Instrument Equivalence across Ethnic Groups

Antonio Olmos
Mental Health Corporation of Denver
Susan R. Hutchinson
University of Northern Colorado

American Evaluation Association Conference
Washington DC., 2002

Importance of Cross-Cultural Measurement Equivalence

- Psychological measurement instruments must provide equivalent measurement across subpopulations if comparative statements are to have any validity
- In the absence of measurement equivalence, the instrument is likely measuring different traits for different groups

How is cross-cultural equivalence manifested?

- Scores from a given measure should be equally reliable for different groups
- The factor structure on a given instrument should be the same for all relevant groups
- Each item on a particular instrument should 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

Measurement Equivalence in a Mental Health Context

- Culture can play a role in diagnosis of psychopathology by:
  - Determining standards of normality
  - Creating personality configurations that may look like pathological in one culture but not in another
- Use of behavior rating scales where rater and ratee come from different cultures may result in biased ratings
  - If so, the scales are not diagnostically valid for those groups, indicating a need to generate separate norms

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 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); 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
- 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

Results – cont'd

- 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) fit comparably among adults for whites and African Americans and less well for Hispanics
  - 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

conclusions – cont'd

- 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