Peer Commentary on 'Measurement as a Dangerous Activity' by Rebecca McCauley

Elizabeth Duncan-MacLeod
School of Audiology and Speech Sciences
University of British Columbia

Donald G. Doehring
School of Human Communication Disorders
McGill University

I applaud Rebecca McCauley's comments about the dangers of measurement in the field of speech-language pathology. While I agree that problems exist with both formal and informal measures used to assess all aspects of speech and language, standardized tests of language disorders in children present the greatest problems.

As a university instructor in speech-language pathology, I need to keep abreast of the latest in testing, and I am continually amazed at the number of new tests available each year in the area of child language disorders. In reviewing these tests, I often am appalled at their lack of sound theoretical base and their inadequate standardization. Even if one were to apply McCauley and Swisher's (1984) criteria for test selection, the most important consideration in evaluating a test would be overlooked–construct validity (Messick, 1980). The theoretical underpinnings of many language tests particularly those designed to assess the language abilities of school-aged children are at best shaky and often nonexistent. While test developers are becoming more aware of these issues and are taking greater responsibility for designing adequate measures (e.g., Test of Auditory Comprehension of Language, Revised Edition, Carrow-Woolfolk, 1985), inadequate tests continue to flood the market and eventually make their way to clinic shelves. While I agree with McCauley that there must be greater communication between test designers and test users, in the meantime, its up to individual clinicians to look critically at new tests as they come on the market and to evaluate carefully their theoretical constructs and standardization procedures before adopting them as part of an assessment battery.

Perhaps we need to place more emphasis on qualitative rather than quantitative measures when assessing language disorders in children. Time consuming as it is, language sample analysis may provide us with a more valid measure of linguistic competence and help to eliminate some of the "dangers" inherent in standardized testing.

E.D.M

Rebecca McCauley very ably describes the salient features of clinical measurement. Measurement is necessary for clinical decisions regarding diagnosis, treatment, termination, and follow up. It is dangerous because measurements may not be perfect. Imperfect measurement procedures can be detrimental to clients, clinicians, researchers, and the profession. Dangers of measurement can be reduced by uniform guidelines, the development of better measures, research into the process of clinical measurement, better communication between test users and developers, and improvement of clinicians' knowledge about the complexities of clinical measurement. The paper is well-reasoned and comprehensive. My comments are intended to extend the discussion in what I hope will be useful directions.

McCauley quite correctly does not restrict measurement to standardized tests, but she does not clearly differentiate between measurement and non-measurement. In defining measurement she includes "those invisible measures that constitute input to clinical judgment" (p. 6). In discussing bad measurement she says "it may appear to be easier to resort to an 'untried clinical impression'" (p. 8). I would be interested to know the exact distinction between invisible measures and untried clinical impressions.

One of the most important points made by McCauley is that the reliability and validity of a measure must be evaluated with reference to its specific application. In this connection, she says that clinicians must provide their own evidence of validity when it is lacking for a specific use. This may not be possible. An alternative to completely abandoning measurement might be to explain to clients and colleagues that this measure, while not yet validated, is the most promising measure currently available to the clinician.

My major comment has to do with validity. McCauley has drawn her examples from the areas of language and learning disabilities. For the measurement of these disabilities, the type of validity that is most relevant is construct validity (Anastasi, 1982). In the absence of an objective criterion such as central nervous system dysfunction, the measurement procedures must be based on theoretical constructs concerning language

and learning disabilities. Language measures are usually derived from well-established theories, but the theories keep changing. It is important for clinicians to be aware of new theories and of the need for new language tests as theories evolve. There is no generally-accepted theory of learning disability. The field is characterized by a changing array of competing explanations of what appear to be a variety of disorders. Speech-language clinicians might best measure only those aspects of learning disabilities that involve language.

Measurement procedures for speech disorders and peripheral hearing disorders tend to be less heavily reliant on construct validity, since there are generally-accepted objective criteria for these disorders. Measures of acquired central auditory disorders can be validated by objective evidence of lesions of the central auditory nervous system, but there are neither objective criteria nor adequate theoretical constructs for measures of developmental central auditory disorders. As part of clinicians' knowledge about measurement, they should be aware of the status of current theories of the disorders with which they are concerned.

McCauley states that test authors and publishers should work together to develop good measures. Although it is not her intention, readers might infer that the raw materials for test construction are readily at hand. There should be explicit recognition of the need for test authors to avail themselves of the best current theoretical constructs in developing new tests.

McCauley concludes that individual clinicians and the profession as a whole can do a lot to reduce the dangers of clinical measurement. I would identify a more specific cast of clinicians, researchers, theorists, test developers, and professional policy makers, many of whom play multiple roles, who should work together as closely as possible to reduce the dangers and increase the quality of measurement

D.D.

References:

Carrow-Woolfolk, E. (1985). *Test of Auditory Comprehension of Language*, Revised Edition. Allen, TX: DLM Teaching Resources.

McCauley, R. & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*, 338-348.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Reply to Duncan-MacLeod and Doehring

In their commentaries, Duncan-MacLeod and Doehring appear to agree with many of the ideas expressed in "Measurement as a dangerous activity." Moreover, they extend the arguments of that paper by exhorting clinicians to consider strong conceptual bases as central to test adequacy and by adding to my list of practical responses to the dangers associated with measurement. In his commentary, Doehring also seeks clarification on the scope of my definition of measurement. In this brief response, it is my intent to return the favors of friendly agreement and elaboration, and add, I hope, clarification as well.

A point that was regrettably understated in my paper, but cogently made by both Duncan-MacLeod and Doehring, is the preeminence of conceptual adequacy or construct validity for all tests, but particularly for tests in areas such as language and learning disabilities. Both authors point out that the theoretical bases of such instruments are critically important, but often sorely lacking. Doehring warns that the conceptual raw materials for the development of good measures in language and learning disabilities may not be easily identified because of the constant evolution of theories in this area. This major problem notwithstanding, he encourages test developers to use the best available constructs. I would add that test developers should be encouraged to do this even where theoretical volatility can soon make an instrument obsolete. I would also remind us that empirical data makes an important contribution to the validation process. The role of both theory and empirical data in test validation are spelled out quite clearly by Messick (1989, p. 13), who says "validity is an integrated evaluative judgment of the degree to which *empirical evidence and theoretical rationales* [emphasis mine] support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment."

In their respective commentaries, Duncan-MacLeod and Doehring add to my list of actions clinicians should take to reduce the dangers of measurement. For her part, Duncan-MacLeod suggests that a change in emphasis toward qualitative measures (e.g., language sample analysis) and away from quantitative measures (e.g., standardized tests) may be helpful in eliminating the "dangers" inherent in measurement. I agree that such measures have great potential. However, I suspect, as do others (e.g., Faust, 1986), that the rejection of measures with some known, but unsatisfying, degree of validity (the devils we know) and the acceptance of measures with unknown validity (the devils we don't know) may simply affect the measurer's awareness of the dangers and not the dangers themselves.

I can endorse the strategic responses to measurement dangers proposed by Doehring more completely. Doehring suggests that clinicians who cannot feasibly provide their own evidence of the validity of a measure's use for a specific application might explain its promise relative to other available measures, point out its limitations, and, I infer, thereby limit its potential for harm. Further, he suggests that the work of reducing the dangers and increasing the quality of measurement will require considerable work by all of the relevant parties–"clinicians, researchers, theorists, test developers, and professional policy makers." I heartily agree with both of these suggestions.

Finally, I will try to clarify the scope of the term *measurement*, as I use it. Traditional definitions of behavioral measurement (e.g., Lord & Novick, 1968) describe it as a systematic procedure in which numbers or categories are assigned to individuals in a way that reflects the real world relationships among the individuals with regard to the attributes being measured. Doehring suggests that by including "those invisible measures that constitute input to clinical measurement" or, similarly, "untried clinical impression(s)" in my consideration of measurement, I may have drawn the boundaries too large to be meaningful. Perhaps. My reason for doing so was to introduce the role of the human measurer more fully into our consideration of measurement. We use measurements as summaries of an individual's behavior or characteristics that may serve as input to decision making of one sort or another. Therefore, I would submit that, in order to fully understand both measurement (input) and clinical decision making (processing), we should rigorously study "measurements" more broadly defined to include the less systematic descriptive procedures used by human measurers. In summary, I would say that we must consider the tangled relationship of measurer and measurement, if we are to improve clinical practice. A broader definition of measurement is surely one way of achieving this end.

R.M.

References

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17,* 420-429.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement.* 3rd edition. New York: American Council on Education and Macmillan Publishing Company.