

Clinical and Research Implications of Revising Psychological Tests

Marshall L. Silverstein
Long Island University

Linda D. Nelson
University of California, Irvine

This article integrates those of other contributors to this special section, "Methods and Implications of Revising Assessment Instruments," to underscore important conceptual factors to consider when undertaking test revisions. These considerations include determination of when test measures have become sufficiently understood to be incorporated in a test revision, cohort effects, revision of administration formats and test instructions, and comparisons of performance levels across test versions. The discussion of these factors also takes into consideration clinical practice and educational implications of making a transition to revised test versions.

The revision of a major psychological test is an effort that is intended to have an influence extending approximately one generation. However, there exist few guidelines in the literature regarding clinical and research adaptations to incorporating new editions of tests, despite the fact that planned revisions of tests have been undertaken for well over 50 years. Although revisions are more frequent now, to our knowledge there has not been a systematic examination of considerations involved in revising tests. Our concluding article represents an attempt to synthesize the views of the other contributors to this special section, adding several of our own.

Because revisions emerge out of a need to enhance the psychometric characteristics of tests, including updating of norms, what a test assesses conceptually is not often reconsidered apart from its psychometric properties. Further, incorporating views about socio-cultural factors and cohort effects such as progressive increments in intellectual capabilities (the so-called Flynn effect) has implications extending beyond achieving superior psychometric or measurement sophistication. In addition, facilitating comparisons between existing versions of tests and their revised editions (e.g., equating test scores) is rarely anticipated. Consequently, the implications of converting to a revised version require greater attention concerning clinical examinations and longitudinal research that study performance levels over time. There is also room for more carefully addressing how to go about revising test formats and instructions, particularly when there exists more than one format of a test in common use, for example, tests of verbal fluency or word list learning.

Our discussion emphasizes conceptual concerns that merit consideration when one creates new test editions. We begin with a discussion of the other contributions to this special section, organized by broad themes to emphasize the links among the articles. We thus stress integrating factors that are usually considered on their own merit but that may be seen to better advantage as congruent ideas. Accordingly, we discuss the articles grouped by the following ideas: (a) establishing when test measures and the psychological concepts underlying tests have become sufficiently understood to be incorporated in a test revision, (b) considering cohort effects such as the Flynn effect and sociocultural factors, and (c) reconceptualizing the psychological meaning of tests and their principal measures.

In our second section, we address the problem of standardization of test instructions and administration formats to be considered when tests are revised. We include here a comment about the matter of updating tests such as the Wechsler intelligence scales that are often used in conjunction with tests that are not themselves revised very often. We also provide a discussion of implications of test revision on longitudinal studies and comparability of test measures during the transition from one version to another.

Discussion of the Special Section

When Is a New Measure or Concept Sufficiently Understood to Be Incorporated in a Test Revision?

Strauss and Spreen (2000) provided a discussion of the fundamental question underlying much that is crucial to this special section by asking, in effect, why tests should be revised. They suggested that a revision provides a better means to answer questions about individuals' functioning than the preexisting version of the instrument. Thus, tests should be revised when their principal constructs are better understood and when levels of performance can be reliably measured or discriminated. Reise, Waller, and Comrey (2000) discussed a different but nevertheless related point about the psychometric purity of a test. They argued that some good reasons for revising tests result from matters such as the internal consistency reliability of test scores, generalizability of factor structures across groups, and the persistence of questions

Editor's Note. Stephen N. Haynes served as action editor for this article—L.D.N.

Marshall L. Silverstein, Department of Psychology, Long Island University; Linda D. Nelson, Department of Neurology, University of California, Irvine.

Correspondence concerning this article should be addressed to Marshall L. Silverstein, Department of Psychology, Long Island University, C. W. Post Campus, Brookville, New York 11548. Electronic mail may be sent to msilver@liu.edu.

about the nature of a test's factor structure and, therefore, its construct representation. Reise et al. provided a reminder that psychological constructs and the psychometrically derived factors that measure such factors are not static phenomena; their point is no less compelling than the more clinically driven considerations for revising tests raised by Strauss and Spreen.

Strauss and Spreen (2000) and Reise et al. (2000) also did not lose sight of the fact that examiners need to be clear about the research and clinical questions they want tests to answer. We extend their argument by posing a question that still remains: What criteria can be applied to incorporating promising scales that have emerged since a previous test revision to determine whether these have become clarified well enough to become part of the new test edition? Stated differently, it may be worthwhile considering whether there is a role for adopting scales for heuristic purposes when tests are revised before such scales' psychometric properties and clinical significance are well understood.

To illustrate this point, it cannot be said with certainty that the clinical meanings of some scales derived from the California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987), for example, such as the proactive inhibition effect on learning List B and the learning slope calculation, are sufficiently established to have been included in the Wechsler Memory Scale—3rd ed., (WMS—III; Wechsler, 1997b) even as supplementary measures. Further, it is not clear whether scales such as List B recall and learning slope measure learning or retention; consequently their inclusion on the WMS—III, an instrument principally assessing memory, may be premature. (This view may perhaps be modified shortly, because the CVLT is another example of a test that at the time of this writing is nearing its own first revision.) However, variables such as these are of interest for what they may imply about the accelerating understanding of the exciting developments in cognitive science.

For this reason alone, these scales are sufficiently promising or important to retain heuristically, at least for additional study as supplementary measures. It is possible to await the evidence for validity, as long as the scales at issue are reliable. Naturally, it should be made clear that preliminary measures such as List B recall or learning slope are not of the same degree of conceptual significance as subtests such as Logical Memory or Digit Span. Except for perseverative responses, the same caveat is largely true about most of the variables on the Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993).

What seems necessary here is for the considerations affecting a given decision to be made explicit by test authors and publishers, rather than making it appear as if a concept or measure had achieved a validated meaning. Psychologists may not always know the definitive answers at the point of a revision, but most can live with the fact that any revised test is always by its fundamental nature a work in progress. Over time, of course, promising measures or scales must progress beyond an exploratory, heuristic stage to achieve the level of psychometric substantiation that Strauss and Spreen (2000) and Reise et al. (2000) have noted.

The Flynn Effect and Sociocultural Considerations

In addition to validation of test measures and the psychometric refinement of factor structures on tests, several contributors to this special section (Tulsky & Ledbetter, 2000; Strauss & Spreen,

2000) commented on the so-called Flynn effect (Flynn, 1984, 1999). The Flynn effect refers to regularly observed increases in IQ scores over time. This cohort effect is usually observed in relation to intelligence, and it corresponds to higher levels of educational attainment across successive generations. It also has important bearing on virtually all tests of cognitive function, although it has not received much attention in the fields of cognitive or neuropsychological assessment as a concept, despite the periodic renorming of select tests that typically corrects for this type of cohort effect.

We would add to this discussion that Flynn (1984, 1999) was himself puzzled by the increase in IQ across a relatively narrow band of just 2 decades, during which time college board aptitude scores declined. Some of our colleagues in education fields are also puzzled as they contemplate what is signified by the overreliance on the ubiquitous spell-check assistance or pocket calculator among even some of our more intellectually gifted youth of today. Many will nod to themselves knowingly, but are left to ponder what to make of this. Whether motivational factors or sociocultural indications of the times lurk behind the facts Flynn documented, the relationship among cohort effects, IQ, and recalibrated norms is not a simple matter. Regardless of what we come to believe IQ or index scores tell us about cognitive abilities or educational standards, in consideration of the Flynn effect and its implications for revising tests, there can be little doubt that revisions should update norms to reflect what they must, as long as we are careful not to delude ourselves about what we believe these updated norms really mean.

Priorities also need to be established about what considerations are most germane in planning revisions. Some of these are sensitive issues, such as the question about how best to incorporate ethnic or sociocultural variables in forthcoming versions of standard tests. Okazaki and Sue (2000) argued that studying different cultural groups introduces new variables to the process of test standardization. Thus, moderator variables such as acculturation level add a whole new dimension to test revision. There is not a clear consensus about oversampling ethnic groups (and if so, which ones, possibly at the expense of other ethnic groups), nor is there general agreement about whether such a decision should be determined by population demographics or whether ethnic populations should have separate norms. Inevitably, however, the question of limited resources, time, and priorities arises in this context, and psychologists may anticipate lively debate about the extent to which psychometric, test development, and item reliability concerns should emphasize adjustments for sociocultural factors.

As the Flynn effect demonstrates, even relatively straightforward adjustments reflecting shifts in cohort effects need to be better understood before one uncritically accepts the position that upgrading norms reflects accelerated intellectual competence. The same caveat applies equally to revising tests while incorporating adjustments for sociocultural factors. The overriding concern needs to remain focused on the clinical and scientific value of correcting inapplicable and sometimes inequitable norms. Calling attention to sociocultural variables is a valuable reminder of one of several crucial areas one should keep in mind when determining what is ultimately pressing or urgent when tests are revised, as Okazaki and Sue suggested. Of all the objectives of test revisions, the relative importance of any of a number of types of inequities that should be remediated remains to be determined wherever

these occur, whether at the level of items, subtests, or misapplication of norms to ethnic groups or minority populations. Ultimately, however, scientific and clinical considerations must remain foremost.

Revising Tests and Rethinking What They Measure

By periodically reexamining how mental functions are measured when tests are revised, there is also an opportunity to consider important new ways and sometimes old ways of thinking about tests, items, and constructs. For example, Knowles and Condon (2000) and Adams (2000) presented good examples of how the progress resulting from test revision leads to better instruments and theories as well as rediscovering value in earlier, discarded views. This often occurs when starting out to study a phenomenon: Our hunches may be correct, but the phenomenon or its nuances are not yet sufficiently understood. Nor is the instrumentation to measure it available or perfected.

Adams (2000) noted that undertaking a revision of a major test permits revisiting the fundamental assumptions underlying the test when it was originally developed. Nowhere is this as compellingly evident as in the case of comparing measures of intelligence and achievement. For example, Sternberg (1999) argued that intelligence and achievement tests are more similar than they are different from each other. This position stands in contrast to the view that intelligence tests assess abilities that predict academic achievement. Thus, a central concept underlying the prognostic significance of intelligence itself is being called into question once again, as it has been repeatedly over many decades. By revisiting the concept during a planned revision, test developers decide on new research data to collect, significant changes in domains represented, and new conditions of test use. Test revision thus presents new opportunities to reexamine such questions, particularly when revisions incorporate plans for conforming with ancillary measures such as achievement.

In observing how progress in one field may affect the developments occurring in another field, Knowles and Condon (2000), in their informative discussion of contemporary psychometric theory, showed how modified measurement approaches have made new developments possible in the field of psychological testing. They demonstrated that psychometric theory, too, has its own internal debates and pressing issues. It may come as a surprise to learn that psychometricians can become as passionate as clinicians about what tests measure. Indeed, the call for "new rules" of measurement that are "statistically sophisticated compared with classical test theory" (Embretson, 1996, p. 348) is provocative. Perhaps the specific area of factor analysis is more settled, although Reise et al. (2000) noted that there remain persistent questions about conceptual approaches, including the number of factors to extract and retain for rotation, the arbitrariness of the "eigenvalue greater than 1.0" rule, and the overlap between factor analysis and principal-components analysis.

Most fields struggle with nagging methodological and conceptual issues, and Knowles and Condon (2000) succinctly summarized some of the central issues of what appears to be a fundamental change in rethinking measurement theory. They also provided an interesting discussion of the role of context in test revision, illustrating a perspective about the level of individual items that will be informative for clinicians, who typically have

something quite different in mind when a term like *context* is mentioned in connection with assessment and the testing situation. There is evidently room for both molecular and molar levels of understanding when one approaches an endeavor such as revising tests.

Test Revisions: Other Considerations

Next we consider two issues that also deserve attention in thinking about revising tests. The first matter is that of the wide variation that persists in the formats and scores of important tests that are commonly used in conjunction with instruments that have undergone extensive revision of norms, instructions, and scales. The second point we address concerns a problem for which there exists little discussion: comparing test scores across test versions (e.g., when a new version has been introduced subsequent to a baseline assessment point). We consider this issue primarily from the standpoint of longitudinal research.

Unstandardized Administration Formats, Scores, and Test Instructions

The first issue concerns a problem in cognitive assessment that seems to affect neuropsychological testing more prominently than it affects other fields. Neuropsychological test batteries typically include many tests. Some of these tests are revised periodically, thus benefiting from psychometric and normative enhancements, such as those seen on the new Wechsler Adult Intelligence Scale—3rd ed. (WAIS—III; Wechsler, 1997a) and the WMS—III. However, other important tests have not been revised, and there is also little consensus about their formats, administration procedures, and even what their major scores should be. We refer here to instruments such as tests of attention, word list learning, and verbal fluency.

For example, there are several formats of cancellation tests and the continuous performance tests of attentional dysfunction. The popular Stroop (1935) method, now enjoying a resurgence, is another good example of an instrument with strong roots in experimental psychology for which little more than the basic color-word conflict or interference concept is constant among the many formats available for its study or clinical application. It may seem unresourceful to devote effort to revising some tests while other important tests with unstandardized administration formats, frequently used in conjunction with the major tests that do undergo revisions, are left behind. This constrains both the interpretation of the research literature and comparisons of findings on clinical examinations.

Admittedly, the issue surrounding these and other tests is not primarily one of revision. However, the fact that there is so little standardization of administration formats compromises the advantages of revisions of major tests such as the WAIS—III and WMS—III. The issue is sufficiently problematic that it deserves to be considered with the same attention accorded to concerns about revising tests.

There is little that is more frustrating than reading empirical reports that are based on different versions of the same test or undertaking a clinical evaluation of a patient when a previous examination used a different or perhaps idiosyncratic version of an important test. One of the first handbooks of comparative norms

(Mitrushina, Boone, & D'Elia, 1999) for different versions of neuropsychological tests reflects the extent of the problem. The point in time at which a major test is scheduled for revision is also an excellent time to contend with this matter of standardized test administration formats. By judiciously selecting one of several standard versions of tests such as the Stroop or by selecting tests of functions such as verbal fluency or letter cancellation, psychologists may clear a path for establishing such versions as accepted standards, particularly if conorming is used simultaneously.

In addition to the lack of consensus or standardization of test formats, a similar concern applies to unstandardized measures as well as how to calculate a score even when there is general agreement about what a particular score or measure should be. Conceptually defining a measure is not the same thing as operationalizing scoring criteria. One of the best examples of a good remedy is the important standardization of the WCST (Heaton, 1981). One of Heaton's significant contributions, in addition to establishing the 128-card format as the standard, is the clarity and rigorous precision of the now-accepted definitions of the test's major measures.

But the WCST standardization effort is an exceptional instance that has not, unfortunately, had enough of an influence on other tests of neuropsychological functions, even 19 years after the first edition of the manual was published (Heaton, 1981). The WCST has fulfilled its promise to achieve a standardized administration of this now-popular test, its revision (Heaton et al., 1993) successfully emphasizing refinement of norms and the meaning of the WCST's major variables. This represents a different view of what a test revision can mean; it is, at the very least, a different type of revision than that of the WAIS-III. In this respect, Heaton has done for the WCST what Exner (1993) did for the Rorschach in the field of personality assessment. As a result of Exner and his colleagues' achievements, standardizing the Rorschach administration procedure enabled revisions of the Comprehensive System to concentrate on more precise, fine-tuned clarity of its primary variables and higher order cluster search strategies (Exner, 1995; Weiner, 1998). The same shift in emphasis is now possible for the WCST. This makes possible such accomplishments as a better understanding of the various perseveration response measures to help clarify which of these variables should best be considered the measure of choice.

Further, the matter of test instructions and their modification with a test revision has rarely been addressed. Although this issue may not appear to belong with a discussion of revising test items, scales, and norms, many clinicians realize that patients often miss some fine or subtle distinctions contained in instructions, especially when clinicians are examining patients with limited intelligence, cerebral dysfunction, narrowed attention, or quite simply, little motivation to listen carefully. To claim that an instruction like "What might this be?" emphasizes perceptual processing while minimizing fantasy or association may be an arguable contention, and to expect that patients can clearly differentiate a picnic scene from a yard scene on the WMS-III is likewise debatable. Similarly, the degree of strict adherence to instructions, flexible paraphrasing, or teaching a task requires greater clarity. For example, a substantial number of patients usually require considerable teaching, reorienting, and reminding on difficult tests such as the WCST and the Trail Making Test—Part B (Reitan & Wolfson, 1985). Examiners may also be of different temperaments concern-

ing whether, how often, and how persistently to repeat the instruction to "Tell me everything you can remember . . ." on WMS-III Logical Memory when evaluating patients with compromised speech or verbal fluency, and examiners may wonder whether it is permissible to add something as seemingly innocuous as "Anything else?," which could affect scoring in varying ways. Thus, phrasing of instructions, clarifying latitude of paraphrasing, and the use of prompts or probes (while simultaneously realizing that patients do not hear and understand instructions in a precise and exacting way) are all additional matters that can profit from further attention when tests are revised.

The WAIS-III made a good start by beginning to address this issue. If for no other reason than forcing attention to the matter, standardizing administration formats of selected important tests and unambiguously clarifying these tests' principal measures are worthwhile matters to include as part of a broadened concept of test revision.

Research and Clinical Comparisons of Performance Levels Longitudinally

Many studies using psychological tests are cross-sectional in design, including research that uses test measures as dependent variables to investigate questions other than test performance as the primary measure of interest. Several studies that use longitudinal designs involve pretest-posttest comparisons, in which a test or scale is not expected to change. Some of these designs require either lengthy follow-up intervals of several years or multiple points in time at which measures will be obtained.

Follow-up studies of this type are often among the most important in the field because they have the potential to produce findings of considerable interest owing to the lengthy follow-up interval. Other types of studies that may not necessarily be longitudinal in design but that can require years to complete include studies investigating conditions of low prevalence, necessitating long periods of time to recruit suitable patients to enter a protocol; studies of multiple variables, requiring large numbers of participants to ensure reliability if multivariate designs are to be used; and collaborative multisite studies, in which patients enter protocols at different times. Studies like these are expensive to conduct, and they also require careful planning, perseverance, and attention to methodological problems of their own, not the least of which are participant attrition and diligent control of potential confounds by intervening variables.

One of the characteristics of such study designs is that measures can be planned in advance and controlled with relative precision. However, it can become quite burdensome for longitudinal studies to contend with revised editions of test measures when these are introduced, particularly when it may not be possible to anticipate the form a revision will take, when it will occur, and how comparable the measures will be at the different time periods. This necessitates a decision at some point between starting and completing the protocol: Does the investigator switch to the revised edition and abandon the data on participants who were administered the prior edition, does the study continue through to completion but with missing data, or are all participants retained but with the use of potentially imprecise calculations based on estimated, adjusted, or "equated" scores? There are no clear-cut answers; thus, in the absence of guidelines, the different solutions

investigators arrive at only muddy the waters and consequently do not promote satisfactory scientific progress or communication. It should also not be overlooked that a similar problem exists when one conducts clinical assessments of progression of illness that emphasize differentiating changes in function from chronic, residual effects of an illness.

One potential solution to both the clinical assessment problem and the research decision of retaining data on participants who entered a study when a preexisting test version existed may be developing strategies for equating test scores. It may be a worthwhile effort to consider this type of solution of comparing test scores between preexisting test versions and their revisions, at least for some limited or specific uses that do not drastically alter their intended meaning.

At transition points during the conversion to a revised form of a test, it may be desirable for investigators to reserve subsamples for counterbalanced administrations of both versions to supplement equating subtest scores. In addition to conserving valuable research resources while collecting data, such approaches offer potential utility in other ways. One instance is in preserving the ongoing value found in the work of older important studies, some of which have even been considered landmarks in various fields of research. The merits of certain of these classics should be able to outlive the fact that preexisting versions of tests were used, which may be a minor concession in view of the broader importance of such major studies for advancing the conceptual development of various research areas.

Equating test scores in this fashion is one solution to a pragmatic problem. By itself it is an imperfect solution, which is probably why it has received little attention, but it is an approach that may have merit if combined with other attempts to demonstrate comparability of test scores across revisions. Equating test scores is not a panacea, but neither is it an approach that is without value or a strategy to be disdained. It enhances flexibility of measurement when used cautiously as a supplement to other approaches to comparing scores on more than one occasion after a revision has been introduced.

Planning for Future Revisions

Butcher (2000) suggested that it is important to devote time to accommodate new editions and to keep pace with the fact that revisions of tests are becoming a fact of life, and a welcome one at that. We may take confidence in knowing that the task is not insurmountable. Butcher (2000) and Tulsy and Ledbetter (2000), in this special section, have pointed the way for the field, and they speak with the voices of experience and confidence about how to go about conducting revisions with the understanding that tests should be revised periodically. Tulsy and Ledbetter demonstrated convincingly that a horse put together by a committee need not resemble an elephant but rather may be a better, more powerful, and ultimately more satisfying horse. Butcher addressed the matter of changing from a "beloved and relied-upon test" (p. 268) to an improved, updated version. The time is at hand to think hard about formulating guidelines for introducing revisions of tests.

Most people are familiar with the wisdom of deferring the use of new software upgrades with the suffix suffix ".0" in the title, as it is common knowledge that trying them is rewarded with bugs, crashed programs, and a host of headaches that replace the initial

surge of excitement after a new program comes on the market. Versions ending in ".01," ".02," and higher provide greater assurance that problems that could not have been anticipated at the outset have been solved. Fixing software bugs may be easy enough, but correcting comparable bugs discovered after a major revision of a test has been issued is by no means a simple matter. Major tests must await many years to be restandardized; therefore, careful anticipatory planning for an instrument with an expected minimum shelf life of 10 to 15 years taxes the foresight of the best of us. Test creators do not usually have updates other than minor error corrections or an additional page to append to a manual.

Another matter that is often unaddressed concerns educational and clinical practice implications of implementing the transition to a revised test. As the past has demonstrated, every revision of a psychological test eventually replaces its predecessor, not just because it is new or necessarily because it is better, but for the simple reason that it becomes the standard. If that does not occur immediately, it will surely come to pass within a few years. The sooner the conversion is made, the better it will be for familiarizing psychologists with the idiosyncracies of a new instrument, applying its benefits to patients, and evaluating its influence on effect sizes and power if the revised version will be incorporated into ongoing studies.

This point may seem so obvious as to be needless to note. However, some clinicians are just beginning to become aware of the existence of the WAIS-III, others have not yet studied or implemented the new instrument (although intending to do so), and still others are taking a wait-and-watch approach before converting. Some graduate programs continue to teach the WAIS-R, although the number has been decreasing in the nearly 3 years since the WAIS-III revision was published. Indeed, the entire matter of graduate and postgraduate education concerning the transition from existing test versions to their revisions has been virtually ignored. Test publishers do not typically orient faculty who teach assessment courses about major changes apart from overviews at national or regional meetings and information brochures. In addition, not only must faculty unlearn old habits as they learn the new instruments, they have to build up experience using new versions to be able to teach the revised instruments with a sophisticated understanding of their nuances and to provide comparisons with the instruments the revised tests are replacing. We may also wonder how many graduate programs make provisions in busy and already overcrowded curricula to orient advanced graduate students who have already completed the assessment courses about revisions that have appeared since taking the courses, so that they may be appropriately prepared prior to undertaking internship and externship placements (Silverstein, 1996). In addition to busy clinicians who must make time to study and acquire experience with new versions of major tests, internship faculty may have even less time than academic faculty to learn the instruments.

From a practical standpoint about implementing test revisions, considerations such as those we have enumerated above are rarely discussed, and guidelines for facilitating such educational exchanges do not exist. It is probably not enough to simply say that transitions to revised test editions should be made as soon as possible, particularly because people do not usually relish having to modify well-learned or perfected habits and deal with that intangible quality of what it means to have a test and its various nuances under one's skin. One immediately apparent example

concerns mastering the subtleties of scoring the WAIS-III Vocabulary, Similarities, and Comprehension subtests. Few people give up an old shoe without a struggle, much like Butcher's (2000) discussion concerning a "beloved and relied-upon test" (p. 268), and educators are no exception. Thoughtful reflection about the educational matters that we have raised should be emphasized to minimize what needs to be done to facilitate the transitions that attend learning and understanding a new version of a test in depth.

If investigators and clinicians remained with the previous versions of important tests, they would have to contend with the problem of working with findings from an instrument that might be considered outdated in a few years. However, converting to the revised versions raises questions about what investigators would do with the frequently substantial numbers of research participants who had already entered ongoing studies, some of which also involved a considerable amount of data collected on instruments other than the tests now reintroduced as revisions. In addition, psychologists are uncertain about the distributions of some revised tests until the tests have been proven. The confusion only intensified when Flynn (1984, 1999) subsequently reported substantial cohort differences in measured IQ ("massive gains" was Flynn's, 1984, term) across just 1 decade. Researchers are thus justifiably unsure of what is a prudent path to follow. Some advise waiting to see what further research clarifies, but one implication of Flynn's findings was that there may not be time to wait before deciding whether to embrace the revision.

A body of literature to inform decisions about this type of transition is either lacking or insufficient. Journal editors rely on the cumulative wisdom of many reviewers for guidelines, but comments about the transition to revised versions or the psychometric issues of combining versions to conserve resources are infrequent regarding the methodological choices that investigators face. Therefore, this special section came about, originating with a pragmatic question, but one with important theoretical undertones, and coming at a point in time when revising major test instruments has become more sophisticated, frequent, and expectable. The time is at hand for attending to the conceptual, educational, and pragmatic concerns that follow test revisions in the same way that psychometric advances have traditionally guided the need for revising tests.

References

- Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12*, 281-286.
- Butcher, J. N. (2000). Revising psychological tests: Lessons learned from the revision of the MMPI. *Psychological Assessment, 12*, 263-271.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test manual*. San Antonio, TX: The Psychological Corporation.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.
- Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (3rd ed.). New York: Wiley.
- Exner, J. E., Jr. (Ed.). (1995). *Issues and methods in Rorschach research*. Mahwah, NJ: Erlbaum.
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.
- Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*, 5-20.
- Heaton, R. K. (1981). *Wisconsin Card Sorting Test manual*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual*. Odessa, FL: Psychological Assessment Resources.
- Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12*, 245-252.
- Mitrushina, M. N., Boone, K. B., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.
- Okazaki, S., & Sue, S. (2000). Implications of test revisions for assessment with Asian Americans. *Psychological Assessment, 12*, 272-280.
- Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287-297.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery*. Tucson, AZ: Neuropsychology Press.
- Silverstein, M. L. (1996). Teaching the Rorschach and learning psychodiagnostic testing: A commentary on Hilsenroth and Handler (1995). *Journal of Personality Assessment, 66*, 355-362.
- Sternberg, R. J. (1999). Ability and expertise: It's time to replace the current model of intelligence. *American Educator, 23*, 10-13.
- Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12*, 237-244.
- Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.
- Tulsky, D. S., & Ledbetter, M. F. (2000). Updating to the WAIS-III and WMS-III: Considerations for research and clinical practice. *Psychological Assessment, 12*, 253-262.
- Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale (WAIS-III)*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler Memory Scale (WMS-III)*. San Antonio, TX: The Psychological Corporation.
- Weiner, I. B. (1998). *Principles of Rorschach interpretation*. Mahwah, NJ: Erlbaum.

Received February 7, 2000
 Revision received May 9, 2000
 Accepted May 12, 2000 ■