6. Testing and the Oscar Buros Lament: From Knowledge to Implementation to Use

James V. Mitchell Jr.
Buros Institute of Mental Measurements

Follow this and additional works at: http://digitalcommons.unl.edu/burostestingissues
Part of the Educational Assessment, Evaluation, and Research Commons

http://digitalcommons.unl.edu/burostestingissues/8

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 Social and Technical Issues in Testing: Implications for Test Construction and Usage by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
The field of measurement can be conceptualized as having three different but interrelated aspects. First of all, it is a science or a body of knowledge concerned with the development of theory and methodology and with the identification and confirmation of generalizations governing interrelationships among variables appropriate to its content. Measurement theory and its application to measurement problems are important contributors here. Second, it is an applied science or technology concerned with the development of products that represent a useful application of such a science or body of knowledge. For the field of measurement, test development and validation are important exemplars. Third, it is a body of information concerned with why, when, and how these products are used, and the results of such use, in the practical measurement setting for which they were typically intended. This sequence of interrelated aspects of measurement, from knowledge to implementation to use, is the conceptual foundation for much of what follows.

Within such a context as that just described, the Buros Institute of Mental Measurements has always played a unique role. The science of measurement or measurement theory has not been one of the Institute’s chief concerns, although the Institute is often an indirect beneficiary of such contributions. However, the Institute has had major involvement with the evaluation of test products, the products of an applied science, and with the education of test users in the more effective selection and use of those products. Because of the nature of this involvement, the Institute has had a perspective on the three separate aspects of the field of measurement that is not typical of those representing only the singular aspects of the continuum. It is this unique perspective of the Institute that will serve as the distinguishing feature of the discussion to follow.
The purpose of this discussion is to evaluate critically the contributions and progress made in these separate, but interrelated, aspects of measurement: knowledge, implementation, and use. The theme of this discussion is that the greatest progress has been made in our knowledge, lesser progress in implementation, and the very least progress in selection and effective use. The implication of the discussion is that there is a pressing need to redress the imbalance that has developed.

MEASUREMENT THEORY AND KNOWLEDGE

No one can accuse the field of measurement of being static. Ferment seems to be the rule. With this ferment has come new theories and models, controversy that sometimes yields as much light as heat, new understanding, and some fresh perspectives. Although it typically seems that activity has been greater than results, the results themselves show evidence of progress. Two of the more recent reviews of test theory (Subkoviak & Baker, 1978; Weiss & Davison, 1981) both devoted considerable attention to criterion-referenced testing, latent-trait theory, and issues of test bias. Another recent review devoted entirely to latent-trait theories (Traub & Wolfe, 1981) described the promise of latent-trait theories in their application to educational measurement but also issued a caveat about work to date and needed precautions. The overall impression obtained from these reviews is that criterion-referenced testing, latent-trait theory, and test bias have received the attention deserved from an able group of professionals and that some relevant problems have been addressed, development has occurred, and progress has been and will continue to be made. A similar reassurance is felt with the more central role accorded to construct validity evidence in all areas of testing, the attention given to problems with minimum competency testing and the setting of standards, and the development of adaptive testing in relation to its needed theoretical underpinnings. The influence of cognitive psychology on testing has also been beneficial and holds important promise for the future. All in all, psychometric theory and knowledge seem to be active, developing, productive enterprises that will continue to furnish strong and supportive bases for the technology of testing and the wise selection and effective use of tests. The foundation is promising; whether its promise will be paralleled by equal promise in the technology or applied science it supports, or in the intelligent utilization of that technology by its consumers, is the critical question to which we now turn.

TEST TECHNOLOGY AND THE CHALLENGE OF IMPLEMENTATION

In comparison with the relatively strong showing of psychometric theory and knowledge, the application of that theory and knowledge to the development and validation of commercially published tests has produced mixed results at best. In
The Seventh Mental Measurements Yearbook (1972) and again in The Eighth Mental Measurements Yearbook (1978), Oscar Buros, after describing the “crusading” or “missionary” objectives of the Yearbooks, complained that:

Our success in attaining the last five missionary objectives has been disappointingly modest. Test publishers continue to market tests which do not begin to meet the standards of the rank and file of MMY and journal reviewers. At least half of the tests currently on the market should never have been published [Buros, 1972, p. XXVII; 1978, p. XXXI].

These are harsh words; yet as one who has followed Oscar Buros as Institute director and editor of the Yearbooks, it is difficult to find fault with his statement even now. The situation is a curious mixture of positives and negatives. On the one hand, there is little doubt that some of the major test publishers employ extremely able measurement specialists who have had much impact, for example, on translating new developments like latent-trait theory into practice in the construction of new tests. On the other hand, there is much of the cottage industry ambience to the test publishing business, and there are many test publishers who are simply test authors distributing their own tests or very small test publishers with single or extremely limited test offerings or book or instructional materials publishers who have acquired a few tests and publish them in a manner almost incidental to their major interest and thrust. Of the 496 test publishers that are listed in Tests in Print II (Buros, 1974), it is startling to discover that over one-half, or 58%, have only a single test listed; 75% have three or fewer tests listed; and 85% have five or fewer tests listed. The 58% who have but one test listed account for only 11% of the tests published. The 85% who have one to five tests listed account for only 16% of the tests published. Although Buros may have missed tests published by some companies, the Buros reputation for accuracy cannot be denied, and the overall impression is doubtlessly correct. On the other end of the continuum, where the large test companies predominate, a mere 1.4% of the publishers are responsible for publishing 26% of the tests! Teachers of measurement looking for strikingly skewed distributions need look no further. With a publishing field as skewed and fragmented as this, there is little wonder that Oscar Buros often despaired about the likelihood of improved quality control.

Quality Issues in Test Publishing

Limitations of size and resources are quite likely to influence quality control despite the efforts of a small test entrepreneur to meet or exceed minimal standards and produce a professional product. One president of a small operation lamented that:

We are a very small cooperative venture with quite limited resources. For this reason we have as yet not been able to move to a professional finish on the
However, in spite of typos and home-grown typing each of the rough drafts gives ample information to permit an assessment of the instruments. They have continued to prove themselves in actual use. For this reason I am forwarding additional copies of the forms and manuals (rough or otherwise). None of the manuals are “finished.” We will revise them as information and funds permit.

This is an instance where the spirit is willing but the funds are weak. There are other instances where the markedly skewed distribution of sizes and resources of test publishers reported earlier seem to be accompanied by a parallel marked skewness in the demonstration of psychometric savvy. The president of one test publishing company, after expressing considerable resistance to our request for complimentary test materials for review, stated that the company:

was highly critical about present methods used for determining the reliability and validity of a psychometric tool. For example, often the concept of concurrent validity is used to determine if a particular test is a valuable tool. Actually what this means is that one or the other tool is unnecessary because they are virtually measuring the same thing. If the correlation is not significant, we know that we are measuring some aspect of behavior not currently being tapped. Buros, however, chose to use this lack of correlation as a reason to reject or criticize a test.

Aside from the fact that Buros let the reviews and reviewers speak for themselves, the statement contains much that would cause concern if not apoplexy among contemporary measurement specialists. Another company divides its tests into those that have validity evidence and those that do not. One wonders what kind of reassurance this provides to its clients!

Some Evidence on Test Quality
If we move from the level of specific examples to the more generic, it is regretfully true that there are still a surprising number of tests that are published without reliability evidence, validity evidence, or norms. When this occurs, it has been and will continue to be the practice of the Buros Institute to point out this critical lack in the descriptive entry accompanying the reviews in the Mental Measurements Yearbook. A small descriptive study was recently conducted by Institute personnel to determine how often these critical data were lacking. The results are not encouraging. They showed that 22% of the tests listed in The Eighth Mental Measurements Yearbook (1978) were without any reliability data whatever; 8½% had no validity data whatever; 7% had neither reliability nor validity data; and an additional 1% had neither reliability nor validity data for certain parts, levels, or editions. Another 5% had no reliability data for certain scores, and 9% had no reliability data for certain grades, subtests, or forms. All together, some 41% of the tests listed in The Eighth Mental Measurements Yearbook were lacking reliability and/or validity data in some important respect.
Tests in the areas of reading, vocations, and speech and hearing were the worst offenders.

The data for norms were somewhat better but still not encouraging. Of the tests listed in *The Eighth Mental Measurement Yearbook*, 11% had no norms whatever. Another 3% had no norms for certain scores, and 8% had norms only for certain subtests, forms, or parts of the standardization population. One percent had no description of the normative population, and for 4 percent the norms consisted only of means and standard deviations. All told, some 28% of the tests listed in *The Eighth Mental Measurements Yearbook* were inadequately normed in some important respect.

It should not be concluded that the 41% of tests lacking in validity and/or reliability data or the 28% lacking in normative data were the result of very rigorous criteria applied by the Buros Institute. As a matter of fact, any kind of correlation coefficient would usually serve to remove the accusing statement for either reliability or validity, and the situation for normative data was equally charitable. The standards for declaring such inadequacies in the descriptive entries were minimal at best, and still many of the tests listed in the 8th *MMY* made an unhappy showing. If 41% of the tests listed in the 8th *MMY* were lacking in validity and/or reliability data and 28% were lacking in normative data, was Oscar Buros far wrong in asserting that at least half of the tests currently on the market should never have been published?

**Some Affirmations**

To consider the implementation of test theory and knowledge in actual test products is a frustrating exercise in the reconciliation of opposites. On the one hand, one observes the amazing rapidity with which a complex development like latent-trait theory has been seized by the test constructor and incorporated into instruments like the British Ability Scales; on the other hand, one observes 41% of the tests in the 8th *MMY* lacking in the simplest kinds of reliability and validity data. Test manuals seem to be improving and more technical manuals are being offered, many of them well conceived and executed; yet there are still commercially published tests that have no manual, an inadequate manual, or instructions for administration masquerading as a manual. *American Psychologist* (Glaser & Bond, 1981) issued a special edition on testing that provides abundant evidence of continuing progress and sophistication in the field of measurement and its application; yet there are some reading and personality tests and diagnostic inventories whose authors appear never to have seen the inside of an elementary measurement text. Because of the makeup of the testing industry, such contradictions are likely to exist for the foreseeable future.

In the face of such contradictions one could argue a good case for applying some minimum competency criteria to the testing industry itself! In any event, it seems clear that the number of poor or marginal tests could be substantially reduced if a climate of opinion could be created for both test developers and users
that would ensure a severe fiscal disadvantage for the test author or publisher who did not meet certain minimal criteria. Specific problem areas are summarized below.

1. Proliferation of Tests. There is a finite amount of money that will be spent on tests, especially with current economic conditions and current attitudes toward testing. Under these circumstances we must do whatever we can in the future to ensure that it will be in the best interests of test authors and publishers, reputationally and fiscally, to publish far fewer tests but much better tests. This was the rallying cry of Oscar Buros for over 40 years, and the years have not diminished its truth or urgency. The proliferation of tests continues unabated, however, and the best defense seems to be that of educating people to be more discriminating test users. Obviously the Institute of Mental Measurements has a critical role here and so do the teachers of measurement. But the amount of money still spent on poor and marginal tests, and the startling amounts of money acquired from the sale of such tests, suggest that we are probably losing ground rather than gaining.

2. Missing Reliability Information. The fact that 22% of the tests in the 8th MMY were without reliability data is alarming and absolutely without justification. We have to find better ways to prevent or discourage a test author or publisher from publishing and accepting payment for an instrument that suffers from such a basic deficiency. Consumer protection for a gullible testing public is far behind consumer protection in other areas.

3. Inadequate Validity Evidence. It was reported earlier that some tests are published without any validity evidence. More often, however, validity evidence is insufficient and flimsy and offered more as a ritual than to make a firm case. We have reached a point in measurement where many measurement specialists feel that all or most validity evidence is properly subsumed under the concept of construct validity. The determination of construct validity requires the marshaling of a comprehensive and integrated set of evidence that is no less demanding than the scientific method itself. We should increasingly insist that test authors and publishers meet these more comprehensive criteria of validity evidence. There is a long way to go from flimsy, halfhearted evidence offered as ritual to construct validity evidence meeting the basic tenets of construct definition and validation in scientific method. This further requirement, however, could be very beneficial in encouraging improvement in the quality of commercially published tests and further reducing the number of poor and marginal tests.

4. Publishers' Claims vis-à-vis Validity Evidence. Measurement professionals should increasingly insist that test authors and publishers bring test validity and putative test benefits into a more reasonable relationship with one an-
other. Often it seems that modest to weak validity evidence is offered but is somehow shunted aside into insignificance by an attitude and aura that implies far more benefits emanating from the test than is justified by the evidence. Many examples could be offered, but a case in point is the Common Examinations of the National Teacher Examinations (NTE). A review of seven studies relating the Weighted Common Examination score with ratings given by principals and supervisors during the first year of teaching revealed a median correlation of .11. Although attenuation could be a factor here, particularly with respect to the criterion, the evidence is hardly encouraging. But the publisher can and does maintain that the NTE is a measure of academic preparation only, and thus the validity issue can be at least partially sidestepped. The public most likely assumes that effective teaching is a simple function of knowledge attained, cares and understands little about the technical aspects of validity issues, and thus uncritically accepts the NTE into its belief system as a guardian of teaching standards. The practical result is that 50% of U.S. teachers college graduates took the NTE in 1980–1981 and nine states now use the NTE as part of the teacher certification process. An overstatement of test benefits, either explicit or implicit, in the face of weak evidence and a public inclination to believe, will not serve us well at a time when test critics are mounting new and more knowledgeable attacks on the industry and the profession. The tendency to promote test utility despite weak validity evidence is surely an obstacle to better understanding and another potential source of public backlash as well.

THE BOTTOM LINE: THE SELECTION AND EFFECTIVE USE OF TESTS

If the application of test theory and knowledge to the development and validation of commercially published tests has produced some mixed results, the actual use of tests in practical settings has departed even further from the ideal. In the Introduction to The Eighth Mental Measurements Yearbook, Oscar Buros (1978) defined five objectives of the Yearbook, in his own inimitable manner, as his "crusading" objectives. The three crusading objectives that related to users of tests were as follows:

1. To foster in test users a greater awareness of both the values and limitations involved in the use of standardized tests.
2. To suggest more discerning methods to test users of arriving at their own appraisals of tests in light of their particular values and needs.
3. To make test users aware of the importance of being suspicious of all tests—even those produced by well-known authors and publishers—which are not accompanied by detailed data on their construction, validation, uses, and limitations [p. XXXI].
As reported earlier, Buros felt that his success in attaining all his crusading objectives, including these three, was "disappointingly modest." It could be of some use now to take each of these objectives and see what they highlight with respect to current standards and practices of test usage.

In relation to the first objective, what can be said about the level of awareness of the rank-and-file test user about the values and limitations of current standardized tests? Buros felt that we have gone through too many periods of "unwarranted optimism" about standardized tests (Buros, 1978, p. 1973). Although some segments of the public may have unwarranted optimism and a lack of appreciation about the limitations of standardized tests, there is some recent evidence that this is not true of teachers and administrators in the public schools.

In a study reported by Salmon-Cox (1981), it was found that teachers, when questioned about how they assessed the progress of their students, most frequently mentioned "observation" as their principal tool. Test scores served a merely confirmatory role to observation; a child's classroom performance, as observed, was given more credence than a test score. In another report in the same series, Resnick (1981) summarized the Salmon-Cox results by suggesting that: "Tests are, quite simply, a natural feature of the U.S. educational environment; it appears that teachers and administrators have adjusted to their presence, neither desiring much benefit from them nor suffering much distress as a result of them [p. 624]."

This certainly seems to suggest rather strongly that teachers are not overly impressed with standardized tests or ignorant of their limitations. They may even be hard pressed to appreciate their values. Unwarranted optimism about tests surely exists, but it is not likely to be found in the rank and file of teachers who must administer the tests and interpret the scores.

In relation to the second Buros objective, what can we say about the methods test users employ in their appraisals of tests? It is difficult to find helpful or trustworthy data on this question, but it seems safe to say that there has been little improvement in the sophistication of methods used to select tests. Perhaps there is a more general understanding of how achievement test objectives and content should match curriculum objectives and content, and perhaps some large school districts with testing offices use the more "discerning methods" referred to by Buros. But despite all the efforts of teachers of measurement and the Buros Institute, test appraisal and selection in the field has still far to go before it becomes the cautious, systematic, methodologically sound process that measurement specialists want it to be.

In relation to the third Buros objective, concerned with the "suspicious" attitudes test users should have in the absence of data on test construction, validation, uses, and limitations, the best available evidence seems to indicate that many test users may not be interested enough to be suspicious. This conclusion, obviously, is quite congruent with the Resnick (1981) quotation reported earlier. If they are interested enough to exercise some careful judgment or show...
some suspicion, that judgment or suspicion seems quickly allayed by the cosmetic assurance of face validity evidence that seems compelling to many who have the strong will to believe in the absence of substantive evidence. "If it looks good, use it," is not a consciously palatable slogan to most people, but it must be an unconscious determinant for many people in the selection of tests or we wouldn't observe so many poor tests being purchased. At the Buros Institute we are continually amazed at how much money a poor test can make. For example, we received word some months ago that one such test, with little to commend it, was responsible for sales amounting to 5 million dollars in 2 years.

Perhaps part of the problem here is that the criteria for determining whether a test is useful or not are all bound up with that esoteric body of thought called psychometric theory, which is available and valued by the specialist but seems downright forbidding and scary for those uninitiated or of uncertain understanding. If a person looks for a new car, the criteria for what constitutes a good car are reasonably within reach and understandable. For tests those criteria are enmeshed in a scientism that for some people might as well be mysticism, with a jargon that seems sufficiently repelling to some to justify ignoring it. Is it any wonder, then, that it is the face validity features of a test that can so often commend the test's use to a potential purchaser and just as often mislead that purchaser after use to believe that the test did in fact yield the results desired?

Perceptions of the General Public about Tests

It is probably in that vast body called the general public where the threat of misunderstanding about tests is greatest and where a little suspicion, or at least a questioning attitude, might be a good thing. Resnick (1981) reports on a 1979 Gallup Poll that indicated that 81% of those polled thought that standardized tests were "useful" or "somewhat useful," with only 17% thinking they were "not too useful." Yet it is this same general public that is likely to be least informed and most confused about testing. Such confusion, lack of information, or evident misinformation has become a critical factor with such issues as bias in testing, minimum competency testing, and evaluation of the public schools. A vague conviction that something is useful combined with a lack of specific understanding about its most appropriate uses and interpretations and no conception of its limitations is a recipe for social disaster. Testing in the public domain has become such a social disaster. One feels it keenly when called upon, as I have been, to participate in briefings to the public about the proper uses and the limitations of tests and testing. One feels it keenly again when two federal district judges in California and Illinois reach diametrically different judgments about whether standardized intelligence tests discriminate against black children, with little evidence that either one of them had an adequate understanding, or cared to obtain such an understanding, of the psychometric issues involved (Larry P. v. Riles, 1979; PASE v. Hannan, 1980).
It is clear that the opportunity for the general public to raise its level of understanding about testing is even more limited than it is for public school personnel or people in business and industry. As professional people with both a moral as well as professional responsibility for our field, I do not believe we can ignore the public’s need for greater understanding of testing without even graver social consequences in the future. If continuing education and lifelong learning are to be as important as some higher education specialists think, I suggest that we do our part to ensure that increased understanding about tests and testing is promoted as a critical component of such lifelong learning. How that is to be done is an issue that deserves the very careful consideration of every person in measurement.

Vocational Tests for Business and Industry

An area of special concern about test usage is the area of vocational tests for business and industry. Recently the Buros Institute conducted a study on who purchased the Mental Measurement Yearbooks, and we were surprised to find that the group that purchased the most yearbooks was not education but business and industry, which accounted for almost half of the yearbooks sold. We are grateful for that, because it has often appeared to us that it is tests for business and industry, among all others, that are most likely to be promoted with very strong promises in the face of little or no evidence that the tests can deliver on those promises. Such ambitious and poorly substantiated claims sorely need the antidote that critical reviews from the Yearbooks can provide. Many tests in business, particularly those in the management area, involve elaborate conceptual schemes, sometimes associated with training programs, that are magnificent in their aspirations and complexity and attractiveness to would-be true believers. Such conceptual schemes would constitute ideal settings for obtaining construct validity evidence, but you can bet your entrepreneurial dollar that there is little effort to do that in the great majority of cases. It would likely prove too embarrassing. What happens instead is that these tests for business and industry are among the most serious offenders when it comes to the simplest kinds of validity evidence, let alone construct validity evidence, and we have found that 57% of the “Vocations” tests listed in The Eighth Mental Measurements Yearbook were lacking reliability and/or validity data in some way that was important for test use.

Test use in business and industry, of course, is coming under the increasingly heavy fire directed toward tests in general. As a result of this double vulnerability stemming from inadequate psychometric evidence and potential criticism or even litigation, some test publishers show resistance to providing the Buros Institute with the complimentary copies of tests needed for review purposes. Fortunately, they remain a distinct minority. One test publisher, for example, was reluctant to provide complimentary copies of his tests for fear that the reviews of these tests
“‘might be used as ‘authoritative’ evidence in a discrimination suit.’” He then went on to say that:

No test’s technical report is so comprehensive or so perfect that it cannot be adversely criticized.

It seems to me that we have an obligation to our test users to avoid providing plaintiffs with ready-made attack weapons which appear to have the prestige of the Buros Institute behind them.

What interesting questions this raises, especially in relation to the concerns with “obligation” raised in this letter. In a recent article on professional standards in testing Novick (1981) pointed out that “There are generally three participants in the ability testing process: the institution, or test user, which requires the test for some decision-making purpose; the test producer, which develops, markets, and/or administers and scores the test; and the test taker, who takes the test by choice, direction, or necessity [p. 1035].” Any reasonable set of professional standards would have to take into consideration issues of obligation to all three of these parties, but particularly to the test taker, who is still the least powerful of the three. The Buros Institute has an obligation to be fair to all three parties involved while providing consumer protection to the test user and the test taker. Although no test is so perfect that it cannot be criticized, it is only the nonexistence or glaring inadequacy of reliability or validity data that can furnish the ready-made attack weapons referred to in this letter, and under such circumstances it is the test producer, not the Buros Institute, that has fashioned the weapons and handed them over to the attacker. The best defense for the test user is to select tests that are well-constructed and validated and that can stand the light of day and not to rely on test companies that have an understandable but misplaced motivation to protect the user from test inadequacies that would be avoided altogether by not using the test.

Test Advertising

A very great influence on test selection and usage is test advertising, and it is test advertising that constitutes one of the greatest current concerns of the Buros Institute. It was reported earlier that Oscar Buros was concerned about “unwarranted optimism” about tests; it is in test advertising that “unwarranted optimism” reaches its peak. Good and poor tests alike are subjected to advertising claims that cannot be substantiated. The influence of such advertising is considerable, and the situation now is no different than it was in 1968 when Oscar Buros, in a presentation to the Association for Measurement and Evaluation in Guidance, reported the following:

At present, no matter how poor a test may be, if it is nicely packaged and if it promises to do all sorts of things which no test can do, the test will find many
gullible buyers. When we initiated critical test reviewing in *The 1938 Yearbook*, we had no idea how difficult it would be to discourage the use of poorly constructed tests of unknown validity. Even the better informed test users who finally become convinced that a widely used test had no validity after all are likely to rush to use a new instrument which promises far more than any good test can possibly deliver [p. 94].

The appeals to gullible buyers still ring loud and clear. A diagnostic-prescriptive reading program is described as “so effective a system that it’s been known to actually improve reading level by one year in only 11 to 12 one hour lessons!” A personality inventory is described as “the quintessential assessment tool for the 80s and beyond—the wave of the future among diagnostic instruments.” The same kind of extravagant advertising mania also affects scoring and interpretive services. A reviewer of several of the scoring and interpretive services for the Minnesota Multiphasic Personality Inventory expressed his strong concern about the advertising for these services in the following excerpted comments from *The Eighth Mental Measurements Yearbook* (Adair, 1978):

In reviewing the several scoring services for this yearbook, the writer was impressed with a curious dichotomy that appears to exist between the professional psychologist who is obliged to uphold the ethics of the profession and the entrepreneurial psychologist who is obliged to make a profit in order to maintain a position in the market. . . . The dilemma of whether to uphold professional ethics or to make a profit is seen most vividly in the promotional literature of the several services. . . . The literature of promotion takes on a Madison Avenue-like quality where caveats are included in the fine print [p. 940].

Examples could be multiplied endlessly. The sins of advertising claims are so numerous that the Institute may well consider sending out test advertising to be reviewed critically right along with the tests themselves. The issue of extravagant and unfounded test advertising claims must receive much greater attention in the next revision of the *Standards for Educational and Psychological Tests*. In the face of such claims the major agents for consumer education and protection are the *Standards*, Buros Institute publications, and a few beleaguered measurement teachers. In terms of current standards of test selection and use and the continued gullibility of the test-buying public in relation to extravagant test advertising claims, even the best efforts of all of these are apparently not enough to change the situation as much as it desperately needs to be changed.

**CONCLUSIONS AND RECOMMENDATIONS**

This chapter has been quite different from the others with which it appears because of its concern with the interrelationships and current status and development of the entire measurement continuum as it encompasses knowledge, implementation, and use. The latter two elements are the historic concerns of the
Buros Institute. The conclusion that seems apparent from the evidence discussed is that the theory and knowledge base of measurement is strong and evolving, the implementation of that knowledge base in developed products has brought tremendous variety and very mixed results, but that the selection, use, and interpretation of tests has been fraught with major difficulties and some unfortunate social consequences. It is my strong conviction that although professionals in measurement are usually most identified with the first or possibly second element of this measurement continuum, they have a strong professional obligation to be alert to and to join with others to take action against the continuing serious offenses and mistakes that take place through ignorance at the level of test usage. Professional support for the *Standards for Educational and Psychological Tests*, especially as those *Standards* relate to test use, is one example of responsible professional concern and action in this area. But in view of the extent of the abuses and the strength of the need, it is not enough. The Buros Institute and the *Standards* cannot do it alone.

**A Call to Action**

What, then, can be done to stimulate substantial improvement in the selection, use, and interpretation of tests (and perhaps, as a consequence, make it unprofitable to publish poor tests)? I submit that it will require nothing less than an organized campaign, launched and sponsored by NCME or the same consortium that produced the *Standards*, that would increase substantially the public understanding about testing concepts; the values and limitations of tests; and the selection, use, and interpretation of tests. Perhaps some funds could be obtained, most likely from private philanthropies in this day and age, that could help support such a campaign. Of what would such a campaign consist? The following are illustrative:

1. *Convention Programming.* In our professional conventions (NCME, AERA, APA, etc.) there should be more discussion of what practical steps could be taken to improve the selection, use, and interpretation of tests. Symposia could be organized on the topic. Although the 1980 NCME meeting featured some useful examples of this kind of programing (Beck & Stetz, 1980; Crocker, 1980; Yeh & Herman, 1980), generally there is far too little of this done at the present time. Practitioners often feel isolated at professional conventions. What a fine opportunity this might provide for greater dialogue among the theory and knowledge oriented and the practitioners. Benefits could be twofold: the development of ideas for improved test usage and the increased recognition by participants of their responsibility for what happens in testing at the grass-roots level.

2. *Education of the Public.* There is much talk these days about how the U.S. population is changing, how people are developing new careers and interests, and how there is more need than ever before for the implementation of a
Continuing education has become an important topic and need. Why shouldn’t increased understanding of measurement concepts, tests, and testing be considered an important component of continuing education or a lifelong learning program—or indeed—for citizenship education itself? As a part of such continuing education the following kinds of projects might be implemented:

a. Public television could sponsor a series of TV programs on measurement concepts and contemporary testing practices. To stimulate interest some dispassionate discussion of contemporary testing issues and controversies could be intermixed with the foundational learning of concepts. The success of the program, “Who’s Keeping Score?”, which included parts of NIE’s Minimum Competency Clarification Hearing, suggests that much more could be done with the media to promote greater understanding of testing in the general public. Further prospects should be actively explored.

b. Many continuing education programs offer “minicourses,” typically with continuing education credit, that are designed to accomplish short-term objectives focused on the development of basic understandings, skills, or interests. Why shouldn’t measurement people develop and offer such short courses not only to principals and teachers but also to the general public? When a local or national testing controversy develops, why shouldn’t minicourses be developed to help the public better understand the real issues involved and the knowledge bases for intelligent decision making?

c. Perhaps the Buros Institute should develop a short pamphlet describing useful procedures and criteria for selecting a test and using and interpreting it properly. Such a pamphlet could be sold to the public at minimal cost and could also be included in the introduction to The Mental Measurements Yearbook. This might also be a useful project for NCME. A pamphlet of this kind would have to be much shorter and more readable than the Standards for Educational and Psychological Testing, which is tedious and forbidding reading at best.

d. As suggested earlier, it may well be appropriate for the Buros Institute to send out advertising as well as the test itself for review. Some bad press in response to extravagant claims might at least temper those claims and motivate those involved toward more recognition of their responsibility for their advertising as well as for their product. “Truth-in-packaging” is a desperate need in testing.

e. More ways should be found to reward and reinforce those test authors and publishers whose products represent high standards of construction and validation. The professional organizations provide this kind of recognition for researchers; why shouldn’t test authors and publishers receive a parallel form of professional recognition? The development of a good test is a very difficult and painstaking process, and its achievement should be professionally acknowledged. The Buros Institute would like to participate in a program with such an emphasis on the positive. Perhaps our reviewers could nominate tests that they
judged to be exceptional exemplars of test construction and validation, and members of our National Advisory Committee could select one or two tests from each area deserving of special commendation. Or perhaps the professional organizations would wish to provide such recognition. There are possible pitfalls in such an undertaking, of course, but a few minor risks may have to be accepted in order to accomplish what is considered just and motivating for test authors and publishers and beneficial for the field and for test users.

f. There are strong professionals in the test publishing organizations, and they are doubtlessly professionally and personally interested in being part of an organization that subscribes to the highest standards of test authorship and publication. Individuals who join professional associations are often subject to a collective code of ethics promulgated by the association. Perhaps an association like NCME should have institutional as well as individual memberships, and both individuals and organizations should be subject to such a code of ethics. Part of that code could cover professional responsibilities relevant to test development, validation, and advertising. A test publishing company that joined the professional association would have to make a written and signed commitment to the code of ethics and could indicate in its advertising that it had done so. But if any members of the professional association, or a duly constituted professional ethics committee, uncovered evidence of code violation by a test publishing organization, constitutionally defined steps could be taken to conduct a hearing in accordance with rules of evidence and ultimately, if necessary, to take action ranging from mild reprimand to ouster from the professional association. This too is a rather radical suggestion, but the epidemiology of the disease seems to require radical cures.

Scientists of any kind, whether they be natural, physical, or social scientists, are increasingly being called upon to recognize the moral and ethical implications of their work. Yet there is a tendency for many professionals in measurement to focus on the theoretical and knowledge bases of their field and to lose sight of what is going on at the levels of implementation and use. It is our business at the Buros Institute to be aware—and sometimes painfully aware—of what is going on at these levels. We recommend that other measurement professionals and social scientists direct more attention to such grassroots issues, encourage their wider discussion, and join with the Buros Institute of Mental Measurements in seeking more effective solutions to these problems than we have ever had in the past. The social utility and reputation of a professional field may hang in the balance.

REFERENCES


