Psychological Assessment
1996, Vol. 8, No. 4, 354-359

1040-3590/96/$3.00

Test Construction in the 1990s: Recent Approaches Every Psychologist Should Know

Mark D. Reckase
ACT

The article summarizes the current state of the art in test construction and contrasts it with previous conceptual models, some of which are wrong or misleading. In particular, new methodologies for item selection and review are presented as well as current thinking on the specification of technical characteristics of tests.

The construction and interpretation of psychological tests has been a standard part of the curriculum for psychology majors and students in graduate programs in psychology since the days of James McKeen Catell. However, information on the construction of psychological tests that appears in psychological and educational training curricula and techniques that are put in practice as part of psychological research are often not state of the art, mainly because it takes some time for information to diffuse from the research literature and from the internal documents of testing organizations to the sources used for instruction. This diffusion process is probably no worse in psychology than it is in other areas—consider how long it has taken advances made by the space program to influence everyday life. Automobile technology is quite different now from the way it was 20 years ago as a result of advances in computer technology and materials from the space program, but the changes took quite a long time to be put in place.

The purpose of this article is to identify some important changes that have taken place in the process of test development over the past 20 years. The changes are highlighted by contrasting the descriptions of the process from sources spanning the period since the 1930s, starting with a book on intelligence testing by Pintner (1931). Overall, the intent is to summarize the current state of the art in test construction methodology and to point interested readers toward references to techniques that might be new to them.

To provide some structure to the article, the test development process has been divided into four parts. They include descriptions of (a) content, (b) statistical specifications, (c) item selection, and (d) test review processes. Although these topics are not truly mutually exclusive, they do provide a relatively concise structure for discussing test construction practice.

Defining Constructs and Contents to be Assessed

Early specifications for the content to be assessed by tests was minimal. Pintner (1931) discussed the item development process as follows:

> First the kinds of items will be chosen. Some tests have one kind of item, others have several kinds, such as Opposites, Analogies, Number Completion and so on. More kinds of items can be chosen than the experimenter may wish finally to retain. A new type of item, that has not been used in intelligence tests, may be included. Having determined the kinds of items, the items themselves are assembled. About twice as many as may finally be used are collected, and they are arranged in a guessed order of difficulty. (p. 128)

Clearly, no construct definition or content specification was used at that time, although it is likely that Pintner had a notion of the item types and content that would be used based on experience. The items were selected mainly on the basis of results from empirical analyses that showed that the item scores correlated with teacher ratings or age.

Later works, such as that of Greene, Jorgensen, and Gerberich (1944), included the concept of an item domain. In the context of achievement tests, these authors indicated that "Care should be taken to sample course content widely and impartially in the selection of materials for a test" (p. 162). By 1966, three general types of test specifications were noted in the first edition of the *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association, 1966, p. 12). The first is a generalization of the test domain concept as "how an individual performs at present in a universe of situations that the test situation is claimed to represent." The second relates to predicting criterion measures: "The test user wishes to forecast an individual's future standing or to estimate an individual's present standing on some variable of particular significance that is different from the test." The third type of test specification relates to defining a psychological construct: "The test user wishes to infer the degree to which the individual possesses some hypothetical trait or quality (construct) presumed to be reflected in the test performance."

By 1974, all of the ways of defining what was to be assessed by a test were combined into one basic idea, that of appropriate inferences—"What can be inferred about what is being measured by the test?" (American Psychological Association, 1974, p. 25). Construct validity was considered the central concept under which all others were subsumed.

Current best practice in defining the content specifications for a test is an elaboration of the concept of a construct. This elaboration has been accomplished from two different perspec-

An earlier version of this article was presented at the 103rd Annual Convention of the American Psychological Association, New York, August 1995.

Correspondence concerning this article should be addressed to Mark D. Reckase, Assessment Innovations, Development Division, ACT, P.O. Box 168, Iowa City, Iowa 52243-0168.

tives. Embretson (1985) considered the constructs that are measured by a test to be all of the variables that determine the item responses. She stated (Embretson, 1985, p. 4): "Tests cannot be designed by their substantive properties unless the variables that determine item responses have been explicated. In essence, understanding item responses requires a theory of the variables that are involved in item performance." Thus, to develop the specifications for a test, Embretson argued that a test designer should have a good understanding of the variables that are the target of the assessment and the types of test items that will yield responses that represent different levels of those variables. This approach to test design is consistent with the theoretical measurement position that the item response vector provides a representation of underlying causes of performance (Krantz, Luce, Suppes, & Tversky, 1971).

The second approach has been presented very clearly by Millman and Greene (1989). Here a domain of content or skills is hypothesized, and the goal in producing a test is to provide a generalizable sample of tasks from the domain. A classic example of this approach was presented by Bock (1996). He described an assessment of the ability to spell words taken from a book of commonly misspelled words. One approach to producing a test related to this domain is to sample words randomly from this book and then to ask the examinees to spell them correctly. The proportion correct score is an estimate of the total proportion of words from the book that can be spelled correctly. In this case, variation in performance over the entire domain is the construct that is being assessed.

Although these two approaches to test development, estimating levels of underlying variables and estimating proportions of domains, now dominate current practice, other approaches do exist. For example, an approach called the *dust bowl empiricist approach* selects items that are able to distinguish between groups of examinees, such a psychotics and nonpsychotics. A large pool of items is constructed and administered to the two groups of examinees. Those items that perform differently for the two groups are put on the test. There may not be any theoretical rationale for differences in performance when the test is constructed, but research is performed after test construction to explain why the test works.

Yet another approach to test development is to construct an index from a set of items with no pretense that the resulting scores represent a domain or an underlying trait. Bollen and Lennox (1991) gave an example of this approach by constructing an index called *exposure to discrimination.* The index is a function of race, sex, age, and disabilities, but there is no suggestion that there is a trait related to discrimination or a domain to be sampled. All that these authors have suggested is that the variable that is constructed is a useful indicator of the likelihood of a person's having been been subjected to discrimination. It does not represent a trait dimension because each of the input variables shows differing levels of discrimination when used as independent variables. This approach is a clear example of a "construct" (i.e., a variable that is created or constructed by the psychologist).

Implementation of the Concepts

All of the approaches to test development share the need to begin with a rationale for the assessment. Current standards (American Psychological Association, 1985) suggest that this rationale be called a construct and that empirical research needs to be done to demonstrate that the test provides information about the construct. The construct might be spelling knowledge, exposure to discrimination, ability to perceive spatial relationships, or something else.

Following the specification of the construct, an approach to assessing the construct needs to be specified. The selection of an approach is guided by answers to questions such as the following four:

1. Is the goal of the test to differentiate among examinees along a real or imagined trait? If so, the items need to be selected that will be related to, or represent levels of, performance on the trait.
2. Is the goal of the test to differentiate among the examinees according to their level of mastery of the skills and knowledge in a domain? In this case the items need to be selected to represent the characteristics of the domain.
3. Is the goal of the test to distinguish between well-defined groups, but without regard to a theory? In this event, which is often called *empirical keying,* items should be selected that are responded to in maximally different ways by examinees in the groups.
4. Is the goal of the test to provide an artificial index that does not seek to generalize beyond the items used? If so, the test constructor will select items that seem reasonable and define the index.

In all these cases, tests still have to be validated to ensure that their scores can be properly interpreted. For each, however, the validation information will be slightly different. In the first case, validity information indicates that the items are related to the trait; in the second, validity is established by demonstrating that the assessment items span the domain. In the third approach, validity information needs to support the view that the items distinguish between the groups. The fourth approach requires that the index be useful and that it have practical meaning related to its name or label.

To ensure appropriate trait or domain coverage, test specifications typically include a table that indicates the number of items that need to be produced with particular characteristics. If a domain model is used, the table will specify the breadth of the domain and the distribution of item types throughout the domain. In general, this approach requires a stratified sampling plan.

Similarly, for the trait approach, a table is produced that indicates the types of items that should be related to the trait and the number of each type that is desired. Of course, there should be a rationale for components of the table. For example, there are many different kinds of item types that can be used to measure skills in spatial relations (e.g., unfolding and folding boxes, rotating figures, matching tasks, hidden figures). A rationale is needed for the selection of each type of item, and if more than one type is used, for why the number of items of each type was selected.

The development of tests still constitutes an art. However, the trend has been to make it a very thoughtful art form, with a clear specification of constructs and a rationale for design decisions.

Statistical Specifications

Pintner's (1931) description of the test development process includes three tryouts of the items. The first effort was used to collect data to determine whether items should be dropped from consideration. He stated (Pintner, 1931, p. 129) that as part of this phase, "Items that are not valid or items of equal difficulty not required [are] discarded." Validity in this case had to do with the relationship of item score to age, grade in school, or teacher judgments of intelligence. It is interesting that Pintner seems to have suggested that more than one item at a particular difficulty level is unnecessary.

By 1944, Greene, Jorgensen, and Gerberich (1944) had provided more guidance about statistical specifications for tests. They stated

> Some test authorities prefer approximately equal numbers of items at all levels from very easy to very difficult, while others prefer to use a few easy and a few difficult items but to have the majority near the 50 percent difficulty level. They are in general agreement, however, that the test as a whole should have about 50 percent difficulty for the average pupil. (p. 78)

They also recommended dividing the analysis group into halves on the basis of total score, indicating high and low scorers, and then comparing the percentage correct for each item for each half. They then noted that "a failure to register the real differences in the ability of the two groups is probably serious enough to warrant the elimination, or at least the revision, of the item" (Greene, Jorgensen, & Gerberich, 1944, p. 81).

By 1970, Cronbach (1970) had presented a much more sophisticated understanding of the relationship between statistical specifications for a test and other characteristics. For example, he stated, "By selecting items suitably, we can change the shape of the score distribution to some extent, flattening the central hump, producing two humps, etc." (Cronbach, 1970, pp. 99–100). He also indicated that for tests based on a domain sampling approach, items should not be excluded on the basis of discrimination criteria because multiple dimensions might be included in the domain. Selecting on the basis of discrimination criteria might reduce domain coverage. Finally, Cronbach indicated that a test does not have a single standard error of measurement for a particular examinee sample but that the standard error differs as a function of point on the score scale.

The classic work that provides methodology that allows the statistical characteristics of a test to be fully under the control of the test constructor is Lord and Novick's (1968). In this amazingly complete volume, the authors showed that the internal consistency reliability of a test is maximized by selecting highly discriminating items of difficulty .5 for items with no chance of guessing the correct response. However, maximizing reliability results in score distributions that are very odd shaped and possibly undesirable.

They also made it clear that there is no "true" distribution for a test or a construct. Any shaped distribution can be produced by the test constructor through careful selection of test items. Further, the use of test items as "building blocks" that can be used to develop tests with desired measurement properties was finally explicated. Items were shown to provide good measurement information over a limited range of ability or skill, and methods for selecting items to cover extended ranges of skills were described. The presentation of item response theory in the chapters by Birnbaum (1968) provided a clear conceptual framework for understanding the use of items to construct tests.

Although Lord and Novick (1968) provided the theory, Lord (1980) provided a fairly accessible guide to practice. In an impressively brief book, Lord showed the range over which items provide useful information as described in an information function. Further, the way that item information functions sum to define the precision of the test is clearly specified. Figure 1 provides an example of the relationship between item information and test information. The information functions for each of five items are presented. The "total" test information is computed by summing the five item-information functions. Because the standard error of measurement at a particular score point can be estimated by 1 over the square root of the test information at each point on the scale, the information functions also to show how to control test precision.

To totally control the error distribution for a test, Lord (1980) suggested producing a target information function that has high points where high test precision is required and lower points where precision is less important. Test construction is accomplished by selecting test items that will yield a test information function that matches the target.

Implementation of the Concepts

The current state of knowledge about the interrelationship of item characteristics and test characteristics allows the test constructor to sculpt a test score distribution into any form that is desired if enough items are available and their characteristics have been estimated through a reasonable tryout. In a sense, items are like children's plastic building blocks. With enough creativity, and the right selection of "blocks," any desired set of test characteristics can be produced. If a test for selection of scholarship candidates is desired, a score distribution that has positive skewness can be produced to spread out the high-scoring candidates. Alternatively, if a diagnostic mathematics test is desired, precision can be concentrated at the low range of ability, resulting in a negatively skewed score distribution. To take advantage of current test development technology, the following questions need to be answered.

Conceptual framework. What is the conceptual framework for the test: domain sampling, trait estimation, empirical prediction, or index construction?

1. If the conceptual framework is domain sampling, is the domain well defined and is the goal to estimate the percentage of mastery of the entire domain? If so, a sampling plan is needed to provide technical specifications. If either one or the other of the foregoing conditions does not hold, then the test constructor must decide whether there is a desired shape for the score distribution, or whether there are specific areas of the score scale where higher precision is desired than at other areas.

2. If trait estimation is the goal of the assessment, it must be determined whether the trait is unidimensional or whether it is composed of a composite of dimensions. In the former case, statistical specifications need to indicate that items should be related to the hypothetical trait. If the construct is based on a

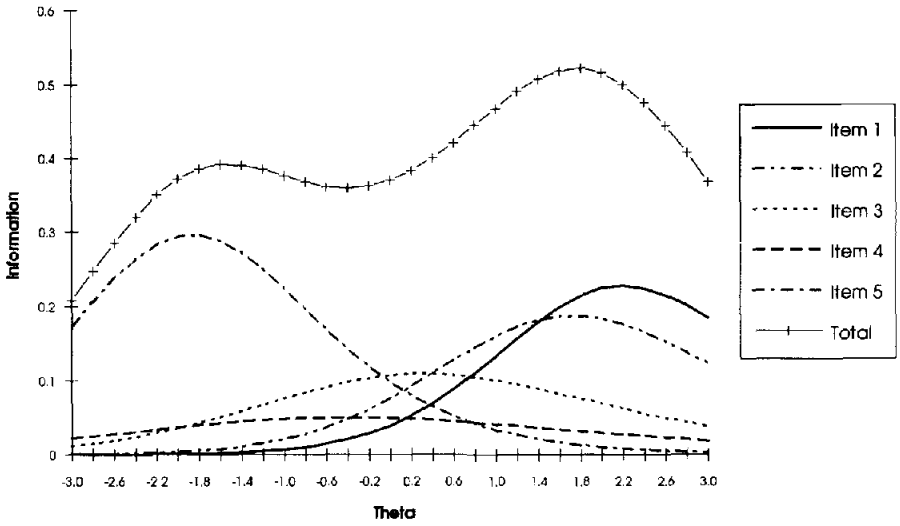

Figure 1. Item and test information curves for a five-item test.

composite of dimensions, such as for measures of mathematics aptitude that might require a combination of mathematics skills, then the approach to ensuring coverage of the dimensions needs to be specified.

3. If prediction or assignment to groups is the goal of the assessment, good criterion measures are needed to guide item selection. Are well-defined groups available? Is there a highly reliable criterion variable?

4. If the goal is to construct an index, the key is to find items that are not too highly correlated that can be combined in a meaningful way. However, without an underlying theory, it may be difficult to interpret the values of the index.

Statistical characteristics. Are there desirable statistical characteristics for the test? Some of the possibilities include a minimum level of reliability for a particular sample of examinees, desired score distributions, required precision at percentile points of the score distribution, form of the test information function, and so on.

1. To meet high internal consistency reliability requirements, tests need to either be long or to have fairly high interitem correlations. It is easier to achieve high reliability with items of about the same level of difficulty than with a spread of difficulty, but such tests do not provide equal precision over a wide range of abilities.

2. If items are evenly spread in terms of proportion correct, a unimodal, near-normal score distribution will result. However, any other shape for a score distribution can be attained through careful selection of test items.

3. The standard error at a point on the score scale can be made as small as desired by adding more items that provide maximum information at that point. This can most easily be done in an item-response-theory (IRT) framework, but tests can also be constructed to maximize precision at a point by picking items with particular p values.

4. Target information functions are a good method for helping test designers to think about the characteristics that they desire in a test. Ranges of the score scale requiring high precision have high levels for the target information curve.

Item Selection Process

The item selection process described in Pintner (1931) was basically intended to screen out nondiscriminating items and then to select items to have an average difficulty of about .5. Whether to use a spread of difficulty or to have the difficulty of the items clustered around .5 is the test developer's choice. Current item selection processes are much more targeted and elaborate. As noted, item selection is now similar to constructing a complex object from component parts. The items must meet content and statistical criteria at the same time, and the entire test must meet all of the requirements related to its use, including domain coverage or trait definition, reliability or information requirements, and validity. Many standardized tests have several levels of content and other specifications, including cognitive complexity, further statistical requirements for the items, and the need to produce multiple forms that can be used interchangeably. Under these complex circumstances, the item selection process may take weeks of effort after the items have been pretested and refined to achieve the required balance incorporating all of the desired test features. This process has become so complex that it may be impossible for a test developer to keep simultaneous track of the match to specifications for all of the needed item characteristics as test construction proceeds. The result is that either some of the characteristics are ignored, the test development process becomes an extended trial-and-error exercise, or computer software is developed to help manage the process.

The use of computers to select test items to match test specifications is one of the newest innovations in the test development process. To use these procedures, the content and statistical specifications are encoded in a way that can be communicated to the computer programs. This encoding process has the added advantage of forcing the specifications to be made absolutely clear and explicit. In addition to the formal encoding of specifications, a number of the procedures require the specification of a goal for optimization. For example, the goal might be to construct the most reliable test possible that meets all of the

requirements of the specifications. Alternatively, the goal might be to match a target information function as closely as possible given all of the other requirements.

Many researchers are now working on the optimal item selection problem, but the articles by Adema and van der Linden (1989), Armstrong, Jones, and Wang (1994), and Stocking, Swanson, and Pearlman (1993) provide a broad selection of the approaches that are being used. Complete understanding of these methods requires background in operations research methodology, so the details of the procedures will not be discussed here. However, user-friendly software is being developed that will make item selection based on multiple content and statistical criteria a relatively easy task. Major testing companies are already using software of this type.

Test and Item Review

Pintner (1931) included no provision for the review of the test or items by any individuals external to the test development process. It can be inferred that the test author would review the items for ambiguity or too great a similarity between items, but this would be an informal process. Later texts on test development such as that by Greene, Jorgensen, and Gerberich (1944) provided guidelines for reviewing test items to remove ambiguity and cluing, but no other types of reviews are suggested.

Current test construction procedures include extensive guidelines for review. The books on item development by Osterlind (1989) and Roid and Haladyna (1982) contain full chapters on content, technical, and stylistic reviews. Testing organizations have also developed extensive review procedures to ensure that items do not advantage or disadvantage subgroups or an examinee population through inclusion of irrelevant item features or through offensive language or situations (Educational Testing Service, 1987). Current practice includes the following levels of review.

Editorial review. The goal of this review is to ensure that the language of the test question is clear and unambiguous. Full test forms are reviewed to ensure that one test question does not answer another one or closely duplicate an item.

Content review. Items are reviewed to assure that they have a correct answer or that they can be interpreted according to the goal of the assessment. They are also reviewed to determine whether they assess the content described in the content-specifications document.

Statistical review. Items are reviewed to determine whether they seem to be functioning properly as measurement tools. This includes checks on discriminating power and difficulty as well as statistical checks on differences in functioning for different subgroups of the population. These later analyses are usually labeled *DIF analysis* (differential item functioning). A good summary of DIF procedures is given in Holland and Wainer (1993).

Sensitivity review. Individuals sensitive to differences in cultural groups are asked to review items to determine whether they are fair to all of the groups and whether they contain offensive stereotypes or language.

For high-stakes standardized tests such as the ACT's college admissions tests, items are reviewed at well over a dozen processing points. Numerous independent reviews are performed both by individuals and committees of external reviewers. Certainly all assessment instruments are not reviewed so thoroughly, but all instruments need to be independently reviewed by at least a few persons other than the test's authors; concerns for fairness should be considered.

Summary

The goal of this article has been to provide a summary of the state of the art in test construction methodology. Although full coverage would require several books, the intent here is to help every psychologist become familiar with the basic issues of test construction, both historically and at present. The basic overview and reference sources will provide a framework for psychologists to use as a guide to new instrument development. Test development has progressed to the point of being less art but not quite science. Although many formalized procedures are now available, informed judgment is a major component in the process.

This article has provided references to many resources that should be helpful for the person who wishes to produce or make intelligent use of high-quality assessment instruments. These references indicate what is currently known about the process, but the comparison to Pintner (1931) demonstrates that over time the sophistication of the process has greatly increased. The research and development work on the test development process is ongoing, and there is every expectation that 30 years from now the process will be vastly improved. Persons who are involved in the test development process need to take the time to stay informed about new advances as they occur.

References

Adema, J. J., & van der Linden, W. J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics, 14,* 279–290.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: Author.

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington, DC: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Armstrong, R. D., Jones, D. H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational Statistics, 19,* 73–90.

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1996). *Domain-referenced reporting in large-scale educational assessments.* Commissioned paper to the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment, Stanford, CA.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Educational Testing Service. (1987). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3–17). Orlando, FL: Academic Press.

Greene, H. A., Jorgensen, A. N., & Gerberich, J. R. (1944). *Measurement and evaluation in the secondary school.* New York: Longmans, Green, & Co.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. 1. Additive and polynomial representations.* New York: Academic Press.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 335–366). New York: American Council on Education and Macmillan.

Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer.

Pintner, R. (1931). *Intelligence testing: Methods and results.* New York: Henry Holt.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17,* 167–176.

Received May 31, 1996
Revision received July 24, 1996
Accepted July 24, 1996 ■

New Editors Appointed, 1998–2003

The Publications and Communications Board of the American Psychological Association announces the appointment of five new editors for 6-year terms beginning in 1998.

As of January 1, 1997, manuscripts should be directed as follows:

- For the *Journal of Experimental Psychology: Animal Behavior Processes,* submit manuscripts to Mark E. Bouton, PhD, Department of Psychology, University of Vermont, Burlington, VT 05405-0134.
- For the *Journal of Family Psychology,* submit manuscripts to Ross D. Parke, PhD, Department of Psychology and Center for Family Studies–075, 1419 Life Sciences, University of California, Riverside, CA 92521-0426.
- For the Personality Processes and Individual Differences section of the *Journal of Personality and Social Psychology,* submit manuscripts to Ed Diener, PhD, Department of Psychology, University of Illinois, 603 East Daniel, Champaign, IL 61820.
- For *Psychological Assessment,* submit manuscripts to Stephen N. Haynes, PhD, Department of Psychology, University of Hawaii, 2430 Campus Road, Honolulu, HI 96822.
- For *Psychology and Aging,* submit manuscripts to Leah L. Light, PhD, Pitzer College, 1050 North Mills Avenue, Claremont, CA 91711-6110.

Manuscript submission patterns make the precise date of completion of the 1997 volumes uncertain. Current editors, Stewart H. Hulse, PhD; Ronald F. Levant, EdD; Russell G. Geen, PhD; James N. Butcher, PhD; and Timothy A. Salthouse, PhD, respectively, will receive and consider manuscripts until December 31, 1996. Should 1997 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 1998 volumes.