Measurement as a Dangerous Activity

Rebecca J. McCauley
Department of Communication Sciences and Disorders
University of Vermont

Preface

The following article appeared in the Spring, 1988 issue of Hearsay *(pp 6-9), the journal of the Ohio Speech and Hearing Association. It is being reprinted with the permission of the author and the editor of* Hearsay. *The commentary that follows the article was not part of the original publication but was solicited by the* JSLPA/ROA *editorial staff for this issue. Appreciation is extended to Wayne Secord, editor of* Hearsay, *and Rebecca McCauley for allowing us to reproduce this most interesting and provocative article.*

As an experimental psychologist turned speech-language pathologist, measurement has been a matter of continuing fascination to me. When I began clinical studies in speech-language pathology, I moved from a scientific environment in which many observations were used to generate tentative conclusions. Thus, 30,000 reaction-time measures of same-different judgments to pairs of American English vowels by 20 normal adults allowed me to make a very minor contribution to the existing "facts" of speech perception. Upon beginning clinical training in communication disorders, I entered a new scientific environment in which behavioral measures (which at the time appeared less exhaustive than my experimental measures) were applied to participants who were sometimes far less cooperative than those who served in my dissertation experiment. Yet the resulting measurements frequently served as the basis for extensive statements regarding a client's communicative competence. I was intensely interested in the specific characteristics of these new measures and in their powerful applications.

In this short space, I will share some of what I have learned about tests and other measures in speech-language pathology and some of what I have learned about test use after having spent time in the role of clinician and test user. Specifically, I will explain why I have come to believe that (1) the measurement process involved in clinical activities is dangerous but inevitable; (2) neither test developers nor users are solely responsible for these dangers; and (3) there are at least four steps that may help minimize these dangers for clinicians and their clients. Although I draw my examples from the areas of language and learning disabilities, I believe that my comments generalize beyond the domain.

Measurement and Clinical Action

In the behavioral sciences, measurement has traditionally been defined as the systematic assignment of numbers or categories to behaviors as a means of representing properties of those behaviors or underlying properties of the individuals exhibiting the behaviors. There is a tendency to equate measurement with standardized tests, particularly norm-referenced standardized tests. However, other types of measures are in fact more common in clinical settings. For example, nonstandardized but commercially available instruments, informal clinician-developed measures, and even those invisible measures that presumably constitute input to clinical judgement come into play in almost all client-clinician interactions.

Clinically, measurement provides a basis for action: We use our understanding of an individual's behavior (obtained by measurement) to guide us in decisions at every step of clinical contact diagnosis, treatment, termination of treatment, and follow-up. Thus, at each step, the quality of clinical action depends upon the quality of measurement. It is this irrevocable connection between measurement and action that makes measurement dangerous. Furthermore, it is the professional demand that clinicians act that makes measurement and the potential problems it entails inevitable, as well as dangerous.

Measurement as Dangerous Activity

In measurement theory, the nature of "dangers" in measurement are typically discussed using the terms *validity* and *reliability*. Simply put, validity is the degree to which a measure reflects the property it is intended to reflect. Reliability is the degree to which a measure will be consistent when the time of measurement, the person doing the measuring, or some other variable changes. Each of these properties of a measure is not intrinsic to the measure itself but rather must be examined anew for each new purpose to which the measure is put. In the field of communication disorders, therefore, reliability and validity (as well as other important test characteristics) must be examined for each clinical decision. Specifically, an instrument may be quite valid and reliable when used to assess a particular behavior for one type of client (for example, a child from a wealthy family), yet be quite invalid and unreliable when used to assess the same behavior

for another type of client (for example, a child from a low-income family). A measure's adequacy, therefore, must be regarded as specific to a particular clinical decision for a particular client.

The most obvious and serious consequences of inadequate measurement are those that result in a client's receiving inappropriate or, at a minimum, inefficient clinical management. What specific kinds of outcomes can result from errors in measurement? At an initial assessment, for example, an individual may be inappropriately identified as normal in communication functioning or may have strengths and weaknesses misidentified. In treatment, a client's progress may be under- or overestimated. Finally, at the termination of treatment, continuing communication problems may be missed or underestimated. These are just some of the negative consequences that can result from errors in measurement.

Less obvious dangers from measurement may affect the individual clinician who makes the measurement error usually by using an unsuitable measure or by misusing an appropriate one (McCauley & Swisher, 1984b). Although wasting one's own time (as well as the client's) as a result of a flawed clinical decision is an unsavory prospect, still more negative personal repercussions exist. For one, there is the moral burden of failing to serve the client.

An additional threat to the individual clinician is the possibility of litigation undertaken because he or she makes an error in measurement during assessment or treatment of a client. Although this is not an imminent danger for clinicians, its seriousness demands attention. One example of such potential vulnerability taken from the area of language-learning disabilities is the use of tests lacking evidence of validity for handicapped populations. In a recent paper, Fuchs, Fuchs, Benowitz, and Barringer (1987) looked for evidence of validity for 27 tests used in special education, including receptive vocabulary tests and achievement tests used by speech-language pathologists. They observed that many of the tests provided "scant technical data" (p. 266), echoing similar previous criticisms (e.g., Berk, 1984; McCauley & Swisher, 1984a). In their paper, Fuchs and his colleagues noted the widely recognized obligation of test users to provide evidence of validity where it is lacking for a specific use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). They also warned about the potential for liability associated with administering tests when such information is lacking. Thus, penalties for errors in test use may potentially enter the legal as well as the clinical arena.

Although problems associated with measurement have grave implications for individual clients and clinicians, they also have very serious implications for the profession as a whole. Two types of dangers associated with measurement problems fall in this category: (1) threats to our stature as an emerging profession and (2) threats to our ability to conduct valid research on test-defined groups of communication-disordered individuals. The threats to our stature come from our failure as a profession to provide clinicians with sufficient training and guidance regarding measurement issues and practices. This failure is reflected, for example, in the failure of the American Speech-Language-Hearing Association either to develop specific guidelines regarding measurement or to adopt the Standards for Educational and Psychological Testing proposed by three professional groups with long histories of attention to measurement in the applied behavior sciences (American Educational Research Association et al., 1985). Work is clearly underway in the profession to remedy the current dearth of information regarding measurement. For example, there are increasing numbers of articles in our professional journals dealing with test development and use, and increasing continuing education presentations dealing with these topics. Nonetheless, more must be done if the field is to be recognized as well trained in measurement.

Dangers associated with problems in measurement also threaten efforts to advance knowledge of communication disorders. One obvious need is for adequate and comprehensive tools to use in defining specific clinical populations. This is particularly important in the identification and description of individuals with language disorders. Thus, a recurring theme among those researchers interested in identifying homogeneous subgroups of language learning disabled children is the need for adequate measures. For example, in a report to the U.S. Congress on learning disabilities, Tallal (1987) writes:

> if specific developmental language disorders are ever to be prevented, we must first begin by understanding their etiology (causes) and prevalence in the population and then develop appropriate treatments. This can never be accomplished if uniformly applied inclusionary and exclusionary criteria for diagnosis are not established and their use enforced. *This will rely ultimately on the development and uniform use of well-standardized, reliable and valid testing procedures.* [emphasis added] (p. 63)

Who's to blame for inadequacies in test development and use? On the face of it, it might seem that the responsibility for test development rests solely with test developers that is, test authors and publishers and that responsibility for test use rests solely with test users that is, clinicians and program administrators. Yet it isn't as simple as that. Paradoxically, although test users and developers share in these responsibilities to some extent, the ultimate responsibility for errors made by the test developer or user falls upon test users because

they are the final arbiters of what actions finally occur in response to instances of measurement.

Test Developers Aren't Solely to Blame

Ideally, test authors and publishers work together to accomplish the many steps involved in the preparation of a well-developed standardized test. These steps include all of the following: (1) developing a standard, scientifically based test content; (2) specifying procedures for administration, scoring, and interpreting the instrument; (3) conducting research designed to examine the measurement characteristics of the instrument for intended populations and purposes; and (4) documenting the results of that research in the test manual. These steps are as technically demanding as they are numerous. Therefore, it is not surprising that not all instruments are well-developed.

Errors made by test developers can be as blatant as failing to collect adequate information to demonstrate an instrument's adequacy for any purpose or using very small sample sizes in their ostensible "norms." Errors in test development can also be relatively subtle. Thus, for example, critically important pieces of technical information, such as the age and gender of normative groups, are much harder to find in some test manuals than are nebulous accolades for the instrument. Marketing, it appears, sometimes takes precedence over test development.

Yet taking the other party's point of view (a position almost always incompatible with blame-laying), there may be little to justify the huge expenditures of time and money required to produce even a moderately well-developed test. The ubiquity of poorly developed tests on clinic shelves surely argues against the consumer's equation of value with psychometric adequacy. As businesses, then, test developers may see no direct incentives from the test-using consumers to justify the truly awesome expenditures of time and money involved in the development of a standardized instrument for even one clinically relevant population much less for the many important subpopulations for which a test might be used. In short, test consumers must accept at least some of the blame for the quality of available instruments.

Test Users Aren't Solely to Blame Either

In order for a given measure to provide an adequate basis for clinical decision-making, clinicians must correctly perform a number of activities: (1) selecting the specific measurement tool to be used, (2) administering and scoring it, and (3) interpreting it in light of other available information. Because the clinician is in practice responsible for each of these steps, she or he is also responsible for the final adequacy of the measurement process. However, just as test developers aren't solely to blame, neither are individual clinicians solely to blame for errors in the measurement process.

As an example, let's consider an unavoidable error in measurement that arises during the selection of a measurement tool. When a clinician chooses a particular measurement tool be it a standardized test, informal measure, or clinical observation technique many factors affect the selection process: the appropriateness of the tool for the measurement purpose and for the client, the availability of the tool, even the cost of competing techniques in terms of money, administration time, and complexity of interpretation required. This mix of practical and theoretical considerations certainly presents a difficult task for a clinician. However, even if technical adequacy of a measurement tool is the only selection criterion, there are many content areas in which well-developed tools are simply not available and many for which there is controversy as to what would constitute a well-developed tool. Thus, for example, clinicians are largely at a loss for a comprehensive description of a given client's semantic system.

Additional, unavoidable errors in measurement may come from subtle psychological phenomena that obscure our appraisal of the quality of the measurements we make. Because the quality of clinical actions is irrevocably tied to the quality of measurement, clinicians (who are, of course, interested in helping their clients) need to invest a confidence in those measurements, which may at times be blinding. "I acted upon this measurement information; therefore, it must have been of high quality." Thus, good clinicians may nonetheless like bad measurement because it will appear to serve the required function it will provide a basis for clinical action. Unless the erroneous clinical decision is self-correcting (as treatment decisions sometimes are), the clinician may not have the opportunity to learn from past mistakes.

In fact, bad measurement may make clinical decision making appear easier, by painting reality as black or white rather than the frustrating shades of gray that so often constitute our current understanding of a behavior or trait. Thus, for example, it may appear to be easier to act solely on an overall test score than to consider the limitations of the test in terms of the specific kinds of tasks used. Similarly, it may appear to be easier to resort to an untried "clinical impression" than to learn more about the intricacies of measurement and tests. However, the dangers are great.

What Can be Done to Reduce the Dangers

Measurement is inherent in all clinical and research activities in which a decision must be made about a complex behavior. Thus, it is just as lively an issue in speech-language pathology as it is in experimental psychology. In clinical speech-language pathology, however, measurement is particularly in-

triguing because of its connection with clinical actions intended to help a communication-disordered individual. In response to the dangers of measurement described above, at least four steps can be taken. Some of these steps may appear to demand action that is beyond the grasp of any single clinician; however, in fact, each requires the support of all speech-language pathologists.

First, there is a need for greater professional and research attention to measurement of all types both formal and informal. The notion of clinical measurement as synonymous with norm-referenced standardized tests and clearly defined diagnostic sessions ignores the importance of measurement to every clinical activity. Instead, the continuity of measurement issues across all clinical, as well as research, pursuits needs to be recognized for progress to be made. Thus, for example, in language-learning disabilities, improvements in research **and** in clinical practice hinge upon the development of more comprehensive and adequate measurement tools.

Second, increased research into the actual processes by which measurements are made and used in clinical decision-making is needed. Most of what I have written here is based on my "sense" of what clinicians do as they make and use measurements, rather than on empirical data concerning what they actually do. Surely, increased attention to the clinical decision-making process and the data upon which it is based as a topic for research could add considerably to the quality of clinical practice. This type of research is flourishing in fields such as special education (e.g., Huebner, 1987), psychology (e.g., Dumont & Lecomte, 1987), and medicine (e.g., Boshuizen, Schmidt, & Coughlin, 1987). It should be aggressively pursued in ours as well.

Third, with regard to standardized tests, test consumers need to increase their constructive communications with test developers. At the very least, this could do much to improve the quality of information available for test users when they consider alternative measures. For example, something as simple as standardization of test manual format would facilitate objective comparisons of tests without entailing crippling expenses; at the same time, it may be an easy concession for test developers to grant. Certainly unified action of the profession as a whole will yield the largest possible benefits.

Fourth, and most importantly, individual clinicians can steadily work at improving their own knowledge and implementation of measurement principles. I have found that this is not an easy task, but a rewarding one. Presentations at national and state conventions and articles in *Language, Speech, and Hearing Services in Schools* and *Asha* as well as other ASHA journals are more and more frequently addressing these issues in ways that are accessible to nonexperts. Several of the sources cited in the reference list (e.g., Anastasi, 1976; Salvia & Ysseldyke, 1981) can also help provide a good beginning. Finally, discussion of specific instruments with fellow test users can provide valuable information and fuel continued interest.

In conclusion, although measurement is an inevitable and dangerous aspect of clinical practice, there is much that individual clinicians and the profession as a whole can and should do to reduce these dangers.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Berk, R. (1984). *Screening and identification of learning disabilities.* Springfield, IL: Charles C. Thomas.

Boshuizen, H., Schmidt, H., & Coughlin, L. (1987, April). *On-line representation of a clinical case and the development of expertise.* Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Dumont, F., & Lecomte, C. (1987). Inferential processes in clinical work: Inquiry into logical errors that affect diagnostic judgments. *Professional Psychology: Research and Practice, 18,* 433-438.

Fuchs, D., Fuchs, L., Benowitz, S., & Barringer, K. (1987). Norm-referenced tests: Are they valid for use with handicapped students? *Exceptional Children, 54,* 263-271.

Huebner, E.S. (1987). Teacher's special education decisions: Does test information make a difference? *Journal of Educational Research, 80,* 202-205.

McCauley, R., & Swisher, L. (1984a). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34-42.

McCauley, R., & Swisher, L. (1984b). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49,* 338-348.

Salvia, J., & Ysseldyke, J. (1981). *Assessment in special and remedial education* (2nd ed.). Boston: Houghton-Mifflin.

Tallal, P. (1987). Appendix on Developmental Language Disorders. In the Interagency Committee on Learning Disabilities, *Report to the U.S. Congress.*