Guidelines and Best Practices for Culturally Competent Evaluations

Prepared for The Colorado Trust by

Kiki Sayre

September 2002

THE COLORADO TRUST
A GRANTMAKING FOUNDATION

The mission of The Colorado Trust is to advance the health and well-being of the people of Colorado.

The Colorado Trust
1600 Sherman Street
Denver, CO 80203-1604
303-837-1200
Toll free 888-847-9140
Fax 303-839-9034
www.coloradotrust.org

Please cite this report as: The Colorado Trust. *Guidelines and Best Practices for Culturally Competent Evaluations*. Denver, CO: The Colorado Trust; 2002.

Contents

Preface

The Colorado Trust has been committed to evaluation as a core component of its initiatives for more than a decade. The Trust funds initiative evaluations using external evaluators to examine how well the foundation is meeting its objectives, how the initiative process could be improved and, more recently, whether the grantees funded in the initiative have the anticipated impact on targeted groups.

As the state's population continues to become more ethnically diverse, The Trust has become concerned that evaluators need additional tools so that they can employ appropriate strategies, gather valid data across diverse groups of people and show sensitivity to the diverse populations being served. In particular, with an increased focus on individual program outcomes through evaluations, The Trust wants to be sure that measures being used are valid and reliable across multiple ethnic groups.

The Colorado Trust asked Lydia M. Prado, Ph.D., and Antonio Olmos-Gallo, Ph.D., both with the Mental Health Corporation of Denver and the University of Denver Department of Psychology, to conduct two half-day workshops for evaluators. The purpose of the workshops was to introduce the topic of cultural competence in evaluation to researchers, and to specifically respond to the question, "How do we know the measurement tools we use are equivalent across multiple ethnic groups?" The research and personal experiences of these two researchers, based on their investigations of mental health issues within a variety of cultural communities, provided tangible and poignant examples of cultural competency issues in evaluation.

This report is a summary of those workshops. The intended audience for this report — as for the workshops — is health and social science researchers engaged in evaluations of programs serving ethnically diverse populations. In addition, people who use evaluation results to make programmatic decisions will find this information useful. Like all technical reports, there are concepts and terms used that may be foreign to some readers; however, a number of these terms have been included to help ensure that the statistical analysis necessary to conduct a culturally competent evaluation can be carried out.

Cultural Competence — A set of academic and interpersonal skills that allow individuals to increase their understanding and appreciation of cultural differences and similarities within, among and between groups. This requires a willingness and ability to draw on community-based values, traditions and customs and to work with knowledgeable persons of and from the community in developing focused interventions, communications and other supports.

— Office of Substance Abuse Prevention, United States Department of Health and Human Services, 2001

Issues of Cultural Competence

It is important when working cross culturally to use an evaluation model that is meaningful and relevant to the specific target population. People cannot assume that the set of rules they are operating under will necessarily apply to everyone. Yet, said Dr. Lydia Prado of the Mental Health Corporation of Denver and the University of Denver Department of Psychology, "Diverse populations tell me this happens all the time. A ubiquitous 'white standard,' or majority culture, is often used to measure, assess and evaluate non-traditional, underserved or marginalized populations. One person of color referred to it as 'cultural imperialism.' In situations where there is a cultural clash, you will often see people withdraw."

She described that "people feel passionate about their own experiences and want evaluators to understand that their experience of the world is different from yours. They expect you to be respectful and to acknowledge the rules you are operating by. People appreciate an attitude of 'learning from' rather than 'learning about' — an attitude that validates and values multiple perspectives."

In most evaluations conducted in the United States, the strategies, analytic techniques and conceptual measurements used come from the "white culture." Evaluations that impose rules from the majority or dominant culture on people of different cultures may be restricted by a number of factors, such as conceptual mismatches, language barriers, different values and differences in the meaning and manifestation of emotions. For example, the same behavior observed — or not observed — in people from different cultures can mean different things.

Example: Anger is a trait that correlates highly with adolescent alcohol abuse in the Caucasian population, whereas in the American Indian population the expression of anger inversely correlates with alcohol abuse. Within this population, anger is a positive factor that can keep adolescents in school and help them stay motivated to improve the lives of their community members.

Example: Research on marital communication involved a coding system that observers used to record conflicts in couples' interactions. Observers who were not of Asian heritage observed no conflicts among Asian couples. However, an observer who was brought in from the Asian community perceived numerous indications of conflicts that

those outside the culture were unable to detect.

"Unless you sit at the table and know the rules the group operates by, you could develop an evaluation — or intervention — that misses the point," said Dr. Prado. "You need to start off with a model that is informed in order to come up with data and meaningful recommendations that will be helpful and relevant to the community you are working in."

Evaluators working with diverse cultures need to understand a number of different variables within those cultures, including the country of origin, level of acculturation, sociocultural issues and environmental factors. Said Dr. Prado, "To find the right questions to ask, you need to have a lot of specific cultural knowledge from that group and understand its worldview. Only then can you know the right questions and the right procedures and analytic strategies to use for that group."

Even when researchers feel they have adequate knowledge, it is important to ask, and not assume, that the knowledge is adequate.

Often, qualitative research can provide a context for interpreting quantitative results. Because cultures adapt in myriad ways, it is important to view behaviors in their cultural context. Behaviors that initially appear maladaptive from one cultural perspective may, in fact, be understood as adaptive — although not necessarily effective in the long-term — from a different cultural perspective. The reason for the adaptation can result from something other than was expected.

Example: The choice to leave school prematurely has been considered by some researchers to be a calculated move of an adolescent who has made a decision for herself that is psychologically empowering.[1] From this perspective, school failure can be seen as a survival strategy adopted by teens who, feeling devalued, bored and unacknowledged, attempt to resist a school system they experience as both disrespectful and irrelevant to their lives. "It is not adaptive to adopt a system that does not value you, and it is adaptive to maintain some distance from that system in service to oneself," explained Dr. Prado.

Evaluators also need to be aware of the wide range of heterogeneity within a cultural group. "It is important to know when it is appropriate to aggregate or collapse data within a cultural group," Dr. Prado pointed out. Oftentimes, ethnicity alone does not form a meaningful basis upon which to evaluate a group of people.

Example: A study to determine the abstinence rate for alcohol among women found that Latina women had a 68% abstinence rate. However, when the numbers among the Latina women were analyzed according to their country of origin, the results varied significantly. For foreign-born Central American women, the rate was 74%; for Mexican women, it was 71%; for Cuban American women, 48%; and for Puerto Rican women, 45%. Additionally, when data were analyzed according to how long the women had been in the United States, the rates of alcohol use soared. The Latina data, reanalyzed by country of origin and level of acculturation within the United States, were much more meaningful and provided a more sound basis for developing specific interventions.

Historical Context and "Master Statuses"

From early in the twentieth century until the 1960s, most research conducted cross culturally in the United States used a standard cultural deviance model. If other groups displayed differences from the majority group, the behavior was considered deviant. Beginning in the mid-1960s, research was based on a cultural equivalence model and differences across groups were attributed primarily to socioeconomic status and differences in the allocation of resources (e.g., private education).

Today, researchers are developing culturally variant models that acknowledge various cultural adaptations to a majority cultural context. This approach assumes that specific behaviors in different groups are strengths and can be considered adaptive. Therefore, evaluators are faced with the challenge of determining to what degree their understanding of behaviors in different cultural contexts is well informed and meaningful, and to what degree current evaluation models apply to diverse ethnocultural groups.

Research suggests that well-informed, culturally competent evaluators understand that race and gender act as "master statuses" in this country — the primary criteria that people use to make judgements about each other. Because race is a master status, racism and discrimination are realities for people from diverse groups. Evaluators need to be aware of the dynamics of differences manifested in our society and understand the experience of being considered the "other" group.

"Racism can be considered a system of advantage that operates daily in people's lives," explained Dr. Prado. "According to the Surgeon General's report on race, culture and ethnicity, just being a person of color puts people at risk for health problems. In the South, for example, racism and discrimination are tied to high blood pressure."

When Working Cross Culturally, Consider:[2]

- To what degree do current evaluation models apply and respond to diverse ethnocultural groups?
- How can ethnicity be defined in a more meaningful manner?
- When should pre-existing instruments be used in evaluation and when is it best to adapt such measures or construct new ones?
- How can results be interpreted in ways that best reflect the lives of the participants being studied?
- What types of evaluation strategies come closest to addressing the needs of those communities that have historically lacked social and economic capital and access to political power?

Instrument Equivalence Across Ethnic Groups

Following Dr. Prado's discussion of the importance of culturally competent evaluations, Antonio Olmos-Gallo, Ph.D., with the Mental Health Corporation of Denver and the University of Denver Department of Psychology, described a process for determining whether an evaluation instrument has cross-cultural equivalence — or measures the same thing across subpopulations.

To explain why instrument equivalence is important in this context, Dr. Olmos-Gallo quoted Drasgow and Kanfer of the University of Illinois at Champaign, who wrote, "Standardized psychological measurement instruments must provide equivalent measurement across subpopulations if comparative statements are to have substantive import. Without equivalent measurement, observed scores . . . are not directly comparable."

Dr. Olmos-Gallo explained that if evaluators cannot guarantee that the data they are collecting from different ethnic groups have the same meaning across different ethnicities, then they do not know what they are measuring. In addition, each item on a particular

Guidelines to Support Cultural Competency in Evaluations[3]

Define the population precisely — Understand a group's country of origin, immigration history, sociopolitical status, level of education, rules and norms. Without a clear understanding of the group's background, it is best to develop a community advisory group.

Develop collaborations with the target population — Community members need to be involved in program planning and implementation. Define the pertinent evaluation questions at the outset.

Encourage buy-in — Know the community well and understand the pressures and external constraints operating among the population. State the goals of the evaluation team and determine the goals of the people being evaluated. Describe how the data will be used. Conduct interviews in a location that is comfortable to the group and without bias**.**

Provide timely feedback and results in clear, useful formats conveyed through culturally appropriate methods — Ask those involved how best to disseminate the information. For example, data from an evaluation conducted among an American Indian population in New Mexico was shared in a "give-back" ceremony using storytelling and visuals, with no written material.

Consider acculturation and biculturalism in interpretation and utilization of data — Acculturation measures are often linear and one-dimensional. Bicultural adaptation — or the adoption of some majority culture attitudes and practices coupled with the retention of ethnic group cultural practices and identity — is now considered a more useful measurement. Cultural identity can be bicultural, or even tricultural. People generally do not lose one culture to gain another.

Know when to aggregate the within-group data from a heterogeneous sample and still maximize external validity — Conduct within-group analyses that consider groups independently of each other to ensure that important data is not overlooked. Only aggregate the data if convincing similarities can be found.

Avoid deficit model interpretations — Abandon stereotypes and models that measure diverse groups against a monocultural standard.

instrument should mean the same thing to people from different cultural groups. If this is not the case, the evaluators will not know what they are measuring and will not be able to accurately say that one group has more or less of something than another. "We are faced with the classic conundrum of trying to match apples and oranges," said Dr. Olmos-Gallo.

Dr. Olmos-Gallo explained how to ensure that instruments have the same meaning across different cultures. In his experience, equivalence can be achieved by using a three-step test. The first step involves invariance reliability scores, which means that the reliability scores of the measure are the same across different populations. Reliability tests check the consistency of the responses. The second step involves factorial invariance, which means that the factor structure of a given instrument is the same across different groups. If an instrument measuring depression has two subdomains in one culture (i.e., according to the instrument's creators, depression can be represented by two factors), it should also have the same two subdomains for all cultural groups. The final — sometimes more stringent — step is based on item-response theory, which is conducted to assess differences in item parameters. Dr. Olmos-Gallo explained that this process should be tested in both self-report instruments, such as surveys, as well as in observational checklists.

"There used to be the belief that behavior was the same across different cultures and

therefore we should not be concerned about the effects of culture, but that has changed," said Dr. Olmos-Gallo. "Now it is believed that culture plays a critical role in determining what is considered normal behavior, thus creating personality configurations that may look normal in one culture but be considered as pathological in another culture."

Example: It is considered impolite in most Latin-American cultures to look a person of authority, such as a doctor or a priest, directly in the eye. In America, that same Latin-American person exhibiting such behavior may be considered to be withdrawn.

To illustrate the three-step process, Dr. Olmos-Gallo applied the instrument equivalence test using the Colorado Client Assessment Record (CCAR) — an instrument often used in the health field — as it applied to an evaluation gauging levels of depression in children from three different ethnic groups enrolled in Denver area health centers.

In the test of reliability, the aim is to check the consistency of the responses. For clinical

Invariance in reliability scores

- Technically, a reliability is $\mathbf{r^2}$, therefore, can be treated as correlation scores and use a Z-score to test differences between groups:

$$Z_{reliability\_white} = \frac{1}{2}\left(\ln(1-reliab\_white) - \ln(1-reliab\_white)\right)$$

$$SE_{white-af\_am} = \sqrt{\left(\frac{1}{\#white-3}\right)\left(\frac{1}{\#af\_am-3}\right)}$$

$$Ztest_{white-af\_am} = \frac{Z_{reliab\_white} - Z_{reliab\_af\_am}}{SE_{white-af\_am}}$$

instruments, and most other types of instruments, the statistic of choice is internal consistency, or Cronbach's alpha. Dr. Olmos-Gallo tested the reliability estimates across different populations by comparing the scores using a Z-test.
In Dr. Olmos-Gallo's experience, the test of reliability needs a minimum of 30 subjects in every group; however, the larger the sample size, the more valid the results.

Example: While Dr. Olmos-Gallo's instrument to measure depression was based on observation, he gave an example of how language issues can lead to reliability problems when using questionnaires. When an Asian woman was asked if she felt blue, she interpreted the question to mean, "Do you feel like you are about to pass out from holding your breath?" In this case, the phrase did not mean the same thing across different ethnic groups. Therefore, the scale's reliability for the Asian-American group should be low compared to other groups.

The factorial invariance test addresses whether the items comprising a particular measur-

ing instrument operate in the same way across different populations. This test can be performed using different statistical procedures:

- First, determine that the same number of factors have the same items associated with them (i.e., the questions in an instrument always load the same factors).
- Second, determine that the factor loadings are identical for every item and every factor.
- Third, make sure that variances and covariances among the factors are the same across populations.
- Finally, ensure that the residuals for every item are the same across populations.

Example: Lack of factorial invariance was found in a standard measure for social deviance after evaluators noticed an unusual finding in the American Indian adolescent population. "I fight" was loading on a different factor from what had been observed when the instrument was first developed. Further investigation found that the American Indian adolescents were interpreting the phrase to mean, "I fight for what I believe is right" rather than "I have physical quarrels."

Item response theory is often used by testing companies for aptitude and achievement tests to determine the difficulty of the items. It can also analyze instruments with rankings (i.e., Likert scales). To test item equivalence, Dr. Olmos-Gallo used Rasch modeling to determine if the items had the same level of difficulty across different cultural groups.

Best Practices for Culturally Competent Evaluation[4]

Develop specific cultural knowledge. Know the relationship between variables and behaviors in the group being evaluated. Only when the norms and values are clearly delineated can they be given proper consideration.

Explicitly examine the theoretical framework that is the foundation of your research. Communicate clearly your own values, beliefs, approach and world view as the evaluator. Acknowledge and address how these may differ from the perspectives of the group to be evaluated. Whenever possible, have someone on the evaluation team who has knowledge and understanding of the group being evaluated.

Define and measure ethnicity in a meaningful manner. To the degree possible, also define and measure key constructs, such as socioeconomic status, that are known to covary with ethnicity. If you suspect there is variability within a group, find out if other characteristics have an impact on the data. Measure the elements and factors that may covary to determine whether it is ethnicity or some other factor. If other factors are involved, the socioeconomic status or additional factors need to be measured along with race and ethnicity..

Choose measures that are appropriate for all the ethnic groups in your study and/or check those measures you use for their equivalence across groups. Make sure the instrument you are using has cross-cultural equivalence. Do not assume factors correlate across different groups.

Make sure your analyses reflect study questions and that you have sufficient power to get accurate answers. The goal is to accurately interpret the experiences of particular groups of people in order to minimize errors throughout the study. For this reason, the evaluation team needs to be involved from the beginning of the implementation stage.

Interpret results to reflect the lives of the people studied. Have someone with knowledge of the particular group analyze the data alongside the evaluators in order to point out variables that should be considered.

What does the Rasch analysis tell us about rating scales?

It provides "difficulty" indices for each item that indicates how relatively easy or difficult it is for a given individual to receive a high rating.

- On a clincial scale to measure some type of dysfunction, for example, an "easy" item suggests that even for clients with low levels of dysfunction, they have a high probability of receiving a high rating.
- A "difficult" item would require much higher levels of dysfunction in order to receive a high rating.

One of the multiple advantages of the Rasch model is that it charts the distribution of the items by difficulty and the distribution of those tested by ability.

In the case of clinical measures, however, Rasch modeling will help determine the level of dysfunction of the individual. In the context of cultural equivalence, it is important to see whether the items can discriminate similar levels of difficulty/dysfunction across different populations. Large differences across items may point to potential biases in the ratings. To evaluate the statistical significance of the differences across items, Dr. Olmos-Gallo suggested Z-tests.

Example: Educational Testing Services (ETS), the organization that develops standardized academic tests such as the SAT, discovered a question on one of its tests that discriminated against students from the Latino culture. The question asked for a similar association to the words strawberry-red. The correct answer was lemon-yellow. For many Latinos, however, lemons are green in their native countries.

The measurement equivalence test for Dr. Olmos-Gallo's Colorado Client Assessment Record instrument found that variabilities in the instrument across different populations suggested a slight bias in how children from different cultures were rated with the instrument. In particular, the factorial invariance test failed in its most stringent test — the residuals were not the same across populations. Similarly, it was found that Rasch models were not the same. Significant differences occurred across items for some of the groups.

In this example, the measurement equivalence test was administered after the instrument was developed. Dr. Olmos-Gallo pointed out that evaluators may instead choose to use the tests as the measures are developed. Otherwise, in order for an instrument to have cross-cultural equivalence, the evaluators may need to modify it using an informed, community-based process.

Notes

1. Robinson T, Ward JV. A belief in self far greater than anyone's disbelief: Cultivating resistance among African American female adolescents. *Women, Girls, and Psychotherapy*. 1991;187-103.

2. Cauce AM, Coronado N, Watson J. Conceptual, methodological, and statistical issues in culturally competent research. In: *Promoting Cultural Competence in Children's Mental Health Services*. Baltimore, Md: Brookes Publishing Company; 1998:305-329.

3. Huang L.N. Data through a cultural lens: Practices to support culturally competent use of evaluation data. *Data Matters*. 2000;3-8.

4. Hernandez M, Isaacs M. Promoting cultural competence in children's mental health. In: *Promoting Cultural Competence in Children's Mental Health Services*. Baltimore, Md: Brookes Publishing Company; 1998:1-25.