Two and One-Half Decades of Leadership in Measurement and Evaluation

BRUCE THOMPSON

This article explores the role of professional organizations, such as the American Association for Counseling and Development's (AACD) Association for Measurement and Evaluation in Counseling and Development (AMECD) division, and the possible futures of such groups. It is suggested that AMECD and similar groups are experiencing a similar set of social and professional pressures. Aspects of the missions of creating and disseminating measurement and evaluation expertise are discussed. It is suggested that the explosion in knowledge has created new tensions involving special challenges for members of AMECD and related groups. Nevertheless, some remarkable recent accomplishments can be readily identified, and are noted here.

The Association for Measurement and Evaluation in Counseling and Development (AMECD), a division of the American Association for Counseling and Development (AACD), celebrated its silver anniversary at the 1989 annual AACD meeting. Sheeley and Eberly (1985) provided a thorough review of the history of AMECD. As AMECD approaches its third decade of existence, now may be a good time to revisit the role of professional organizations such as AMECD and to look toward the possible futures of such groups.

AMECD essentially constitutes an interest division more than a discrete occupational specialty. The current membership in AMECD, as of December 1990, was 1,593. Thus, AMECD is the smallest AACD division. Current membership is greater than the total of roughly 1,400 that occurred in the early 1980s, but is appreciably less than the peak of 3,249 reached in 1968 (Sheeley & Eberly, 1985). This membership profile over the years is strikingly similar to that of the National Council on Measurement in Education (NCME) (Lehmann, 1990).

The stated purpose of AMECD is the following:

> to provide leadership, advice, and counsel on matters related to measurement and evaluation in counseling and development by (a) providing a forum for the discussion of ethical, social, and technical issues related to measurement and evaluation; (b) improving the standards of professional service in the field of measurement and evaluation; (c) promoting better communication between the producers of measurement and evaluation devices and consumers; (d) promoting greater technical competency in the use and interpretation of measurement and evaluation devices. (AMECD, 1986)

Thus, AMECD is "about the business" of both developing and disseminating knowledge involving measurement, evaluation, or assessment.

THE CONTEMPORARY MILIEU AND TECHNICAL GROUPS

It is interesting to note that self-study and reflection about role and mission are currently occurring in a variety of related groups, although of course such reflection has also occurred previously (Goldman, 1972; Prediger, 1980). Apparently, the social and professional forces influencing AMECD also have an impact on other groups involved in the creation and dissemination of technical expertise.

For example, Jaeger (1990) recently reflected on the salience of American Educational Research Association (AERA) Division D to its members, noting that "since so many Division D members also hold membership in some other AERA division, I have heard speculation that Division D is the primary professional home of few Association members, and the secondary affiliation of many" (pp. l–2). This view is remarkably similar to Sheeley and Eberly's (1985) position that AMECD "remains as the division within AACD with the largest number of members who indicate that their primary affiliation is in another division" (p. 437).

The ongoing self-study in many such groups is reflected in the recent editorial in an NCME outlet titled "NCME: Where Is It Going?"

> Given these wide-ranging and very variable uses of educational assessment, where should NCME fit in? Should the association speak to issues and practices at every level of the enterprise, or should it focus periodically only on particular issues at particular levels? (Nitko, 1990, p. 2)

And American Psychological Association (APA) Division 5 President Susan Embretson (1990) recently argued that "although the potential contributions of quantitative methods to psychology are greater than ever, there are significant reasons to believe that the methodological underpinnings of psychology are being eroded" (p. 1).

These treatments give the impression that various professional groups find themselves undergoing similar transitions. Mehrens (1990) offered empirical evidence that various groups are indeed somewhat related in their self-perceived missions. Mehrens, who is a past president both of NCME and of AMECD and a former member of AACD Council, surveyed 31 leaders from AMECD, NCME, APA Division 5, and AERA Divisions D

and H. He asked respondents to indicate the perceived relevance of 33 topics to their groups.

Mehrens (1990) found that all five groups ranked test fairness within their top-10 lists. One topic was in the top-10 lists of AMECD and three other organizations: computer-assisted instruction. One topic was in the top-10 lists of AMECD and two other organizations: standardized achievement testing. One topic was in the top-10 lists of AMECD and one other organization: aptitude testing. The remaining topics in the AMECD top 10 were not ranked as highly by the other four organizations.

Five topics were not in the top-10 lists of any of the five organizations. All the topics were technical in nature: factor analysis, mathematical models of behavior, needs assessment, norming procedures, and theories of intelligence.

SOME OF THE MANY RECENT AMECD ACCOMPLISHMENTS

An argument that AMECD and related groups are undergoing transition does not mean that AMECD has been unable to make important contributions in fulfilling its missions. Some of the many recent accomplishments warrant acknowledgment.

1. *Offering National Leadership Across Organizations.* Belying AMECD's somewhat small size and limited budgets, the division has had noteworthy impacts on national "joint committee" projects involving multiple organizations (e.g., AMECD, AACD, APA, AERA). Examples include the Joint Committee on Educational Evaluation projects on personnel evaluation (see Jay Millman's book review in the October 1989 issue of *Measurement and Evaluation in Counseling and Development)* and the Joint Committee on Testing Practices. AMECD members who have been prominent in these activities include (at risk of inadvertently leaving someone out) Esther Diamond, Jo-Ida Hansen, Alan Robertson, and John Stewart. These joint committee projects have had and will continue to have important impacts on professions and on society more generally.

2. *Offering National Leadership to the Profession.* AMECD has also had dramatic impacts on the counseling profession more specifically. As AMECD representative to the Council for Accreditation of Counseling and Related Educational Programs (CACREP), a corporate affiliate of AACD, Joe Kandor has had substantial impacts in his role as CACREP chair. Nicholas Vacc and Pat Elmore have made very visible contributions to the AACD Committee on Testing, and AMECD members were all justifiably proud when an AMECD past president, Jane Myers, was elected president of AACD for 1990-1991.

3. *Disseminating Skills.* In addition to representing members in the councils of the profession, AMECD has been active in sharing the knowledge and insights of our members.

- The AMECD journal, edited by William Schafer, has continued to be recognized for excellence. More than 850 libraries subscribe to the journal. The regularly scheduled computer software reviews and the "Methods, Plainly Speaking" articles presenting measurement topics in readily accessible terms (e.g., Baldwin, 1989; Fish, 1988; Webb, Rowley, & Shavelson, 1988)—and used as class handouts by some faculty—have both been especially popular.
- Joe Ciechalski has done an outstanding job in bringing *AMECD Newsnotes* to the membership. The newsletter he has edited is informative and timely. Among other noteworthy contents, the newsletter regularly includes the useful and considered test reviews compiled by the Committee on Screening and Career Guidance Instruments, chaired by Robert Bauernfeind.
- AMECD has repeatedly sponsored very popular training sessions at the AACD annual meeting. For example, the 1989 Professional Development Institute on the Myers-Briggs Type Indicator (MBTI) was the most popular Professional Development Institute (PDI) offered during the 3 days before the convention began, and the two MBTI PDIs in 1990 were the second and fifth most popular. Similar turnouts occurred for the 1991 AMECD PDI on the MBTI.

Three recent developments merit special attention. First, the AMECD board has begun to nominate annually members for the various AACD awards. In 1990 AMECD members were delighted when a past president, Jo-Ida Hansen, won the AACD Research Award for her impressive work on the Strong Interest Inventory.

Second, AMECD annually presents an Exemplary Practices Award. The 1991 winner was Frank Womer, for his outstanding contributions in bringing all sorts of folks together in the annual Michigan Testing Conference. Now, a generous anonymous donor has created a trust to generate interest to provide a cash stipend for the award winner, beginning in 1992. The trust has been named after Don Hood, in recognition of this AMECD past president's contributions to the federal court in Dallas with respect to desegregation. Don's combination of affability, dedication, and expertise represents the very best of what AMECD has to offer.

Finally, a host of AMECD members labored long and hard in preparing the Responsibilities of Users of Tests (RUST) policy statement, now adopted by AMECD and AACD, and published in the AACD *Guidepost.* Now, Chronicle Guidance (P.O. Box 1190, Morovia, NY 13118-1190) has generously begun disseminating this important statement as their publication *CGP Professional #P90-22 File 20.*

THE EXPLOSION IN KNOWLEDGE CREATION

All of these accomplishments have been realized during an explosion in knowledge about both theories of psychological processes and methods for testing elaborate models. The joint elaboration of both new theory and new methodology has led to the dawning of a new day that will change much of what we think and do. Examples of methodological developments include the evolution of generalizability theory (Eason, 1991; Webb, Rowley, & Shavelson, 1988), of latent trait measurement (McKinley & Mills, 1989), and of covariance structure analytic methods (Baldwin, 1989). These developments have led to new views of assessment (e.g., Legg & Algina, 1990). With respect to knowledge creation, Embretson (1990) noted, "Rapidly expanding developments in quantitative psychology have greatly increased the potential to test and estimate effects of complex psychological theories and applied hypotheses" (p. 1).

Although many new challenges involving the explosion of knowledge creation continue to confront professional groups such as AMECD, the dissemination of "old" insight itself remains daunting. To make this point concretely, let me point out five "old" insights that have not yet permeated contemporary practice. These are insights involved in evaluating measurement

reliability, in measuring validity, or in testing substantive hypotheses.

First, *an inappropriateemphasis continues to be placed on statistical significance testing.* As Meehl (1978) noted, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false" (p. 822). Thus, Hays (1981) noted the resulting inescapable conclusion that "virtually any study can be made to show significant results if one uses enough subjects" (p. 293).

Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of participants, then conduct a statistical test to evaluate whether there were a lot of participants, which the researchers already know because they collected the data and know they're tired. This tautology has created considerable damage regarding the cumulation of knowledge about counseling (Thompson, 1988).

The sine qua non of science is the cumulation of evidence across studies, but significance tests regrettably do not inform the researcher regarding the likelihood that results will replicate (Carver, 1978). The future will bring more use of computer-intensive resampling strategies that do inform this judgment, such as cross-validation or Tukey's jackknife procedures (Thompson, 1989).

Efron's "bootstrap" logic is particularly appealing. It involves randomly resampling participants into hundreds of different configurations—when results replicate over many different combinations of participants, the researcher has more confidence that the results will reoccur in future work. Lunneborg (1987) has offered an excellent first installment of microcomputer programs in the bootstrap rubric.

It has been refreshing to witness the recent cascade of articles on this issue appearing in the *American Psychologist,* with authors like Kupfersmid (1988), Rosnow and Rosenthal (1989), and Cohen (1990). But it remains to professional organizations to continue to disseminate the word about the limits of statistical significance testing.

Second, *the generalizability theory approach to evaluating measurement error continues to be too infrequently applied.* Eason (1991) and Webb et al. (1988) provided readable treatments of this important theory.

Generalizability theory helps us to see that tests are not reliable or unreliable; rather, data have these characteristics, albeit data generated on a given measure administered with a given protocol to given participants on given occasions. This is not just an issue of sloppy speaking—the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice.

As Rowley (1976) noted, "It needs to be established that an instrument itself is neither reliable nor unreliable" (p. 53). Sax (1980) explained:

> It is more accurate to talk about the reliability of measurements (data, scores, and observations) than the reliability of *tests* (questions, items, and other tasks). Tests cannot be stable or unstable, but observations can. Any reference to the "reliability of a test" should always be interpreted to mean the "reliability of measurements or observations [i.e., a particular set of data] derived from a test." (p. 261)

If an achievement "test" for 12th graders were "reliable," would it be reliable for 1st graders? Would the test be reliable for all 12th graders?

One important implication of the realization that reliability inures to data (rather than tests) is that reliability should generally be explored whenever data are collected. And we always need to thoughtfully and explicitly explore whether the data in hand were collected on a sample similar to the samples used in previous reliability studies with a given measure.

Generalizability theory also helps us to see that sources of measurement error may not overlap and also that they can interact to create additional measurement error. Classical reliability statistics (e.g., KR-20, coefficient alpha) are inherently limited in telling us about the types and magnitudes of measurement error sources. For example, if an internal consistency reliability estimate for data is .9, the test-retest reliability is .9, and the equivalent forms reliability is .9, can we be reasonably certain that the measurement protocol has yielded reliable data?

We can make such a presumption if and only if (a) the measured sources of error variance perfectly overlap and also if (b) the measured sources of error variance do not interact with each other to create additional sources of unreliability. Classical test theory does not readily evaluate these two independent premises. That is why generalizability theory (Eason, 1991) has come to be viewed as so useful.

Finally, generalizability theory allows us to compute different kinds of reliability coefficients for the different kinds of decisions we wish to make (Webb et al., 1988). For instance, if three participants all complete two forms of a state counseling licensure exam, and their scores on one form are 6, 7, and 10 and their scores on the other form are 3, 4, and 7, the classical equivalence form coefficient for these data is 1.0. If the passing score is 7, will *examinees* care which of the equivalent forms they take? The classical equivalent form's reliability coefficients says no. But this situation requires a reliability coefficient for an "absolute" (Eason, 1991) decision. Generalizability theory provides different reliability estimates for different applications, but classical theory does not.

Third, *multivariate methods continue to be too infrequently used in both measurement and substantive research.* There are two reasons why multivariate methods are so important. *Multivariate methods can be used to control the inflation of Type I "experimentwise" error rates.* As Huberty and Morris (1989) noted, "Whenever multiple statistical tests are carried out in inferential data analysis, there is a potential problem of 'probability pyramiding' " (p. 306). For example, if 14 dependent variables are tested in a single study and the variables are uncorrelated (or if 14 of the 15 omnibus hypotheses in a balanced four-way ANOVA are tested), the probability of making a Type I error somewhere in the study (called "experimentwise error") will be the following:

$$1 - (1 - .05)^{14} = 51.2\%$$

This is true even though each of the 14 tests is evaluated at the .05 alpha level (called the "test-wise" error rate), as explained by Fish (1988). The problem is that the researcher will know a Type I error is likely somewhere but will not know which of the statistically significant effects are real and which are illusory.

But paradoxically, although the use of several univariate tests in a single study can lead to too many hypotheses being spuriously rejected, as reflected in inflation of an experimentwise error rate, it is also possible that the failure to use multivariate methods can lead to a failure to identify statistically significant results that

actually exist. Fish (1988) and Maxwell (in press) both provide data sets illustrating this equally disturbing possibility, a possibility suggesting that multivariate methods are also often vital in behavioral research because *multivariate methods best honor the reality to which the researcher is purportedly trying to generalize.* Thompson (1986) noted that the reality about which most researchers wish to generalize is usually one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects" (p. 9). Because significance testing and error rates may not be the most important aspect of research practice, this is the primary reason why multivariate methods are so often vital.

Fourth, *stepwise methods* tend to be used too frequently and incorrectly. Snyder (1991) offered an excellent treatment of the three major problems with stepwise methods. As she explained, the computer packages incorrectly compute the denominator in their stepwise *F* tests. Also, stepwise methods are prone to distortions from sampling error. Finally, stepwise methods do not inform the researcher regarding the importance of variables (Huberty, 1989). For example, if a researcher has 10 predictor variables and variables A and B are entered on the first two steps, it is still perfectly possible for variables C and D to constitute the best predictor set of size two. In short, "A large proportion of the published results using this method probably present conclusions that are not supported by the data" (Cliff, 1987, pp. 120-121).

Fifth, *the inappropriate use of ANCOVA seems to be the rule rather than the exception.* As Keppel and Zedeck (1989) repeatedly and emphatically argued, ANCOVA is appropriate only for use in conjunction with randomly assigned groups (e.g., pp. 455, 456, 466, 478-479, 480). Keppel and Zedeck (1989) cogently explained why. They also repeatedly and emphatically pointed out the importance of the homogeneity of regression assumption. They noted:

> It is somewhat depressing to note that while all statistical methodology books continue to stress [would that this were true] the conclusion that ANCOVA should not be used in quasi-experimental designs, misapplications of the procedure are still committed and reported in the literature. (p. 482)

Anyone contemplating an ANCOVA should seriously consider the trenchant arguments made by Campbell and Erlebacher (1975).

THE FUTURE CHALLENGE CONFRONTING PROFESSIONAL GROUPS

Nitko (1990) suggested that today there is "an almost bewildering array of assessment usages at every level of the educational enterprise: Educational assessment results seep into everyone's minds" (p. 2). But, as the previous discussion suggests, the evolution of theory and methodology have made measurement, evaluation, and assessment increasingly technical.

Certainly, these advances can have very important benefits for our clients and our students. For example, we classically focus achievement tests at the mean ability of the group of examinees, so that measurement error will be minimized, on the average. But measurement error is not equally distributed throughout a normal curve of abilities. Rather, there is more measurement error in the two extremes of the distribution. Thus, a reliability coefficient of .9 does not mean that we have highly and equally reliable ability estimates for all the participants.

But we now live in a day when we can give different examinees different pools of items focused, on the average, at the ability of each individual examinee rather than the average ability of the group. We can do this efficiently with microcomputers. And we can do this in ways that scores will still be comparable with each other (see the April 1990 special issue of *Measurement and Evaluation in Counseling and Development).*

Thus, professional groups such as AMECD now live in a time when expertise has become increasingly technical. And the rate of growth in new insight is explosive. Even the most dedicated professionals find it virtually impossible to keep up with new (or even past) developments.

Some persons understandably come to seek respite in a rationalization that common sense is all that is required to be effective in practice. Unfortunately, this view denies clients the absolute best that the profession can offer, and that is what our clients deserve.

But it is increasingly incumbent on professional groups to communicate technical insight in ways that are as clear and as painless to master as possible. And it is also incumbent on professional groups to bring together both theoreticians and practitioners so that they can communicate, exchange views, and learn from each other. These challenges are not easy, but they are unavoidable if we are to act responsibly.

AMECD is well positioned to confront these challenges. The preponderance of the members have advanced degrees. There is heavy representation of counselors, of counselor educators, and of research and measurement specialists. Just as the AMECD journal publishes informative "Methods, Plainly Speaking" pieces, useful "In the Field" pieces (such as the valuable discussion on the MBTI by Mary McCaulley [1990], who personally worked with Isabel Myers) are also available to readers. Thus, AMECD provides a forum that allows people to talk and to face the challenges that lie ahead.

REFERENCES

Association for Measurement and Evaluation in Counseling and Development. (1986). *Bylaws of AMECD.* Washington, DC: Author.

Baldwin, B. (1989). A primer in the use and interpretation of structural equation models. *Measurement and Evaluation in Counseling and Development, 22,* 100-112.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 1, pp. 597-617). Newbury Park, CA: Sage.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review. 48* (3), 378-399.

Cliff, N. (1987). *Analyzing multivariate data.* San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45* (12), 1304-1312.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Embretson, S. (1990). A message from the Division 5 president. *The Score, 13* (3), 1, 4, 13.

Fish, L. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development, 21,* 130-137.

Goldman, L. (1972). Tests and counseling: The marriage that failed. *Measurement and Evaluation in Counseling and Development, 4,* 213-220.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart and Winston.

Huberty, C. J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analysis. *Psychological Bulletin, 105,* 302-308.

Jaeger, R. (1990). A note from the veep. *AERA Division D Newsletter, 1* (1), 1-2.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs.* New York: Freeman.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43,* 635-642.

Legg, S., & Algina, J. (Eds.). (1990). *Cognitive assessment of language and math outcomes.* Norwood, NJ: Ablex.

Lehmann, I. J. (1990). The state of NCME: Remembering the past, looking to the future. *Educational Measurement: Issues and Practice, 9* (1), 3-10.

Lunneborg, C. E. (1987). *Bootstrap applications for the behavioral sciences* (Vol. 1). Seattle: University of Washington.

Maxwell, S. (in press). Recent developments in MANOVA applications. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2). Greenwich, CT: JAI Press.

McCaulley, M. (1990). The Myers-Briggs Indicator: A measure for individuals and groups. *Measurement and Evoluation in Counseling and Development, 22,* 181-195.

McKinley, R. L., & Mills, C. N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 71-135). Greenwich, CT: JAI Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

Mehrens, W. A. (1990). NCME and its counterparts: Unions, intersections, and complementation. *Educational Measurement: Issues and Practice, 9* (1), 22-25.

Nitko, A. J. (1990). NCME: Where is it going? *Educational Measurement: Issues and Practice, 9* (1), 2.

Prediger, D. J. (1980). The marriage between tests and career counseling: An intimate report. *Vocational Guidance Quarterly, 28,* 297-305.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal, 13,* 51-59.

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2nd ed.). Belmont, CA: Wadsworth.

Sheeley, V. L., & Eberly, C. G. (1985). Two decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 63,* 436-439.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), *Advances in educational research* (Vol. 1, pp. 99-106). Greenwich, CT: JAI Press.

Thompson, B. (1986, November). *Two reasons why multivariate methods are usually vital.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN.

Thompson, B. (1988). A note on statistical significance testing. *Measurement and Evaluation in Counseling and Development, 20,* 146-148.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22,* 2-5.

Webb, N., Rowley, G., & Shavelson, R. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21,* 81-90.

Bruce Thompson *is a department head and professor in the Department of Educational Psychology at Texas A&M University and an adjunct professor of community medicine at Baylor College of Medicine. Although Thompson was president of AMECD during the 1990-1991 academic year, the views presented in this article are not necessarily intended to reflect the formal positions of the division. Correspondence regarding this article should be sent to Bruce Thompson, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225.*