

Issues in Advancing Diversity Through Assessment

William E. Sedlacek

Five issues that affect assessment with diverse populations are presented and discussed. The "quest for the golden label" issue concerns problems with labels rather than operational measurement. The "three musketeers" issue is the tendency to believe one measure can work equally well for all. The "horizontal research" issue refers to the tendency to do nondevelopmental research on nontraditional groups. Sampling problems are discussed under the "bias is bias" issue. The "I'm okay, you're not" issue is the tendency to train few assessment specialists who are knowledgeable about diversity issues.

The title "Issues in Advancing Diversity Through Assessment" suggests an important focus that is recent in the counseling profession: using assessment as a proactive tool to initiate change rather than in a more passive, defensive way to ensure that someone is not being harmed. The Association for Assessment in Counseling (AAC) has developed a new document on multicultural assessment standards (Prediger, 1993), which is a compilation from many sources, including publications of the American Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the National Council on Measurement in Education (NCME), and two divisions of ACA: AAC and the Association for Multicultural Counseling and Development (AMCD).

MULTICULTURAL ASSESSMENT STANDARDS

The standards presented in Prediger (1993) are grouped by major assessment tasks: selection of instruments (content), selection of instruments (norming, validity, and reliability), administration and scoring, and use/interpretation of results. Standards are presented, and then a reference for the source is provided under each task.

For example, under content, AAC recommends that potentially insensitive content or language be avoided. Another content area noted is the need to determine if bilingual groups should be tested in either or both languages or tested first for bilingualism or language dominance. The need to design research to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups is also noted.

Standards concerning norming include ensuring that appropriate norms have been developed for the population to be tested. A particularly important point related to validity is the need to provide evidence of both actuarial and clinical criterion-related validity for all populations for which a test is used. A standard on reliability noting the importance of checking reliability as an instrument translated from one language or dialect to another is discussed.

In the area of administration, one standard recommends considering the effects of examiner-examinee differences in ethnic and cultural background, attitudes, and values. Also, linguistic modifications recommended by test publishers should be described in detail in the test manual.

Major differences between norms and actual test takers were recommended to be taken into account in interpreting scores. It was further recommended that counselors be employed who know how race, culture, ethnicity, and so forth may affect assessment scores.

There are many other standards noted in the Prediger (1993) document. It may be a good check list, which assessment instrument developers and users could employ to ensure that multicultural issues have been appropriately considered in a given assessment procedure.

Although Prediger (1993) covers many useful standards, they are a molecular presentation of points relating to diversity as they affect assessment. What about the broader issues, however, that cause us to consider these standards? What about the ambiguous areas that are not covered by a specific standard? I believe that a discussion of some of the issues that cause the need for these standards may help put them in a context and provide a greater understanding of the problems involved. Although there are many diversity issues that could be addressed, five are covered in the remainder of this article. The first issue is called the "quest for the golden label." The idea is that if the right label could just be found, our problems related to diversity would be solved.

THE QUEST FOR THE GOLDEN LABEL

There is often much confusion about how *diversity* is defined, what groups should be included, and what terms should be applied to those groups. Questions arise such as the following: Should gays and lesbians be included in our conceptualizations of diversity? Can *Black* and *African American* be used interchangeably? Is the whole issue just a matter of being politically correct? Although labels and terminology are important and contain many symbolic and connotative elements, those concerned with assessment can cut through some of the ambiguity by helping to operationally define terms. Differences or similarities among groups can be demonstrated through measurement.

Westbrook and Sedlacek (1991) studied the labels used to describe what they called "nontraditional students" in the Education Index since the 1950s. Terms have varied from a focus on *acculturation* in the 1950s to *disadvantaged* in the 1960s, to *culture-specific differences* in the 1970s, to *multicultural* in the 1980s. *Diversity* could be added as the probable term for the 1990s. Although these terms may suggest different approaches to the groups discussed, operationally, the same people may be being discussed: those with cultural experiences different from those

of White middle-class men of European descent, those with less power to control their lives, and those who experience discrimination in the United States. Does it make sense, however, to include such variables as sex, sexual orientation, or athletic status as aspects of cultural experience?

Here is where some principles of assessment can be applied. As one example of how this issue might be addressed, measures have been developed in two areas and have been applied to many different groups. The first is the Situational Attitude Scale (SAS), which measures prejudicial attitudes toward different groups including racial-cultural groups (Balenger, Hoffman, & Sedlacek, 1992; White & Sedlacek, 1987), women (Minatoya & Sedlacek, 1983), older people (Schwalb & Sedlacek, 1990), gays and lesbians (Washington, 1993), people with physical disabilities (Stovall & Sedlacek, 1983), and athletes (Engstrom & Sedlacek, 1991).

The SAS is based on an experimental-control methodology. An experimental form or forms of a questionnaire are developed that contain situations depicting a certain reference group (e.g., African American) in a variety of situations that might generate prejudice toward them among others (e.g., a new family member moves next door). The control form of the SAS contains the same situations with no reference to African Americans or any other group. If respondents are randomly assigned to forms and then mean responses are compared to a series of semantic differential scales to each situation, any significant difference in response would be the result of the only difference between the forms: the term *African Americans*. This method allows for a validation of the effects of a stimulus term and has been demonstrated for a number of forms in the references noted. Generally, reliability estimates range from the .70s to the .90s for different forms, using several different reliability estimates. The references cited earlier provide more detail on particular forms of the SAS.

Using the SAS, one can evaluate whether the first criterion to determine if a group is nontraditional has been met: if the group is a target of prejudicial attitudes by others.

The second measure that may be used in assessing nontraditionality is the Noncognitive Questionnaire (NCQ; Sedlacek, 1989), which is designed to assess attributes that are not typically measured by other instruments and which may be common ways for persons with nontraditional experiences to show their abilities. The NCQ is discussed further under another issue. If the NCQ systematically correlates more with the success of members of a given group (e.g., African Americans) than it does for traditional White men, then the group meets the second criterion for inclusion as a nontraditional group. The NCQ has been shown to have predictive validity for a number of groups, and scale reliability estimates range from the .70s to the .90s using test-retest and Cronbach alpha methods. (See Boyer & Sedlacek, 1988; Sedlacek, 1989, 1991; Sedlacek & Adams-Gaston, 1992; Tracey & Sedlacek, 1984, 1985, 1987, 1988, 1989; White & Sedlacek, 1986.)

In summary, groups as different as athletes and older people may show their diversity in different ways. Nevertheless, there may be some similarities in the variables underlying their problems and in the ways they cope with a traditional system that was not designed for them.

There are many who would disagree with these conclusions or the definition of *nontraditional* presented earlier. For example, Helms (1992) saw visible racial-ethnic groups (VREGS) as having some distinct problems and characteristics. Also Sue, Arredondo, and McDavis (1992) expressed concern that the term *cross-cultural* can be defined so broadly that it allows counseling professionals to avoid dealing with what they see as the four major minority groups in our

society: African Americans, American Indians, Asian Americans, and Hispanics-Latinos.

The concern here is not that one conception or another is chosen as best but that decisions are based on measurement principles and empirical evidence wherever possible. It is possible that relevant variance can be identified based on nontraditionality and more specific group memberships as well.

THE THREE MUSKETEERS

The rallying cry of "all for one and one for all" is one that is often used in developing fair and equitable measures. The interpretation of how to handle diversity can often be to hone and fine-tune measures so they are equally valid for everyone (see Helms, 1992). If different groups have different experiences and different ways of presenting their attributes and abilities, however, it is unlikely that we could develop a single measure or test item that would be equally valid for all. If there is a focus on results rather than on intentions, it could be concluded that it is important to do an equally good job of assessment for each group, not that there is a need to use the same measures for all to accomplish that goal. Equality of results, not process, is the desired outcome.

Sternberg's (1985, 1986) work on intelligence might prove instructive. He suggested that there are three kinds of intelligence. *Componential* intelligence is the ability to interpret information in a hierarchical and taxonomic fashion in a well-defined and unchanging context. People who do well on standardized tests such as the American College Test (ACT) or the Graduate Record Examination (GRE) have this type of intelligence. *Experiential* intelligence involves the ability to interpret information in changing contexts: being creative. Standardized tests do not measure this type of intelligence, according to Sternberg. Sternberg termed his third type of intelligence *contextual*; it has to do with the ability to adapt to a changing environment and the ability to handle and negotiate the system.

If Sternberg's (1985, 1986) types of intelligence are applied to what is typically done in college admissions, for example, there is a heavy concentration on componential intelligence. Work in assessing nontraditional variables with the NCQ suggests that nontraditional people tend to show their abilities more often through experiential and contextual intelligence (Boyer & Sedlacek, 1988; Sedlacek, 1989, 1991; Sedlacek & Adams-Gaston, 1992; Tracey & Sedlacek, 1984, 1985, 1987, 1988, 1989; White & Sedlacek, 1986). Much of this is out of necessity because nontraditional people must learn to be "bicultural" and examine issues from different perspectives, and to be able to negotiate a system that was not designed for them. Having long-range goals, a self-concept that includes how the system views you, and an ability to handle racism are some of the scales on the NCQ. *Institutional racism* is defined as the negative consequences that accrue to a member of a given group because of the way a system or subsystem operates in the society (e.g., college admissions) regardless of any other attributes of the individual (Sedlacek, 1988). Thus, if there is a concentration on componential intelligence, less valid assessments will be done for nontraditional persons than for those with more traditional experiences in the system, which would be an example of institutional racism.

It is not that componential intelligence is not important to nontraditional people; it may be that experiential and contextual abilities may be prerequisite (Westbrook & Sedlacek, 1988). If one is struggling with racism in the system, time and energy may not be available for one to show componential talents.

The point illustrates that there is a need to think of measures differentially in order to achieve equitable assessments for all. There is probably a classic oxymoron here in thinking that one can assess

diversity of experience with a single measure. Notice that the arguments presented are positive and proactive. Lowering standards of assessment is not being advocated. The suggestion is to develop and use the most valid measures one can for all groups that can be operationally defined.

HORIZONTAL RESEARCH

The point could be made that because there seems to be an increasing literature on assessments with nontraditional groups; isn't this showing great process? The answer seems to be a little yes but mostly no. Unfortunately, the most common paradigm for a study that involves nontraditional persons is probably "We are doing the study anyway, so let's add some (e.g., Blacks, women, Hispanics) to it and see how they look." The nontraditional participants often seem to be "throw ins." The study or the measures weren't designed with them in mind, but they are included anyway. Motives for doing this probably vary, from being politically correct to having a genuine interest in a given nontraditional group. As stated earlier, however, results, not intentions, are what count.

If researchers continue to use this methodology over and over again in different settings, the same results are achieved over and over again. The research isn't developmental or sophisticated. It's not vertical or inductive; it's not going anywhere. It also tends to give a false sense of security or "self-righteousness" by allowing for the number of studies to be counted, and the reaching of an erroneous conclusion. For example, this is what seems to have happened in college admissions testing. Samples of available nontraditional students have been drawn on campus after campus; their scores have been correlated with grades, and conclusions have been reached. Studies have been built on theory and earlier research less often; thus, new measures have not been developed, and higher-order constructs have not been utilized.

The impression, however, should not be left that this applies only to admissions testing or higher education. While I was recently attending a presentation by a colleague who was developing a new instrument in the career area, an African American graduate student was also attending and was introduced to the colleague, and the student indicated an interest in career issues for African Americans. When asked if he had any information on how useful the instrument was for African Americans, the colleague said that he had not checked that out yet, but indicated that it was a good idea. The point was then made that if some of the constructs he was measuring differed for African Americans, he might have excluded some important issues, such as handling racism, which might need to be included in the development of an instrument that was useful with African Americans. He admitted that he had never thought of that and that perhaps we could talk further. The student's presence is mentioned to illustrate a point that is made later about how people are trained to do assessments.

The previous episode seems to be a typical situation. The colleague is a highly respected, well-trained, humanistic person. He would not intentionally engage in racism, but the effect of what he was doing was just that. It is probable that his new instrument will not be as valid for African Americans as it will be for others, and he seems to have been caught in a horizontal research problem. There was no developmental or sophisticated thinking or planning relative to diversity that was taking place.

BIAS IS BIAS

Any good methodological reference suggests guarding against loss of participants in doing assessment research and in developing norms for instruments (e.g., Mehrens & Lehmann, 1991). In practical situations, however, there are almost always missing or incomplete data, and the

assumption is commonly made that the missing participants do not cause the sample to be biased. When the missing participants are from nontraditional groups, however, that assumption frequently may be incorrect.

When I took my first undergraduate course in statistics, the instructor pointed out the distinction between statistical bias, "a consistent overestimate or underestimate of a parameter," and prejudice or social "bias" of some sort. This may not have been a good idea. Much of the work done in assessment has become so esoteric and molecular that many tend to forget, or perhaps never think of, the larger implications of the bias issue.

An example of sampling bias was documented by the institutional research office at a university. To develop regression equations for use in student selection, data were included for all students whose Scholastic Assessment Tests (SAT, formerly Scholastic Aptitude Test) scores and high school grades were in the university's data base. The equations resulting from this process were found to predict equally well for all races. Probing further, however, it was noted that students with incomplete data were not included in the equation-generating data set, and students in special programs were less likely to be included. The rationale was that the number of exclusions was small and that the researchers were using all the data available.

Analyses of the data from the missing groups showed that the relationships among predictors and criteria were very different for these students. On closer inspection, it was discovered that students for whom the SAT and high school records did not predict as well were less likely to have data in the student data base, either because students didn't turn it in, or because they didn't take the SAT and were in a special program that generated its own data. Such students were also much more likely to be nontraditional. In this example, the loss of participants resulted in more sampling bias for nontraditional students than for traditional students. The regression equations produced on the biased data set were used to select students for admission for several years. Consequently, in this case, sampling bias resulted in the use of selection procedures that were less valid or invalid for many nontraditional applicants. Sampling bias that invalidates assessments for nontraditional persons is common. For example, the problem of differential quality of samples when studying African Americans in career development studies was noted by Carter and Cook (1992).

Sampling bias can be reduced by checking the characteristics of population members with missing data, asking what the effects might be on people with similar characteristics if the missing data are ignored, and working to get data from all members of the sample or population. If it is not possible to get complete data, extraordinary efforts should be made to get data on a well-drawn sample of those missing, so that the extent of the sampling bias can be estimated. Parameter estimates for the population can then be modified, or, at the very least, caution can be given so that the biased equations or norms are treated as more fallible. This should result in less rigid applications of the results of assessment studies to people for whom they may not be appropriate. As discussed in the next section, however, the lack of knowledge of nontraditional populations by assessment specialists often prevents a sophisticated examination of variables on which sampling bias may occur.

I'M OKAY, YOU'RE NOT

At this time there are few assessment specialists who are adequately trained in developing measures for use with nontraditional populations (Helms, 1992). There are many graduate programs that turn out methodologists with good technical and quantitative skills, but few of these

graduates have even a rudimentary appreciation of racial and cultural variables, let alone any sense of how to measure them. Conversely, those most interested in cultural and racial variables tend to have little training or interest in quantification or assessment. As a result of the shortage of well-trained professionals working in the area, there are few assessment techniques available for work with nontraditional groups (Helms, 1992).

Differences in career interests may explain the lack of quantitatively oriented programs designed to work on issues relevant to nontraditional persons. Holland's (1985) research on vocational preferences suggests that investigative and realistic types tend to be interested in assessment and quantification, whereas social types tend to be interested in issues related to culture, race, and nontraditionality.

Helms (1992) believed that the relatively small number of VREG members of APA (3.1% and no psychometricians; see Russo, Olmedo, Stapp, & Fulcher, 1981) requires that those trained under the Eurocentric tradition in testing must develop the needed culturally sensitive measures.

Obviously, requiring those interested in assessment to take courses in multicultural issues and requiring those with multicultural interests to take measurement courses is one answer. This, however, already takes place in many training programs, and it is not nearly enough to change much behavior in the profession. There need to be more opportunities for collaboration in instrument development through grants, workshops, and so forth, as well as theses, dissertations, and everyday professional activities.

For example, in the situation discussed earlier involving the African American graduate student and the colleague, a collaboration among us was begun. The student could have used the experience, and the colleague needed some input. Without the suggestion that they work together, the student may not have been able to identify a problem with the measure, and the issue of use of the measure for nontraditional populations may have never come up for the colleague. Simply putting the two together without some further effort, however, may not have reduced the skepticism of the student or the reluctance of the colleague.

SUMMARY

Although there are many useful standards available for evaluating content, validity, reliability, and use of assessment instruments that would improve their applicability to diverse populations (Prediger, 1993), there are some larger issues that also need to be addressed in promoting diversity through assessment.

Five issues were presented and discussed. The first was the idea that finding just the right label to use to describe groups that are diverse would solve the problems that diversity may present. This was called the "quest for the golden label" issue. The "three musketeers" issue concerned the tendency to believe that developing a single measure that is equally valid for all groups is the best way to ensure fairness, as opposed to using different measures for different groups. That much research involving diverse populations is done in a nondevelopmental way was called the "horizontal research" issue. The problems of sampling bias in studies of diverse populations was called the "bias is bias" issue. The "I'm okay, you're not" issue involved a discussion of the tendency to train few assessment specialists who are knowledgeable about diversity issues.

DISCUSSION

There seem to be many things ACA could do to address the issues raised earlier. More publications, either free standing or part of the *Journal of*

Counseling & Development or *Guidepost*, are recommended. Also, regular and preconvention programs could be supported, including more jointly sponsored by AAC and AMCD, perhaps involving even more divisions. Obtaining grants and contracts from or with ACA to provide such training for AAC members or with knowledgeable AAC members working with others all seem worthwhile. Like the national deficit, it took a long time to develop problems in this area, but the sooner efforts are begun to solve them, the better off counseling professionals will be.

Issues have been discussed that work against promoting diversity. These issues are in no sense independent. Some are easier to reduce or control than are others, but suggestions have been offered regarding how attempts can be made to remedy each problem. What is done about these problems and many others that are undoubtedly present in assessment procedures may be conceptualized in terms of the classical statistical error model: Type I versus Type II. If some of the approaches discussed in this article are implemented and if the hypotheses I have proposed are false, a Type I error would be made, and considerable time, effort, and money would be wasted. Conversely, if the hypotheses are true and conclusions are made that differences based on nontraditionality are not present or at least are not worth investigating further, there would be a Type II error, but there would also be the risk of producing an extremely negative series of outcomes for many groups by perpetuating biased approaches to testing and measurement. I believe the very least that can be done, if there is a commitment to diversity in the counseling profession, is to tolerate the possibility of a Type I error and increase the power of currently used procedures by working to eliminate the problems presented and discussed here.

REFERENCES

- Balenger, V. J., Hoffman, M. A., & Sedlacek, W. E. (1992). Racial attitudes among incoming White students: A study of ten-year trends. *Journal of College Student Development*, 33, 245-252.
- Boyer, S. P., & Sedlacek, W. E. (1988). Noncognitive predictors of academic success for international students: A longitudinal study. *Journal of College Student Development*, 29, 218-222.
- Carter, R. T., & Cook, D. A. (1992). A culturally relevant perspective for understanding the career paths of visible racial/ethnic group people. In H. D. Lea & Z. B. Leibowitz (Eds.), *Adult career development: Concepts, issues and practices* (pp. 192-217). Washington, DC: National Career Development Association.
- Engstrom, C. M., & Sedlacek, W. E. (1991). A study of prejudice toward university student-athletes. *Journal of Counseling & Development*, 70, 189-193.
- Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist*, 47, 1083-1101.
- Holland, J. (1985). *The Self-Directed Search*. Odessa, FL: Psychological Assessment Resources.
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Ft. Worth, TX: Holt, Rinehart and Winston.
- Minatoya, L. Y., & Sedlacek, W. E. (1983). The Situational Attitude Scale toward Women (SASW): A means to measure environmental sexism. *Journal of the National Association for Women Deans, Administrators, and Counselors*, 47(1), 26-30.
- Prediger, D. J. (1993). *Multicultural assessment standards: A compilation for counselors*. Alexandria, VA: American Counseling Association.
- Russo, N. F., Olmedo, E. L., Stapp, J., & Fulcher, R. (1981). Women and minorities in psychology. *American Psychologist*, 36, 1315-1365.
- Schwab, S. J., & Sedlacek, W. E. (1990). Have college student attitudes toward older people changed? *Journal of College Student Development*, 31, 127-132.
- Sedlacek, W. E. (1988). Institutional racism and how to handle it. *Health Pathways*, 10(9), 4-6.
- Sedlacek, W. E. (1989, Fall). Noncognitive indicators of student success. *Journal of College Admissions*, 1(125), 2-9.

- Sedlacek, W. E. (1991). Using noncognitive variables in advising nontraditional students. *National Academic Advising Association Journal*, 2(1), 75-82.
- Sedlacek, W. E., & Adams-Gaston, J. (1992). Predicting the academic success of student-athletes using SAT and noncognitive variables. *Journal of Counseling & Development*, 70, 724-727.
- Sternberg, R. J. (1985). *Beyond IQ*. London, England: Cambridge University Press.
- Sternberg, R. J. (1986). What would better intelligence tests look like? In *Measures in the college admissions process* (pp. 146-150). New York: The College Entrance Examination Board.
- Stovall, C., & Sedlacek, W. E. (1983). Attitudes of male and female university students toward students with different physical disabilities. *Journal of College Student Personnel*, 24, 325-330.
- Sue, D. W., Arredondo, P., & McDavis, R. J. (1992). Multicultural counseling competencies and standards: A call to the profession. *Journal of Counseling & Development*, 70, 477-486.
- Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement and Evaluation in Guidance*, 16, 171-178.
- Tracey, T. J., & Sedlacek, W. E. (1985). The relationship of noncognitive variables to academic success: A longitudinal comparison by race. *Journal of College Student Personnel*, 26, 405-410.
- Tracey, T. J., & Sedlacek, W. E. (1987). Prediction of college graduation using noncognitive variables by race. *Measurement and Evaluation in Counseling and Development*, 19, 177-184.
- Tracey, T. J., & Sedlacek, W. E. (1988). A comparison of White and Black student academic success using noncognitive variables: A LISREL analysis. *Research in Higher Education*, 27, 333-348.
- Tracey, T. J., & Sedlacek, W. E. (1989). Factor structure of the Noncognitive Questionnaire-Revised across samples of Black and White college students. *Educational and Psychological Measurement*, 49, 637-647.
- Washington, J. E. (1993). *An investigation of attitudes of heterosexual identified resident assistant towards students based on the sexual orientation of the student*. Unpublished doctoral dissertation, University of Maryland, College Park.
- Westbrook, F. D., & Sedlacek, W. E. (1988). Workshop on using noncognitive variables with minority students in higher education. *The Journal for Specialists in Group Work*, 13, 82-89.
- Westbrook, F. D., & Sedlacek, W. E. (1991). Forty years of using labels to communicate about nontraditional students: Does it help or hurt? *Journal of Counseling & Development*, 70, 20-28.
- White, T. J., & Sedlacek, W. E. (1986). Noncognitive predictors of grades and retention for specially admitted students. *Journal of College Admissions*, 3, 20-23.
- White, T. J., & Sedlacek, W. E. (1987). White student attitudes toward Blacks and Hispanics: Programming implications. *Journal of Multicultural Counseling and Development*, 15, 171-182.

William E. Sedlacek is the assistant director of the Counseling Center and a professor in the Department of Counseling and Personnel Services at the University of Maryland, College Park. Portions of this article are from an invited address to the Association for Assessment in Counseling at the American Counseling Association national convention, March 15, 1993, Atlanta, GA. Correspondence regarding this article should be sent to William E. Sedlacek, Counseling Center, Shoemaker Building, University of Maryland, College Park, MD 20742.