Journal of Substance Abuse Treatment 36 (2009) 25 – 43

Journal of Substance Abuse Treatment

Regular article

Measurement and data analysis in research addressing health disparities in substance abuse

Ann Kathleen Burlew, (Ph.D.)[a,*], Daniel Feaster, (Ph.D.)[b], Mary-Lynn Brecht, (Ph.D.)[c], Robert Hubbard, (Ph.D., M.B.A.)[d]

[a]*Department of Psychology, University of Cincinnati, Cincinnati, OH 45221, USA*
[b]*Center for Family Studies, University of Miami School of Medicine, Miami, FL 33136, USA*
[c]*Integrated Substance Abuse Programs, University of California, Los Angeles, Los Angeles, CA 90025, USA*
[d]*Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC 27705, USA*

Received 14 December 2007; accepted 11 April 2008

Abstract

This article describes concrete strategies for conducting substance abuse research with ethnic minorities. Two issues associated with valid analysis, measurement and data analysis, are included. Both empirical (e.g., confirmatory factor analysis, item response theory, and regression) and nonempirical (e.g., focus groups, expert panels, pilot studies, and translation equivalence) approaches to improve measures are described. A discussion of the use of norms and cutoff scores derived from a different ethnic group along with the effects of the ethnicity of the interviewer or coder on measurement is included. The section on data analysis describes why the use of race-comparison designs may lead to misleading conclusions. Alternatives to race-comparison analysis including within-group and between-group analyses are described. The shortcomings of combining ethnic groups for analyses are discussed. The article ends with a list of recommendations for research with ethnic minorities. © 2009 Elsevier Inc. All rights reserved.

Keywords: Ethnic minorities; Health disparities research; Measurement equivalence; Valid analysis; Data analysis

1. Introduction

Ethnic minorities are projected to account for one third of the U.S. population as soon as 2010 and almost half of the population by the year 2050 (U.S. Bureau of the Census, 2000; U.S. Bureau of the Census, 2004). Although ethnic minorities report similar rates of drug use as non-Hispanic Caucasians, the available data suggest that the consequences of drug use are more negative for ethnic minorities (e.g., higher rates of drug-related involvement in the criminal justice system [Iguchi et al., 2002], higher rates of HIV infection associated with drug use [Center for Disease Control and Prevention, 2001; CDCP, 2002; Galea & Rudenstein, 2005]).

The available information makes a convincing case for the need for more research on the efficacy of substance abuse treatments among ethnic minorities. For example, a recent review by Szapocznik, Prado, Burlew, Williams, and Santisteban (2007) only identified one substance abuse treatment intervention with demonstrated efficacy for Hispanic adolescents (Brief Strategic Family Therapy [BSFT]) and only one for African American adolescents (Multisystemic Therapy). Other data suggesting that ethnic minorities are less likely to participate in substance abuse treatment (Substance Abuse Mental Health Services Administration [SAMHSA], 2002) and are less likely to describe the drug treatment experience as helpful (Heron, Twomey, Jacobs, & Kaslow, 1997; Longshore, Grills, & Annon, 1999) also support the need for more research on substance abuse treatment for ethnic minorities. Nevertheless, only limited

* Corresponding author. Department of Psychology, University of Cincinnati, Cincinnati OH, 45221, USA. Tel.: +1 513 708 9009; fax: +1 513 556 1904.

E-mail address: rkburlew@juno.com (A.K. Burlew).

0740-5472/08/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsat.2008.04.003

research is available on psychotherapies and interventions with ethnic minorities (Hall, 2001).

The inclusion of more ethnic minorities in treatment studies alone may not be sufficient to inform public policy adequately on drug treatment for ethnic minorities (Miranda, 1996). Instead, research on ethnic minorities may warrant special attention because the theoretical models, assessment instruments, and even methodological procedures developed on non-Hispanic Caucasian populations may need to be examined to ensure their appropriateness for ethnic minority samples. Sue, Kurasaki, and Srinivasan (1999) described how cultural considerations must be taken into account at every stage of the research project. Moreover, the impact of cultural factors may warrant consideration both in research samples that only include a specific ethnic minority group and heterogeneous samples that include ethnic minorities. The goal of this article is to provide some concrete strategies for improving substance abuse research with ethnic minorities.

Although the expected increase in diversity presents new challenges, diversity also provides new opportunities for researchers to contribute to our understanding of substance abuse treatment among ethnic minorities. Moreover, rather than conceptualizing diversity as a burden or potential confound, we conceptualize the diversity in large, multisite clinical trials such as the National Institute of Drug Abuse (NIDA) Clinical Trials Network (CTN) as a tremendous opportunity to increase the number of questions examined and articles published from data sets. The data collected from clinical trials have the potential to address a number of important issues beyond their original purpose such as whether treatments work differently for specific subgroups of an ethnic minority group or whether the mechanisms (i.e., mediators) of how treatments work are different for specific ethnic groups. The baseline data may be a rich resource for understanding precursors or correlates of drug use for ethnic minorities. In addition, longitudinal studies of the control group may be useful for understanding the natural course of drug use among ethnic minorities.

Despite the widespread call for more appropriate research on substance abuse with ethnic minorities, few guidelines are available to assist research teams in translating the mandates into practice. The guidelines of the National Institutes of Health (NIH) address two broad areas—inclusion and valid analysis. Yet, much of the existing information focuses on inclusion (e.g., recruitment and retention) rather than valid analysis. This article is intended to serve as a guidance document on the issue of valid analysis in substance abuse research.

This article addresses two major issues associated with valid analysis—measurement and data analysis. However, before discussing these two topics, a brief discussion is included of two documents that play an important role in shaping the mandate for more effective substance use treatment research on ethnic minorities: NIH Guidelines on the Inclusion of Women and Ethnic Minorities and the NIDA Strategic Plan for Reducing and Ultimately Eliminating Health Disparities.

2. NIH guidelines on the inclusion of women and ethnic minorities

In 1994, NIH established a mandate on the inclusion of women and ethnic minorities in research and the valid analysis of the data collected on those groups. The guidelines were later updated, and the most recent amendments were added in 2001. The main elements of the NIH Guidelines for the Inclusion of Women and Minorities as Subjects in Clinical Research are outlined in Table 1. The guidelines apply to all clinical trials. To eliminate uncertainty, the NIH defined clinical trials as pharmacologic, nonpharmacologic, and behavioral interventions "usually involving several hundred or more human subjects" designed to compare an experimental intervention with a standard or control aimed at generating scientific evidence to change health policy or altering the standard of care (Hohmann & Perron, 1996, p. 853; NIH, 2001). According to Hohmann & Perron, 1996, the guidelines especially apply to effectiveness research because external validity is a critical concern in these studies. Researchers are encouraged to review these guidelines (http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm) as well as the companion document, *Questions and Answers concerning the 1994 NIH Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research* (http://orwh.od.nih.gov/inclusion/incloutreach.html; NIH, 2001), for more details.

The NIH guidelines on the inclusion of minorities and women define valid analysis as one in which "participants are assigned to groups in an unbiased manner, assessment of

Table 1
Elements of the NIH guidelines on the inclusion of women and minorities

Element	Description
Inclusion	"NIH must ensure that women and members of minorities and their subpopulations are included in all human participant research, not just clinical research."
Valid analyses	"For Phase III clinical trials, NIH must ensure that women and minorities and their subpopulations are included so that valid analyses of differences in intervention effect can be achieved."
Cost	"NIH will not allow cost as an acceptable reason for excluding these groups."
Outreach	"NIH must initiate programs and support for outreach efforts to recruit these groups into clinical studies."

Note. Based on excerpts from Hohman and Parron (1996).

outcomes is unbiased, and unbiased statistical analyses are used to estimate intervention effects" (NIH, 2001). That operational definition led to our decision to include both measurement and data analysis in this article on valid analyses with ethnic minorities.

3. The NIDA Strategic Plan for Reducing and Ultimately Eliminating Health Disparities

In 2001, the NIH and each of its institutes developed strategic plans for addressing health disparities. The NIDA Strategic Plan for Reducing and Ultimately Eliminating Health Disparities emphasized NIDA's commitment to conducting research to increase the amount of information available on substance abuse among ethnic minorities. The key elements of the NIDA strategic plan are summarized in Table 2.

4. Measurement

Adequate measurement is essential to eliminating health disparities (Ramirez, Ford, Steward, & Teresi, 2005) and to conducting good substance abuse research (Widaman & Reise, 1997). Cultural norms and practices influence the relevance of specific constructs for a particular group, the range of behaviors and responses that are indicators of the construct, and even how individuals understand and interpret items intended to assess the constructs (Hughes & Dumon, 1993). For that reason, the adequacy of a measure for one culture or subculture does not guarantee the adequacy or appropriateness of that measure for another cultural group (Bravo, 2003). The cultural appropriateness of the measure is a concern for researchers who use measures standardized on non-Hispanic Caucasian samples to conduct research on ethnically diverse samples or on samples of a particular ethnic group. Despite the importance of considering the characteristics of the group to whom the measure is to be administered, this issue is frequently overlooked. For example, in a study of the 18 available treatment outcome studies of adolescent substance abusers, (Strada, Donohue, and Lefforge 2006) reported that the only reference to the cultural appropriateness of the measures was the translation of several measures into Spanish. Obviously, the cultural appropriateness of a measure is a larger issue than whether the measure is in the participants' primary language.

Ensuring measurement equivalence may not be such a problem in urine drug screens or other biological measures of substance use. However, measurement equivalence may be a more relevant issue for self-report measures (Shillington & Clapp, 2000) such as alcohol expectancies (Randolph, Gerend, & Miller, 2006) or addiction-related problems (Brodey et al., 2004). Moreover, few substance abuse studies only include measures of substance use. Rather, studies frequently include other measures such as externalizing or internalizing disorders, antisocial behaviors, self-esteem, or family interaction patterns that are often conceptualized as either predictors, correlates, or consequences of substance use. Inequivalence on such measures may lead to misleading conclusions about the etiology of substance abuse, its effects, or appropriate treatments.

The next section begins with a discussion on measurement equivalence. Several other topics associated with measurement are discussed briefly after the section on measurement equivalence. These include the use of norms/cutoff scores developed on a different group and the effect of the race/ethnicity of the interviewer and the race/ethnicity of the observer on ratings of behavior or other characteristics.

4.1. Measurement equivalence

There are some obvious reasons why group differences may occur on a specific measure that may not necessarily reflect group differences on the underlying trait. Some common reasons are presented below along with examples provided by Bravo (2003) and others:

1. *The situations mentioned in a measure may apply to one group but not another.* An item may be inappropriate for a group if it asks about experiences

Table 2
NIDA strategic plan for reducing and ultimately eliminating health disparities

Element	Description
Date of adoption	2001
Areas emphasized in plan	"Over the next five years, NIDA will strive to 1) Improve our understanding of the incidence and causes of drug abuse and addiction in all racial/ethnic groups recognizing the diversity by gender, SES, and other factors within racial/ethnic populations, (2) strengthen and expand the community and research infrastructure for conducting research within racial/ethnic populations, (3) improve prevention and treatment for racial/ethnic groups at highest risk for addiction and medical consequences of drug use and addiction, and (4) widely disseminate information on drug use and the disease of addiction in racial/ethnic communities identifying best approaches to prevention and treatment." (p. 4)
Goals for treatment research	"Goal One: Increase the number of treatment research studies that focus on racial/ethnic differences and improve dissemination of the study results." (p. 9) "Goal Two: Determine the factors that contribute to differences, if any, experienced by racial/ethnic minority populations in access to services and outcomes of treatment in managed care and other service systems. In addition, assess the impact of welfare reform on substance abuse services provided to ethnic minorities especially minority women." (p. 10)

Adapted from the NIDA Strategic Plan for Reducing and Ultimately Eliminating Health Disparities (2001).

that are not relevant for the specific group. For example, Rogler, Malgady, & Rodriguez (1989) describe how an item asking who makes vacation plans may not be an appropriate indicator of shared decision making between spousal partners in a low-income ethnic minority sample if they cannot afford a vacation.

2. *Various cultural groups may differ in the connection between specific behaviors and the underlying trait.* If two cultural groups differ in their beliefs about the appropriateness of a specific behavior, then the link between that behavior and the underlying trait may differ as well. For example, although family functioning is frequently included as a protective factor in substance abuse research, Bravo (2003) points out that Puerto Rican families may not share the beliefs of some other groups that a child's input into his or her own discipline is an effective parenting strategy. Therefore, in a study of the effects of family therapy on improving outcomes for substance-abusing adolescents and their families, it may be inappropriate to use an item about child input into discipline in a measure aimed at assessing family functioning in Puerto Rican families.
3. *Differences in the opportunity structure may lead to differences in the manner in which certain traits are manifested.* Cultural groups may differ in the extent to which they have the opportunity to participate in certain behaviors or activities. For example, some of the items in the Child Behavior Checklist, which is used to assess social competence, ask about participation in activities (e.g., involvement in sports, hobbies, and organizations) that may require financial resources. Such items may not provide an accurate indication of social competence in low-income children who may be forced by financial hardship to demonstrate their social competence in different ways (e.g., leadership in informal groups). Therefore, in this case, researchers evaluating the outcomes of a prevention or treatment program for substance-abusing adolescents may need to consider the cultural appropriateness of the items used to assess the adolescent's functioning.

Similarly, questions about employment may also illustrate the effect of the opportunity structure on responses to self-report questions. For example, standard questions about employment may be phrased in such a way that the items do not elicit the “off the books” employment (e.g., handiwork, household repair, or yard work) that some low-income individuals may obtain. If so, the response of the low-income individual to such items may not provide a complete picture of work-related activity.

4. *Group differences in circumstances may result in differences in the meaning of a specific behavior.* It is even possible that the meaning of a behavior could differ so much across groups to the point that the same behavior may be adaptive in one group even if pathological in another. For example, Nichol, Padilla, and Lucio (2000) describe a study conducted by Walker and colleagues in which low-income African American youth were more likely than their non-Hispanic Caucasian peers to endorse an item about the attention paid to ensuring that doors and windows were locked when leaving home. Nichol et al. (2000) point out that although this response may be associated with the increased probability of phobias or unusual fears in other groups, concern about the safety of open doors and windows may be a realistic response to the increased danger in the environment that some youth experience in high-crime neighborhoods.

These are all examples of why it is important to ensure measurement equivalence. Broadly speaking, measurement equivalence is achieved when a measure assesses the same underlying concept across groups. When researchers use a measure developed in one cultural group with another cultural group or to compare two cultural groups, two assumptions are being made implicitly about the equivalence of the measure across groups: (a) the measure evokes the same conceptual frame of reference in both groups and (b) the association between the measure and some other criterion thought to have a conceptual relationship to the variable is similar across groups. For example, cross-group similarities in the relation of a measure of stress to a criterion measure of anxiety supports the equivalence of the stress measure. Berry (1980) and Brislin (1993) assert that equivalence is a prerequisite condition when using an assessment tool developed in one culture in a different culture.

Test bias occurs when “an existing test does not measure the equivalent underlying psychological construct in a new group or culture as the test measured within the original group in which it was standardized” (Allen & Walsh, 2000). Knight and Hill (1998) point out that cultural bias, a form of test bias, may lead to misinformation when unintended systematic variance is present due to factors that vary across cultures or subcultures. Moreover, in that case, it is difficult to determine whether observable group differences on scores are attributable to true differences in psychological processes or to measurement issues (Knight and Hill, 1998).

Vandenberg and Lance (2000) and Woehr, Arciniega, and Fowler (2003) argue that the assumptions about equivalence are actually testable. The next section includes a description of some of the strategies for assessing the equivalence of measures across groups.

The sections below describe both empirical and nonempirical strategies that may be useful for conducting substance abuse research. The objective is not to suggest that a conscientious investigator must use all of these approaches in any one study. Rather, the objective is to present information so the researcher can select the most appropriate strategies for a particular study.

4.1.1. Empirical strategies for assessing equivalence

Historically, researchers have used empirical techniques associated with classical test theory (CTT) such as reliability and validity to evaluate a measure (Crocker & Algina, 1986). However, violations of measurement equivalence may jeopardize accurate interpretation just as much as problems associated with reliability or validity (Vandenberg & Lance, 2000). Therefore, researchers with either ethnic minority samples or diverse samples that are including more than one ethnic group in a study may find it useful to include analyses that directly address whether the measure is operating similarly across groups. Clinical trials may be rich with opportunities for conducting equivalence studies because the sample sizes may be larger and may include sufficient numbers of ethnic minority participants.

Vandenberg and Lance (2000) conducted a review of 14 projects that used confirmatory factor analyses (CFAs) to test for different types of equivalence. Most used CFA to assess the adequacy of the factor structure and then conducted a multigroup analysis to compare the change in fit in a series of nested models as constraints to additional parameters were added (e.g., similar factor loadings) to a model in which no constraints were imposed on that specific parameter. The two groups are considered to be similar on a particular parameter if adding the constraint does not diminish the fit of the model.

The test of the equality of covariance matrices assesses whether the relationships among the measured variables (e.g., items) of a scale are similar across groups. One approach to examining the equality of covariance matrices is to compare the pattern of relationships among items. Vandenberg and Lance (2000, p. 17) suggest that similar covariance matrices across groups make a strong case for measurement equivalence. However, Woehr et al. (2003) point out that it is conceptually possible that meaningful group differences are present on other parameters even if the covariance matrices are similar. This is analogous to the situation in which a multivariate test of several outcomes fails to reject the null, but nevertheless, several of the individual dependent measures do show significance and the null would have been rejected had the test been more specific than an omnibus multivariate test. Therefore, Woehr et al. are not convinced that equivalent covariance matrices across groups alone are sufficient evidence of measurement equivalence. Rather, they advocate for conducting additional tests of equivalence even if the analyses suggest similar covariance matrices across groups. Byrne (1998) agrees with Woehr et al. regarding the limited utility of the test of the equality of covariance matrices because subsequent tests of specific parameters have demonstrated noninvariance when the omnibus tests suggested equivalence and the reverse. Therefore, our discussion of the assessment of equivalence will describe commonly used tests other than the test of the equality of covariance matrices.

Vandenberg and Lance (2000) described five measurement invariance tests and three structural invariance tests in the review mentioned above. They add that there are other equivalence tests beyond these eight. However, it would be unusual to conduct even all of the eight invariance tests described in the article. The general consensus is that configural (invariant factor structure) is the most serious source of measurement nonequivalence.

4.1.2. Test of configural equivalence: is the factor structure the same across groups?

Configural equivalence is essential to measurement equivalence (Woehr et al., 2003). Tests of configural equivalence examine whether the factor structure of a measure is equivalent across groups. The first example below is a study in which the configural equivalence of an attention deficit/hyperactivity disorder (ADHD)-related measure, a measure sometimes included in substance abuse research, was tested by comparing the factor structure of African Americans and non-Hispanic Caucasians.

The configural equivalence of the IOWA Conners Rating Scale for African American and non-Hispanic Caucasian children

Reid, Casat, Norton, Anastopoulos, and Temple (2001) examined whether the factor structure of the IOWA Conners Rating Scale, a measure of ADHD behaviors, was similar for African American and non-Hispanic Caucasian children. Separate confirmatory factor analyses were conducted for the African American and non-Hispanic Caucasian boys. The fit indices indicated that a two-factor model was a well-fitting model for both African American and non-Hispanic Caucasian boys. Next, the researchers analyzed the fit of a multigroup model that assumed the same factor structure for African American and non-Hispanic Caucasian boys. The findings suggested that the factor structure was similar for African American and non-Hispanic Caucasian boys. Similar analyses were conducted for the girls.

A study by Crockett, Randall, Shen, Russell, and Driscoll (2005) illustrates a way to proceed if the findings do not reveal configural equivalence. The variable in that study, depressive symptoms, is also used frequently in substance abuse research.

The configural equivalence of the Center for Epidemiologic Studies Depression Scale among Anglo Americans and Mexican Americans

Crockett et al. (2005) used CFA to examine whether the factor structure of the Center for Epidemiologic Studies Depression Scale (CES-D) was similar to the factor structure in the original sample by Radloff (1977) in each of four groups of youth, Anglo American, Mexican American, Cuban, and Puerto Rican, participating in the Add Health national study (Udry, 1998). Their first step was to examine the fit of the original four-factor structure (i.e., negative, positive, somatic, and interpersonal) within each group rather than in a multigroup analysis. The fit indicators revealed a good fit for the four factor solution for Anglo Americans and Mexican Americans but not for the Cuban or Puerto Rican groups. Next, additional analyses were conducted on the two groups (i.e., Cubans and Puerto Ricans) for whom the original four-factor structure was not a good fit. First, the researchers assessed whether differences in acculturation may explain the poor fit for those two groups. Since no actual measure of acculturation was available in the data set, generational status was used in a multigroup CFA as a proxy of acculturation to see if the four factor solution might fit for second and higher but not first-generation adolescents. However, the analyses did not support the hypothesis of generational differences in the fit of the four-factor solution. Next, CFA was used to test the fit of a three-factor solution reported elsewhere in the literature for Latino adults. That model also yielded a poor fit. Finally, exploratory factor analysis was used to generate an alternative model for Cuban American and Puerto Rican youth. An alternative four-factor solution emerged that demonstrated a better fit for Puerto Rican adolescents and a five-factor solution for Cuban American adolescents. These findings suggest that any analyses that focus specifically on either of these two ethnic groups should consider the alternative factor structure.

Meredith (1993) proposed a classification system for labeling the level of equivalence once configural equivalence has been established. The levels of equivalence in the Meredith classification system include weak, strong, and strict equivalence. These types of equivalence are discussed and illustrated in the sections that follow.

4.1.3. Establishing weak equivalence

Meredith (1993) suggests that a measure has weak invariance across groups if the factor loadings are the same across groups. Vandenberg and Lance (2000) and Woehr et al., 2003 refer to invariant factor loadings as metric invariance. The study by Crockett et al. (2005) described above also provides an illustration of metric invariance.

The metric equivalence of the CES-D among Anglo Americans and Mexican Americans

Since configural equivalence is essential to establishing measurement equivalence, Crockett et al. (2005) only conducted tests of metric invariance in the two ethnic groups (Anglo Americans and Mexican American) that demonstrated configural equivalence. The configural model might be conceptualized as the baseline model. The change in fit of a second multigroup model that constrained the factor loadings to be equal across ethnic groups was examined next. Chi-square statistics and other fit indices indicated that adding the constraint of similar factor loadings diminished the fit. However, follow-up tests suggested partial invariance. Therefore, the researchers tested for partial invariance by sequentially freeing the specific factor loadings with the largest group differences. After freeing the factor loadings for three items, subsequent analyses revealed that freeing additional factor loadings would not improve the fit any further. The findings suggested that three CES-D items (i.e., "thought your life was a failure," "felt fearful," and "enjoyed life") may account for group differences in the factor loadings. This finding would suggest that these three items are operating differently in the two ethnic groups. A research team using the CES-D may then want to adjust for those differences in samples that include both Anglo Americans and Mexican Americans.

4.1.4. Establishing strong equivalence

Strong equivalence is supported if the item intercepts along with the factor loadings are invariant across groups. The item intercepts are the mean item score in the sample when the value of the latent construct is zero. Conceptually, this is the mean level of endorsement of the particular item that would exist within the ethnic group if the ethnic group showed a score of zero on the latent factor. Vandenberg and Lance (2000) and Woehr et al. (2003) refer to invariance on the item intercepts as scalar equivalence. The absence of scalar equivalence may not be as serious a problem as the absence of configural or metric invariance (Woehr et al., 2003). However, the absence of scalar equivalence would have implications for normative cutoffs based on simple sums of the means. An example of scalar equivalence is provided later in the section on Item response theory.

4.1.5. Establishing strict equivalence

Meredith's criteria for strict equivalence includes invariant error variance along with invariant factor loadings (metric equivalence) and invariant item intercepts (scalar equivalence; Meredith, 1993). Vandenberg and Lance (2000) refer to the test of error invariance as a test of invariant uniqueness. In this case, the mean item scores across ethnicities when the latent factor is zero are similar, and the variability in the items less than the intercept and the loading multiplied by the value of the individual's latent factor are similar across ethnicities. The study by Widaman and Reise (1997) illustrates the examination of strict equivalence on a smoking measure.

The strict equivalence of a smoking scale among males and females

Widaman and Riese (1997) addressed the extent to which a scale of attitudes and behaviors about smoking demonstrated invariance across gender groups participating in the 1993 Monitoring the Future Survey. Eleven items were assumed to be associated with four latent variables (perceived coolness of peer smokers, perceived insecurity of peer smokers, attitudes toward smoking, smoking behavior). The research team used a multigroup CFA model to establish similar factor structure (configural equivalence). Weak and strong equivalence were demonstrated next by demonstrating that the fit did not decrease by adding constraints to the factor loadings and then the item mean intercepts in follow-up multigroup models. However, the difference in statistical fit did decrease significantly when constraints to the error terms were added to the model. The research team concluded that the measure had weak and strong but not strict factorial invariance.

The absence of invariant uniqueness suggests that the reliabilities differ across groups. If nonequivalence is present, the researcher can compare the coefficient alphas for each sample to determine the factors with group differences on reliabilities.

4.1.6. Other empirical tests of equivalence

4.1.6.1. Other CFA tests. Woehr et al. (2003) and Vandenberg and Lance (2000) describe other CFA tests of measurement equivalence beyond the ones described above. Three of the more common are the tests of invariant factor variance, means, and covariances. Equivalence on tests of invariant factor variances suggests that the distributions on the factor scores are the same across groups. Nonequivalence on tests of invariant factor variances may occur if the range of scores on the factors differs across groups. Nonequivalence on this parameter could suggest that one group has a more restricted range of responses than the other group. Cultural differences in response style may provide one explanation for nonequivalence on tests of invariant factor variances, Two examples from the literature may be helpful for illustrating the source of this type of invariance. Bachman and O'Malley (1984) found that African American youth were more likely than non-Hispanic Caucasian youth to use the extreme response options (strongly agree or strongly disagree instead of agree or disagree) on survey-type items. Similarly, it has been reported that the tendency to say yes or to indicate agreement within some Latino cultures is so prominent that a term actually exists for it—*si-ismo*. Clearly, cultural differences in response styles may account for what may appear to be ethnic group differences on self-report measures. Such differences on the items may lead to scalar nonequivalence as well as to the lack of equivalence on the factor means and variances. If the CFA findings suggest measurement nonequivalence on factor variances, the researcher can follow up by conducting a Levene test on the factor scores to identify the factors that account for the group differences in variance.

Nonequivalence on tests of invariant factor means suggests that two groups differ on the mean factor scores. In this case, the researcher might address a poor fit by conducting a *t* test of the factor score means to identify the factor means with group differences. Finally, nonequivalence on tests of invariant factor covariances suggests that the groups differ on the relationships between factors. If the CFA suggests a poor fit on this parameter, the researcher can calculate the correlations among the factor scores to determine which correlations differ across groups.

4.1.6.2. Item response theory. Item response theory (IRT) is another approach to assessing equivalence. Simply stated, IRT matches the two groups statistically on the underlying trait and then examines whether the relationship between the items and the underlying trait is similar across groups. Cooke, Kosson, and Michle (2001) argue that IRT may make a more convincing case than CFA for scalar equivalence because IRT considers not just group similarities in item means but also whether group differences are present in the relation of specific measures (e.g., items) to the underlying trait. IRT accomplishes this by specifying the full distribution in the process of specifying the relationship between items and traits. In particular, IRT uses a distribution to describe the categorical or ordered nature of items and how that maps onto the underlying trait. Distributions for this type of data typically jointly estimate the mean and variance (the variance is a function of the mean). Therefore, IRT is really combining the idea of scalar invariance with aspects of item uniqueness invariance (holding constant the level of the latent trait).

The distributional approach of IRT causes it to examine aspects of the shape of the distribution that relate the item to the underlying trait. Two issues related to the performance of items across groups are relevant. First, extremity refers to the score on the overall trait above which a specific item is likely to be endorsed. This is sometimes referred to as the maximum inflection point. A difference in extremity (i.e., inequivalence) would be suggested if the item is likely to be endorsed with just an average score on the overall trait in one ethnic group but is unlikely to be endorsed in the other ethnic group unless a high score is obtained. Second, the steeper the slope, the more discriminating (i.e., relevant) the item is for determining the score on the overall trait. Differential item functioning is said to occur when the plotting of the relationship of the item to the overall trait (i.e., the Item Characteristic Curve) suggests that an item is more discriminating or more extreme in one ethnic group than in another.

IRT studies of equivalence are similar to CFA in the use of a multigroup approach that compares the goodness of fit of a model that imposes constraints on a parameter to a model without those constraints. A chi-square is used to evaluate the difference in the G^2 statistic (ΔG^2) for the slope and extremity parameters. Group differences in slope suggest that the item may be more discriminating for one ethnic group than another. The potential bias suggested by differences in slope is more serious than differences in extremity.

The study below by Cooke et al. (2001) illustrates the use of IRT for assessing scalar equivalence on the Psychopathy Checklist—Revised. Antisocial personality disorder (APD) is frequently included in substance abuse research. The potential negative impact of APD on substance abuse treatment outcome has been discussed elsewhere (Alterman, Cacciola, & Rutherford, 1993).

The use of IRT to examine the scalar equivalence of the Psychopathy Checklist—Revised between African Americans and non-Hispanic Caucasians

Cooke et al. (2001) were interested in determining the scalar equivalence of the Psychopathy Checklist-Revised (PCL-R) across African Americans and non-Hispanic Caucasians. They argued that psychopathy might be conceptualized differently within the two groups perhaps because African American psychopaths may not share the appraisal and response modulation deficits evident in prior research on White psychopaths. Since configural equivalence is a necessary but not sufficient condition for establishing scalar equivalence, CFA analyses were conducted first that revealed similar factor structures. Similarly, metric equivalence was demonstrated in a second CFA model in which the factor loadings were constrained along with other parameters. To assess scalar equivalence directly, two constraints were imposed in the IRT, equal slopes and equal extremity scores. The fact that imposing these two constraints did not change the G^2 significantly supported equivalence.

If a significant G^2 suggests group differences in the slope, then it is possible to identify the item(s) responsible for the ethnic group differences by adding the item constraints one at a time, beginning with the item with the smallest difference in slope. Researchers could then decide how best to address those differences (e.g., remove the items, transform the items) in subsequent analyses.

Our focus is on CFA rather than IRT because more guidance is available on using CFA on multifactorial models (Raju, Laffitte, & Byrne, 2002). However, those interested in reading more about the use of IRT to examine equivalence are encouraged to review a second article by the Hare team (Bolt, Hare, Vitale, & Newman, 2004), which includes an excellent description of the underlying concepts associated with IRT along with a demonstration of applying IRT in a multigroup study.

4.1.6.3. Functional equivalence. A test for functional equivalence is similar to the test of invariant factor covariance. However, the question in functional equivalence is whether the groups differ on the relationship of the scale or factor to some measure that is external to the scale. A measure is said to have functional equivalence if it has similar antecedents, correlates, and precursors across groups. Miyamoto, Hishinuma, & Nishimura et al. (2001) assessed the functional equivalence of a measure of self-esteem, a variable commonly included in research on substance abuse (Malcolm, 2004; Otsuki, 2003). However, the study described below is somewhat different from the ones reviewed earlier because it illustrates the use of regression to test for functional equivalence.

The functional equivalence of the Rosenberg Self-Esteem Scale for Hawaiian and non-Hawaiian males and females

The goal of this study was to assess the equivalence of the Rosenberg Self-Esteem Scale (RSES) among four ethnic/gender groups of adolescents: Hawaiian (including part Hawaiian) males and females, non-Hawaiian (non-Hispanic Caucasian, Filipino, Japanese, Hispanic, and mixed/two or more) males and females attending school in Hawaii. The RSES, one of the most popular measures of self-esteem, was developed in 1965 with an original sample that was primarily non-Hispanic Caucasian and later with an African American sample. Although Miyamoto et al. assessed other types of equivalence as well, only their analyses aimed at assessing functional equivalence are discussed here. Equivalence is supported if the relation of the RSES to some external (other) measure is the same across groups. The Major Life Events Checklist (MLE; Andrews, Lewinsohn, & Hops, 1993) was selected as the external predictor. First, Miyamoto et al, conducted separate regression analyses for each of the four groups. The predictor variables were the scales of the MLE, and the outcome variable was RSES for each regression. Similar regression slopes across groups supported equivalence. Then, Miyamoto combined all groups for the next regression analysis. The analysis conducted with the combined sample examined whether gender or ethnicity altered the relationship between RSES and the MLE predictors by adding both MLE × Gender and MLE × Ethnicity interaction term to the regression model. Similar findings on the separate regressions as well as the finding that the interaction terms did not contribute to predicting RSES were used to suggest that the regression slope linking MLE to RSES was similar across groups.

Two points about this approach are noteworthy. First, although this team opted to use regression to test for functional equivalence, they could have conducted CFA in Structural Equation Modeling to test these same relationships. Second, a challenge of this particular approach is to identify an external variable with demonstrated equivalence across groups. Otherwise, it would not be possible to determine if group differences in the relationships are due to group differences in the measure being assessed or to group differences on the external variable.

Although the main purpose of a clinical trial might be to test the efficacy of the intervention, secondary analyses that conduct the tests described above may provide valuable information on differences in the way in which various measures operate across groups. Researchers may advance the field by publishing separate articles that report the equivalence or nonequivalence of a measure across groups.

4.1.7. Nonempirical strategies

The strategies discussed so far are empirical tests for establishing the equivalence of a measure. However, nonempirical strategies may be useful to the researcher as well. Two nonempirical strategies are described in this section: the use of focus groups, expert panels, or pilot studies and the use of strategies for establishing translation equivalence.

4.1.7.1. Focus groups, expert panels, and pilot studies. It is possible to gather some very useful information during the planning stages regarding the equivalence of measures just by asking representatives of a group to critique or pilot test the measure. Focus groups (Knight & Hill, 1998, Ramirez et al., 2005), expert panels (Ramirez et al., 2005), and pilot studies (Beauvais and Trimble, 1992) are both useful for this purpose. Obviously, pilot studies could be empirical in nature. However, we opted to include pilot studies in the set of nonempirical strategies to emphasize that much information can be obtained even if the researcher does not subject the data to empirical analysis. For example, Beauvais and Trimble (1992) described how an item as innocent as "Other boys like to play with me" had a double meaning among a specific group of young people. The researchers only recognized this problem when they received some unprintable responses! Focus groups, expert panels, or pilot studies may be particularly helpful for detecting such subtle differences in meaning earlier.

4.1.7.2. Translation equivalence. It has become quite common for researchers to translate measures into the preferred language of the respondents. The developers of the BSFT went one step further by placing the English translation right next to the Spanish on those measures to enhance understanding among bilingual participants who may be more comfortable with certain words or phrases in English but other content in English. However, translation equivalence is essential to such procedures. Brislin indicated that translation equivalence is present when the scale is accurately translated into another language in a manner that promotes linguistic equivalence. Allowing members of groups who may have English as a second language to select the language for any paper-and-pencil instruments is the preferred strategy. However, it is important that the versions of the measure be identical. Butcher (1996) described a four-step process used to translate the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) and other measures into different languages. These steps include the following: (a) using two or more translators who are knowledgeable about both languages to translate the measure independently, the translators then collaborate on any differences to decide on a final version; (b) use of an additional translator to translate the measure back to its original language; (c) a comparison of the back-translated version with the original version; (d) field testing of the translated version to determine the response to the measure in the translated language. In addition to these four essential steps, Butcher (1996) also pointed out that some additional steps may be taken including (a) a comparative study of the adequacy of the American norms; (b) the development of new norms specific to the second culture; and (c) research with the new version to test its validity with the targeted population.

Translation (linguistic) equivalence may even be an issue when an ethnic group speaks the dominant language because of cultural differences in idiomatic nuances (Okazaki and Sue, 1995; Ramirez et al., 2005). For example, some American Indians and Alaska Natives may only have been exposed to reservation English (Allen, 1998). Similarly, Dunnigan, McNall, and Mortimer (1993) described a group of Laotian Hmong adolescents who spoke sufficient English to complete their academic requirements. Nevertheless, they interpreted the common meanings of words and metaphors so differently that it shaped their responses on psychological measures. Accordingly, the adequacy of standardized instruments used to assess their mental health was questionable along with the measurement equivalence across groups. Moreover, it is important to note that different cultural groups may still have some linguistic differences although they speak the same language. For example, Spanish-speaking groups may vary in the Spanish terms they use. Hence, it is important to be sure that the translation is appropriate for the particular Spanish-speaking group who will be using the Spanish version.

4.2. Issues in using norms and cutoff scores developed on other samples

The use of scores derived from a standardization sample of primarily non-Hispanic Caucasians to interpret the scores of ethnic minority participants in a separate study is another measurement concern. The appropriateness of norms and cutoff scores derived on one group may need to be examined before being applied to a different group. This issue is related to our earlier discussion of measurement equivalence. That is, using the data from one group to interpret the scores of another ethnic group may be a particular concern if the scalar, factor mean, or factor variance equivalence on the measure have not been demonstrated for the two groups.

Typically, clinical trials for substance abuse research do not use norms to assess outcomes. However, there are several reasons to include a discussion on the use of norms in a discussion of substance abuse research with ethnic minorities. First, multisite clinical trials may yield ideal data sets for assessing the appropriateness of applying cutoff scores to a second ethnic group. Second, it would not be inconceivable to use scores derived from norms as an inclusion/exclusion criterion or to determine the proportion of participants who are no longer in the clinical range on an outcome measure. Moreover, clinical trials provide a rich opportunity to examine whether cutoff scores based upon norms derived from samples of other groups are appropriate for a specific ethnic minority group. Therefore, the use of norms and cutoffs may be a relevant discussion for clinical trials.

The MMPI-2 is a good example of a measure for which the interpretation of the score is based on norms. Before illustrating the issues associated with the use of norms derived from samples of other groups to set cutoffs, it may be helpful to provide some background on the development of the MMPI-2 norms. The representativeness of the original standardization sample has been challenged because it consisted largely of non-Hispanic Caucasians residing in or near Minnesota. The developers addressed that concern in the MMPI-2 standardization sample by being careful to include ethnic minorities in proportion to their representation in the population based upon the 1990 census data. It is not unusual for scale developers to rely on a similar strategy to enable them to make the case that the measure can be used universally. In the case of the MMPI-2, this approach meant that various ethnic minorities were included in the standardization sample but in small numbers. For example, the male MMPI-2 standardization sample included 933 non-Hispanic Caucasians (82%), 35 Hispanics (3.1%), 126 African Americans (11.1%), 38 American Indian and Alaska Natives (3.3%), but only 6 Asian Americans (.5%). One obvious concern is whether norms based on a sample with so few Asian Americans, Hispanics, and American Indians and Alaska Natives can be validly used to interpret the scores of members of those ethnic groups. Two examples that illustrate different issues are described below.

4.2.1. Example 1

A hypothetical example with the MMPI-2 might be useful for illustrating the potential shortcoming of applying norms that were primarily derived from one group to another ethnic group. Suppose a researcher designed a substance abuse intervention for males who are not yet demonstrating any evidence of internal distress. To operationalize internal distress, they decide to use a clinically significant ($T \geq 65$) score on the Psychathenia (Pt) scale of the MMPI-2 as an exclusion criterion. An Asian American male is screened with the MMPI-2 for potential participation, and his raw score on the Pt scale is 21.

The means for the non-Hispanic Caucasian and Asian American males in the standardization sample were fairly different on several scales including the Pt scale (Hathaway and McKinley, 1989; see Table H-1). To keep the example simple, let us ignore the other ethnic groups for the moment and imagine that the standardization sample only included the 933 non-Hispanic Caucasians (M = 11.04) and the 6 Asian Americans (M = 14.33) in the MMPI-2 standardization sample. In that event, the overall mean would have been 11.06.[1] The obvious first point here is that the overall mean used to calculate the uniform T score would be much closer (i.e., almost identical) to the mean for the non-Hispanic Caucasian group because a much higher proportion of the sample was non-Hispanic Caucasian.

Next, imagine that the MMPI-2 had been normed in a region of the world that was largely Asian American to the point where it would be feasible to have the reverse proportions, 6 non-Hispanic Caucasians and 933 Asian Americans. In that event, assuming the same group means, the corresponding overall mean would have been 14.30, almost identical to the mean for the Asian American sample.

If the mean of the first hypothetical standardization sample that consists of predominately non-Hispanic Caucasians (overall mean = 11.06) is used to establish the cutoff, then the T score associated with the potential Asian American participant's raw score of 21 would be approxi-

[1] The actual mean for the male MMPI-2 is 11.24 on the Pt scale. The difference is due to the inclusion of other ethnic groups in the standardization sample not just non-Hispanic Caucasians and Asian Americans. Nevertheless, the heavy influence of the Caucasian sample of 83.7% is still clearly evident in the actual overall mean.

[2] For the purpose of this example, we used the formula provided by Hathaway and McKinley (1991) to calculate the T score. That formula is $T = 50 + [10\ (X - M)]\ /\ SD$, where X equals the raw score. We used the standard deviation of the majority group (Caucasians).

mately 66.[2] Since his score is above the cutoff for clinical significance (T = 65), he would be excluded from participation. However, if the potential participant's T score were derived from a sample of individuals primarily from his own ethnic group (the sample that is predominately Asian American with a mean of 14.30), then his T score would only be 60 and he would be eligible for participation.[3]

4.2.2. Example 2

One might argue that there are legitimate reasons why ethnic minorities may have higher (more pathological) scores on the attributes measured by the MMPI-2. Hence, it may be expected that their scores are higher on certain scales. However, it is not uncommon for the mean scores of an ethnic minority group to be lower than the dominant group even on the MMPI-2. In fact, Asian American females have lower scores (M = 10.00) than non-Hispanic Caucasian females (M = 12.27) on the Pt scale and six other clinical scales of the MMPI-2. Example 2 illustrates the problems that can occur in that scenario when using norms standardized on another group. Suppose we are again using the same cutoff score for clinical significance (T = 65) to exclude females from participating in the same substance abuse intervention, a potential Asian American female participant has the same raw score (21) as her male counterpart in the first example. Again, let us ignore the other ethnic groups for this example and assume that the standardization sample only included the 1,184 non-Hispanic Caucasians (M = 12.27) and 13 Asian Americans (M = 10.00) in the actual MMPI-2 female standardization sample. The combined mean of 12.25 would be very similar to the non-Hispanic Caucasian mean for females. The T score associated with the hypothetical participant's raw score of 21 would be 63, and the female participant would meet the criterion for participation.[4] However, suppose the scale had been normed in a region of the world where it would be appropriate to include 1,184 Asian Americans and 13 non-Hispanic Caucasian females. In that case, the overall mean (10.02) would be much closer to the Asian female mean.[5] When compared with her own cultural group, the T score associated with her raw score of 21 (T = 72) would have been above the clinical cutoff of T = 65, and the female participant would have been excluded.

These two examples illustrate several important points. First, including a small, albeit proportionate, representation of an ethnic minority group in the standardization sample does little mathematically to change the overall norms from the norms for the majority group. Consequently, if the resulting scale mean is then used to interpret the scores universally, in effect, the norms for the majority group are being used to interpret the scores for the ethnic minority group. Second, the use of norms developed on another cultural group to interpret an individual's scores runs the risk of either labeling individuals as *deviant* when, in fact, their scores are quite similar to other members of their own ethnic group (Example 1) or labeling other individuals as *normal* when their scores are similar to the normative sample for another ethnic group but very different from the scores of other members of their own ethnic group (Example 2).

This issue is certainly not limited to the MMPI-2. Instead, the same issue may apply to other variables frequently included in substance abuse treatment studies such as cognitive functioning. For example, an issue of *The Clinical Neuropsychologist* (volume 19, issue 3/4) is devoted to the issue of the use of normative data in interpreting the cognitive functioning of older African Americans.

Obviously, one strategy for addressing this issue is to establish alternative cutoff scores for specific ethnic groups. The advantages and disadvantages of alternative norms and cutoff scores have been debated elsewhere (Manly, 2005). It may be outside of the scope of most clinical trials to establish alternative norms for specific cultural groups. However, even if the original researcher does not have the time or resources to examine whether the cutoff scores published in the literature are appropriate for a specific ethnic minority group, the data collected in rigorous multisite clinical trials may be ideal for other researchers to conduct and publish secondary analyses that examine the appropriateness of published clinical cutoffs for specific ethnic groups. IRT as well as receiver operator characteristics may be useful techniques for those researchers who decide to take on this daunting challenge.

This issue promises to become more critical as the U.S. population becomes more diverse. Ironically, this issue should not just be viewed as a concern that is only relevant to ethnic minority group members. Rather, if the predictions regarding the increases in the proportion of the U.S. population that are ethnic minority are realized, then it would certainly be conceivable that pockets of the United States would be predominately ethnic minority in the future. In those situations, even non-Hispanic Caucasians could eventually be in the unenviable situation of being judged by norms developed primarily on other groups.

4.3. The effects of the ethnicity of the interviewer and the ethnicity of the coder on measurement

Previous research suggests that another measurement issue may be the effects of the ethnicity of both the interviewer and the rater on the data. A very early study on this issue by Hatchett and Schuman (1975) revealed that respondent concerns about offending an interviewer from another race may jeopardize the accuracy of the information obtained in the interview. However, the conclusions from studies since then have been inconsistent. The findings from

[3] The same formula described above was used to calculate this T score. However, the standard deviation of 7.15 provided by Hathaway and McKinley (1991) for the Asian male sample on the Pt was used.

[4] The SD for the Caucasian female sample (6.89) was used to calculate this T score.

[5] The SD for the Asian female sample (5.03) was used for this example.

several earlier studies suggest that ethnic differences between the respondent and the interviewer primarily matter when the content is race related (Anderson, Silver, & Abramson, 1988; Cotter, Cohen, & Couter, 1982; Finkel, Guterbock, & Borg, 1991). However, other researchers have reported the effects of the ethnicity of the interviewer in content areas that did not appear to be race related such as willingness to share a secret with an interviewer (Dunkerley, 1997), willingness to disclose a history of child abuse (Dunkerley & Dalenberg, 1999), and the type of stories generated on a projective task (McClelland, 1974). Multisite clinical trials represent a special unique opportunity to conduct secondary analyses that examine the relation of the ethnicity of the interviewer or rater to the findings.

The use of coders or raters raises issues regarding the role of race or ethnicity in coder ratings. Several studies in both social and clinical psychology are useful for understanding how ethnicity may influence coder ratings. The first is a study on eyewitness identification by Hutchens (2003) in which participants were asked to identify a stranger whom they had only seen once. Both African American and non-Hispanic Caucasian participants more accurately identified a stranger of the same versus a different race. A second group of studies is on person perception. A study by Eberhardt, Goff, Purdie, and Davies (2004) demonstrated the role of ethnicity in influencing how individuals interpret a target's role in a social situation (Eberhardt et al., 2004). In addition, Weathers, Frank, and Spell (2002) found that individuals perceive the emotions of members of their own ethnic group more accurately than they perceive the emotions of members of another ethnic group. This pattern has been evident when individuals are asked to interpret emotions based upon facial expressions or other prosodic (i.e., pitch, loudness, and rate of speech cues; Weathers et al, 2002) or other nonverbal cues (Bailey, Nowicki, & Cole, 1998). Another study demonstrates how ethnicity influences clinical diagnosis. Similarly, Reid et al. (2001) examined teacher ratings of African American and non-Hispanic Caucasian schoolchildren on ADHD behaviors, another set of behaviors commonly studied in substance abuse treatment or prevention studies. Both African American and non-Hispanic Caucasian teachers rated the children on the IOWA Connors. The findings revealed that teachers from the same ethnic group as the student rated the child lower in ADHD behaviors than teachers from the other ethnic group. That is, African American children were rated higher on ADHD behaviors by non-Hispanic Caucasian teachers. Similarly, non-Hispanic Caucasian students were rated higher on ADHD behaviors by African American teachers. Loring and Powell (1988) reported similar findings in clinical ratings. They presented the same written case scenario to four different groups of psychiatrists. One group was told the patient was an African American male. The other three groups were told that the patient was either a non-Hispanic Caucasian male or female or an African American female. The psychiatrists were to use the information provided to render a diagnosis. The case was more likely to be diagnosed as paranoid schizophrenic when described as an African American male than when the same case was described as any of the other groups.

The same ethnic biases evident in eyewitness identification, person perception, and clinical diagnosis may influence coder ratings of individual or family variables that may be included in substance abuse research studies as well. In a study by Gonzales, Cauce, and Mason (1996), African American or non-Hispanic Caucasian coders were asked to observe videotaped interactions of African American mothers and daughters. Their results demonstrated that observers who were from the same ethnic group as their targets (African Americans in this case) rated the parent–child interactions as less controlling and conflictual than ethnically different (non-Hispanic Caucasian) observers. Moreover, the ratings by the observers from the same ethnic group as their targets (African American raters) were more similar to how the African American mothers and daughters perceived their interactions themselves than the ratings of ethnically different raters (non-Hispanic Caucasian raters). Interestingly, all observers participated in the same intense training to educate them on styles of parent–child interactions within African American families. The authors speculate that the racial differences between observers may have been even more apparent without this training.

The point here is not to label ratings by one group as better or more accurate than ratings by the other group. The point is that researchers may contribute to existing research by considering the ethnicity of the interviewer or the coder rather than by disregarding it. Generalizability Theory represents one promising approach for examining the effects of the ethnicity of the interviewer or the rater. Specifically, the approach builds upon CTT (obtained score = true score + error) by identifying and estimating the sources of error that contribute to the obtained scores (Strube, 2000). Generalizability Theory is particularly appropriate for observational coding because it offers an approach to assess the various sources of measurement error such as the ethnicity of the target or coder that may affect the observed scores (Hoyt & Melby, 1999).

It is not uncommon for family treatment studies of substance-abusing adolescents to include observer ratings of family interaction patterns. For example, Jose Szapocznik is currently leading a study in the NIDA CTN (0014) that is evaluating the efficacy of BSFT. The outcomes include both substance abuse and improvement in family functioning. Varda Shoham (R01 DA017539) is conducting a platform study of the same BSFT study that trains coders to use the Structural Family Systems Rating system to assess family interaction patterns of both the BSFT and treatment-as-usual groups at the beginning and end of BSFT treatment. Based on a broad body of research suggesting that the ethnicity of the coder may influence the ratings of target persons or families, the team is planning a secondary study to examine this issue.

Larger clinical trials may provide a particularly rich arena to study these issues better. However, to do so, it is important for research teams to create the opportunity for secondary studies on this issue by recording the ethnicity of the interviewer and the coder routinely even if the research team has no immediate plans to examine race/ethnicity of the interviewer/rater in the design.

5. Appropriate analytic strategies for research on substance abuse treatment for ethnic minorities

In addition to measurement, valid analysis requires a careful plan to conduct the analyses in a manner that yields the best information about the efficacy of the intervention for ethnic minorities. The ideal approach of conducting studies that target sufficient numbers of a specific ethnic group and analyzing those data separately is not always feasible. Moreover, the inclusion of even sufficient numbers of ethnic minorities in the sample is not adequate if no analyses are conducted to understanding the outcomes for specific ethnic groups (Hall, 2001). For example, the review by Szapocznik et al. (2007), which was mentioned earlier, identified only two randomized clinical trials that specifically examined the efficacy of the substance abuse treatment for either Hispanic or African American adolescents, However, substantial numbers of Hispanics or African Americans were included in several other clinical trials that demonstrated the efficacy of other substance abuse treatments. Yet, because separate analyses were not conducted for the ethnic minority group, no strong conclusions could be made about the efficacy of the treatment for the specific ethnic minority group.

Other researchers address ethnicity by comparing ethnic minorities on outcomes or other variables. Such designs are frequently referred to as race-comparison designs. This section begins with a discussion of the pitfalls of a reliance on race-comparison designs and then describes alternatives to race-comparison designs. The section ends with a discussion of the pitfalls of combining ethnic groups for the analyses.

5.1. The pitfalls of reliance on race-comparison analytic strategies

The strategy of addressing ethnicity by comparing two or more ethnic groups on the outcome variable or some other variable is quite common. One review by Graham (1992) revealed that 72% of articles in American Psychological Association journals that included African Americans were race-comparison studies. This section discusses several concerns associated with a race-comparison analytic approach.

The first concern with a race-comparison approach is that this approach ignores the within-group differences for each ethnic group. As mentioned earlier, each of the four ethnic groups mentioned in the NIH mandate on the inclusion of women and ethnic minorities has important subgroups with meaningful cultural differences that may affect substance abuse. For example, research on Hispanic adolescents suggests that U.S.-born (compared with foreign-born) Hispanics (Vega et al., 2002) and Cuban Americans (compared to Mexican American or Puerto Rican; Delva et al., 2005) demonstrate higher levels of substance abuse.

Both the presence and absence of mean differences may result in inappropriate next steps or conclusions. The absence of mean differences may lead the researcher to feel justified in combining ethnic groups with the non-Hispanic Caucasian groups for subsequent analyses. However, that decision may cause the researcher to overlook some important findings (e.g., group differences in the pattern of relationships among variables) that would be more evident in a more comprehensive analysis of the data. The presence of mean differences may tempt the researcher to overinterpret the meaning of the findings. For example, Donald Campbell (cited in Gonzales et al., 1996) suggested that this analytic practice of using mean differences to compare groups may contribute to ethnic stereotyping.

The third concern is that a race-comparison design may be especially vulnerable to misleading conclusions if the measures are not equivalent across groups. Michaels, Barr, Roosa, and Knight (2007) argue that strong measurement equivalence (equal factor loadings and item mean intercepts) is a prerequisite for meaningful mean comparisons to be made across groups. Otherwise, as Vandenberg and Lance (2000) point out, “if one set of measures mean one thing to one group but something different to another group, a group mean comparison may be tantamount to comparing apples and spark plugs” (p. 9).

Even if the measure is equivalent across groups, a fourth concern for researchers is that many samples in substance abuse clinical trials are convenience samples. This reality raises two issues. First, convenience samples may not be representative of the groups from which they are drawn. For example, the population of treatment presenting drug-abusing African Americans may be very different from the population of drug-abusing African Americans and certainly different from the population of African Americans in this country. A second issue is that race/ethnicity may be a proxy for other differences between two samples. For example, previous research indicates that African Americans and non-Hispanic Caucasians who present for substance abuse treatment differ substantially on demographic, psychiatric, and type of drug use (Morgenstern & Bux, 2003; McKay, Lynch, Pettinat, & Shepard, 2003). Therefore, a focus on ethnic differences may mask the potential reality that the differences are attributable to demographic (e.g., gender, age, socioeconomic status, educational, geographic location [e.g., rural vs. urban]), social (e.g., family variables, cultural/racial experiences), or behavioral (e.g., type of drug used) differences.

The possibility of site effects in large, multisite, substance abuse treatment studies may represent still another concern (Wakim, 2006) with race-comparison designs. Site effects may reflect the demographic, social, or behavioral differences mentioned above. For example, if the Hispanics in a multisite trial were recruited from a low-income treatment center that served rural clients who misuse drug X but the non-Hispanic Caucasians were recruited from a treatment center that served suburban, middle class clientele that misuse drug Y, it may be difficult to separate statistically the effect of ethnicity from the effect of the differences in income, geographic location, or the type of drug used.

Of all the criticisms levied against race-comparison designs in clinical trials, perhaps the most important one is whether that approach addresses what the treatment and scientific communities want to know—does the intervention work for this group? In intervention studies, it is not unusual for researchers to be interested in whether the intervention works for a specific ethnic group but to design the study to address a different question—does it work as well for members of the ethnic group as for non-Hispanic Caucasians. In Fig. 1, suppose Group 1 is an ethnic minority group and Group 2 includes non-Hispanic Caucasians, the dashed line indicates the cutoff (a score of 40) between the abnormal (below the line) and the normal (above the line) ranges. The hypothetical design includes a treatment and control group for both ethnic groups. If the focus is on the difference between the two ethnic groups that received the treatment, it would be easy to overlook that the mean score for the members of the ethnic minority that received treatment is in the normal range, but the mean for the members of the ethnic minority control group is in the abnormal range. Although ethnic group differences are present between the two treatment groups, the most important finding may be that the mean scores for both groups who received the experimental treatment were in the normal range. The larger point here is to consider that the examination of group differences is only a preliminary step at best rather than an adequate examination of the role of ethnicity in the research project.

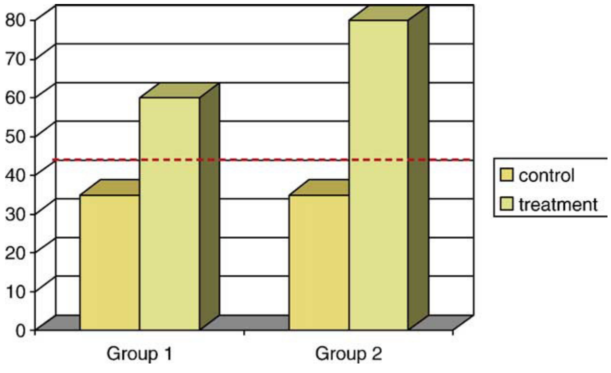

Fig. 1. Hypothetical illustration of the treatment results of two groups.

5.2. Alternatives to race-comparison analytic strategies

A number of important questions can be addressed in heterogeneous data sets other than race-comparison questions. Some of these questions are described below.

5.2.1. Examining the efficacy of the treatment for a specific ethnic minority group

As mentioned in the previous section, one of the most useful questions is whether the intervention works for a specific ethnic group. Unfortunately, several reports in the past decade reveal that few if any evaluations of treatment efficacy meet the criteria required to demonstrate their efficacy for ethnic minorities (Chambliss et al., 1996; Hall, 2001; Szapocznik et al., 2007). The common strategy of combining ethnic minorities with other groups for the analyses contributes substantially to this shortcoming. Moreover, Strada et al. (2006) found that only one study in their review of 18 different adolescent treatment studies examined the response to treatment for a specific ethnic group.

Sample size is a common barrier for addressing this issue particularly in secondary analyses of studies with diverse populations. The ideal strategy would be to have sufficient power to compare the outcomes of ethnic group members in the experimental and control conditions. Obviously, that will not always be possible. Hence, several options are worth considering. First, it may be possible to calculate the effect size even with small samples. In addition, it may be more feasible to design a study with a sufficient sample to examine treatment effects for one specific ethnic group than for all four ethnic groups identified in the NIH guidelines. Therefore, the researcher may want to target one ethnic group for a specific study, select community treatment programs accordingly, and power the study to consider treatment effects for that ethnic group as well as for the entire sample. In that event, it may be useful to include ethnicity as one of the variables for urn randomization to ensure that the ethnic minorities are distributed fairly evenly across conditions.

Even the NIH guidelines on the inclusion of women and minorities acknowledge that it may not be feasible to require every trial to “provide high statistical power” for all four ethnic groups (NIH, 2001, p. 4).

Instead, researchers may be forced to adjust their analyses of the findings for ethnic minorities to the sample sizes in their data sets. Therefore, researchers are referred to a book by Rick Hoyle (1999) titled *Statistical Strategies for Small Sample Research*. Doctor Hoyle asked a number of leading statisticians to write chapters describing potential statistical strategies for small samples of 150 or fewer cases. Their topics include multiple imputation, reduction of measurement error, effect size, bootstrapping, cross-tabulated data, structure analysis mediation, and factor analysis. Hoyle cautions that these strategies are certainly not perfect—one “cannot make a silk purse from a sow’s ear!” (Hoyle, 1999).

Nevertheless, imperfect strategies may still add to what can be examined from a particular data set.

5.2.2. Alternative within-group questions

An obvious drawback of race-comparison studies is that such studies ignore important within-group differences. Therefore, an alternative to racial comparison research is to design studies that consider within-group questions. At least three types of meaningful within-group issues are worthy of study within each ethnic group. The first type might address the diversity within various ethnic groups by examining how one's status on any of a number of characteristics might moderate relationships among the study variables. One potential moderator variable is subgroup membership. For example, the fact that the Census Bureau reported 562 federally recognized tribal nations, villages, and corporations among American Indian and Alaska Natives (Bureau of Indian Affairs, 2002) suggests that combining all these subgroups into one group for statistical analysis ignores the considerable information that could be acquired by examining the within-group diversity. Similarly, the finding reported by Turner, Lloyd, and Taylor (2006) of higher levels of substance abuse and delinquency among U.S.-born versus foreign born Hispanic youth suggests that nativity may be a second useful within-group characteristic. A third characteristic worthy of within-group study is acculturation or racial identity. Previous research suggests that both acculturation (Turner et al., 2006) and racial identity (Caldwell, Sellers, Bernat, & Zimmerman, 2006) may be associated with drug use and response to treatment (Want, Parham, Baker, & Sherman, 2004). A fourth important within-group characteristic is geographic location. For example, in a prevention study, Brody, McBride Murry, Kim, and Brown (2002) revealed that the environmental stressors facing low-income African American residents in rural areas may differ from the stressors that low-income African Americans in urban settings encounter, perhaps due to the restricted access to employment, transportation, public recreation, and medical or mental health care services in rural areas. These differences may impact family functioning and adolescent outcomes. Finally, education, socioeconomic status, and gender are other important characteristics that may be the basis of within-group studies.

In addition to moderator variable studies, a second type of within-group studies might address mediator relationships or the mechanisms that explain the relationship between one variable (e.g., treatment type) and another (e.g., outcome). Another study by Brody et al. (2006) provides a good example. Their findings revealed increased parental involvement as a mediator variable that explained why adolescents who participated with their families in the Strong African Americans Families Program, a substance abuse prevention program, had more positive outcomes than the adolescents in the control condition.

Studies of engagement and retention in treatment and other interventions represent a third important type of within-group questions. Those ethnic minorities who need substance abuse treatment the most may be less likely to participate. For example, according to the National Study on Drug Use and Health, only 15.3% of African Americans who need treatment actually receive it (SAMHSA, 2002). The higher dropout rates among those who do participate in substance abuse treatment may reflect the fact that fewer ethnic minorities in treatment describe the experience as helpful (Heron et al., 1997; Longshore et al., 1999). For all these reasons, it would be quite useful to design large-scale treatment outcome studies to include the examination of engagement and retention within specific ethnic minority groups. The work by Jose Szapocznik and his team at the Center for Family Studies at the University of Miami is a good model for work in this area (Coatsworth, Santisteban, McBride, & Szapocznik, 2001; Santisteban et al., 1996; Santisteban et al., 2005; Szapocznik et al., 1988) because they designed their substance abuse clinical trials studies to include engagement and retention.

5.2.3. Alternative between-group questions

The most common between-group questions are the race-comparison questions described earlier (e.g., does group X differ from group Y on a specific outcome variable).

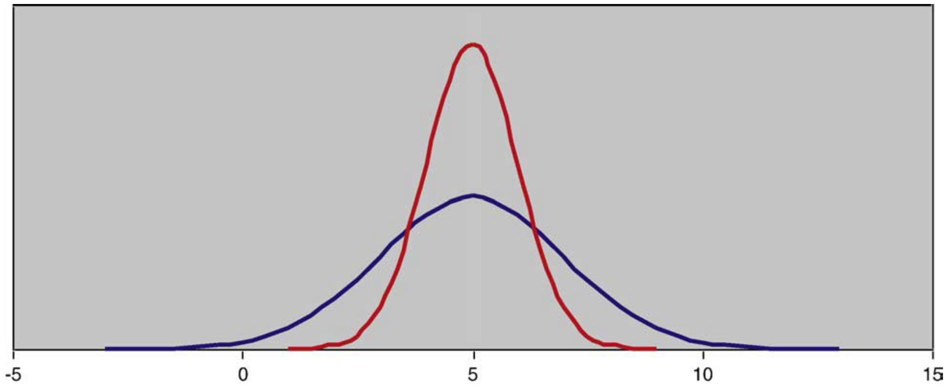

Fig. 2. Hypothetical example with two groups with the same mean but differing in the distribution of the scores.

However, many data sets may provide opportunities to examine other interesting between-group questions that may not be as vulnerable to the same limitations discussed above for the race-comparison question. For example, it may be interesting to learn that group differences exist in variability on the outcome variable (Matsumoto, 1994). It is possible that the means of ethnic groups X and Y on a specific outcome variable (e.g., number of drinking days) are quite similar but that most of the cases in ethnic group X are clumped close to the mean, whereas the cases in ethnic group Y are more dispersed (see Fig. 2). Assuming measurement equivalence has been established, the research team might then conclude that the intervention has uniform effects for group X. However, for group Y, the findings might suggest that the intervention worked very well for one subgroup of ethnic group Y but did not work well for another subgroup of that same ethnic group. Such a finding might suggest the need for future research to determine the characteristics associated with the varied outcomes in ethnic group Y. Limiting the analysis to the examination of group differences on the mean would not have revealed this interesting finding.

Identifying ethnic group differences in the mediators between treatment and outcomes may also be useful for understanding treatment outcomes among ethnic minorities. Sometimes, these approaches are referred to as mediation in the context of a moderated effect (Tein, Sandler, MacKinnon, & Wolchik, 2004). For example, McKay et al. (2003) found that self-efficacy and self-help participation were more important mediators for explaining the relationship between substance abuse treatment type and outcomes for non-Hispanic Caucasians than for African Americans.

5.3. The pitfalls of combining ethnic groups for analyses

The small numbers of ethnic minorities in many samples sometimes tempt the researcher to combine various ethnic minorities into one group for the analysis. Although this strategy may yield a larger sample of ethnic minorities, the disadvantages appear to outweigh this advantage. Fig. 3 illustrates this situation. Suppose a researcher collected information on annual drinking days using a sample of 120 that includes 60 non-Hispanic Caucasians (Group 1) and 60 ethnic minorities—15 from each of four ethnic groups (Groups 2 to 5 in Fig. 3). Due to the small numbers of each ethnic group, the researcher is tempted to combine all the ethnic minorities into one group. The mean number of drinking days per year for the non-Hispanic Caucasian sample is 30. The overall combined mean for the four ethnic groups is also 30. However, the mean values of ethnic Groups 1 and 2 are less than 30 (M = 10 and M = 20, respectively), ethnic Group 3 is equivalent to the non-Hispanic Caucasian sample, and the mean for ethnic Group 4 (M = 60) is much higher than the non-Hispanic Caucasian group. Obviously, combining the small groups in this hypothetical study may lead to the conclusion that the non-Hispanic Caucasian and ethnic group samples are reporting similar numbers of drinking days when, in reality, some important differences may exist among the four ethnic groups.

Some justify the strategy of combining ethnic groups as a preliminary approach. However, it is questionable how that approach is helpful because it would be unclear how to interpret the results from such an approach. For example, in a treatment study, if group differences are present between the majority group and the combined ethnic groups, it would be unclear whether the differences are attributable to all the ethnic groups or just some of them. Conversely, as Fig. 3 illustrates, a finding of no group differences between the combined ethnic minority sample and the non-Hispanic Caucasian sample may be masking the fact that certain ethnic groups may differ from the non-Hispanic Caucasian group (e.g., Groups 1 and 4 in the figure), although ethnic Group 3 does not. In either event, it seems that the

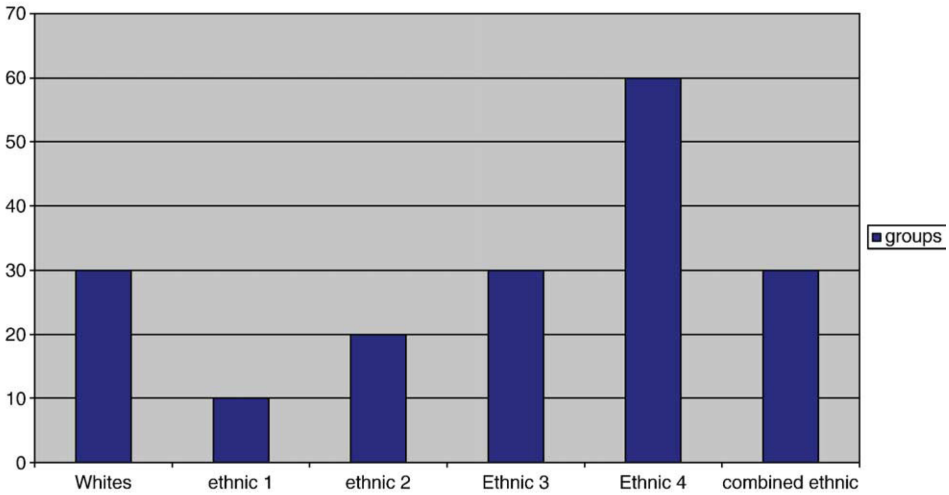

Fig. 3. Hypothetical scores for Whites, separate ethnic groups, and combined ethnic groups.

researcher would still need to follow up with additional analyses comparing the majority group to specific ethnic groups to even understand the larger finding between the majority group and the combined ethnic groups. If that is the case, the benefit of beginning with a comparison of the combined ethnic group to the non-Hispanic Caucasian group is not very clear.

6. Summary and recommendations

Policies by NIH, NIDA, and others mandate that inclusion and valid analyses be addressed in research with ethnic minorities. Inclusion, although important, is not sufficient. More research on substance abuse among ethnic minorities that considers valid analysis are needed. The objective of this article was to address two issues that may improve the valid analysis of research on ethnic minorities: measurement and data analysis. The following recommendations are suggested as strategies for conducting valid research with ethnic minorities:

1. Measures standardized on other populations need to be evaluated to determine their appropriateness with a specific ethnic group. CFAs, IRT, and regression are all tools that may be useful empirical approaches for assessing measurement equivalence. In addition, pilot studies, expert panels, and focus groups may play an important role in assessing the appropriateness of a measure.
2. Researchers are encouraged to review the characteristics of the standardization sample before blindly applying cutoff scores to an ethnic minority group.
3. Clinical trials provide rich opportunities for examining the effects of the race of the interviewer or the race of the coder (rater) on measurement.
4. Race-comparison designs may not address the most important issues for understanding treatment effects for ethnic minorities and may even lead to misleading conclusions. Ideally, studies are needed that focus on the efficacy of specific interventions with specific ethnic minority groups. In addition, a number of alternative within-group (e.g., moderator and mediator variable studies, studies of engagement, and retention) or between-group studies (group differences in variability, groups differences in mediators and moderators) may provide more information for understanding treatment issues in ethnic minority samples.
5. Sample sizes may be a daunting challenge for understanding treatment efficacy within ethnic minority samples. The researcher may address this issue by targeting specific ethnic minority groups, evaluating effect sizes when the sample is not large enough for other analyses, and considering the use of other statistical techniques appropriate for small samples.
6. Comparing a combined group of various ethnic minorities to a non-Hispanic Caucasian sample may ignore important group differences among various ethnic minorities and, consequently, lead to misleading conclusions about treatment effects.

Acknowledgment

Preparation of this manuscript was supported by NIDA cooperative agreements DA013732, DA0137320, and DA13045 and NIDA grant DA015004.

References

Allen, J. (1998). Personality assessment with American Indians and Alaska Natives. Instrument consideration and service delivery style. *Journal of Personality Assessment*, *70*, 17−42.

Allen, J., & Walsh, J. (2000). A construct-based approach to equivalence: Methodologies for cross-cultural/multicultural personality assessment research. In R. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessments* Mahwah, NJ: Lawrence Erlbaum.

Anderson, B., Silver, B., & Abramson, P. (1988). The effects of the race of the interviewer on race-related attitudes of Black respondents in SRC/CPS national election studies. *Public Opinion Quarterly*, *52*, 289−324.

Andrews, J. A., Lewinsohn, P. M., & Hops, H. (1993). Psychometric properties of scales for the measurement of psychosocial variables associated with depression in adolescence. *Psychological Reports*, *73*, 1019−1046.

Alterman, A., Cacciola, J., & Rutherford, M. (1993). Reliability of the Revised Psychopathology Checklist in substance abuse patients. *Psychological Assessment*, *5*, 442−448.

Bailey, W., Nowicki, S., & Cole, S. (1998). The ability to decode nonverbal information in African American, African, Afro-Caribbean, and European American Adults. *Journal of Black Psychology*, *24*, 418−431.

Bachman, J., & O'Malley, P. (1984). Yea-saying, nay-saying, and going to extremes: Black White differences in response style. *Public Opinion Quarterly*, *48*, 491−509.

Beauvais, F., & Trimble, J. (1992). The role of the researcher in evaluating American Indian alcohol and other drug Abuse treatment programs. In Orlandi (Ed.), *Cultural competence for evaluators: A guide for alcohol and other drug abuse prevention practitioners working with ethnic/racial communities* Rockville, MD: OSAP.

Berry, J. W. (1980). Introduction to methology. In H. C. Triandis, & J. W. Berry (Eds.), *Handbook of cross-cultural psychology*, *vol.2*. (pp. 1−28). Boston: Allyn & Bacon.

Bolt, D., Hare, R., Vitale, J., & Newman, J. (2004). A multigroup item response theory analysis of the Psychopathology Checklist—Revised. *Psychological Assessment*, *16*, 155−168.

Bravo, M. (2003). Instrument development: Cultural adaptations in ethnic minority research. In G. Bernal, J. Trimble, A. Burlew, & F. Leong (Eds.), *Handbook of racial ethnic minority psychology* Newbury Park, CA: Sage.

Brislin, R. W. (1993). *Understanding culture's influence on behavior*. New York: Harcourt Brace.

Brodey, B., Rosen, C., Brodey, I., Sheetz, B., Stinfeld, R., & Gastfriend, D. (2004). Validation of the Addiction Severity Index (ASI) for internet and automated telephone self-report administration. *Journal of Substance Abuse Treatment*, *26*, 253−259.

Brody, G., McBride Murry, B., Gerrard, M., Gibbons, F., McNair, L., Brown, A., et al. (2006). The Strong African American Families Program: Prevention of youths' high risk behavior and a test of model for change. *Journal of Family Psychology*, *20*, 1−11.

Brody, G., McBride Murry, V., Kim, S., & Brown, A. (2002). Longitudinal pathways to competence and psychological adjustment among African American children living in rural single-parent households. *Child Development*, *73*, 1505−1516.

Butcher, J. N. (1996). Translation and adaptation of the MMPI-2 for international use. In J. N. Butcher (Ed.), *International adaptations of the MMPI-2* (pp. 3−46). Minneapolis: University of Minnesota Press.

Bureau of Indian Affairs (2002, July 12). Indian entities recognized and eligible to receive services from the United States Bureau of Indian Affairs. *Federal Register*, *67*.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum.

Caldwell, C., Sellers, R., Bernat, D., & Zimmerman, M. (2006). Racial identity, parental support, and alcohol use in a sample of at-risk African American high school students. *American Journal of Community Psychology*, *34*, 71−82.

Center for Disease Control and Prevention. (2001). *Division of HIV/AIDs prevention, HIV/AIDs surveillance report* (vol. 13). Atlanta, GA: CDC 2001. Available at: http://www.cdc.gov/hiv/sats/hasr1302.htm.

Chambliss, D., Sanderson, W., Shoham, V., Johnson, S., Pope, K., Crits-Critoph, P., et al. (1996). An update on empirically validated therapies. *Clinical Psychologist*, *49*, 5−18.

Coatsworth, J. D., Santisteban, D. A., McBride, C., & Szapocznik, J. (2001). Brief strategic family therapy versus community control: Engagement, retention, and an exploration of the moderating role of adolescent symptom severity. *Family Process*, *40*, 313−332.

Cooke, D., Kosson, D., & Michle, C. (2001). Psychopathology and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist—Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment*, *13*, 531−542.

Cotter, P., Cohen, J., & Couter, P. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, *46*, 278−284.

Crocker, L., & Algina, J. (1986). *Introduction to classic and modern test theory*. Belsworth, GA: Wadsworth.

Crockett, L., Randall, B., Shen, Y., Russell, S., & Driscoll, A. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo Adolescents: A national study. *Journal of Consulting and Clinical Psychology*, *73*, 47−58.

Delva, J., Wallace, J. M., O'Malley, P. M., Bachman, J. G., Johnston, L. D., & Schulenberg, J. E. (2005). The Epidemiology of alcohol, marijuana, and cocaine use among Mexican American, Puerto Rican, Cuban American and other Latin American eight-grade students in the United States: 1991–2002. *American Journal of Public Health*, *95*, 696−702.

Dunkerley, G., & Dalenberg, C. (1999). Secret-keeping behaviors in Black and White children as a function of intereviewer race, racial identity, and risk for abuse. *Journal of Aggression, Maltreatment, & Trauma*, *2*, 13−35.

Eberhardt, J., Goff, P., Purdie, V., & Davies, P. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*, 876−893.

Finkel, S., Guterbock, T., & Borg, M. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *Public Opinion Quarterly*, *55*, 313−330.

Galea, S., & Rudenstine, S. (2005). Challenges in understanding disparities in drug use & its consequences. *Journal of Urban Health*, *82*, iii5−iii12.

Gonzales, N., Cauce, A., & Mason, C. (1996). Interobserver agreement in the assessment of parental behavior and parent adolescent conflict. *African American mothers, daughters, and independent observers. Child Development*, *67*, 1483−1498.

Graham, S. (1992). "Most of the subjects were White and middle class": Trends in published research on African Americans in selected APA journals, 1970–1989. *American Psychologist*, *47*, 629−639.

Hall, G. (2001). Psychotherapy research with ethnic minorities: Empirical, ethical, and conceptual issues. *Journal of Consulting and Clinical Psychology*, *69*, 502−510.

Hatchett, S., & Schuman, H. (1975/1976, Winter). White respondents and race-of-interviewer effects. *Public Opinion Quarterly*, *39*, 523−528.

Hathaway, S. R., & McKinley, J. C. (1989). Minnesota multiphasic personality inventory-2: Manual for administration and scoring. Minneapolis, MN: University of Minnesota Press.

Heron, R., Twomey, H., Jacobs, D., & Kaslow, N. (1997). Culturally competent interventions for abused and suicidal African American women. *Psychotherapy*, *34*, 410−424.

Hohmann, A., & Perron, D. (1996). How the new NIH guidelines on inclusion of women and minorities apply efficacy trials, effectiveness trials, and validity. *Journal of Consulting and Clinical Psychology*, *64*, 851−855.

Hoyle, R. (1999). *Statistical strategies for small sample research*. Thousands Oaks, CA: SAGE Publications.

Hoyt, W., & Melby, J. (1999). Dependability of measurement in counseling psychology: An introduction to Generalizability Theory. *The Counseling Psychologist*, *27*, 325−352.

Hughes, D., & Dumon, K. (1993). Using focus groups to facilitate culturally anchored methodology. *American Journal of Community Psychology*, *21*, 727−746.

Hutchens, S. (2003). Eyewitness identification and attention: Bizarreness and own-race bias. Paper presented at the 50th annual convention of the Southwestern Psychological Association. April 17, 2003.

Iguchi, M., London, J., Forge, N., Hickman, L., Fain, T., & Reihman, K. (2002). Elements of well-being affected by criminalizing the drug user. *Public Health*, 146−150.

Knight, G., & Hill, N. (1998). Measurement equivalence in research involving minority adolescents. In V. McLoyd, & L. Steinberg (Eds.), *Studying minority adolescents* Mahwah, NJ: Lawrence Erlbaum and Associates.

Longshore, D., Grills, C., & Annon, K. (1999). Effects of a culturally congruent intervention on cognitive factors related to drug-use recovery. *Substance Use & Misuse*, *14*, 1223−1241.

Loring, B., & Powell, A. (1988). Gender, race, and *DSM-II*: A study of the objectivity of psychiatric diagnostic behavior. *Journal of Health and Social Behavior*, *29*, 1−22.

Malcolm, B. (2004). Evaluating the effects of self esteem on substance abuse among homeless men. *Journal of Alcohol and Drug Education*, *48*, 39−61.

Manly, J. (2005). Advantages and disadvantages of separate norms for African Americans. *Clinical Neuropsychologist*, *19*, 270−275.

Matsumoto, D. (1994). *Cultural influences on research methods and statistics*. Pacific Grove, CA: Brooks/Cole.

McClelland, L. (1974). Effects of interviewer-respondent race interactions on household interview measures of motivation and intelligence. *Journal of Personality & Social Psychology*, *28*, 392−397.

McKay, J., Lynch, K., Pettinat, H., & Shepard, D. (2003). An examination of potential sex and race effects in a study of continuing care for alcohol- and cocaine-dependent patients. *Alcoholism: Clinical and Experimental Research*, *27*, 1321−1323.

Michaels, M., Barr, A., Roosa, M., & Knight, G. (2007). Self esteem: Assessing measurement equivalence in a multiethnic sample of youth. *Journal of Early Adolescence*, *27*, 269−295.

Miranda, J. (1996). Introduction to the Special Section of Recruiting and Retaining Low Income Latinos in Psychotherapy Research. *Journal of Consulting and Clinical Psychology*, *64*, 848−850.

Miyamoto, R., Hishinuma, C., Nishimura, S., et al. (2001). Equivalences regarding the measurement and constructs of self-esteem major other life events in an Asian/ Pacific islander sample. *Cultural and Ethnic Minority Psychology*, *7*, 152−163.

Morgenstern, J., & Bux, D. (2003). Examining the effects of sex and ethnicity on substance abuse treatment and mediational pathways. *Alcoholism: Clinical and Experimental Research*, *27*, 1330−1332.

National Institutes of Health. (2001). NIH guidelines for the inclusion of women and minorities as subjects in clinical research. http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm.

Nichol, D., Padilla, J., & Lucio, E. (2000). Issues in the cross-cultural adaptation and use of the MMPI. In R. Dana (Ed.), *Handbook of Cross-Cultural/Multicultural Personality Assessment.* New Jersey: Lawrence Erlbaum Associates, Inc.

Otsuki, T. (2003). Substance abuse, self esteem and depression among Asian American adolescents. *Journal of Drug Education*, *33*, 369−390.

Okazaki, S., & Sue, S. (1995). Methodological issues in assessment reserach with ethnic minorities. *Psychological Assessment*, *7*, 367−375.

Radloff, L. S. (1977). The CES-D Scale: A self report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385−401.

Raju, N., Laffitte, L., & Byrne, B. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517−529.

Ramirez, M., Ford, M., Steward, A., & Teresi, J. (2005). Measurement issues in health disparities research. *Health Research and Educational Trust*, *40*, 1640−1657.

Randolph, K., Gerend, M., & Miller, B. (2006). Measuring alcohol expectancies in youth. *Journal of Youth and Adolescence*, *35*, 939−948.

Reid, R., Casat, C., Norton, J., Anastopoulos, A., & Temple, E. P. (2001, Winter). Using behavior rating scales for ADHD across ethnic groups: The IOWA Conners. *Journal of Emotional and Behavioral Disorders*, *9*, 210−219.

Rogler, L. H., Malgady, R. G., & Rodriguez, O. (1989). *Hispanics and mental health: A framework for research*. Malabrar, FL: Kriger.

Santisteban, D. A., Dillon, F., Mena, M. P., Estrada, Y., & Vaughan, E. L. (2005). Psychiatric, family, and ethnicity-related factors that can impact treatment utilization among Hispanic substance abusing adolescents. *Journal of Social Work Practice Addictions*, *5*, 133−155.

Santisteban, D. A., Szapocznik, J., Perez-Vidal, A., Kurtines, W. M., Murray, E. J., & Laperriere, A. (1996). Efficacy of intervention for engaging youth and families into treatment and some variables that may contribute to differential effectiveness. *Journal of Family Psychology*, *10*, 35−44.

Shillington, A., & Clapp, J. (2000). Self report stability of adolescent substance abuse: Are there differences for gender, ethnicity and age. *Drug and Alcohol Dependence*, *60*, 19−27.

Strada, M., Donohue, B., & Lefforge, N. (2006). Examination of ethnicity in controlled treatment outcome studies involving adolescent substance abusers: A comprehensive literature review. *Psychology of Addictive Behaviors*, *20*, 11−27.

Strube, M. (2000). Reliability and Generalizability Theory. In L. Grimm, & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (ch. 2) (pp. 23–36). Washington, DC: American Psychological Association.

Substance Abuse Mental Health Services Administration. (2002). National Study on Drug Use and Health. Department of Health and Human Services, Office of Applied Studies.

Sue, S., Kurasaki, K., & Srinivasan, S. (1999). Ethnicity, gender, and cross cultural issues in clinical reseach. In P. Kendall, J. Butcher, & G. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (pp. 54−71). Hoboken, NJ: Wiley & Sons.

Szapocznik, J., Perez-Vidal, A., Brickman, A., Foote, F. H., Santisteban, D., et al. (1988). Engaging adolescent drug abusers and their families into treatment: A strategic structural systems approach. *Journal of Consulting & Clinical Psychology*, *56*, 552−557 (Reprinted Annual Review Addictions Res. Treatment, 1991, 331-336).

Szapocznik, J., Prado, G., Burlew, A. K., Williams, R., & Santisteban, D. (2007). Drug abuse in African American and Hispanic adolescents. *Annual Review of Clinical Psychology*, 77−105.

Turner, R. J., Lloyd, D. A., & Taylor, L. J. (2006). Stress burden, drug dependence and the nativity paradox among U.S. Hispanics. *Drug Alcohol Dependence*, *28*, 79−89.

Tein, J., Sandler, I., MacKinnon, D., & Wolchik, S. (2004). How did it work? Who did it work for? Mediation in the context of a moderated prevention effect for children of divorce. *Journal of Consulting and Clinical Psychology*, *72*, 617−624.

Udry, J. R. (1998). The national longitudinal study of adolescent health (Add Health), waves I and II, 1994–1996. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

U.S. Bureau of the Census. (2000). Projections of the resident population by race, hispanic origin, and nativity: Middle series, 1999–2100. (online). www.census.gov/population/www/projections/popproj.html (October 2001).

U.S. Census Bureau. (2004). (internet release date, March 18, 2004). U.S. Interim Projections by Age, Sex, Race and Hispanic Origin. http://www.census.gov/ipc/www/usinterimproj.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4−70.

Vega, W. A., Aguilar-Gaxiola, S., Andrade, L., Bijl, R., Borges, G., et al. (2002). Prevalence and age of onset for drug use in seven international sites: Results from the international consortium of psychiatric epidemiology. *Drug Alcohol Dependence*, *68*, 285−297.

Wakim, P. (2006). Differences in treatment effect between sites in multi-site clinical trials on substance abuse treatments. Presented at the Steering Committee of the NIDA Clinical Trials Network, October 18, 2006, Seattle, Washington.

Want, V., Parham, T., Baker, R., & Sherman, M. (2004). African American students' ratings of Caucasian and African American Counselors varying in racial consciousness. *Cultural Diversity and Ethnic Minority Psychology*, *10*, 123−136.

Weathers, M., Frank, E., & Spell, L. (2002). Differences in the communication of affect: Members of the same race versus members of a different race. *Journal of Black Psychology*, *28*, 66−77.

Widaman, K., & Reise, S. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention. Methodological advances from alcohol and substance abuse research* Washington, DC: American Psychological Association.

Woehr, D. J., Arciniega, L. M., & Fowler, O. (2003). Measuring Work Ethnic: An Examination of the Measurement Equivalences and Spanish Versions of the Multidimensional Work Ethic Profile. Proceedings of the 2003 European Applied Business Research Confernece, Venice, Italy, June 2003.