2002

Gender and Ethnic Differences in Adolescent Self-Esteem in Alcohol and Other Drug Use Research: A Rasch Measurement Model Analysis

Joseph E. Trimble, PhD
Western Washington University, joseph.trimble@wwu.edu

Eldon R. Mahoney
Western Washington University

NIAAA Research Monograph No. 37

ALCOHOL USE AMONG AMERICAN INDIANS AND ALASKA NATIVES

Multiple Perspectives on a Complex Problem

Edited by:
Patricia D. Mail, Ph.D., M.P.H.
Suzanne Heurtin-Roberts, Ph.D., M.S.W.
Susan E. Martin, Ph.D.
Jan Howard, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health
National Institute on Alcohol Abuse and Alcoholism
6000 Executive Boulevard
Bethesda, MD 20892

2002

Chapter 9

Gender and Ethnic Differences in Adolescent Self-Esteem in Alcohol and Other Drug Use Research: A Rasch Measurement Model Analysis

Joseph E. Trimble, Ph.D., and Eldon R. Mahoney, Ph.D.

KEY WORDS: Native American; self-esteem; adolescence; gender differences; ethnic differences; problem behavior theory of AODU (alcohol or other drug [AOD] use); statistical modeling; specificity and sensitivity of measurement; alcohol use test; psychosocial AODU identification and diagnostic method; cultural sensitivity

A few years ago the senior author of this chapter was invited to attend several evening meetings of American Indian parents, community leaders, and alcohol and other drug (AOD) use specialists held at a community center on a nearby reservation. The meetings were initiated in response to an alarming increase in AOD-related problems occurring among many of the community's youth. Over the course of these meetings, parents and community leaders offered many suggestions for dealing with the problems; these suggestions often led to lengthy and sometimes heated debates. As one can imagine, there was considerable anger expressed at those who provided AODs to young people and at the physical and psychological damage created by the youth while in their intoxicated states. For many residents, though, defining the problem and its solution was straightforward: the young people had a problem with their self-esteem, and AOD abuse

J.E. Trimble, Ph.D., is a fellow at Harvard University, Radcliffe Institute for Advanced Study, Henry A. Murray Research Center, 10 Garden St., Cambridge, MA 02138. He is also professor of psychology at the Center for Cross-Cultural Research, Department of Psychology, Western Washington University, Bellingham, WA 98225–9089. E.R. Mahoney, Ph.D., is a professor in the Department of Sociology, Western Washington University.

would decline if they felt better about themselves. The belief that most youth, especially those of ethnic minority background, engage in AOD use because of problems with their self-esteem is a common one in many communities across the country—including many American Indian communities.

Fleming and Manson (1990) conducted an extensive evaluation of the characteristics and effectiveness of 18 American Indian AOD prevention programs. Ninety-four percent of the community-based programs emphasized primary prevention activities (i.e., activities developed to prevent a health-related problem from occurring among those who may be at risk). Some of these activities involved the use of educational materials, promotion of Indian identity and building self-esteem through cultural events, and the use of self-help groups. Fleming and Manson also asked their respondents to identify those factors that placed Indian youth at risk for AOD use. Eighty-eight percent singled out poor self-esteem and parental abuse of alcohol as the greatest contributors to high risk. The respondents also identified additional contributing factors, including use of drugs by peers and friends; abuse, neglect, and family conflict; sexual abuse and emotional and psychological difficulties; previous suicide threats or attempts; and alienation from the dominant culture's social values. The researchers also asked their respondents to identify factors that presumably prevented one from using and abusing AODs. Protective factors identified by respondents included a well-defined spiritual belief system, a positive sense of self-worth, ability to make good decisions about personal responsibilities, and the ability to act independently of the influences of others. The respondents also believed that one's friends and peers who act in healthy and responsible ways could serve as models for at-risk youth.

Owan, Palmer, and Quintana (1987) surveyed nearly 420 schools from Head Start to the secondary school level with large American Indian enrollments and 225 different tribal groups who were receiving grant support for AOD abuse projects from the Indian Health Service. Both the school and community respondents indicated that AOD abuse education was a major priority, followed by a concern for building self-esteem and developing effective coping and decision-making skills. Owan and colleagues drew some important conclusions that emphasize the need for "early intervention to combat alcohol and substance abuse among Indian youths" (p. 71). They also emphasized the point that Indian youth need strong families to promote positive self-esteem, identity, and values. "Weak families," they argued, "produce uprooted individuals susceptible to 'peer clusters' prone to alcohol and substance abuse" (p. 71).

Moving from the lay and community setting to the research setting, there has been intensive study on the etiology of AOD use, and some of this research has addressed the relationship between self-esteem and AOD use. However, our knowledge

of measurement topics, particularly those that occur with culturally unique populations, leads us to contend that there are problems associated with the measurement of self-esteem and that the problems may stem from different cultural understandings of the self-esteem concept. There are issues, too, concerning the relationship between self-esteem and AOD use among adolescents that deserve research attention. However, this chapter focuses on the measurement and analysis of a particular self-esteem scale rather than the relationship between self-esteem and AOD use.

To set the stage for the findings of the Rasch model analysis of the self-esteem scale reported in this chapter, we first discuss the *self* construct. Consideration is given to the concept of cultural equivalence and measurement. Following a detailed description of the measurement process and its relationship to Rasch model statistical analysis, we report the results of our analysis and discuss the findings.

SELF-ESTEEM: CONCEPTUAL AND METHODOLOGICAL ISSUES

Self-esteem has been defined in a number of ways by theoreticians and researchers (see Wylie 1979). The definition developed by Rosenberg (1965) is one of the more widely accepted ones:

> When we speak of self-esteem, then, we shall simply mean that the individual respects himself, considers himself worthy; he does not necessarily consider himself better than others, but he definitely does not consider himself worse; he does not feel that he is the ultimate in perfection, but, on the contrary, recognizes his limitations and expects to grow and improve. Low self-esteem, on the other hand, implies self-rejection, self-dissatisfaction, self-contempt. The individual lacks respect for the self he observes. The picture is disagreeable, and he wishes it were otherwise. (Rosenberg 1965, p. 31)

The measure of self-esteem discussed and subsequently analyzed in this chapter follows Rosenberg's definition; however, we go beyond this definition to focus on the way people from different ethnic groups attempt to define themselves and the categories they use to do so.

Self-Esteem as a Master Status

Central to the theoretical notions of the development and maintenance of self-esteem is that the statuses one occupies in society are major sources of self-esteem. Locations in the social structure that serve as "master statuses" are salient for individuals because these statuses tend to influence all aspects of their psychosocial existence. Gender and ethnicity are two of the prominent master statuses in the United States and likely elsewhere where emphasis is placed on individualism and ego-centered development. This may especially be the case for adolescents; because they have not yet taken other locations in

the social structure (e.g., occupation, educational attainment, marital status), they may be very dependent on these master statuses for their sense of self.

Academic and intellectual interest in the self construct and its relationship with adolescent development likely began with theoretical discussions and subsequent research concerning the basic tenets of psychoanalysis and, more specifically, ego psychology. Additionally, interest in the self was advanced by Carl Rogers through his early writings on the construct and its importance in directing the goals of his well-known client-centered counseling approach. Indeed, the self construct captured the attention of numerous 20th century researchers, theoreticians, and clinical practitioners; moreover, many educators built elementary and secondary curricula on the premise that self enhancement should be the goal of learning outcomes. Literature on the self construct is very extensive and comprehensive, with many academic disciplines represented in the studies and debates. A thorough review of the construct no doubt would cover several volumes, and such a review is beyond the scope of this chapter.

In this chapter, we are primarily interested in self-esteem matters that occur among culturally unique adolescents and the patterns that emerge for gender and ethnic statuses. Although the literature on this topic is extensive, only a handful of empirically based literature citations include the topic of AOD use with gender and ethnic variables.

Ethnic and Gender Patterns

The research findings on gender and ethnic differences in adolescent self-esteem are far from uniform. Dukes and Martinez (1994) concluded that "the impact of racism and sexism on the self-esteem of members of minority and dominant groups remains controversial" (p. 105). The work of these researchers (Martinez and Dukes 1987, 1991; Dukes and Martinez 1994) also suggests that conceptualizing ethnicity and gender as separate master statuses may be an oversimplification. Introducing the concept of "ethgender," these researchers found differences in self-esteem level as a function of various ethnicity and gender combinations. This suggests, of course, that master statuses may be additive in their influence on self-esteem.

The discussion of gender and ethnic differences in self-esteem takes place, however, in the context of a dubious measurement of self-esteem. While the most popular measure has been the Rosenberg Self-Esteem Scale (Rosenberg 1979), numerous other instruments have been used, some of which are variants of Rosenberg's original scale. Many critics have suggested that the inconsistency in self-esteem research findings may reflect differences in the way the construct is measured (Gray-Little and Applebaum 1979; Wylie 1979; Dorgan et al. 1983; Trimble 1987). The utilization of these various measures has taken place in the context of inadequate evaluation of the instruments, including the Rosenberg instrument. These instruments, when evaluated,

are usually based on outmoded measurement models that define a measure as adequate when it contains virtually none of the characteristics needed for scientific activity. We explore this point in more detail later in the chapter. For the moment, though, we strongly contend that before any attempt can be made to address the question of gender, ethnicity, and "ethgender" differences in self-esteem, we must understand the measurement of self-esteem in individuals differing in gender and ethnicity.

Problem-Behavior Theory and Substance Use

For the past 30 years considerable attention has been devoted to exploring the correlates of personality variables with AOD use patterns. Although many of the published articles are nontheory based, they are at least implicitly grounded in the *problem-behavior theory* developed by Jessor and Jessor (1977). The conceptual framework of the theory consists of antecedent-background variables, social-psychological variables, and social behavior outcome variables. In the social-psychological variable domain, the theory holds that self-esteem is an essential element in the personal belief structure along with social-criticism, alienation, and locus of control. Jessor and Jessor maintained that "the preservation of high self-esteem serves as a barrier to engaging in deviance" (1977, p. 21). Thus, one's sense of self-esteem can be negatively influenced if one engages in nonconforming or problem behaviors. So, according to their theory, if adolescents heavily use psychoactive drugs and alcohol, then their sense of self-esteem must be low; conversely, if self-esteem is high, then there should be no need to engage in problem or deviant behaviors.

In the past two decades numerous AOD use researchers have attempted to identify the correlates of AOD use patterns with a variety of social and psychological constructs. In part guided by the tenets of problem-behavior theory, much of the research effort has focused on self-esteem, especially as it relates to adolescents. Results have been uneven and discrepant. Schroeder, Laflin, and Weis (1993) maintained that "regardless of the definition or measure of SE [self-esteem] used, no sizeable relationship between SE and drug use has been found" (p. 659). Moreover, they contended that the inconsistent research findings can be attributed to (a) measurement of AOD use, (b) presence of confounding variables in the research design, (c) inferring causality from correlation data, (d) statistical problems stemming from inflated research design error, (e) misinterpretation of findings, (f) failure to report strength of association indices, (g) reporting insufficient statistical information, and (h) procedures and scales used to measure self-esteem.

Moore, Laflin, and Weis (1996) used the "social deviance" model, a variant of problem-behavior theory, to test the relationship between self-esteem and cultural norms. As predicted in their problem statement, their results failed to support the model. They concluded that "consid-

eration of the respondents' cultural norms does not reveal a relationship between [self-esteem] and tobacco, marijuana, alcohol, and/or drug use" (p. 539).

In summary, researchers have not found any consistent relationship between self-esteem and AOD use. Yet work in this area continues despite the criticisms. A cursory review of literature published since 1993, the year Schroeder and colleagues published their critique, indicates that the number of published articles extends well into the hundreds.

Although considerable attention can be devoted to the way the self-esteem construct is conceptualized, measured, and interpreted, we contend that the way self-esteem scales are analyzed, especially when used with culturally different populations, may be problematic. To illustrate, later in this chapter we present a series of analyses using Rasch modeling and measurement procedures to show that three ethnic groups may be responding to and interpreting a common set of self-esteem items differently; consequently, differential response patterns may be attributed to ethnic and cultural orientations of the respondents. We then present a measurement model that specifies the characteristics of scientific measurement and analyzes the measurement of self-esteem typically used in AOD use studies among adolescents differing in gender and ethnic identification. Findings from our analysis can assist researchers in understanding how ethnicity and gender status influence psychosocial scale items.

In the next section, we provide summary information about the issues associated with the development and use of psychosocial scales for cultural-specific and cultural-comparative research. Debates abound regarding the influence of one's worldview on understanding and interpreting standardized tests and psychosocial scales (for reviews, see Berry 1969; Berry and Dasen 1974; Irvine and Carroll 1980). Moreover, many cross-cultural psychologists contend that "comparing elements from differing societies leads to inadmissible distortions of reality" (Kobben 1970, p. 584). The anthropologist Goldschmidt (1966) equated this contention with what he called the Malinowskian Dilemma; that is, "every culture [must] be understood in its own terms, that every institution be seen as a product of the culture within which it developed. It follows from this that a cross-cultural comparison of institutions is essentially a false enterprise, for we are comparing incomparables" (p. 8). Cultural-comparative research using instruments such as self-esteem scales may be fraught with problems of "incomparability" and thus may lead researchers to draw conclusions about a finding that may not be valid or justified. To avoid these possibilities, attention must be given to the concept of cultural equivalence in measurement studies.

CULTURAL EQUIVALENCE AND MEASUREMENT

Use of standard assessment scales and tests across cultures is filled with numerous problems and concerns,

which have been pointed out by Irvine and Carroll (1980), Irvine and Berry (1983), and Lonner and Berry (1986), among others. The problem of cultural equivalence or comparability is the most common theme that runs through the literature on cultural-comparative research. Considerable attention has been given to this important issue (see especially Berry 1969; Brislin et al. 1973; Berry and Dasen 1974; Poortinga 1983; Trimble et al. 1983; Malpass and Poortinga 1986).

Definitions

Cultural equivalence refers "to the problem of whether, on the basis of measurements and observations, inferences in terms of some common psychological dimension can be made in different groups of subjects" (Poortinga 1983, p. 238). Most cross-cultural researchers agree that cultural equivalence can be examined by giving attention to the following concepts: functional equivalence, linguistic equivalence, conceptual equivalence, stimulus equivalence, and metric equivalence.

Embedded in the notion of equivalence is the fundamental tenet that comparisons between groups require that a common, if not identical, process exists; stretched to the extreme, the notion holds that a universal process must exist to demonstrate and assess comparability. Consequently, to achieve functional equivalence two or more behaviors must "pre-exist as naturally occurring phenomena" that are related or identical to a similar problem or circumstance; the behaviors serve a similar function for both groups (Berry 1969, p. 122).

Linguistic equivalence exists when the translated content of survey or questionnaire items exhibits identical meaning when applied to two or more cultures (Prince and Mombour 1967). Conceptual equivalence exists when constructs are mutually intelligible and meaningful across ethnocultural groups; that is, "subjects have an equal understanding of the meaning of behavior or of concepts pertaining to behavior" (Malpass and Poortinga 1986, p. 66). Often cross-cultural researchers include stimulus equivalence and response equivalence in discussions about conceptual equivalence, since the equivalence of meaning of both terms is a necessary prerequisite for cultural comparative research.

Metric equivalence or scale (scalar) equivalence (Poortinga 1975) "exists when the psychometric properties of two (or more) sets of data from two (or more) cultural groups exhibit essentially the same coherence or structure" (Berry 1980, p. 10). Of the five equivalence types, metric or scalar equivalence has received the least amount of attention, perhaps because it is the most technical and poorly understood. Yet for the psychometrician it may be the most important concern. Before a measure can be used in cultural comparative research, it must first meet standards within the groups; then and only then can it be used between two or more groups.

Metric or scalar equivalence actually involves two separate but related forms of equivalence. Poortinga (1983) pointed out that scale equivalence involves the "equality of scaling units across groups" (p. 248); equality

emerges from the discovery that the statistical relationships among the dependent variables are similar for all groups. Metric equivalence, however, is concerned with the relative stability of the variables across the research experience. In addition, Drasgow (1987) offered the term *measurement equivalence* as a variant of metric equivalence to refer specifically to the constancy with which traits are measured among different subpopulations. Unlike the other forms of equivalence, metric, scalar, and measurement equivalence depend on response outcomes and, therefore, can only be determined after data have been collected and analyzed.

Analysis of data to test the existence of metric equivalence typically relies on the use of multivariate statistical routines. Initially, when cultural equivalence emerged as an issue in cultural-comparative research, researchers relied on principal components and factor analyses. Strength of the factor-based scales for the respective groups serves as partial criteria. Factor solutions also can be expanded to include congruence coefficients and related manipulations to isolate the nature of the equivalence. Windle and colleagues (1987) and Nishimoto (1986), for example, used factor solutions to examine the metric equivalence of personality scales administered to Asian and non-Asian populations. In both studies the factor solutions did not differ. However, the item composition and thus the factor meanings did vary.

Statistical Approaches To Assess Equivalence

A few cross-cultural researchers also recommend use of covariance structural modeling (e.g., LISREL) or variants of confirmatory factor analysis to test for metric equivalence (Poortinga 1983). There are limitations associated with the use of exploratory factor models; the advances in confirmatory factor modeling, however, appear to overcome these limitations. Some researchers recommend a form of latent trait analysis, especially when the scale contains binary scores. The Rasch (1960/1980) one-parameter model can be used, but Irvine and Carroll (1980) remind us that the model should be used "along-side traditional models as part of another method of looking at the same data" (p. 210).

The use of item response theory (IRT) to assess metric equivalence has produced interesting findings. Ellis and colleagues (1993) used IRT to test the equivalence of the Trier Personality Inventory, originally developed for use in West Germany. The differential item functioning (DIF) index showed that subsequent retranslations of original inventory items reduced the overall content and reduced error due to translations. Bontempo (1993) also used IRT on an individualism-collectivism scale to demonstrate the efficacy of the procedure and to test for translation bias. Both lines of research show promise for using IRT to assess equivalence of translated scales and tests.

The use of factor analysis in psychometric research and testing equiva-

lence is not without criticism (Kline 1983). Although some of the arguments are compelling, a discussion of this debate is beyond the scope of this chapter. Nonetheless, three critical points should be made: (1) factor solutions rarely fit the data completely in cultural-comparative research, primarily because of nonrandom measurement and translation error and unspecified conceptual contributions to the obtained weights; (2) factor solutions are suggestive; and (3) data should be, at a minimum, at the interval level. Most scales and inventories use binary or ordinal level response categories with presumed equality of the numerical distances between the alternatives; distortions can exist, thus eroding the strength of the correlation coefficients. Kim and Mueller (1978) pointed out that in a sense "variables with limited categories are . . . not compatible with factor analytic models." The most forceful of the critics is Duncan (1984), who considers factor analysis to be a failure in the measurement field because, among other points, "we . . . see nothing more than a 'correlational' science of 'inexact constructs'" (p. 207).

Rasch modeling and analysis is a powerful alternative to factor analysis in assessing the properties of tests and psychosocial scales. According to Linacre (1996), "factor analysis is confused by ordinal variables and highly correlated factors. Rasch analysis excels at constructing linearity out of ordinality and at aiding the identification of the core construct inside a fog of collinearity" (p. 470).

In the next section, we provide a detailed summary of the major properties and elements associated with Rasch analysis. The description is intended to provide the reader with background information to assist in understanding our approach to the subsequent analysis of the self-esteem scale selected for use in this chapter.

THE MEASUREMENT PROCESS AND ADEQUATE MEASUREMENT

Rasch modeling is a stochastic approach developed by the Danish mathematician Georg Rasch for the analysis of test responses and variations of ordinal observations. From the sums of the observations, Rasch analysis constructs linear measures of person abilities and item difficulty along with measures of precision (reliability) and accuracy (fit) indices. As originally conceived, the Rasch model specifies that each useful test response is an outcome of the probabilistic linear interaction between a person ability measure and an item difficulty measure (Rasch 1960/1980; Wright 1994). It should be noted that the Rasch model is not a data model; as Wright (1988) stated, "You may use it with data, but it's not a data model. The Rasch model is a definition of measurement, a law of measurement" (p. 32).

Andrich (1988) pointed out that Rasch analysis is an evolving statistical approach that challenges a data-dominated approach to how science should be done and which problems are useful. Rasch approaches have been gaining acceptance in the scientific community,

although there has been and continues to be resistance from sectors of the test development community. In this section we describe various elements necessary to understand the Rasch approach and its corresponding statistical features.

Measurement Elements

Following Wright and Masters (1982) and Andrich (1988), measurement in science is defined as consisting of the following elements: variables, unidimensionality, differences of degree and differences of kind, and item-free person measurement.

Variables

The central requirement for scientific observation is the ability to assess the magnitude or quantity of a property of interest. When the magnitude of a property has been operationalized, it is a variable. When the variable has been constructed, the property of interest can be measured. This measurement results in a numerical value of quantity or magnitude of the property on which arithmetic operations can be meaningfully (and ethically) performed. These numerical values have specified mathematical properties and are not arbitrary.

Unidimensionality

Any phenomenon can be characterized by many different properties. In the construction of a variable we identify a single property that can be mapped on a *single real number line*, which forms the property continuum. When this mapping on a single real number line is possible, the variable is unidimensional. If we consider more than one variable at a time, each will have a value on a different single real number line and the analysis is multidimensional.

Differences of Degree and Differences of Kind

Only when our measurement is unidimensional can comparisons of units of analysis be made in terms of degree to which the property is possessed. When the difference between units of analysis is not one of degree, it is one of kind. Unless one can be reasonably certain that the observations are all of one kind, comparisons cannot be said to be comparisons of degree, and the measure lacks validity. When the set of observations constituting the measure are unidimensional, they are all of the same kind. Further, when a measure is unidimensional, it possesses the characteristics of concatenation, invariant comparison, and group invariance.

A set of observations that are unidimensional, and thus have single real number line values, can be concatenated or linked together with arithmetic operations since the observations have (a) real rather than arbitrary values and (b) are all of the same kind. These real number values are possible only by knowing the location of the observation on the single real number line constituting the measurement continuum. The scores produced by a measurement can only be a function of the degree of the property they represent and no other property of the object or person being measured.

When we construct a measurement, each observation must remain stable in its value on the real number line regardless of what or who is being measured. The value of the observation must therefore be group (e.g., age, gender, ethnicity) invariant or sample free. When a measure does not have group invariance, it lacks measurement validity.

Item-Free Person Measurement

Items or empirical observations used to generate a score indicating the amount of the property of interest must be capable of measuring persons regardless of what particular subset of items is being used. If, in fact, items are located at reliably known points on the single real number line, a person score can be generated regardless of which specific items are used. It is item-free person measurement and person-free item calibrations that define fundamental measurement (Andrich 1988).

The Measurement Model

The analysis uses the Rasch measurement model (Rasch 1960/1980; Wright and Masters 1982; Andrich 1988). A Rasch analysis provides precise examination of the extent to which a measure possesses the elements described above. In its most simple form the Rasch model assumes that all items in a measure have the same underlying structure both across individual respondents and across the underlying single real number line continuum. This underlying structure is that the probability of responding in a certain manner to an item is a function of the overall score across all items. The central mathematical operation in the Rasch model is the odds ratio of responding in a certain manner to an item given a total score computed from the response to all other items in the measure.

Individuals with high total scores should have a specific probability of response to items indicating a high score that is different from the probability of response to items indicating a low score for those individuals with low total scores. The value or location of an item and/or a case on the underlying continuum is thus defined by its associated probability. The analysis in this chapter is conducted using the Quest computer program (Adams and Siek-Toon 1993) from the Australian Council for Educational Research. The Rasch model used by Quest is applicable to ordered category response data and is a generalized form of the Masters (1982, 1988) partial credit model (Wright and Masters 1982). Formally stated, the Rasch model for ordered category responses is that the response of person n to item i is represented by item score X_{ni}. This score may take any integer value from $0, \ldots . m_i$. The Rasch model describes the probability of observing a particular score x_{ni} as

$$P(X_{ni} = x_{ni}) = \frac{\exp \sum_{j=0}^{X_{ni}} w_{ij}(\beta_n - \delta_i - \tau_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} w_{ij}(\beta_n - \delta_i - \tau_{ij})}$$

where βn is the degree of self-esteem of person n, w_{ij} is the unit value assigned to response category j of item i, and δ_i and τ_{ij} represent the calibration of item i.[1] In Quest, item parameters are estimated with a joint (UCON) maximum likelihood procedure, with a correction factor (L-1/L) applied after convergence (Adams and Siek-Toon 1993). In all analyses convergence criteria for both case and item estimates are 0.005.

Rasch Model Statistical Analysis

The extent to which a set of items conforms to the Rasch model criteria of adequate measurement is statistically assessed in this analysis by the following elements: item locations or calibrations on the underlying continuum, item precision, continuum coverage, reliability and separation of items and cases, goodness of fit, and group invariance.

Item Locations or Calibrations on the Underlying Continuum

An adequate measure consists of items located along the full range of a single construct continuum. These item calibrations or locations specify the scale value of the item and define the hierarchical order of the items on the continuum. The calibrations are expressed as *logits* (Ludlow and Haley 1995). Since the purpose of an item is to provide information about persons, the logit for an item is the performance level of an item relative to the performance level on the total set of items and total set of persons. This analysis is conducted for each individual in the sample, and the logits are averaged (arithmetic mean) across all respondents. These mean logits thus indicate the average location of an item *for all individuals.* Logits can be calculated for each item and each response category for each item. The logits are thus true interval values generated out of clearly ordinal response categories.

In this analysis the Rasch partial credit model for ordered categories (Masters 1982) is used. The response category logits for each item are in the form of thresholds (Masters 1988). The threshold for a response category is the numerical amount required for an individual to have a 50 percent chance of responding positively to that item-response option and is thus analogous to Thurstonian thresholds (Masters 1988).

Item Precision

The standard errors of item or item-response category calibrations (logits) indicate the precision of the item calibration across all respondents and thus the precision of the item location on the underlying continuum.

---

[1] *The model can be written as a single expression by defining*

$$\sum_{j=0}^{0} w_{ij}\left(\beta_n - \delta_i - \tau_{ij}\right) \equiv 1$$

and for identification, the following constraints are applied:

$$\sum_{j=0}^{m_i} \tau_{ij} \equiv 0 \text{ and } \sum_{i=1}^{I} \delta_i \equiv 0$$

Continuum Coverage

Once calibrated, item-response categories may be examined for their location along the underlying continuum. An adequate measure consists of items that cover the full range of the continuum (logits generally range from −4 to +4). Several adjacent locations on the trait continuum for which there are no item-response categories constitute gaps in the measurement and should be minimized.

Reliability and Separation of Items and Cases

The reliability of the items is the proportion of observed item variance not due to estimation error. The reliability of the cases is the proportion of observed sample variance not due to measurement error. Item separation is the extent to which items are separated on the construct continuum and may be expressed as the number of statistically distinct levels of the variable found in the items. Case separation is the extent to which the cases in the sample are separated on the variable and is expressed as the number of statistically distinct levels of the variable found in the sample (Wright and Masters 1982; Wright 1996). Separation for either items or cases is defined by item or case reliability estimates as

$$\sqrt{\frac{\text{Reliability}}{1 - \text{Reliability}}}$$

Goodness of Fit

The fit of the data to the model is assessed by item fit tests indicating the degree to which the response pattern (across all respondents) fits the expectations of the model. The fit of an item is evaluated by the *infit* (information-weighted fit) statistic. An infit mean square value of 1.0 indicates that the observed response pattern is the expected response pattern under the model. An infit mean square of $1 + x$ indicates x percent more variation (residual) between the observed response pattern and the response pattern predicted by the model. Positive infit values are thus $1 + x$ percent higher than expected by the model. Positive infit occurs because (a) the item is not measuring what is measured by the other items or (b) the item lacks clarity and is differentially interpreted by respondents. Positive infit values indicate unmodeled noise in the measure and therefore represent a challenge to unidimensionality, and thus the validity of the measure.

An infit of $1 - x$ indicates less variation between the model predicted and observed response patterns than would be expected by the model (Adams and Siek-Toon 1993). Negative infit values occur because the item is redundant with other items and thus does not identify content of the variable not identified by other items. Negative infit values thus represent deficiency in the stochastic variability needed for useful measurement (cf. McNamara 1996, pp. 169–179).

Excessively large positive infit is of greater concern than excessively large negative infit because the former reveals invalidity, whereas the latter reveals only inefficiency. To define "excessively large" infit, mean square

values are converted to a standardized form (infit t). A positive t results when the infit mean square is > 1.0; a negative t results when the infit mean square is < 1.0. Misfit is defined here as $t = +/- 2.0$ standard deviations from model expected fit. It is important to emphasize that the fit of data to a model is always a matter of degree. Not only is a perfect fit very unlikely, but the final decision as to an adequate fit must be made by each user of the data in terms of the context of the intended use and how much accuracy is desired. Although these are the generally accepted criteria, the fit criteria used here merely serve as an unambiguous standard for decision making.

Group Invariance

A central requirement of a measure (under any model) is that it must perform the same mathematically regardless of other attributes of the thing being measured. When we identify groups of units of analysis on the basis of one of these attributes, the location of the items on the measurement continuum must remain stable across groups. Group invariance is tested by the goodness of fit between the item calibrations for two or more groups of theoretical relevance in the use of the measure. In the analysis conducted for this chapter, group invariance tests are conducted by pairwise group comparisons of the standardized delta values for each item across gender and ethnic groups. When the variable defining group membership consists of more than two categories, a series of pairwise comparisons of goodness of fit are conducted.

Given the multiple number of comparisons for the group invariance tests examined later in this chapter, it should be noted that invariance as a necessary characteristic of a measure was recognized by both Thurstone and Thorndike (see Englehard 1991, 1992), but their concerns were never institutionalized into listed measurement standards in their field of psychology. Moreover, discussions of invariance in "classical" measurement approaches are centered on factor structure invariance, which is not adequate for evaluating the kinds of responses of interest in this chapter since the factor structure cannot be sample free. Thus, the Rasch approach is the only measurement model in which invariance is a central, routine, and appropriate part of the analysis.

METHOD AND PROCEDURES

Participants

Data for this Rasch model analysis were collected from school records and self-report surveys between the summer of 1989 and the winter of 1991 from three middle school and secondary school adolescent groups composed of self-identified Anglos, American Indians, and Hispanics. The participant pool consisted of youth who were in good academic standing in school (GAS), those who were academically "at risk" (AR), and those who had dropped out of school (DO) and had been out for at least 1 month. Data were collected from

six sites in the western and southwestern parts of the United States.

A total of 3,986 adolescents completed the survey form. Sample sizes for each ethnic group varied according to gender and academic status, as follows: Anglos = 1,119 (571 males and 548 females, with an overall mean age of 16.7 and a standard deviation of 1.1); American Indian = 767 (342 males and 425 females, with an overall mean age of 16.5 and a standard deviation of 1.6); Hispanics = 2,100 (1,180 males and 920 females, with an overall mean age of 16.5 and a standard deviation of 1.2). Sample sizes for each of the academic status conditions were as follows: Anglos, GAS = 355, AR = 325, and DO = 439; American Indians, GAS = 243, AR = 255, and DO = 269; and Hispanics, GAS = 635, AR = 691, and DO = 774.

Materials

The survey was a multiple-scale instrument developed by the staff at the Tri-Ethnic Center for Prevention Research at Colorado State University, using scales that had been developed for previous studies. There were more than 1,000 items in the survey, and it took between 60 and 90 minutes to complete. A seven-item self-esteem scale was selected for use in the measurement analysis. The short scale consisted of the following items: "I like myself," "I am good at games," I am good looking," "I am lucky," "I am proud of myself," "I am intelligent," and "I am able to do things well."

Self-esteem scale items initially were treated with the usual correlational statistical routines to determine their psychometric properties. Cronbach's alpha ranged from 0.78 to 0.85 for each of the three ethnic groups and all groups combined. A principal components factor analysis produced two factors, with the first factor accounting for 68 percent of the variance. Both sets of findings gave the researchers at the Tri-Ethnic Center for Prevention Research considerable confidence in the reliability of the self-esteem items; consequently, the scale has been used in several studies concerning AOD use among ethnic populations.

Procedure

Data for the school-based population were collected during school hours; data for the DO sample were collected at different times of the day depending on the availability of the participant. Survey questionnaires were identified by number only. Upon completion of the survey and in the presence of the field researcher, the survey was sealed in an envelope and immediately mailed in for data entry and processing. Respondents were paid $10 for their participation.

All participants were assured of confidentiality and were asked to sign a "consent to participate agreement" describing the rights and responsibilities of participation; parent consent was obtained for participants under the age of 18. Participants were informed that the survey itself and answers to the survey questions were protected by the U.S. Government's issue of a certificate of confidentiality that guarantees the legal confidentiality of all survey responses.

RESULTS

As indicated earlier in the chapter, this analysis was conducted using the Quest computer program (Adams and Siek-Toon 1993) from the Australian Council for Educational Research. Quest typically is used to construct and validate variables based on dichotomous and polychotomous observations such as Likert-type ordinal scales. The Quest (version 2.1) software program provides Rasch analysis item estimates, case estimates, and fit statistics. Results from our analysis of the self-esteem scale are first presented for the total sample; we then present comparisons by ethnic group, by gender (across all ethnic groups), and finally by ethnicity-gender.

TOTAL SAMPLE

Item Calibrations, Precision, Continuum Coverage, and Separation

As shown in figure 1, the seven items with four response categories fail to maintain response category order across all seven items. The "like self" response "some" has a higher scale value than the "good-looking" response "a lot." With this one exception, the items also maintain their relative scale value order across the three calibrated response categories. The range of scale values is less than ideal, with logit values from –2.78 to 2.04. The truncating of "self-liking" is con-

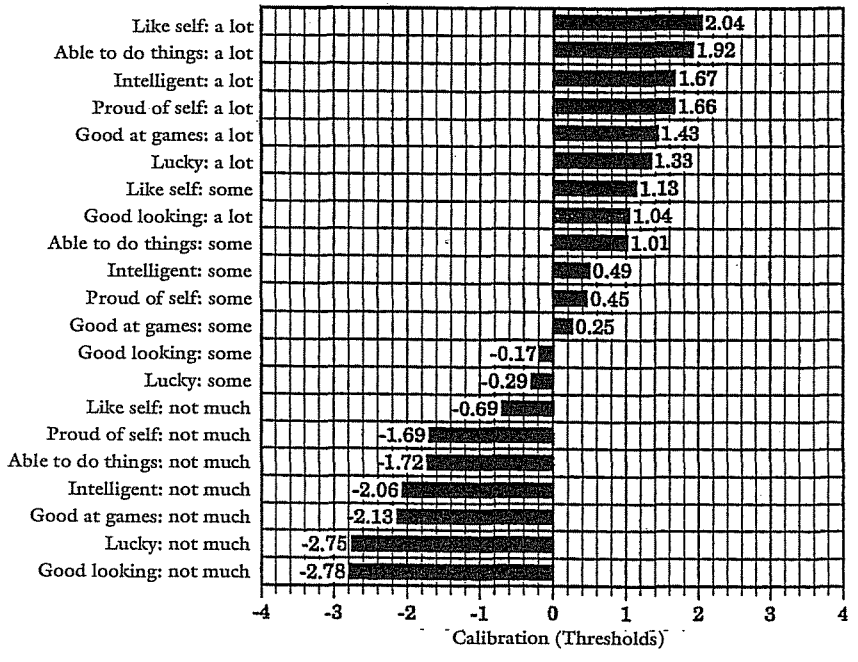

Figure 1. Item-response category calibrations, self-esteem items.

siderably greater at the high self-esteem end of the continuum. It is also apparent that several item-responses share essentially the same location on the continuum. Additionally, the lower end of the scale contains two major gaps in the construct continuum (between "like self: some" and "proud of self: some;" between "good at games: some" and "lucky: some"). Overall, the seven items are not particularly efficient in providing information about adolescent self-esteem.

The item scale values indicate that there are too many items, because some items occupy the same continuum location. In addition to this problem, there are gaps in the continuum and a truncated range of self-esteem scores. Regardless of these problems, the items statistically identify five levels of self-esteem with an item reliability of 0.96. This level of identification, however, is based on all seven items. Final assessment of item separation depends on only those items having acceptable fit.

Item Fit

As shown in table 1, only two of the seven items have acceptable fit with the model. "Good at games" and "lucky" have very high positive infit *t* values, indicating that they are not on the same dimension as the other five items. This is not unexpected, since "good at games" probably reflects athletic ability and "lucky" has no self-evident relationship to positive self-esteem. "Proud of self," "intelligent," and "able to do things" have excessively high negative infit *t* values,

Table 1. Item-Response Category Calibrations, Precision, and Fit, Total Sample.

		Response Category				
Item	Scale Value	Not Much	Some	A Lot	Infit Mean Square	Infit *t*
Like self	Logit	-0.69	1.13	2.04	0.95	-1.5
	SE	0.06	0.13	0.17		
Good at games	Logit	-2.13	0.25	1.43	1.24	8.2
	SE	0.06	0.09	0.13		
Good looking	Logit	-2.78	-0.17	1.04	0.99	-0.6
	SE	0.09	0.07	0.09		
Lucky	Logit	-2.75	-0.29	1.33	1.21	7.8
	SE	0.06	0.07	0.09		
Proud of self	Logit	-1.69	0.45	1.66	0.81	-7.5
	SE	0.06	0.11	0.12		
Intelligent	Logit	-2.06	0.49	1.67	0.92	-3.0
	SE	0.06	0.09	0.13		
Able to do things	Logit	-1.72	1.01	1.92	0.86	-5.0
	SE	0.06	0.12	0.16		

Note: SE = standard error.

indicating that they provide no information not already provided by the other items. The two items having acceptable fit are "like self" and "good-looking," with the latter item defining the construct. Since those items other than "lucky" and "good at games" have face validity in the context of self-esteem measurement, it appears that the perception and assessment of one's physical appearance (i.e., "like self" and "good-looking") are central to adolescent self-esteem.

Ethnic Group Comparisons

Item Characteristics: Anglos

Item fit by ethnic group is shown in table 2. Among Anglo adolescents only "good looking" and "intelligent" have acceptable fit. "Good at games" and "lucky" are, not unexpectedly, off dimension. "Like self," "proud of self," and "able to do things" are redundant items. Therefore, among Anglo adolescents self-esteem consists of considering oneself "intelligent" and "good-looking." The items have relatively low reliability (0.82), resulting in the identification of only two levels of self-esteem (separation = 2.13).

Item Characteristics: American Indians

Among American Indian adolescents, again only two items have acceptable fit, and in this case the items are "like self" and "good looking." "Good at games" and "lucky" again are off dimension, and "able to do things," "intelligent," and "proud of self" are all redundant. Thus, what constitutes the central ingredients of self-esteem is somewhat different than for Anglo adolescents. In this case, the reliability of the items is slightly improved over that of Anglos (0.88), but still only two levels of self-esteem are statistically identified (separation = 2.71).

Item Characteristics: Hispanics

Among Hispanics three items have acceptable fit: "like self," "good-looking," and "intelligent." "Lucky" and "good at games" are again off dimension, and "proud of self" and "able to do things" are redundant. It thus appears that the intention of those who developed the self-esteem scale is most nearly realized among Hispanic adolescents, since self-esteem is defined by those attributes important for both Anglos and American Indians. Moreover, reliability is considerably higher among Hispanics (0.92), resulting in the identification of three levels of self-esteem (separation = 3.39).

Clearly, the components of self-esteem differ by ethnicity, and this particular set of self-esteem items works best among Hispanic adolescents.

Group Invariance

To test for invariance in item location across the three ethnic groups, pairwise item invariance tests were conducted. In this test each item logit (delta) was computed for each ethnic group, and the goodness of fit of the pairs of item calibrations was tested by a chi-square. Alpha was set at 0.01 because a large number of comparisons were conducted. As shown in table 3, the greatest invariance exists

Table 2. Item-Response Category Calibrations, Precision, and Fit by Ethnicity.

Item	Scale Value	Response Category: Not Much	Some	A Lot	Infit Mean Square	Infit *t*
Anglo						
Like self	Logit	-1.19	0.95	2.27	0.88	-2.2
	SE	0.19	0.23	0.40		
Good at games	Logit	-2.19	0.12	1.41	1.28	4.7
	SE	0.16	0.20	0.23		
Good looking	Logit	-2.91	-0.07	1.09	0.95	-0.8
	SE	0.16	0.18	0.21		
Lucky	Logit	-2.78	-0.17	1.20	1.24	4.2
	SE	0.16	0.18	0.21		
Proud of self	Logit	-2.19	0.28	1.70	0.77	-4.5
	SE	0.16	0.20	0.28		
Intelligent	Logit	-1.88	0.77	1.87	0.95	-0.9
	SE	0.13	0.22	0.33		
Able to do things	Logit	-1.88	1.25	2.38	0.87	-2.4
	SE	0.19	0.28	0.41		
American Indian						
Like self	Logit	-0.56	1.34	2.23	0.93	-1.0
	SE	0.16	0.29	0.35		
Good at games	Logit	-2.31	0.39	1.72	1.33	5.4
	SE	0.19	0.18	0.23		
Good looking	Logit	-3.13	-0.37	1.07	1.02	0.3
	SE	0.16	0.16	0.20		
Lucky	Logit	-3.06	-0.36	1.53	1.27	4.7
	SE	0.19	0.16	0.22		
Proud of self	Logit	-1.44	0.46	1.79	0.78	-4.3
	SE	0.16	0.20	0.25		
Intelligent	Logit	-2.47	0.37	1.62	0.81	-3.5
	SE	0.16	0.19	0.25		
Able to do things	Logit	-1.88	1.13	2.02	0.83	-2.9
	SE	0.19	0.24	0.32		
Hispanic						
Like self	Logit	-0.56	1.13	1.88	0.98	-0.4
	SE	0.13	0.18	0.22		
Good at games	Logit	-2.06	0.24	1.36	1.19	4.8
	SE	0.13	0.12	0.13		
Good looking	Logit	-2.63	-0.16	1.02	0.99	-0.2
	SE	0.13	0.12	0.13		

Table 2. *Continued*

Item	Scale Value	Response Category: Not Much	Some	A Lot	Infit Mean Square	Infit *t*
Hispanic *Continued*						
Lucky	Logit	-2.69	-0.31	1.28	1.18	4.9
	SE	0.13	0.10	0.14		
Proud of self	Logit	-1.63	0.52	1.60	0.83	-4.8
	SE	0.13	0.13	0.17		
Intelligent	Logit	-2.03	0.44	1.62	0.95	-1.2
	SE	0.09	0.15	0.18		
Able to do things	Logit	-1.63	0.89	1.75	0.87	-3.3
	SE	0.09	0.18	0.21		

Note: SE = standard error.

between Hispanics and American Indians, where only "good-looking" has a significantly different scale value. Seeing self as "good-looking" indicates more self-esteem for Hispanics than for American Indians. Three items have significantly different scale values between Anglos and American Indians. "Like self" and "proud of self" have higher scale values for American Indians, and "intelligent" has a higher scale value for Anglos. Three items are also significantly different between Anglos and Hispanics. "Proud of self" has a higher scale value for Hispanics, and "intelligent" and "able to do things" have higher scale values for Anglos.

It is important to note that scale values represent the interval value of the difficulty or group salience of the item. (Difficulty, in this case, has to do with an item's relationship to other items and other persons. It implies that a respondent has "difficulty" endorsing an item or set of items.) Thus, endorsing "good-looking" is more difficult for American Indians than it is for Hispanics; endorsing "like self" and "proud of self" is more difficult for American Indians than for Anglos; endorsing "intelligent" is more difficult for Anglos than for either American Indians or Hispanics; and endorsing "proud of self" is more difficult for Hispanics than Anglos.

Gender Comparisons Across All Ethnic Groups

Item Characteristics: Males

Item fit by gender is shown in table 4. For males three of the seven items have acceptable fit ("like self," "good looking," "intelligent"). "Good at games" and "lucky" are not on the

Table 3. Tests of Group Invariance of Items by Ethnicity.

	Scale Values[a] By Ethnicity			Comparisons ($\chi 2$)		
	Anglo	American Indian	Hispanic	Anglo vs. American Indian	Anglo vs. Hispanic	American Indian vs. Hispanic
Like self	0.68 (0.07)	0.99 (0.07)	0.80 (0.05)	9.62 *	2.25	4.59
Good at games	-0.22 (0.06)	-0.07 (0.06)	-0.16 (0.04)	2.95	0.69	1.45
Good looking	-0.63 (0.06)	-0.81 (0.06)	-0.58 (0.04)	4.50	0.48	10.36*
Lucky	-0.58 (0.06)	-0.63 (0.06)	0.57 (0.04)	0.30	0.02	0.63
Proud of self	-0.08 (0.06)	0.26 (0.06)	0.16 (0.04)	14.88*	9.86*	2.00
Intelligent	0.25 (0.07)	-0.16 (0.06)	0.02 (0.04)	19.20*	8.61*	5.63
Able to do things	0.58 (0.07)	0.43 (0.07)	0.34 (0.04)	2.41	8.40*	1.14

[a] Delta values, standard error in parentheses.
*$p < 0.01$.

self-esteem dimension, and "proud of self" and "able to do things" are redundant. The reliability for the items is sufficiently high (0.92), resulting in the identification of three levels of self-esteem (separation = 3.34).

Item Characteristics: Females

As shown in table 4, only two of the items have acceptable fit for females ("like self" and "good-looking"). It is more than noteworthy that self-esteem for females is best defined by physical appearance, since "good-looking" has perfect fit with the model. "Good at games" and "lucky" are again not part of self-esteem, and "proud of self," "intelligent," and "able to do things" are redundant. The items have relatively high reliability for females (0.93), and thus three levels of self-esteem are again statistically identifiable (separation = 3.69).

Group Invariance

Group invariance tests by gender are shown in table 5. Three of the seven items have significantly different scale values for males and females. "Good at games" and "good-looking" have significantly higher scale values for males, and "able to do things" has a higher scale value for females. It is more difficult for males to see themselves as good at games and as good-looking. It is more difficult for females to see themselves as able to do things as well as others.

Ethnicity-Gender Comparisons

Ethnic Differences Among Males

Among males the fit of the items is very similar for Anglos and Hispanics (table 6). In both groups "lucky" and

"good at games" are off dimension, and "proud of self" is redundant. Among Hispanics, however, "able to do things" is also redundant, while among Anglos this item has acceptable fit. Among American Indian males "good at games" and "lucky" are off dimension, and "able to do things" and "intelligent" are redundant. These differences have important implications for the measurement of adolescent self-esteem. While seeing

Table 4. Item-Response Category Calibrations (Thresholds), Precision, and Fit by Gender.

		Response Category				
Item	Scale Value	Not Much	Some	A Lot	Infit Mean Square	Infit *t*
Males						
Like self	Logit	-0.50	1.09	1.77	0.99	-0.2
	SE	0.13	0.20	0.21		
Good at games	Logit	-1.72	0.48	1.28	1.19	4.4
	SE	0.09	0.13	0.17		
Good looking	Logit	-2.69	-0.14	1.19	0.98	-0.6
	SE	0.09	0.13	0.14		
Lucky	Logit	-2.75	-0.31	1.24	1.19	5.1
	SE	0.09	0.11	0.12		
Proud of self	Logit	-1.59	0.47	1.47	0.83	-4.5
	SE	0.09	0.15	0.16		
Intelligent	Logit	-2.00	0.42	1.52	0.95	-1.4
	SE	0.13	0.13	0.17		
Able to do things	Logit	-1.56	0.87	1.65	0.88	-3.0
	SE	0.09	0.16	0.20		
Females						
Like self	Logit	-0.94	1.16	2.33	0.94	-1.6
	SE	0.13	0.16	0.24		
Good at games	Logit	-2.69	0.03	1.58	1.32	7.6
	SE	0.13	0.13	0.17		
Good looking	Logit	-2.84	-0.22	0.94	1.00	0.0
	SE	0.13	0.09	0.14		
Lucky	Logit	-2.75	-0.25	1.42	1.22	5.5
	SE	0.13	0.12	0.12		
Proud of self	Logit	-1.81	0.43	1.85	0.79	-6.0
	SE	0.13	0.13	0.20		
Intelligent	Logit	-2.16	0.58	1.86	0.87	-3.3
	SE	0.13	0.13	0.19		
Able to do things	Logit	-1.91	1.19	2.23	0.83	-4.1
	SE	0.13	0.16	0.24		

Note: SE = standard error.

Table 5. Tests of Group Invariance of Items by Gender.

Item	Males[a]	Females[a]	χ^2
Like self	0.78 (0.05)	0.85 (0.05)	1.15
Good at games	0.00 (0.04)	-0.35 (0.04)	34.42*
Good looking	-0.55 (0.04)	-0.71 (0.04)	8.38*
Lucky	-0.62 (0.04)	-0.52 (0.04)	2.63
Proud of self	0.10 (0.04)	0.16 (0.04)	0.83
Intelligent	-0.02 (0.04)	0.09 (0.05)	3.01
Able to do things	0.31 (0.04)	0.50 (0.05)	8.15*

[a]Scale values are deltas.
*$p < 0.01$.

Table 6. Item Fit by Gender and Ethnicity.

	Males			Females		
Item	Anglo	Hispanic	American Indian	Anglo	Hispanic	American Indian
Life self						
Mean square	0.90	1.01	0.98	0.89	0.96	0.91
Infit t	-1.1	0.2	-0.1	-1.6	-0.7	-1.1
Good at games						
Mean square	1.20	1.14	1.38	1.37	1.29	1.34
Infit t	2.3	2.6	3.4	4.5	4.8	4.2
Good looking						
Mean square	0.97	0.99	0.95	0.96	0.99	1.07
Infit t	-0.4	-0.1	-0.6	-0.5	-0.1	1.0
Lucky						
Mean square	1.27	1.15	1.23	1.16	1.22	1.28
Infit t	3.3	3.3	2.8	2.0	3.8	3.8
Proud of self						
Mean square	0.74	0.86	0.85	0.82	0.78	0.74
Infit t	-3.6	-2.8	-1.8	-2.5	-4.2	-4.0
Intelligent						
Mean square	0.99	0.96	0.82	0.92	0.90	0.79
Infit t	-0.1	-0.8	-2.1	-1.1	-1.7	-3.0
Able to do things						
Mean square	0.92	0.89	0.81	0.84	0.84	0.83
Infit t	-0.9	-2.2	-2.1	-2.2	-2.6	-2.1

Note: Item scale values are deltas.

oneself as intelligent is an important self-esteem item for Anglo and Hispanic males, it is not uniquely important for American Indian males. While ability to do things as well as others is an important self-esteem item for Anglos, it is not uniquely important for Hispanics or American Indians. "Proud of self" is not a unique self-esteem component for Anglos and Hispanics, but it is marginally so for American Indians.

Ethnic Differences Among Females

For females all seven items have the same fit characteristics for Anglos and Hispanics. "Good at games" and "lucky" are off dimension, and "proud of self" and "able to do things" are redundant. Among both Anglo and Hispanic females self-esteem is defined by seeing oneself as good-looking, liking oneself, and seeing oneself as intelligent. Among American Indian females the fit pattern is very similar to that of Anglos and Hispanics, with the exception that "intelligent" is redundant and thus not a unique element of self-esteem.

Comparing Gender Within Ethnic Groups

With regard to item fit, Anglo males and females are the same except for "able to do things," which is redundant for females but is a unique element of self-esteem for males. Hispanic males and females are nearly identical; and among American Indians the two genders also are nearly identical. For American Indians, "proud of self" is clearly redundant for females and marginally so for males. Thus, the only gender difference in items composing self-esteem is among Anglos, where ability to do things as well as others is a unique self-esteem component for males but not females. In summary, testing for gender differences within ethnicity reveals only one item having a significantly different scale value by gender among Anglos and Hispanics.

Group Invariance: Crossing Gender and Ethnicity

Group invariance was examined across all gender and ethnicity groups to evaluate the importance of "ethgender" in self-esteem measurement. This analysis does not include "good at games" or "lucky" because in all gender and ethnic group combinations these two items were not part of the self-esteem dimension. Results of the group invariance tests are shown in table 7.

Among Anglos "like self" has a higher scale value for females. Among American Indians "able to do things" has a higher scale value for females. Among Hispanics none of the five items differs in scale value by gender. Among males "intelligent" has a higher score value for Hispanics than for Anglos or American Indians, and "intelligent" has a higher scale value for Anglos than for American Indians. However, "like self" has a higher scale value for American Indians than for Anglos and Hispanics (see table 3). Among females, "proud of self" has a higher scale value among Hispanics and American Indians than among Anglos; but Hispanic and American

Table 7. Items With Significantly Different ($p < 0.01$) Scale Values for Gender-Ethnicity Pairs.

	Anglo Males	Anglo Females	Hispanic Males	Hispanic Females	American Indian Males
Anglo Females	1:F > M	-	-	-	-
Hispanic Males	6:H>A	7:A>H 5:H>A	-	-	-
Hispanic Females	1:H>A	5:H>A	none	-	-
American Indian Males	1:N>A 6:A>N	7:A>N 5:N>A 6:A>N	1:N>H 6:H>N	6:H>N	-
American Indian Females	1:N>A 3:A>N	5:N>A	7:N>H 3:H>N	none	7:F>M

Note: F = female; M = male; A = Anglo; H = Hispanic; N = American Indian.

Indian females do not differ in the scale value of any of the five items.

CONCLUSIONS AND DISCUSSION

The results of the analysis have several implications for researchers interested in using survey-type scales for the comparison of individuals from different cultural or ethnic groups. Moreover, the results indicate that there are considerable differences in the way gender and ethnic status influence responses to a common set of self-esteem items. The broad implications for the Rasch analysis findings and its relationship to measurement equivalence are not clear and, thus, merit further investigation.

It is not surprising that there is a considerable amount of disagreement in the research on ethnic differences in self-esteem (see Martinez and Dukes 1987, 1991; Schroeder et al. 1993; Moore et al. 1996). Clearly, self-esteem measurement involves appreciably more than summing up responses to a set of questions having face and content validity. Our gender and ethnicity analyses indicate that most of the rather standard self-esteem items are redundant and thus not unique contributors to determining levels of self-esteem. Setting gender and ethnicity differences aside, it is clear from our findings that physical appearance plays a central role in adolescent self-esteem.

When we turn to measuring self-esteem for adolescents from different ethnic groups, the measurement of self-esteem becomes more complicated. Among Anglo adolescents, self-esteem is defined by intelligence and physical appearance. Among American Indian adolescents, self-esteem is defined by liking oneself and physical appearance. Among Hispanic adolescents, self-esteem is defined by liking

oneself, intelligence, and physical appearance. Clearly, what attributes constitute the central elements of self-esteem differ considerably by ethnicity, and any measure not taking this fact into consideration lacks construct validity.

The importance of evaluating the measurement of self-esteem by ethnicity is readily apparent in item invariance across ethnic groups. Recall that Rasch measurement approaches measure differences—differences between persons, between items, and between persons and items—hence, the invariance property is important in understanding outcomes. Any measure of self-esteem that fails to have item invariance across ethnic groups contains built-in bias in self-esteem scores for the different groups. These findings strongly suggest that the current measurement of self-esteem contains a considerable amount of ethnic group bias, and thus lack of construct validity. Not unexpectedly, item invariance is greatest for the two ethnic minority groups and greatest for Anglos compared with American Indians and Hispanics.

The analysis of group invariance of self-esteem items also provides important information about ethnic difference in self-esteem. Endorsing "like self" and "proud of self" is more difficult for American Indians than for Anglos, and endorsing "proud of self" is more difficult for Hispanics than for Anglos. On the other hand, endorsing "intelligent" is more difficult for Anglos than for either Hispanics or American Indians. Between the two minority groups, endorsing "good-looking" is more difficult for Hispanics than for American Indians. The numerous differences in scale value location of what are generally considered indicators of self-esteem emphasize the need for extreme caution in merely summing scores across a set of items and comparing the means of different ethnic groups. Such an exercise will almost invariably result in incorrect conclusions regarding ethnic differences in self-esteem. More than that, though, the exercise probably violates differences in measurement equivalence.

Gender findings add more complications to self-esteem measurement that at least partially contribute to the confusion regardless of gender differences in self-esteem. For males, the attributions "like self," "good looking," and "intelligent" uniquely contribute to the measurement of self-esteem, and "like self" and "good looking" define the self-esteem construct. Such attributions as "able to do things" and "proud of self" are redundant with the above attributions. Among females, however, only the attributions "like self" and "good looking" uniquely indicate level of self-esteem, and "good-looking" clearly defines the self-esteem construct. The attributions of "intelligence," "ability," and "pride" are redundant with "liking oneself" and, more pointedly, "physical appearance." Turning to item invariance, results further reveal the difficulty in merely creating summated scores by gender. It is more difficult for males than females to see themselves as "good-looking," and it is more difficult for males than females to see themselves as "able to do things."

Problems associated with measuring self-esteem are further compli-

cated by the finding that the components, defining elements, and scale value weights of self-esteem items differ by combinations of ethnicity and gender. The finding suggests that the two variables interact in ways that invite further research and analysis in the domain of ethnicity.

In this chapter, ethnicity was treated as a nominal variable where respondents self-identified their ethnic affiliation. Data are available from the project to determine the depth and degree of ethnic identity for the respondents. Within each ethnic group, disaggregated analyses can be performed to determine if the degree to which respondents identify with their self-identified group will create yet another subset of scale items measuring self-esteem (see Trimble 1995). Analyses of combinations by degree of ethnic identity within gender groups may assist researchers in discovering the extent to which each of the sets interact and covary and in further understanding the dynamics associated with studies of the self-esteem. Moreover, it may be that people from specific tribes who are marginally acculturated may view the self-esteem construct very differently than someone from another tribe who shares the same level of acculturative status; the combinations of different statuses and tribal affiliations are staggering and stretch the imagination.

Researchers interested in measuring self-esteem and using scales with ethnic and cultural groups are encouraged to include indigenous (*emic*) items and closely follow the standards associated with measurement and cultural equivalence. Additionally, it is hoped that the findings produced by the Rasch analysis will encourage researchers to use the approach to analyze scales designed for use with ethnic and cultural groups. More important, it is hoped that the use of Rasch analysis will lead to new insights concerning culturally unique psychosocial processes not available through use of the usual psychometric procedures.

Finally, many American Indian communities continue to believe that levels of self-esteem in youth are related to AOD use. However, some of the research on the topic fails to support these beliefs. Perhaps there is a difference between the way researchers conceptualize the self construct and the way it is viewed in many Indian communities; to assess self-esteem, for example, many researchers continue to use variations of Rosenberg's self-esteem scale. Is Rosenberg's theory of self-esteem culturally equivalent to Indian worldviews? Is it culturally equivalent to tribal and band-specific worldviews? If there are differences between the worldviews, would it be possible to develop scientifically sound measures to tap the self-esteem that would permit culturally equivalent comparisons? To collect the information necessary to respond to the worldview equivalent, researchers should use quantitative research analysis techniques. The technique and the research must be conducted in close collaboration with Indian participants who are deeply grounded in their respective tribal lifeways and thoughtways. Answers to measurement and scale construction questions cannot be obtained until this process is completed.

ACKNOWLEDGMENTS

We wish to extend our gratitude to the staff at the Tri-Ethnic Center for Prevention Research at Colorado State University for their assistance and support in writing this chapter. Grant support was provided from the National Institute on Drug Abuse (P50 DA07074) and the National Institute on Alcohol Abuse and Alcoholism (AA 08302).

REFERENCES

Adams, R.J., and Siek-Toon, K. Quest: *The Interactive Test Analysis System.* Victoria, Australia: The Australian Council for Educational Research, 1993.

Andrich, D. *Rasch Models for Measurement.* Newbury Park, CA: Sage, 1988.

Berry, J. On cross-cultural comparability. *Int J Psychol* 4:119–128, 1969.

Berry, J. Introduction to methodology. In: Triandis, H., and Berry, J., eds. *Handbook of Cross-Cultural Psychology, Vol. 2: Methodology.* Boston: Allyn & Bacon, 1980. pp. 1–28.

Berry, J., and Dasen, P., eds. *Culture and Cognition.* London: Methuen, 1974.

Bontempo, R. Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *J Cross-Cultural Psychol* 24(2): 149–166, 1993.

Brislin, R.; Lonner, W.; and Thorndike, R. *Cross-Cultural Research Methods.* New York: John Wiley & Sons, 1973.

Dorgan, M.; Goebel, B.; and House, A.E. Generalizing about sex role and self-esteem: Results or effects? *Sex Roles* 9:719–724, 1983.

Drasgow, F. Study of the measurement bias of two standardized psychological tests. *J Appl Psychol* 72:19–29, 1987.

Dukes, R.L., and Martinez, R. The self impact of ethgender on self-esteem among adolescents. *Adolescence* 29: 105–115, 1994.

Duncan, O. D. *Notes on Social Measurement.* New York: Russell Sage Foundation, 1984.

Ellis, B.; Becker, P.; and Kimmel, H. An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *J Cross-Cultural Psychol* 24(2):133–148, 1993.

Englehard, G. Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. *J Res Dev Educ* 24:45–60, 1991.

Englehard, G. Historical views of invariance: Evidence from the measurement theories of Thorndike, Thurstone, and Rasch. *Educ Psychol Meas* 52:275–291, 1992.

Fleming, C., and Manson, S. *Substance Abuse Prevention in American Indian and Alaska Native Communities: A Literature Review and OSAP Program Survey.* Rockville, MD: Office for Substance Abuse Prevention, 1990.

Goldschmidt, W. *Comparative Functionalism.* Berkeley: University of California Press, 1966.

Gray-Little, B., and Applebaum, M.I. Instrumentality effects in the assessment of racial differences in self-esteem. *J Pers Soc Psychol* 37:1221–1229, 1979.

Irvine, S., and Berry, J., eds. *Human Assessment and Cultural Factors.* New York: Plenum, 1983.

Irvine, S., and Carroll, W. Testing and assessment across cultures: Issues in methodology and theory. In: Triandis,

H., and Berry, J., eds. *Handbook of Cross-Cultural Psychology, Vol. 2: Methodology.* Boston: Allyn & Bacon, 1980. pp. 181–244.

Jessor, R., and Jessor, S. *Problem Behavior and Psychosocial Development: A Longitudinal Study of Youth.* New York: Academic Press, 1977.

Kim, J., and Mueller, C. *Factor Analysis: Statistical Methods and Practical Issues.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills, CA: Sage, 1978.

Kline, P. The cross-cultural use of personality tests. In: Irvine, S., and Berry, J., eds. *Human Assessment and Cultural Factors.* New York: Plenum, 1983. pp. 337–352.

Kobben, A. Comparativists and non-comparativists in anthropology. In: Naroll, R., and Cohen, R., eds. *A Handbook of Method in Cultural Anthropology.* New York: Natural History Press, 1970. pp. 581–596.

Linacre, J. Factor analysis and Rasch. *Rasch Meas Trans* 9(4):470, 1996.

Lonner, W., and Berry, J., eds. *Field Methods in Cross-Cultural Research.* Newbury Park, CA: Sage, 1986.

Ludlow, L.H., and Haley, S. Rasch model logits: Interpretation, use, and transformation. *Educ Psychol Meas* 55:967–975, 1995.

Malpass, R., and Poortinga, Y. Strategies for design and analysis. In: Lonner, W., and Berry, J., eds. *Field Methods in Cross-Cultural Research.* Newbury Park, CA: Sage, 1986. pp. 47–83.

Martinez, R., and Dukes, E. Race, gender and self-esteem among youth. *Hispanic J Behav Sci* 9:427–443, 1987.

Martinez, R., and Dukes, R. Ethnic and gender differences in self-esteem. *Youth and Society* 22:318–338, 1991.

Masters, G. A Rasch model for partial credit scoring. *Psychometrika* 47:149–174, 1982.

Masters, G.N. Measurement models for ordered response categories. In: Langeheine, R., and Rost, J., eds. *Latent Trait Class Models.* New York: Plenum, 1988.

McNamara, T. *Measuring Second Language Performance.* New York: Longman, 1996.

Moore, S.; Laflin, M.; and Weis, D. The role of cultural norms in the self-esteem and drug use relationship. *Adolescence* 32(123):523–542, 1996.

Nishimoto, R. The cross-cultural metric equivalence of Langner's twenty-two item index. *J Soc Serv Res* 9(4):37–52, 1986.

Owan, T.; Palmer, I.; and Quintana, M. *School/Community-Based Alcoholism/Substance Abuse Prevention Survey.* Rockville, MD: Indian Health Service, 1987.

Poortinga, Y. Some implications of three different approaches to intercultural comparison. In: Berry, J., and Lonner, W., eds. *Applied Cross-Cultural Psychology.* Amsterdam: Swets & Zeitlinger, 1975.

Poortinga, Y. Psychometric approaches to intergroup comparison: The problem of equivalence. In: Irvine, S., and Berry, J., eds. *Human Assessment and Cultural Factors.* New York: Plenum, 1983. pp. 237–257.

Prince, R., and Mombour, W. A technique for improving linguistic equivalence in cross-cultural surveys. *J Soc Psychol* 13:229–237, 1967.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedogoiske

Institut, 1960 (Chicago: University of Chicago Press, 1980).

Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press, 1965.

Rosenberg, M. *Conceiving the Self*. New York: Basic Books, 1979.

Schroeder, D.; Laflin, M.; and Weis, D. Is there a relationship between self-esteem and drug use? Methodological and statistical limitations of the research. *J Drug Issues* 23(4):645–664, 1993.

Trimble, J.E. Self-understanding and perceived alienation among American Indians. *J Community Psychol* 15: 316–333, July 1987.

Trimble, J.E. Toward an understanding of ethnicity and ethnic identity, and their relationship with drug use research. In: Botvin, G.; Schinke, S.; and Orlandi, M., eds. *Drug Abuse Prevention With Multiethnic Youth*. Thousand Oaks, CA: Sage, 1995. pp. 3–27.

Trimble, J.; Lonner, W.; and Boucher, J. Stalking the wily emic: Alternatives to cross-cultural measurement. In: Irvine, S., and Berry, J., eds. *Human Assessment and Cultural Factors*. New York: Plenum, 1983. pp. 259–273.

Windle, M.; Iwawaki, S.; and Lerner, R. Cross-cultural comparability of temperament among Japanese and American early and late adolescents. *J Adolesc Res* 2(4):423–446, 1987.

Wright, B. Georg Rasch and measurement. *Rasch Meas Trans* 2(3):25–32, 1988.

Wright, B. Data analysis and fit. *Rasch Meas Trans* 7(4):324, 1994.

Wright, B. Reliability and separation. *Rasch Meas Trans* 9(4):472, 1996.

Wright, B., and Masters, G. *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press, 1982.

Wylie, R.C. *The Self-Concept*. Vol. 2. Lincoln: University of Nebraska Press, 1979.