the average difficulty level of
the conventional test. Haladyna and Roid (1983) recently showed that adaptive
criterion-referenced tests had lower errors of measurement than conventional
tests composed of items randomly sampled from achievement-item domains.

Studies of adaptive testing can be traced back to Cowdon (1946) and Fiske
and Jones (1954) who argued for the cost-efficiency of sequential testing in
which costly assessment items (e.g., medical or physiological measures each of
which is time consuming or intrusive to the examinee) can be evaluated in
sequence, with testing terminated when an underlying parameter (such as propor-
tion correct or probability of illness) can be estimated accurately. Hansen (1969)
presented through evidence that an adaptive, science-achievement test could be
administered via computer with 50% fewer items (but equivalent or better infor-
mation precision) than conventional tests. Weiss (1979) similarly found a cost-
efficiency in which 50% of the length of achievement tests in college biology and
military technical exams were saved by computerized adaptive testing strategies.

Another advantage of adaptive testing is in the breadth of coverage possible
when sequential levels of tests can be designed. In the statewide assessment of
special-education students (Brodsky & Roid, 1977), for example, adaptive tests
have been designed to span the broad range of functioning characterized by
mildly to severely retarded children (e.g., in dressing or eating skills as well as
academic achievement). Adaptive assessment saves time for the examiner who
has many students to individually test and yet provides a scaling such that each
level of the test can be related to a common numerical scale. Cleary, Linn, and
Rock (1968) argued that the breadth of coverage possible with adaptive achieve-
ment tests would prevent the topping out or bottoming out that can occur when
examinees represent a wide range of skill or achievement. Clearly, a test which
tops out has little variance from which differential evaluation of individuals can
be completed. In addition, students or applicants for employment may perceive
adaptive tests as more fair and less punishing because tests are tailored to their
level of competence (Schmidt, Urry, & Gugel, 1978).

Finally, adaptive testing makes possible the implementation of various tech-
nical improvements in the selection of test items. Tests can be specifically
designed to be highly sensitive to abilities or achievement in a narrow range such
as that defined by a mastery criterion or cutting score level (Lord, 1980). In
addition to selecting items based on their difficulty level, the discriminating
power of items or their relationship to external criteria can be used for composing
tests having differential validity for specific individuals. Hansen (1969), Fossum
(1973), and Roid (1969) presented experimental adaptive methods of selecting
items to increase the correlation between a test and external criteria. Small but
important improvements in validity were found by these methods. However, the
potential of such methods is limited to cases in which the cost of testing or the
importance of decision making (e.g., in the assessment of suicide potential) is
significantly high.
With all of the potential advantages of adaptive testing, a very perplexing question for many enthusiasts of tailored tests is, "Why has adaptive testing been so slow to appear as a widely-used method?" McArthur (1984) has pointed out some important reasons for the resistance to adaptive, computerized methods: (a) that the American tendency is not to accept packaged curricula or test-item banks that do not allow local control of content in academic subject matter, (b) educators and other professionals may be concerned that adaptive tests place too much faith in individual items, and (c) there has been an information gap and an absence of widely implemented software that has prevented the development of the knowledge and skill needed by small institutions and individuals to apply the Rasch, 3-parameter, or other IRT models to practical testing problems. Other factors that have contributed to the slow growth in adaptive testing may be the unavailability of large sample sizes required for item calibration (although some Rasch devotees would argue that samples as small as 50–200 may suffice), the complexity of score interpretation in comparison to simple number-correct scoring, the enormous work required to make an IRT-based scale curriculum-referenced (Haladyna & Roid, 1983), and the cost concerns surrounding the hardware and software needs of an individualized, computerized testing system for large numbers of examinees.

The promise of increased precision, breadth of coverage, and efficiency remain, but widespread dissemination of adaptive testing may require future technological developments such as lower-cost respondent keyboards with built-in storage for remote data acquisition. But, more importantly, there remains a need for more practical demonstrations and more time for professionals to learn the terminology and inner workings of IRT methods. These and other future perspectives will be discussed in a later section, "Future Directions."

Research on Computerized Test Administration. Numerous studies over the last decade have contrasted computer-administered and paper-and-pencil versions of widely used tests. Some of these studies have involved straightforward administration of items, without adaption to the examinee, but nevertheless have been important in documenting the similarities and the small number of differences between computer and conventional test administration. Results of these studies are briefly presented in Table 3.1.

The majority of studies reviewed and presented in Table 3.1 showed nonsignificant differences between computer administered and conventional test administrations. This is somewhat surprising for performance vocabulary tests such as the Peabody Picture Vocabulary Test. Elwood (1969) found slightly lower WAIS IQ estimates from an automated administration which could be explained as an elimination of examiner subjectivity as easily as an interference from automated procedures. One study found differences in state anxiety under computerized administration (Lushene, O’Neil, & Dunn, 1974) but another found no differences (Katz & Dalby, 1981). One of the studies (O’Brien & Dugdale,
### TABLE 3.1
Studies Comparing Computer Versus Conventional Test Administration

<table>
<thead>
<tr>
<th>References</th>
<th>Test Name or Type</th>
<th>Examples of Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elwood (1969)</td>
<td>WAIS (Wechsler Adult Intelligence Scale)</td>
<td>High correlation between computer and face-to-face testing with performance IQ lower on automated WAIS. (N = 35)</td>
</tr>
<tr>
<td>Hedl (1971)</td>
<td>Slosson Intelligence Test</td>
<td>Correlation of .75 between computerized and conventional, but higher state anxiety on computer. (N = 48)</td>
</tr>
<tr>
<td>Lushene, O'Neil, and Dunn (1974)</td>
<td>MMPI</td>
<td>Correlations between computer and booklet modes comparable to booklet and card form correlation. (N = 63)</td>
</tr>
<tr>
<td>Scissons (1976)</td>
<td>CPI</td>
<td>Differences between subscale scores between modes particularly for males. (N = 20)</td>
</tr>
<tr>
<td>Klinge &amp; Rodziewicz (1976)</td>
<td>Peabody Picture Vocabulary Test</td>
<td>No difference in IQ or test-retest reliability between modes of testing. (N = 52)</td>
</tr>
<tr>
<td>Elwood &amp; Clark (1978)</td>
<td>Peabody Picture Vocabulary Test</td>
<td>Nonsignificant differences between testing modes on IQ or test-retest practice effects. (N = 76)</td>
</tr>
<tr>
<td>Biskin &amp; Kolotkin (1977)</td>
<td>MMPI</td>
<td>No differences except computer slightly higher on Paranoia scale and lower on &quot;Cannot Say&quot; scale. (N = 165)</td>
</tr>
<tr>
<td>O'Brien &amp; Dugdale (1978)</td>
<td>A questionnaire on personal bathing habits</td>
<td>Tendency for computer responses to be nearer the &quot;honest&quot; end of each scale. (N = 126)</td>
</tr>
<tr>
<td>Johnson &amp; White (1980)</td>
<td>Wonderlic Personnel Inventory</td>
<td>Elderly subjects given training on the computer prior to testing performed significantly higher than those not treated. (N = 20)</td>
</tr>
<tr>
<td>Harrell &amp; Lombardo (1984)</td>
<td>16PF</td>
<td>Multivariate analyses revealed no significant differences between Apple II computer and standard booklet modes of presentation. (N = 80)</td>
</tr>
</tbody>
</table>
3. COMPUTER TECHNOLOGY IN TESTING

1978) suggested increased honesty or openness for the computer administration. Biskin and Kolotkin (1977) also found the expected decrease in number of "Cannot Say" responses on the MMPI, indicating increased willingness to make a response commitment. With computers, the examinee's responses can also be screened during testing to catch any unexpected responses or the double marks common to paper-and-pencil answer forms. However, the differences in personality scale scores found in some studies listed in Table 3.1 indicate the highly specific effects that computerized administration can have.

Test Scoring

Clearly, computer technology has had an enormous effect on the scoring and reporting of results for standardized achievement tests used in the nation's schools. Although school personnel occasionally regret the inevitable delays between testing and receipt of reports, it is truly remarkable that literally millions of answer sheets are processed each year, with detailed reports generated for individual students, classrooms, school buildings, districts, and States. Also remarkable is the proliferation of types of derived scores for educational tests—percentiles, stanines, NCEs, grade-equivalents, and normalized standard scores. With the development of computer algorithms for the direct calculation of normalized standard scores from percentages (Beasley & Springer, 1977), some of the common derived scores can be obtained directly rather than retrieved from lengthy tables.

In psychological testing, computerized scoring allows the derivation of complex scores such as factor scores, Bayesian-derived probability scores for low base-rate behaviors such as suicide (Greist, Gustafson, Strauss, Rowe, Laughren, & Chiles, 1973; Vanderplas & Vanderplas, 1979), item-option weighted scores (Cull & Gill, 1982; Roid, 1983a), profile similarity indexes for the test scores of married couples (Krug, 1983), weighted scores from tailored, adaptive, or multilevel tests calibrated with the 3-parameter model (Lord, 1980; Weiss, 1979), and sociometric ratings from entire classrooms contrasted with self and teacher ratings from individual students (Barclay, 1983). For psychological and vocational tests having complex and numerous scores, such as the MMPI, the Strong-Campbell Interest Inventory, the Tennessee Self Concept Scale, or the 16 Personality Factor Questionnaire, computer scoring provides a richness of interpretive data that could not otherwise be obtained without enormous effort.

The studies listed in Table 3.1 also suggest that, with few exceptions, the published norms of tests may be applicable without adjustment to computerized version of these tests. However, more research is needed on the effects of computerization. It may be that effects on score distributions are slight but highly specific to each instrument, hence, the need for comparability studies similar to those required for alternative forms of tests. The availability of inexpensive home computers and methods of field testing in which a computer is left unat-
tended but responsive to each examinee who voluntarily steps forward in locations such as community centers (McArdle & Kirson, 1983) show promise for the automated collection of norms specific to computerized test administration.

Test Interpretation

Although the interpretation of test results for individuals continues to require first-hand knowledge of the examinee by experienced clinicians, educators, or other professionals, the computer has an increasing role in the processing of test results using statistical and actuarial methods that are complex or time consuming for the professional. When test results are aggregated for classrooms, schools, or other groupings computerized summarization of results seems natural and widely accepted particularly in educational measurement. With the recent advent of computerized interpretation of individual psychological, vocational, employment screening, medical-psychological, special education, and counseling instruments, concern the for bounds of acceptable computerization has increased. The important ethical issues in the clinical use and interpretation of computerized test results are complex, and thorough reviews are available elsewhere (e.g., Bersoff, 1981; Zachary & Pope, 1983).

The best of available test-interpretive computer programs for major vocational and psychological tests have been designed on the basis of empirically validated decision rules and intended for the use of the trained professional who is otherwise experienced with the instrument and its supportive research. Unfortunately, the worst of available programs include the private and subjective narratives of individuals who developed the programs without benefit of empirical studies. The problem facing the field of measurement is to provide criteria for distinguishing between the objective and subjective programs.

Much of the controversy over computerized psychological testing (e.g., Matarazzo, 1983) is based on at least four prominent concerns:

1. that it is questionable whether there are real advantages to computerized interpretation of tests as compared to the clinician working without a computer,
2. that computerized interpretive reports will reach the hands of inexperienced or unqualified individuals who will respond to the halo-effect of objectivity projected by a computerized report,
3. that publishers or developers of computerized interpretive programs may not openly reveal their decision rules for professional review, and
4. that computerized reports (particularly those that cannot be evaluated closely) will not be sufficiently validated.

Each of these concerns are discussed in turn.

Advantages of Computerized Interpretation. The controversy seems to spring from a combination of true belief in the richness of clinical interpretation,
philosophical differences in approach to clinical prediction (Meehl, 1954), as well as an information gap that separates those who are familiar with the inner workings of interpretive programs and the professional consumer. Obviously, the long-standing debate on clinical versus actuarial prediction will continue, with proponents of clinical prediction citing examples of poorly constructed computer reports and proponents of actuarial prediction citing examples of rigorously validated, empirically developed reports. Devotees of empirically based interpretive reports who never intended such reports to replace all clinical judgment do not understand the arguments that say all computerized reports attempt to replace human interpretation. An analogy with statistical computer programs is useful in pointing out that because some users of multivariate analysis of variance, for example, may misapply the underlying general linear model, can the statistical program therefore be condemned? And, more pointly, should a misused multivariate analysis program be outlawed because some users may reach false research conclusions (by violating the assumptions of the program), and all multivariate analyses of variance be conducted by hand? Developers of sophisticated computerized reports would argue that their validated programs include complex calculations and decision rules that approach the complexity of some of the packaged statistical programs in wide use today.

Several practical advantages to computerized interpretive programs seem clear to most developers of such programs. First, accuracy of scoring and retrieval of norms from complex norm tables should provide a measure of quality control (many of those who supervise interns would readily attest to the frequent error rates in psychological test scoring and norms retrieval). Second, the time saved by clinicians who are relieved of hand-scoring and profiling could be invested in additional testing or personal interviews that would supplement and add fidelity to the computerized report. Third, because the decision rules for interpreting multiscale tests are often complex and numerous, it seems illogical to argue that human memory can retain and access all such rules in the same fractions of seconds required by a computer—therefore the computer acts as a memory aid. Fourth, when there is research showing moderator effects on test interpretation (e.g., that certain age groups or ethnic groups have different ranges or patterns of scores), a computerized interpretive report again provides a memory aid which reminds the clinician of such moderator effects. Finally, as is detailed below, there are numerous technical advances in profile analysis and statistical processing of scores that would be impossible to implement in a hand-scoring system without complex calculations by each clinician (and a resulting complexity in the published profile sheets or test manuals).

Unauthorized Use. The concern that unqualified users will be attracted to computerized interpretations seems to be an issue of controlling access to such programs or reports, and the ethical responsibility of test distributors and users. Clearly, any system of controlling access will be imperfect to some degree if an unqualified user is determined to bend the rules to obtain a copy. The screening
currently done by test publishing companies is more extensive than may be apparent to the professional consumer (e.g., qualifications questionnaires, approvals by supervisors, registry procedures that assign user-numbers only to qualified applicants), but, even so, it is not perfect if applicants exaggerate qualifications or lend copies of software to those who were not screened. Also, the frequent practice of honoring an institutional order for test materials from an approved clinic, hospital, or school can result in the placement of test materials in a location where both qualified and unqualified users may gain unauthorized access to computerized reports. Therefore, the Joint Technical Standards (APA, 1985) emphasize the individual responsibility of the users of computerized interpretive reports to be familiar with the research base of such reports, to have test manuals available for reference, and to use appropriate caution in making decisions from these reports. Clearly, both distributors and users (individuals and institutions) need to continue to examine their procedures for allowing access to computerized interpretive reports with an eye to the problems of potential unqualified use.

**Documentation of Decision Rules.** The ethical responsibility of the developer of computerized interpretive reports is to carry out and document validity studies of the underlying decision rules for each program. However, a controversy has developed over that extent to which the inner workings of a commercially-distributed computer report should be exposed. From the view of the pure scientist and individual helping professional, it must seem that all such aids to assessment should be part of the public domain, contributing to the advancement of science and the health and welfare of people. For programs developed by public funds, this view seems entirely appropriate. Also, if a researcher wishes to donate his or her efforts to the social good, it becomes a matter of individual choice to make such a contribution. However, if a private individual or organization has invested years of study, research, computer programming, and other resources into an inventive program, and is not in a position to donate such efforts to free public use there is legitimate concern over protection of one’s proprietary rights. From the developer’s view, there are at least two important issues: (a) the legal and ethical rights of developers who wish to retain their rights to an inventive creation, and (b) the economic realities of producing and distributing computerized reports.

The issue of rights to an inventive creation are clarified by the wordings of laws and regulations for patents and copyrights, and the many formalized procedures established by universities and research centers. Also, case law which develops from the successive decisions of court actions in representative cases (e.g., in disputes over copying of computer software), is relevant here. A discussion of legal issues goes beyond the present review, and the reader would need to consult with recognized legal experts.

The importance of the economic aspects of computerized program development was discussed several years ago by Campbell (1976) when the Strong-
Campbell Interest Inventory became the first major computer-scored test to withhold its scoring weights from public release. Campbell argued that the "pragmatic research scientist" must recognize that research funding to expand and improve an inventory, and to "cover all new issues" (e.g., the new concerns about the ethnic bias of tests for the growing subpopulation of hispanics comes to mind) is not usually supported by public or nonprofit foundation grants. Therefore, if a widely used commercial instrument is to be improved over time, revenue must be protected from the erosion created by copying of tests and the proliferation of competing commercial scoring services. Unfortunately, years of experience have shown publishers that one of the only effective controls for certain tests is to withhold keys, norms, or portions of the interpretive decision rules. The present author would argue that documentation of the validity of decision rules is not incompatible with securing the rights to a program. It should be possible in nearly every case to withhold some central element of scoring- or interpretive-program logic and still document in detail the validity of the resulting report. For example, the research base (including all references to published articles) and even most of the numerical decision rules can be revealed in the documentation of a program without having to publish the entire operating specifications of a scoring/interpretive program.

Another approach to documenting computerized interpretive programs and discussing the validation of such programs is to provide a typology of different categories of programs. A typology of computerized programs in testing would allow developers and users to have appropriate descriptive labels to distinguish one program from another. A typology with examples of how each category of program might be validated is given in the following section.

A Typology of Computerized Interpretive Programs

A proposed typology for programs, which may be useful in labeling and distinguishing among the various commercially available programs, uses four categories (Roid & Gorsuch, 1983): (a) scoring only, (b) descriptive, (c) clinician-modeled, and (d) clinical actuarial. Proper categorization and labeling of commercially offered programs, particularly those for microcomputers, would contribute to informed usage. Each category of program beyond the scoring-only level will be briefly reviewed (the reader interested in scoring programs is referred to the previous section on "Test Scoring").

Descriptive Programs. Once the subtest or scale scores for a test are available, and perhaps presented on a profile, quantitative criteria such as cut-off scores are often applied to describe the test results. Because the computer can store literally thousands of quantitative criteria and descriptive words attached to give the criteria meaning, computerized descriptive programs can be useful to the trained and experienced test user. The simplest of descriptive programs provide phrases such as "above average," "in the gifted range," "indicates mastery of"
this objective,” “in the disabling range,” and “possible organic problems,” to list only a few of hundreds of examples. The more sophisticated descriptive programs use alternative modifiers within the same score range (“average,” “at the mean,” or “typical performance”) so that the report does not become overly redundant and repetitive with the same descriptors used line after line. Some other attributes of the better descriptive programs include: (a) selection of descriptive words based on empirical studies of language, (b) narrative paragraph composition, and (c) statistical description of differences among subtest scores, each of which is briefly described.

Empirical research on the scaling properties of words, modifiers, adverbs, and other verbal phrases have been used to design Likert-type rating scale questions, but may also have a role in the design of descriptive computer reports of tests (Gorsuch, 1982). Hakel (1968) and Lichtenstein and Newman (1967) studied the scaling properties of words and phrases such as “often,” “seldom,” “very likely,” “unlikely,” and “highly improbable.” Altemeyer (1970) studied the equal-interval scaling of sets of adverbs such as “completely, substantially, somewhat” and identified several with good interval scaling properties. More recently, Pohl (1981) analyzed 39 expressions such as “frequently,” “occasionally,” and “seldom” in relation to the anchor referent “sometimes.” These studies provide empirically based standards for composing computer-assembled descriptors of test score intervals.

Some of the currently available microcomputer programs for test analysis have descriptive reports that are highly redundant with wording such as “He is above average on Scale 1... average on Scale 2... average on Scale 3....” Using more of the power of the computer, it is possible to combine sentences into paragraphs, using different modifiers and sentence forms to create more readable reports. Extensive use of this method was recently implemented by Barclay (1983) in a comprehensive computer report for students in elementary classrooms as a means of summarizing multiple indicators of social competence based on self, peer, and teacher ratings.

Some of the true power of the computer comes into play when test scores can be analyzed statistically, such as is done to describe the “scatter” of profile scores on tests such as the Weschler Intelligence Scale for Children—Revised (Kaufman, 1979, pp. 195–209). Figure 3.1 shows a sample of a computerized report for the Luria-Nebraska Neuropsychological Battery (Golden, Hammke, & Purisch, 1980) that implements for the type of analysis of profile scores suggested by Reynolds (1982). Figure 3.1 is intended as a statistically rigorous study of strengths and weaknesses within an individual score profile. The mean of all 14 profile scores in Fig. 3.1 is calculated and printed at the top of the display (mean = 51.89). Then, each profile score is subtracted from the mean and the difference is plotted (e.g., scale M1 was 52.30, and has a difference of +0.41 from the mean of all profile scales). Thus, positive differences are potential “weaknesses” in the profile, (because all scales are keyed in the clinical
### Mean T-Score for Clinical and Summary Scales

<table>
<thead>
<tr>
<th>Scales</th>
<th>Mean</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>41</td>
<td>6.24</td>
</tr>
<tr>
<td>M2</td>
<td>19.32***</td>
<td>3.46</td>
</tr>
<tr>
<td>M3</td>
<td>-1.61</td>
<td>3.61</td>
</tr>
<tr>
<td>M4</td>
<td>-6.34</td>
<td>4.90</td>
</tr>
<tr>
<td>M5</td>
<td>-5.23</td>
<td>8.66</td>
</tr>
<tr>
<td>RH1</td>
<td>-4.75</td>
<td>4.58</td>
</tr>
<tr>
<td>T1</td>
<td>11.44</td>
<td>5.57</td>
</tr>
<tr>
<td>T2</td>
<td>2.25</td>
<td>4.69</td>
</tr>
<tr>
<td>V1</td>
<td>4.61</td>
<td>5.83</td>
</tr>
<tr>
<td>V2</td>
<td>21.95**</td>
<td>6.63</td>
</tr>
<tr>
<td>R1</td>
<td>-11.89*</td>
<td>3.87</td>
</tr>
<tr>
<td>R2</td>
<td>-9.26</td>
<td>5.10</td>
</tr>
<tr>
<td>R3</td>
<td>-6.18</td>
<td>8.49</td>
</tr>
<tr>
<td>R4</td>
<td>-4.39</td>
<td>7.42</td>
</tr>
</tbody>
</table>

**Direction**

- Positive differences are potential strengths.
- Negative differences are potential weaknesses.

**What remains is to test the significance of these differences using information about the reliability and standard error of measurement of each scale. Also, a correction for multiple comparisons is included. As shown in Fig. 3.1, scales M2 and V2 are significant weaknesses and scale R1 is a significant strength for this patient.**

**Profile analysis such as that shown in Fig. 3.1 accomplishes several objectives. Specifically, it:**

1. Displays confidence intervals for observed scores based on the standard error of measurement of each scale.
2. Provides a statistical test of the differences between each scale value and the mean of all profile scale differences.
values as a rigorous test of "scatter," and (c) corrects for the multiple comparisons using a Bonferroni-type correction to the \( t \)-statistic used in testing hypotheses of significant scatter. The methodology for implementing these procedures is given by Davis (1959), Dunn (1961), and Bailey (1977), and most recently Stoline (1983) has provided even more complex statistical procedures with increasingly fine accuracy. Clearly, such statistical operations would be virtually impossible for the clinician or educator who approaches profile-score analysis \textit{bare handed} without the benefit of either extensive tables of pairwise score differences or some form of calculator or computer aid. The practical benefit of such methods is that they replace the subjective reading of profiles which may be characterized by overinterpretation of small differences between profile scales.

\textit{Clinician Modeled.} Another type of computerized interpretation of individual test results can take two general forms: (a) where the program simulates the interpretive decisions of a renowned clinician, and (b) where a statistical model is constructed from studies of groups of expert clinicians and programmed into the computer (e.g., Goldberg, 1970).

Work in progress by the current author is aimed at modeling the clinical interpretations of the \textit{Louisville Behavior Checklist} (Miller, 1981) as provided by its developer, Dr. Lovick Miller. Tape recordings of his actual case interpretations are being studied, and objective decision rules extracted from this rich clinical source. Several cycles of development will be used to produce trial interpretive programs, apply them to actual results on the Checklist, present them for reinterpretation by Dr. Miller (under blind conditions), and validate the fit between the objectively programmed rules and those actually used by this experienced clinician and test developer.

Wiggins (1973), in summarizing the work of Goldberg (1968, 1970) and his colleagues, noted that a statistical model of clinician judgments often can be more accurate than individual clinicians working in isolation. In a classic series of studies (Goldberg, 1970) examined the ways in which clinicians diagnosed psychosis versus neurosis from the standard profile scales of the MMPI. Although clinicians reported that they were combining information in complex interactive ways, the data showed that simple linear regression models were effective in describing how they made diagnoses from the test-score information. When the methodology reflected in these classic experiments are applied to other psychological tests, it becomes an empirical question whether or not a model with acceptable accuracy can be derived.

\textit{Clinical Actuarial.} Following the rationale proposed by Meehl (1954), numerous computerized actuarial systems have been developed for educational (McDermott, 1980, 1982) and psychological tests (e.g., Lachar, 1974). The term \textit{actuarial} is perhaps an unfortunate choice of wording, because it brings to
mind the statistical tables developed by insurance actuaries in the assessment of the likelihood of death or accident, but the term has an historical tradition particularly in MMPI prediction research. McDermott (1982, p. 248–249) recently expanded the definition of actuarial assessment to include a broad array of multivariate statistical procedures useful in making decisions about people on the basis of test and nontest information.

In educational measurement, the work of McDermott (1980) in the identification of students in a variety of special education categories using computer algorithms is an example of sophisticated methodology. McDermott implemented a computer program that examines intelligence, achievement, and adaptive-behavior scores of individual students in order to make quantitative judgments about diagnoses such as learning disability status. For example, test scores from the WISC-R, WRAT-R, and Adaptive Behavior Scales might be input for a student who had particularly low mathematics achievement. The computer program would have stored information on the reliability, standard error of measurement, and intercorrelation of these tests and would calculate the statistical significance of the discrepancy between the WISC-R and WRAT-R scores. The program would then print out the numerical estimates and description of the result. Obviously, such calculations are possible by hand, but with the numerous comparisons possible, the busy evaluator equipped with a computer could make more such comparisons in less time than an evaluator using a hand-calculator.

The exceedingly complex computer analysis provided for the Barclay Classroom Assessment System (Barclay, 1983) is another example. Barclay distilled 25 years of multivariate statistical studies of self, peer, and teacher ratings of elementary students into a computerized interpretive program that provides a narrative, diagnostic, and prescriptive report (up to 100 pages in length for a given classroom) useful for teachers, school psychologists, and other personnel or resource professionals. The logistics of attempting to analyze just the peer data (sociometric choices by each member of a classroom), by hand, would be challenging enough, let alone the integration of self and teacher ratings, and achievement test scores. Clearly, the computer has an inherent value in such applications that cannot be dismissed as extravagant technologizing.

In psychological testing, the most widely used and discussed actuarial programs are probably those of the MMPI. Although some investigators (e.g., Matarazzo, 1983) may wish to see additional evidence, there have been a considerable number of research studies aimed at assessing the validity of the narrative reports generated by computers for the MMPI and other psychological tests. Lushene and Gilberstadt (1972) used independent judges to rate the accuracy of 3,926 statements in 355 computerized reports and found 79% of the statements judged correct and 93% of the reports rated favorable overall. Lachar (1974) studied computerized reports for 1,410 adult patients and found that clinicians rated 107 frequently occurring paragraphs (which appeared 7,555 times in the reports) as accurate 90.3% of the time. In a study of the use of adolescent norms
in computerized MMPI reports, Lachar, Klinge, and Grisell (1976) found that clinicians rated 20% of narratives based on standard adult norms as inaccurate for 100 patients aged 12–17, whereas narratives based on adolescent norms were judged inaccurate in only 10% of the cases.

Another approach to increasing the validity and accuracy of computerized reports based on statistical/actuarial methods, is that of tailoring reports of results on a comprehensive personality inventory to findings from studies with particular patients or applications (Krug, 1982). For example, a computerized report for the 16 PF (Dee-Burnett, Johns, & Krug, 1982) was developed specifically from validity studies in law-enforcement settings. Another report (Krug, 1983) is specifically designed for use in marriage counseling.

Another approach to assuring the accuracy of descriptions generated by computerized reports is the method of replicated correlates used extensively by Lachar and Alexander (1978) on the MMPI, and by Lachar and Gdowski (1979) on the Personality Inventory for Children (Wirt, Lachar, Klindedinst, & Seat, 1977). In this method, clinicians who personally interview each patient are asked to provide detailed ratings using a behavioral and symptom checklist. All patients are then given the inventory which is to be computer scored, and the profile scales from the inventory are plotted for each subject on a standard T-score profile. Each profile scale is divided into elevations or segments such as 80T+, 70-79T, 60-69T, 41-59T, and 40T-. The frequency of each checklist description of the patients is then calculated for each elevation on each scale. High frequency checklist items are called correlates of a given scale. A new sample of subjects is used to replicate the findings, and only replicated checklist descriptors are used in the final computerized report to describe the potential behavior and symptoms of the patient.

Even though there remain examples of undocumented computer programs that provide narrative reports similar to the clinical-actuarial programs described above, the best of the reports provide extensive documentation. Certainly, as a means of combating the aura of objectivity projected by computerized reports, it is essential that detailed documentation of the empirical bases for decision rules and narratives be provided, such as is done in the manual for the Strong-Campbell Interest Inventory (Campbell, 1977) and the book-length monograph by Lachar and Gdowski (1979).

FUTURE DIRECTIONS IN THE APPLICATIONS OF COMPUTERS IN TESTING

In an interesting book on the early history of computer programming languages, Sammet (1969), one of the codevelopers of the COBOL language, used the analogy of the biblical story of Babel to describe the proliferation of languages in computer technology. Similarly, the future of computerized testing and test
interpretation will undoubtedly be characterized by incredible diversity and lack of standardization of procedures. The current proliferation of brands of microcomputers is but one example.

Amidst the diversity that seems to be inevitable in the fast-changing world of computer technology, there are several important forces that may help to unify some of efforts of independent researchers and developers. The developments of CP/M (Kildall, 1982) and the highly acclaimed UNIX operating system (Christian, 1983) promise to increase the machine-independence of computer software and systems, thus making wider distribution of computer programs possible. The developments of memory devices such as the inexpensive floppy disk which will become increasingly miniaturized and higher in capacity, will allow the distribution of testing programs (and translators required to adapt them to particular hardware) that feature exceedingly large item banks or narrative-interpretive material. However, it is likely that a continuing problem into the future will be the compatibility of various types of hardware and software.

This segment of the chapter proceeds with a review of the four major areas introduced earlier: (a) test development, (b) test administration, (c) scoring of tests, and (d) test interpretation. Emphasis is on trends that are in the experimental stage now that seem feasible for wide-spread application in the near future.

Test Development

Item Writing. The availability of microcomputers and the software developed especially for them, has brought some new advantages to the developer of tests and test items. Many of the microcomputer word-processing packages are more sophisticated than those previously available on large mainframe computers. For example, page-oriented editing systems are more efficient than line-by-line editors and widely distributed software is available for correcting spelling, grammar, and the general readability of test items and other material needed for computerized test reports. The future promises more and more aides for the item writer and test developer.

Roid (1984) called for the development of software that would be useful for the automated development of reading comprehension tests keyed to textbooks. Some merging of the methods used in library science to access the keywords in text, and the methods of transforming text into test questions (Bormuth, 1970; Roid & Haladyna, 1982, chapter 6) would seem to be helpful to the publishers or widely used school textbooks. As mentioned earlier, such automated methods may improve the match between teacher and testing as reviewed by Airasian and Madaus (1983). Computerized versions of word lists, such as that compiled by Carroll, Davies, and Richman (1971) for American textbooks would play an important role in such methods.

For years, the field of computer-assisted instruction has been experimenting with methods of helping authors create computerized lessons. Systems called
"authoring aids" have been developed for the PLATO system at the University of Illinois (Alpert & Bitzer, 1970) and at several military CAI installations (Schulz, 1979). These aids help the developer to create common forms of questions such as multiple choice, constructed response, and matching items. For reasons that are not entirely clear, there has been little cross-fertilization between the fields of CAI and standardized testing, but the future promises to see more sharing of ideas between these fields as more and more testing is implemented on small computers used in schools.

Another important area of cross-fertilization between CAI and testing will hopefully be in the definition of programming languages useful for constructing items and implementing computerized testing. Currently, the most widely used programming languages for computerized testing appear to be BASIC and PASCAL, which are general purpose languages. Instead, great advantages would accompany the usage of application languages such as the TUTOR language for the PLATO CAI system. Cory, Rimland, and Bryson (1977) used an IBM 1500 CAI system, which includes the COURSEWRITER language, to develop a battery of information-processing tests. These tests were used by Cory (1977) for predicting job performance.

In the area of perceptual and cognitive experimental psychology, which should contribute more and more to the assessment of memory and cognitive functioning as evaluated by computerized testing in the future, several high-level programming languages have been developed. The LAB-TALK language (Maxwell & Schvaneveldt, 1983) is useful in presenting stimuli, collecting responses, and recording data. Two other examples are the EXPERIMENT WRITER language of Posner and Osgood (1980) and the ARTIST system of Kornbrot (1981). These higher-level languages should serve as a model for the development of applications languages in computerized adaptive testing, particularly in cognitive/perceptual evaluation where complex graphic and multiple-trial stimulus events are used. Perhaps the future implementation of such languages on widely distributed microcomputers will play a key role in making computerized testing more feasible.

**Item Banking.** The future development of sophisticated and miniaturized memory devices, as mentioned earlier should encourage the increasing use of item banking, particularly in the area of criterion-referenced achievement testing. There are recently developed examples of such item banks published for use on microcomputers. One large-scale project (Forster & Doherty, 1978) in the Portland (Oregon) Public Schools has included the development of Rasch-calibrated items numbering 1000 or more in each of the three basic achievement domains, reading, mathematics, and language arts. Work is currently in progress to implement these large item banks on microcomputers tailored to the needs of individual school districts. Haladyna and Roid (1983) showed that adapting the difficulty of mastery tests to the functioning achievement level of students pro-
vided greater measurement precision than random sampling of items from domains, and this is the approach being implemented in the Forster and Doherty (1978) system. Furthermore, computer tailored tests such as these promise to provide a “curriculum-referenced” interpretation to achievement tests (Rentz, 1982; Woodcock, 1982). Since achievement test items can be calibrated for difficulty along the same scale as the estimates of achievement for each student (Wright & Stone, 1979), it is possible to draw a curriculum continuum that maps the specific skills achieved by students at various score levels on a test. For example, in basic arithmetic, items of long division would be more difficult than items of addition. If a test includes, subtraction, multiplication, and division, it is possible to draw a continuum showing where specific skills (e.g., 2-place addition with carrying) lie, and then to reference the total test score of a student to this curriculum-continuum. Woodcock (1982) uses this system for his finely detailed profiles such as those available with computer scoring on the KeyMath Diagnostic Arithmetic Test (Connolly, Nachman, & Pritchett, 1976).

Developing the Psychometric Properties of Tests. Among the many new multivariate analysis methods, only possible on computers, that will help to shape the tests of the future, are two that deserve special attention in the next decade: (a) increasing use of linear structural relationship (LISREL) analyses in the development of evidence for the construct validity of tests, and (b) the use of factor analyses specifically designed for dichotomous items, including the important exploration of the unidimensionality assumption for achievement and ability tests.

A number of multivariate methods known as causal modeling (Bentler, 1980), structural equation modeling (Joreskog & Sorbom, 1981), or path analysis (Wolfe, 1980) show promise for the study of construct validity of tests. Because the study of construct validity (Cronbach & Meehl, 1955; Messick, 1980) involves the comparison between empirical findings and a theoretical nomological network that posits the expected relationships between variables measured by tests and variables in the real world, models of multivariate relationships apply. As LISREL-type programs become more available to a wider circle of test developers, it will be possible to use them increasingly in test development research (e.g., Marsh & O'Neil, 1984). These programs will be useful in demonstrating that the latent variables underlying a test battery are related to the latent variables underlying a series of observations collected by means other than the test being examined. For example, a test battery measuring teacher ratings of students could be examined in relation to known correlates of teacher ratings, such as achievement, parent behavioral ratings, and readiness tests among children entering the first grade.

Another area of new interest requires considerable computer power. With growing interest in item-response theory for achievement and ability tests, has come the increasing concern that some tests do not fit the unidimensionality
assumptions of IRT models. Some investigators have simultaneously questioned the value of traditional indexes of test homogeneity, such as the alpha internal-consistency coefficient, as indicative of unidimensionality (Green, Lissitz, & Muliak, 1977; McDonald, 1981; Smith, 1980). In response to these concerns, several new statistical models allowing for multidimensional tests to be analyzed with IRT-like models have emerged (Reckase, 1979; Stegelmann, 1983; Whiteley, 1981). Considering that Damarin (1970) called for multiple latent-variable models for psychological test item analysis, these developments have great importance for psychologists as well as for educators and other social scientists.

Why should test developers be interested in the debate on unidimensionality? There is a reason perhaps even more important than the concern that an IRT model may not fit the data for an educational test in wide usage. Just as factor analysis has often motivated psychologists to add or subtract certain items or scales from personality and ability tests, because of a desire to measure documented factors, perhaps more frequently, educational test developers will examine achievement tests in order to add items that measure secondary factors above and beyond the single-factors that may have been assumed in the past.

The new developments that make the investigation of multidimensionality possible are the new methods of factor analysis specifically designed for dichotomous items (Green, 1983). Gorsuch and Yagel (1982) recommended two types of factor analysis: (a) the factor of small groups of items called “parcels” as used by Cattell (1956, 1974), and (b) hierarchical factor analysis (Gorsuch & Dreger, 1979) which extracts higher-order factors from the first-level of factors extracted (including potential spurious factors due to binary items). Examples of the application of hierarchical factor analysis are provided by Gorsuch (1983a, chapter 11) and in studies such as Wallbrown, Blaha, and Wherry (1973).

Another approach to factor analyzing dichotomous items is described by Muthen (1978, 1981) who is developing a comprehensive computer program that will be an important addition to the test-developers software collection in the future. Roid (1984) has emphasized the value of these new factor analysis programs in the development of criterion-referenced test items to measure potential multiple dimensions in achievement tests. However, because so many psychological tests and checklists have dichotomous items, new vistas in the explorations of test dimensionality remain for psychological and clinical tests as well as achievement testing. And, clearly, the factor analyses of data matrices having 100 items or more is not possible without increasingly sophisticated computer technology and software, coupled with advances in computer-memory technology.

Test Administration

Perhaps more than any of the other areas discussed, test administration by computer will be the area of most tangible and observable progress in the future. The massive effort to computerize many of the tests that have been developed and published during this century has begun. For those tests that are not amenable to
computerized test administration (such as individually administered tests for young children), efforts are underway to implement computerized scoring and interpretive programs.

Computers promise to expand the range of human responses that can be recorded automatically during test administration. Whereas the technology of automated test scoring was limited by the medium of optical-scanning answer sheets marked by pencil, the new wave of computerized tests will include a wide array of input media. These will include touch-sensitive screens, light-pens and toggle-levers, physiological receptors sensitive to changes in skin response and heart-rate, etc., and even voice-pattern recognizers. After years of development, touch-screens will now be more easily obtained for adaption to the responses of young children and anyone who cannot easily communicate via a typewriter keyboard. Inexpensive touch devices are also available for adaption to existing equipment (Cumming, 1983). Richards, Fine, Wilson, and Rogers (1983) recently reported success with a voice-operated microcomputer system that allows the patient to respond True or False vocally to MMPI items presented on a computer screen, by using voice-pattern analysis methods in the computer.

Similarly, the future will see increasing use of sophisticated test-stimulus displays, following price reductions on complex multimedia equipment. The work of Elwood and Griffin (1972) to administer the full-battery WAIS via tape decks and complex equipment, for example, may give way to computer/video-disc systems such as used by Morf, Alexander, and Fuerth (1981) to administer a picture-preference test. Again, as mentioned earlier in this review, the two worlds of computer-assisted instruction and standardized testing will hopefully meet in the future, to the benefit of both. It would seem that a complex battery such as might be used to diagnose a learning disability would benefit from the sophisticated branching, response-time recording, and graphic/multimedia nature of CAI systems such as PLATO, the IBM 1500 system, or other CAI facilities currently in wide use in military and industrial training.

As discussed in the first section of this chapter on test administration, there has been a surprising lag in the implementation of computerized adaptive administration of tests, which have been possible since the late 1960s (e.g., Hansen, 1969). Perhaps the recent distribution of new item-analysis programs (Wingersky, Barton, & Lord, 1982; Masters, Wright, & Ludlow, 1980) will contribute future development of calibrated item collections needed for adaptive on-line testing. The work of Jensen (1976, 1977) may be very important in the implementation of the 3-parameter model. Jensen (1976) provided estimates of the 3-parameter model that are easily programmed, and easily understood by most psychologists without detailed training in IRT models. Jensen (1977) also provided guidelines for building a good, tailored item bank. As more and more test developers become skilled in the use of such programs, perhaps more adaptively-administered tests will appear such as those being released for use in government agencies and the military (e.g., Schmidt, Urry, & Gugel, 1978). Perhaps also, computerized adaptive testing will emerge more strongly when
general-purpose computer software is available to help test authors prepare an adaptive test for a particular computer. A general software package could connect files of items with files of item-calibrations (e.g., item difficulty estimates) to present some standard methods of testing using adaptive branching. Development of such software may break the log-jam created by the effort required for a test developer to create not only the items (and the field-test statistics) but the test-administration program as well.

There remain numerous technical issues that must be resolved in adaptive computerized testing. First, many educators and psychologists will question the nature of normative comparisons that might be made with adaptive testing. Because each subject may be given a different set of items, there appears to be a certain statistical wizardry in calculating a total score for normative purposes. Only through numerous practical applications of adaptive testing will test users begin to see examples of scoring and methods of interpreting total scores. When IRT models underlie adaptive testing, total scores are estimates of the trait, ability, or achievement continuum assessed by an item pool. As discussed earlier in this chapter, in a process similar to “curriculum referencing” (Haladyna & Roid, 1983; Rentz, 1982), and IRT-based test score can be made interpretable by defining various points along the latent continuum. If such definition can be achieved, the resulting scale can provide both normative and criterion-referenced interpretations. The interpretation is normative if all items have been calibrated on representative samples in which all users have confidence and studies have been conducted to determine the relative number of students expected to score at successive points on the continuum (from which some new type of percentile can be derived). Total scores can be given criterion-referenced interpretations by the anchoring of specific items, skills, or meaningful trait levels along the continuum (e.g., as in the Woodcock-Johnson or KeyMath tests, Woodcock, 1982). Second, several technical issues surround the problem of optimal ways to determine the starting place for adaptive testing. The most sophisticated solution proposed to date for this problem involves the storage of longitudinal records of examinee or student performance on previous tests which can be used to begin subsequent testing. Some adaptive testing systems (noncomputerized), rely on the judgments of teachers in placing the student at an approximate level for beginning testing, followed by readministration of scales that prove to be improperly tailored (e.g., the use of special education teachers to estimate functional levels of retarded students assessed by the statewide assessment survey of Brodsky & Roid, 1977). Clearly, there is much that needs to be done to develop viable solutions to these technical problems with adaptive testing.

Test Scoring

Computer technology opens several new avenues for test scoring. In the past, the research finding that item weighting was usually not necessary (e.g., Stanley & Wang, 1970) gave us little reason to search beyond the basic total score (sum of
a series of item scores) method used for most tests in both education and psychology. Also, for tests that are hand-scored, it is very difficult for users to calculate anything but integer scores for items (zero/one for dichotomous items, 1-5 for five-point Likert items, etc.). With new research emerging on item option weighting (Downey, 1979; Roid, 1983a; Stanley & Wang, 1970), and with continuing interest in factor scales for tests, there is increasing likelihood that computers will play an important role in providing more complex scoring systems for educational and psychological tests.

In educational measurement, the extensive research by Wilcox (e.g., Wilcox, 1981) on answer-until-correct scoring for achievement or ability tests is very promising and could be implemented in sophisticated ways using computer technology. New work on diagnostic scoring for achievement tests (Birenbaum & Tatsuoka, 1982, 1983) promises to allow for the diagnosis of erroneous problem-solving rules used by students. The new multicomponent models (e.g., Sternberg, 1977, 1979, 1981; Whitely, 1977, 1981) for achievement and ability tests of the problem-solving type would require complex scoring procedures because each item performance may entail several cognitive steps each of which may be scored separately.

In both psychological and educational measurement, a promising new method of computerized scoring for norm-referenced tests may prove useful in the future. Because the computer can ask the examinee to give exact demographic facts such as age in months (or this can be retrieved by processing a birth date in relation to the current date of testing), it may be possible to calculate what are called “continuous norms” (Gorsuch, 1983b; Roid, 1983b; Wendler, 1983; Zachary & Gorsuch, 1985). In continuous norming, one or more continuous variables such as age are examined in extensive computer analyses of field-test or normative-data results to discover whether or not a statistical formula can be derived to “fit” the pattern of test parameters (e.g., means, standard deviations) observed across the range of the variable. For example, it is often found with cognitive or skill tests that the mean score on a test increases steadily from ages 5 to 10. The traditional way of norming such tests is to provide separate norm tables for each year or 6-month increments of age. However, as Zachary and Gorsuch (1985) showed on an intelligence battery, the traditional norm table may inaccurately estimate the examinee’s score if the age of the examinee is on the borderline between two adjacent norm tables. In continuous norming, values of test means and standard deviations are smoothed across a full range of age groups, so that estimates can be made at each and every continuous age level rather than in the graded steps implied by the use of printed norm tables.

An example of continuous norming taken from Roid (1983b), for a test in a learning disabilities battery, is presented in Fig. 3.2. Figure 3.2 shows the fitting of a polynomial regression equation to the progression of mean test scores across the age of students (in months). The vertical axis of Fig. 3.2 is the mean test score of an auditory memory test for school children. The horizontal axis of Fig. 3.2 is age in months (from 66 months to 162 months, i.e., 5.5 to 13.5 years).
The plotted points in Fig. 3.2 are either observed (O) or predicted (P) values of mean test score for groups of students at each age level. By drawing a best-fitting regression line through the predicted (P) points one can see that the mean test score steadily increases up to about 140 months, at which point it decreases slightly (due to a possible flaw in the sampling of older students, or some factor related to "topping out" on the test among older students). Because the fit of the regression line is adequate ($R = .92$ between mean score and age), test score means can be estimated for intermediate values of age (e.g., 120.5 months). Also, standard-score norms employing mean estimates derived from the prediction equation shown in Fig. 3.2 can be smoothed across the age span shown in
the Figure. Operationally, the prediction equation would be programmed into a computer, the student's age requested for input, and a predicted mean test score for his or her age group calculated from the equation. In a similar fashion, estimates of standard deviation can be obtained so that normative standard scores are derived for each individual subject.

To determine the applicability of continuous norming, it is necessary to investigate each score on a test across a wide ranging sample and to discover whether or not a statistical formula (e.g., polynomial regression equation or cubic splines) is significantly accurate as an estimation device. Then, the formula or equation would be programmed for a computer-scoring routine. Error of estimation would also be evaluated to display the accuracy of continuous norming (Gorsuch, 1983b) along with the scoring output. Additional research is needed to extend the concept and methodology of continuous norming to linear (e.g., standardized T-scores) and nonlinear (e.g., normalized NCE scores) scoring of tests.

Test Interpretation

Some sophisticated methods of test interpretation become possible when the great memory and logic power of the computer can be carefully used following empirical studies of the links between test score patterns and verifiable behaviors or characteristics of examinees. For example, (R. L. Gorsuch, personal communication, 1983) following a study that demonstrates the discriminating power of a test to distinguish between examinees in various criterion groups (e.g., different psychiatric classifications, various types of dyslexia, etc.), the discriminant function equations derivable from such a study could be programmed into a test interpretive program so as to calculate the probability that an examinee belongs to a given criterion group. This involves only a linear equation with weights for each test score, but is often too complex of a calculation to do by hand, particularly if there are numerous criterion groups to assess.

A promising new technique has been developed by Huba (1985) for the matching of psychological test profiles to prototype profiles of criterion groups. Using multivariate techniques, an individual's vector of profile scores can be matched to a vector of criterion-group means using a chi-square test of goodness of fit. Huba's method is one of the first to take into account, explicitly, the correlations among profile scales.

Another realm in which computerized interpretation of tests may be important is in the establishment of links between two or more tests. For example, a brief test may be used to predict performance on a lengthier test, such as when full-scale WAIS IQ is predicted from a brief intelligence test (Zachary, Crumpton, & Spiegel, 1985). Typically, empirical studies of the brief tests have included a regression analysis in which scores on the longer test are predicted from the brief test. If the regression changes for different subgroups of examinees, as it often
does for different age or ethnic groups, then the prediction of scores on the longer test involves a lengthy series of equations which are difficult to implement by hand. The computer, however, easily calculates any number of such predictions, and can also print confidence intervals and appropriate cautions to consider in evaluating the accuracy of such predictions.

In general, the future should see computer technology assisting professionals in integrating results of diverse test data. In McDermott’s (1982) actuarial system for helping school psychologists diagnose learning-related problems, diverse tests of achievement, cognitive potential, and adaptive behavior are interrelated using statistical information such as the reliabilities of each test and their intercorrelations. Barclay’s (1983) system for analyzing self, peer, and teacher ratings is another example of integrating diverse test results for practical diagnosis and prescription in the schools. As modes of inputting data from diverse tests progresses, and as more and more schools and clinics have their own computers with large memory capacities, more integration of test results and intertest interpretations will be possible.

Barclay’s classroom assessment system (1983) also suggests another important advantage of computer technology in the interpretation of test results—the use of measures of the environment or situational factors in the assessment of individual differences. As Mischel (1979) Moos and Trickett (1974), and Walberg (1980) among others have been emphasizing for a number of years, the environment and changing situations of the individual must be taken into account in educational and psychological assessment. The computer’s ability to analyze patterns of data collected across situations, across time periods, and from diverse sources such as self and observer perspectives, should make the evaluation of person-environment fit (or the lack thereof) more feasible.

SUMMARY

The field of measurement and testing, with its affinity for objective scores and statistical sophistication is a natural breeding ground for the emergence of computer technology in education and psychology. Dedication and attention to detail will be required of those who attempt to implement computerized testing and interpretation, if the facade of objectivity created by such systems is to be backed by empirically-validated procedures. Extensive and clear documentation has always been a somewhat difficult challenge in computer science, and the temptation will continue to be great to create novel programs that are undocumented.

Constant reminders of the probabilistic nature of computerized interpretations, and the errors of prediction inherent in them, will need to be woven into computer-printed reports that otherwise appear to have an aura of authoritative objectivity, especially for users who are less clinically experienced or knowledgeable of the limitations of all imperfect measuring instruments. As with any sophisticated tool, the professional must learn the limits of misinterpretation.
possible with computerized tests and interpretations. Clearly, technology cannot replace the complex evaluations that the experienced educator, researcher, or clinician can bring to the assessment of a child or adult when the individual is observed functioning in his or her environment over a period of time. An area of difficult assessment comes to mind that emphasizes the limitations of technology. It is in the assessment of complex learning disabilities, particularly in the case of the child who may appear to be retarded but whose inconsistent performance includes obtaining a very low or borderline score on an intelligence test and yet the child obtains isolated high scores on various diagnostic tests including adaptive behavior measures. The parents may also report inconsistent behavior, including "flashes of brilliance" which come and go. This child does not match the classical picture of the learning disability student who has average or above-average intelligence with specific achievement delays or underachievement (Gaddes, 1980). A complex pattern of perceptual, communication, and perhaps neuropsychological disorders may interfere with the child's performance on many tests, computerized or conventional. The response-timing features of computerized testing may be helpful in the diagnosis of conditions such as word-finding disorder (German, 1979, 1983), in which the child knows the answer but cannot retrieve it fast enough for a timed test, but the larger problem is that word-retrieval is only one of many potential disorders that interact with standardized test performance. Such disorders can interfere with assessment so that a false picture of the child's true potential is given. Clearly, skilled clinical observation, input from parents or other observers, and recognition of the limitations of testing must be considered in the interpretation of test results in such cases.

Despite its inevitable limitations, computerized testing and interpretation shows interesting promise for the future. The cost-effectiveness of relieving the professional from tedious hours of hand-scoring tests or in calculating various statistical indexes derived from test scores is clearly apparent. The value of computers in the multivariate studies that lead to test refinement and development has been clear for decades. If the fields of measurement and computer-assisted instruction can become even more closely related, there is great promise for the development of tests that use multiple input and output media for presenting test-items (or test-like experimental events such as in the assessment of memory and perception) and for displaying results. As the equipment used in applications such as speech and hearing therapy and biofeedback treatment become linked more and more to computers, new forms of auditory and physiological data will become amenable to computerized interpretation along with more conventional test items and scales.

ACKNOWLEDGMENT

This chapter is dedicated to the memory of Bruce Choppin who encouraged the author to pursue some of the ideas presented here during informal meetings and symposia at the UCLA Center for the Study of Evaluation. Acknowledgments are extended to Manson...
Western Corporation for their support of research and development activities at Western Psychological Services. Many of the major ideas presented in this chapter originated in conversations with Richard L. Gorsuch, Fuller Graduate School of Psychology, Pasadena, California, whose advice is based on a breadth of experience spanning the history of computerization in psychology. Dr. Gorsuch’s contribution is gratefully acknowledged. However, any omissions or errors remain the responsibility of the author. Opinions expressed or future predictions made are those of the author, not the official policy of any corporation or institution.

REFERENCES


