

Instrument Equivalence across Ethnic Groups

Antonio Olmos

Mental Health Corporation of Denver

Aolmos@MHCD.com

Susan R. Hutchinson

University of Northern Colorado

susan.hutchinson@unco.edu

Poster presented at the annual meeting of the American Evaluation Association, November 2002,  
Washington, D. C.

## **Importance of Cross-Cultural Measurement Equivalence**

“Standardized psychological measurement instruments must provide equivalent measurement across subpopulations if comparative statements are to have substantive import. Without equivalent measurement, observed scores ... are not directly comparable” (Drasgow & Kanfer, 1985). In the presence of measurement nonequivalence, findings of differences between individuals and groups cannot be unambiguously interpreted; differences in means could be just as easily interpreted as indicating that different things were measured.

### **How is cross-cultural equivalence manifested?**

Cross-cultural equivalence is manifested in three ways:

- When test scores from a given measure are equally accurate for all subgroups, i.e., reliability coefficients should be the same for all groups;
- When the factor structure on a given instrument is the same for all relevant groups, e.g., if depression has 2 subdomains, it should have 2 subdomains for all cultural subgroups;
- And when items on a particular instrument mean the same thing to people from different cultural groups, i.e., a psychological test that lacks item equivalence is in essence two different tests; one for each cultural group.

### **Instrument Equivalence in a Clinical Context**

In the past, it was often believed that etiology, expression, course, and outcome were universal and independent of cultural factors. Now, it is assumed that culture can play a role in psychopathology by: determining standards of normality and creating personality configurations that may look like pathological in one culture but not in another. "Normal" behaviors in one culture can be classified as “pathological” in another, e.g., dependency in Japan is valued, whereas in America it has negative connotations. Another problem is that culturally different individuals are not adequately represented in the norm groups who serve as the basis for scoring clinical measures. Classifying individuals of different cultures as pathological, based on culturally inappropriate norms, may lead to tragedy. In mental health settings where behavioral rating scales are used, often raters and ratees come from different cultures. If cross-cultural differences result in biased ratings, then the scales are not diagnostically valid for those groups, suggesting the need to generate norms for different cultures. In addition, findings of cross-cultural differences could lead to research to study the factor structure of the measure across different cultures.

### **Methods for Assessing Measurement Equivalence**

- Equality of reliability estimates  
In a mental health context, this would indicate consistency of ratings by a clinician across different ethnic groups
- Factorial invariance  
Involves determining if the factor structure (including number factors, factor loadings, and correlations between factors) is equivalent across groups

- Item response theory (IRT)  
Provides information about characteristics of individual items, including the relative difficulty or ease with which clients are given high ratings by clinicians

### **Purpose of the Study**

- To examine measurement equivalence of a clinical measure of depression across three ethnic groups (White/Caucasian, African-American, Hispanic) in adults and children diagnosed with depression
- To compare findings from three different methods for assessing measurement equivalence

### **Subjects**

- Adults (N = 1,182) and children (N = 778) with a primary diagnosis of major depression, who were clients of a large, urban mental health organization in the western U. S.
- Adults ranged in age between 18 and 65 (M = 40.29, SD = 11.96) and children ranged in age between 6 and 18 (M = 14.56, SD = 2.51)
- Ethnic breakdown for adults: White (n = 607), African American (n = 220), Hispanic (n = 355);
- Ethnic breakdown for children: White (n = 166), African American (n = 213), Hispanic (n = 399)

### **Instrument**

- The Problem Severity Scales from The Colorado Client Assessment Record comprise 18 symptoms or characteristics that are rated by a clinician on a scale of 1 (none) to 9 (extreme)
- The same scales are used for children and adults
- Factor analysis suggests the symptoms fall into two different dimensions: internalizing (10 items) and externalizing (8 items)

### **Data Analysis Procedures**

- Internal consistency reliability based on Cronbach's alpha was estimated separately for each ethnic group among children and adults on the two CCAR dimensions
- Rasch analysis was conducted on the two dimensions for each subgroup to determine appropriateness of the items
- Confirmatory factor analysis was conducted to test the fit of the two-factor model for each subgroup

### **Results**

- Reliability was similar for whites and African Americans for both children and adults on both the internalizing and externalizing dimensions, with reliability for Hispanics consistently lower (see table 1)

**Table 1.** Reliability scores by age (children, adults) and race (White, African American and Hispanic)

	<b>White</b>	<b>African American</b>	<b>Hispanic</b>	<b>ALL CHILDREN</b>
Internalizing	0.8516	0.8084	0.7708	0.8054
Externalizing	0.789	0.8249	0.7485	0.7813
<b>ADULT SCORES</b>				<b>ALL ADULTS</b>
Internalizing	0.8169	0.7965	0.7259	0.7926
Externalizing	0.7393	0.7517	0.746	0.7477

- Rasch analyses identified a number of items that exhibited different “difficulty” levels between ethnic groups (based on standardized mean differences), with more ethnic differences found among adults than children (see tables 2 and 3)
- No two ethnic groups tended to differ any more than any other groups overall, although certain groups differed more on some of the symptoms than on others
- Results of the confirmatory factor analysis revealed that the 2-factor model (based on 15 of the symptoms; see Figure 1) fit comparably among adults for Whites and African Americans and less well for Hispanics (see table 4)
  - Among children the fit work comparably for African Americans and Hispanics and less well for Whites see table 5)
- However, the presence of numerous correlated residuals suggests the possibility of a “halo” effect in providing ratings of symptoms

### **Conclusions**

- The CCAR has comparable reliability for white and African American children and adults, diagnosed with depression indicating clinician consistency in rating symptoms of depression
- Reliability is somewhat lower for Hispanic clients
- The factor structure is generally similar for the three groups indicating that the construct of depression is being measured in a similar way for different ethnic groups
- The halo effect might be more of a problem for Hispanic clients, based on the greater number of correlated residuals for that group
- Results of the Rasch analysis indicate that some of the items are operating differently across the three group
- For items exhibiting different difficulty levels, it may be that clinicians are either overdiagnosing or underdiagnosing particular symptoms for certain ethnic groups

### **Recommendations**

- Ideally, when comparing factor structures among ethnic groups, multiple groups invariance analysis analysis should be used
- In this study, due to substantial nonnormality in the data, this procedure was not possible

- However, there is currently little guidance in the literature regarding use of invariance testing in the presence of nonnormal data; therefore, future research should explore this topic further

### Implications for Program Evaluation

- Evaluators need to be aware of the issue of measurement equivalence when assessing outcomes, particularly when target populations include well-defined subgroups
- Ignoring the possibility of measurement nonequivalence could produce misleading findings in evaluation studies
- Apparent subgroup differences in outcomes could actually reflect measurement artifacts

**Table 2.** Standardized differences in “Difficulty” levels between ethnic groups (based on standardized mean differences) Adults. Internalizing dimension

ITEM	ZWB	ZWH	ZBH
Suicide-Danger to Self	<b>-2.77</b>	<b>-2.81</b>	0.26
Thought Processes	0.17	-0.20	-.31
Cognitive Processes	1.71	1.00	-0.78
Self-Care Basic Needs	1.54	<b>-4.20</b>	<b>-4.69</b>
Attention Problems	0.34	0.80	0.31
Emotional Withdrawal	0.20	1.60	1.24
Role	0.00	0.20	0.18
Anxiety	<b>-2.60</b>	<b>2.60</b>	<b>-2.30</b>
Interpersonal	0.40	<b>-2.20</b>	<b>-2.30</b>
Depression	1.37	<b>3.80</b>	1.72

**Note: bold scores are significant at  $p \leq 0.05$**

*Note:* ZWB = differences between white and African American; ZWH = differences between white and Hispanic; and ZBH = differences between African American and Hispanic

### Difficulty Measures for Adults, Externalizing Dimension

ITEM	ZWB	ZWH	ZBH
Security Management	-0.69	<b>-3.74</b>	<b>-2.59</b>
Manic Issues	-0.47	0.71	1.09
Violence/Danger to Others	<b>3.28</b>	<b>3.18</b>	-0.47
Legal Issues	0.69	1.80	0.78
Socialization	<b>3.80</b>	0.00	<b>-3.36</b>
Resistiveness	0.00	0.24	0.20
Medical/Physical	<b>-5.59</b>	<b>-3.33</b>	<b>2.60</b>
Family Issues	<b>-2.91</b>	1.66	<b>3.80</b>

Note: bold scores are significant at  $p \leq 0.05$

**Table 3.** Standardized differences in “Difficulty” levels between ethnic groups (based on standardized mean differences) Children. Internalizing dimension

ITEM	ZWB	ZWH	ZBH
Suicide-Danger to Self	<b>-5.31</b>	-0.28	<b>5.83</b>
Thought Processes	0.40	0.81	.35
Cognitive Problems	0.50	-0.14	-.67
Self-Care Basic Needs	<b>-2.54</b>	<b>-2.06</b>	.80
Attention Problems	0.28	<b>-2.74</b>	<b>-3.09</b>
Emotional Withdrawal	<b>1.98</b>	<b>3.26</b>	.86
Role	0.99	-1.20	<b>-2.40</b>
Anxiety	-0.71	0.17	1.03
Interpersonal	<b>3.82</b>	.51	<b>-4.12</b>
Depression	<b>3.11</b>	<b>2.74</b>	-1.03

Note: bold scores are significant at  $p \leq 0.05$

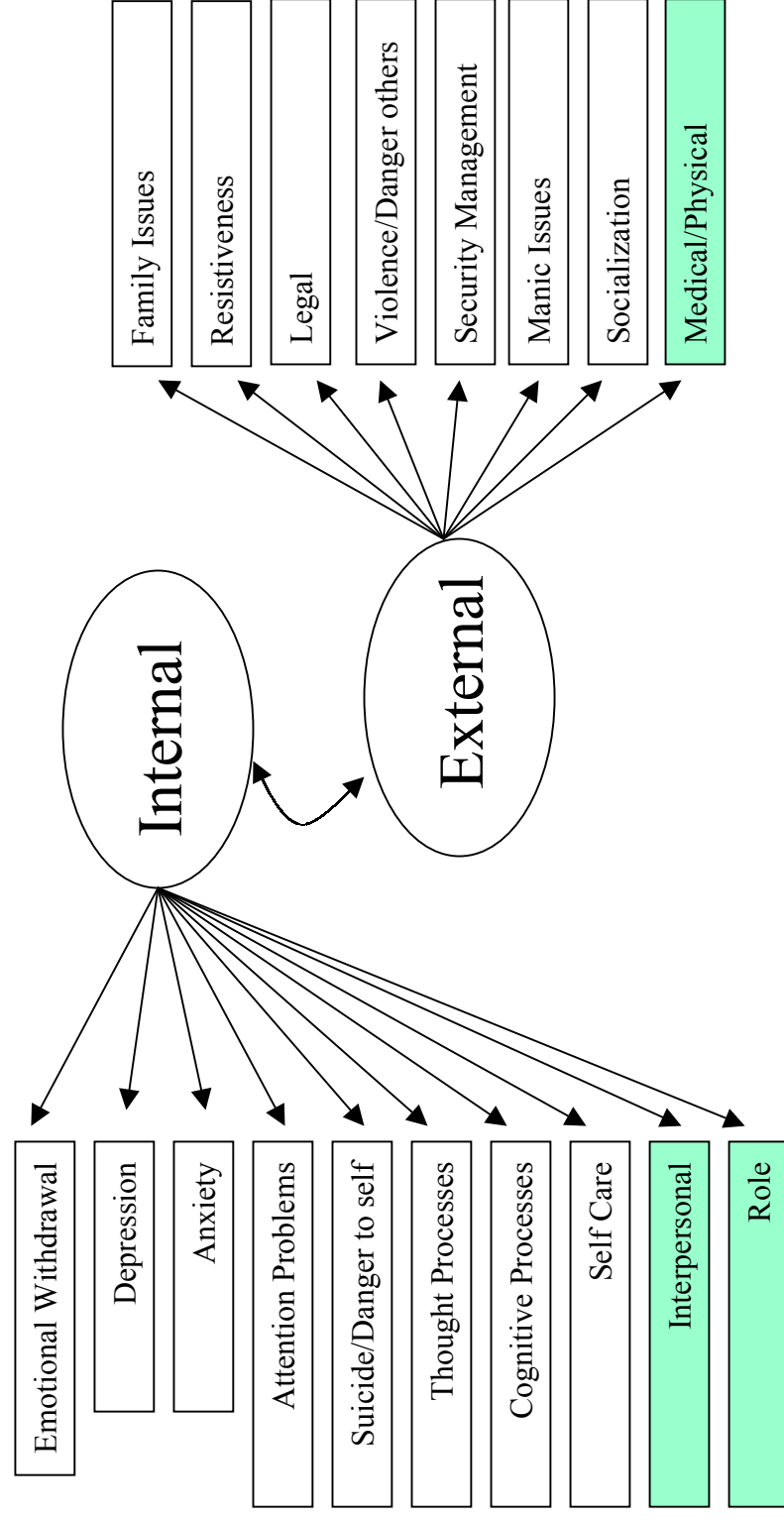
Note: ZWB = differences between white and African American; ZWH = differences between white and Hispanic; and ZBH = differences between African American and Hispanic

Difficulty Measures for Children, Externalizing Dimension

ITEM	ZWB	ZWH	ZBH
Security Management	0.14	-1.03	-1.20
Manic Issues	-1.41	<b>-3.12</b>	-1.56
Violence/Danger to Others	<b>2.12</b>	1.03	-1.54
Legal Issues	-0.28	<b>3.94</b>	<b>4.29</b>
Socialization	-0.94	0.34	1.60
Resistiveness	-1.27	<b>-2.74</b>	-1.20
Medical/Physical	1.41	<b>2.01</b>	0.46
Family Issues	-0.42	-1.20	-0.69

Note: bold scores are significant at  $p \leq 0.05$

Figure 1. Depression Model



Note: shaded items were eliminated after EFA due to poor fit