Culture Bias in Clinical Assessment: Using New Metrics to Address Thorny Problems in Practice and Research

MICHAEL CANUTE LAMBERT[1]

GEORGE T. ROWAN[2]

FREDRICK HICKLING[3]

MAUREEN SAMMS VAUGHAN[3]

[1]The university of North Carolina at Chapel Hill
[2]Michigan State university
[3]University of the West Indies

Agenda

- Brief focus on diverse groups—their challenges and strengths
- Measurement of functioning in diverse groups and concerns associated with historical and contemporary trends
- A look at measurement equivalence within and across diverse groups, as well as corrective efforts
- Focus on content and cultural validity for diverse groups—a thorny issue and how our efforts address this
- Some examples of how we created culture specific measures for different groups
- Introduction to item response theory (IRT) and how we use it to permit culturally appropriate and unbiased cross-group assessment and research
- The complexity of IRT and its strengths

Challenges Diverse Groups Face

- Persons from diverse groups often face tremendous challenges whether they live in industrialized nations such as the U.S. or in lower and middle income nations
- Regardless of challenges they face, such groups develop immense behavioral and emotional strengths
- Challenges experienced could also contribute to stress and stress-related problems that are also group-specific

Measuring Functioning of Diverse Groups in Research and Intervention

- Functioning in diverse groups has been historically compared with persons of European heritage
- Such studies use measures from North America or Europe

Why should we be concerned?

- Our ethical principles as researchers and practitioners require that our clinical and research findings do not create bias for diverse groups
- When diverse groups are compared (e.g., Latinos or African Americans with Euro-Americans) they should reflect true differences or similarities between such groups (Lambert, Ferguson, & Rowan, 2016)

Equivalence

- Language equivalence (often achieved through translation and back-translation)
- Conceptual Equivalence
- Psychometric Equivalence
  - Configural
  - Metric
  - Scalar

Typical Corrective Efforts

- Examining factorial equivalence (often stops here)
- Scalar and metric invariance studies sometimes done
- Such efforts represent a step in the right direction but also problematic
  - Ignores content, cultural, and conceptual validity for diverse groups

Content and Cultural Validity—A Thorny Issue

- Content validity is the pillar on which other forms of validity (including construct validity) rests
  - Rarely if ever achieved without cultural validity
- Specific questions to ask as one evaluates cultural validity
- Measures without appropriate content and cultural validity for diverse groups can adversely impact them by yielding inaccurate research and clinical findings especially in research where groups are compared
- Validly assessing, studying, and understanding functioning in diverse populations require measures that are designed to be culture specific to such populations and their psychometric properties should be estimated for these populations

Addressing the Challenges of Cultural Validity

- Cross-cultural behavioral scientists typically look at this in terms of etic vs. emic approaches
  - Etic theories, theoretical models and measures of constructs from such models are applied universally across cultural groups
  - Emic theories, theoretical models and measures of constructs from such models are applied specifically to each cultural groups
- More recently researchers have noted that either approach is inadequate and proposed integrating the two approaches

Hardin, Robitschek, Flores, Navarro, & Ashton (2015)

Our Approach

- A combination of the joint emic and etic approach
  - Creating measures that reflect cultural validity for each group studied
- Using new metrics to permit cultural specific measurement yet also allow unbiased measurement

Diversity in Specific Socioethnic Groupings

- Acknowledge that diversity exists even within specific ethnic groups (e.g., African American, Latino) yet common cultural characteristics and experiences are common within diverse groups (e.g., history of oppression, culture).
- We argue that persons within any groups of color could have psychological syndromes and strength dimensions that might be similar to those established for other groups
  - Posit that individuals might present their behavioral and emotional strengths and problems in a manner that is inconsistent with content depicted in items from European or North American measures
  - Constructs for certain groups might also be configured differently from those of others

Our Emic Measurement Approach

SOME EXAMPLES

The Caribbean Symptom Checklist

- Addressing cultural and content validity concerns items derived from data acquired from two sources
  - Focus groups that asked the broad question regarding behavioral and emotional difficulties do adults in the general Caribbean population observe in their fellow citizens
  - Studying the presenting problems from clinic records of more than 500 adults in more than a dozen mental health facilities (i.e., 12 outpatient and 3 inpatient facilities)
- Qualitative data derived from above sources examined and themes created
- Test items written from such themes

Conceptual Validity

- Identification of factors using Lengua et al. (2001) recommendation of both rational and empirical approaches
  - Rational approach, where 11 clinicians (i.e., three psychologists, seven psychiatrists, and one clinical social worker) grouped CSC items. According to broad categorical dimensions of the DSM-IV-TR labeled *Antisocial/Aggressive* (setting fires, stoning people) *Anxiety* (heavy Arms/legs, beating/shooting sensation in head) , *Attention Problems* (trouble finishing tasks, restless), *Depression* (feeling dead inside, head feels heavy) Hypomania/*Mania* (too religious, excessive singing), and *Psychosis* (believe obeahed, ganja tea, rolling sensation in head)
  - Ensuring essential unidimensionality, confirmatory conducted on each dimension showed good data to model fit (i.e., CFI and TLI ≥ 0.9, RMSEA ≤ 0.05)
  - Furthermore, full information factor analysis and bifactor analysis showed that items on this multidimensional models selected by clinicians met criteria for conditional independence

The Behavioral and Emotional Assessment for Children of Caribbean Heritage (BEACCH) Youth Self Report

Development of the BEACCH

- Developed with similar methodology as used with CSC
- Emerged from work on the Jamaica Youth Checklist (JYC)
  - JYC comprised of items from the Child Behavior Checklist and items derived from presenting problems listed in clinic records of well over 600 youth referred for clinical service
  - BEACCH contains only items on the JYC that match those on the Behavioral Assessment of African Heritage (BACAH) and items derived from presenting problems in clinic records of more than 600 youth

Derivation of Dimensions for BEACH

- From an early study where a dozen child psychiatrists and child psychiatry residents from UWI department of psychiatry and from a child guidance clinic examined items and sorted them under categories similar to the CSC
  - Since significant overlap exist between items on the BEACCH and CSC, item grouping compared with those of CSC to further calibrate groupings across the two measures permit use in longitudinal studies
- Note that the BEACCH has three forms
  - Parent-Report for ages 6 – 18 (BEACCH-P)
  - Teacher-Report for ages 5 – 18 (BEACCH-T)
  - Adolescent Self-Report for ages 11 to 18 (BEACH-A)

The Behavioral Assessment for Children of African Heritage

A CULTURALLY VALID MEASURE FOR AFRICAN AMERICAN YOUTH

Below is a list of items that describe a person's behavior. **On the left side of each item,** please **blacken** the number **(i.e., 0, 1, or 2)** which shows how true the item is for you. On the **right side of each item,** please **blacken** the number **(i.e., -1, 0, or 1)** which describes how this affects you? **Please rate yourself on all items.**

How true is this for you? 0 = not true (as far as you know) 1 = somewhat or sometimes true 2 = very true or often true		Based on your rating, what effect does this have on your life? -1 = negative effect 0 = no effect 1 = positive effect	How true is this for you? 0 = not true (as far as you know) 1 = somewhat or sometimes true 2 = very true or often true		Based on your rating, what effect does this have on your life? -1 = negative effect 0 = no effect 1 = positive effect
0 1 2	[1] Able to deal with my emotions	-1 0 1	0 1 2	[17] Do good school work	-1 0 1
0 1 2	[2] Accept my own mistakes or failures, accept constructive criticism	-1 0 1	0 1 2	[18] Express pride	-1 0 1
0 1 2	[3] Adjust to difficult problems or changing situations	-1 0 1	0 1 2	[19] Express feelings when appropriate, communicate well, have appropriate social skills	-1 0 1
0 1 2	[4] Appropriately physically active	-1 0 1	0 1 2	[20] Feel safe when I am in safe surroundings	-1 0 1
0 1 2	[5] Ask for and accept help when appropriate	-1 0 1	0 1 2	[21] Feel loved	-1 0 1
0 1 2	[6] Attend school regularly	-1 0 1	0 1 2	[22] Focused when I am working	-1 0 1
0 1 2	[7] Avoid fighting by appropriately	-1 0 1	0 1 2	[23] Follow directions	-1 0 1

The BACAH Measures

- The Behavioral Assessment for Children of African Heritage Measures (BACAH) (Lambert et al., 2005; 2016) developed to address some of the problems presented in the previous slides
- Four BACAH measures were developed (i.e. parent-teacher-, and self-report measures as well as an interview schedule)
  - Parent report assesses ages 4 to 18
  - Teacher report assesses from 5-18
  - Self report assesses 11-18
  - Interview schedule assesses 6-10
- Syndromes derived in similar fashion to those of the CSC and BEACCH
- Normed entirely on reports from 1,465 parents, teachers, and adolescents who reported on Black youth functioning

Using Etic-designed Measures in Unbiased Cross-Group (emic) Assessment?

THE BACAH RESILIENCE SCALE AS AN EXAMPLE

Cross-Informant and Cross-Ethnic Group Assessment

- Cross-ethnic group child/adolescent assessment exact heavy burden on researcher as gold standard is cross-informant assessment
  - Multiformat assessment across socio ethnic groups requires equivalence in scores across informants and ethnicities (note that types of items asked of each informant type vary in content and composition of factor solutions might also vary across informants)
- Our recent study on equivalence of the BACAH Resilience scales across African American and Jamaican Adolescents demonstrate equivalence is possible (Lambert et al., 2016)
- How did we do this?
  - We used item response theory linking to place different and similar sets of items across informants and socioethnic groups on the same scale and metric

Introducing Item Response Theory

- Defined—IRT addresses the probability that informants respond affirmatively to items that match the adolescent's trait (labeled θ)
- IRT requires that researchers ensure that certain assumptions are met
  - Appropriate dimensionality (IRT full information factor analysis)
  - Conditional independence (IRT bifactor model)
  - Most appropriate IRT model is selected
- IRT used to identify items with significant differential item functioning (DIF)—i.e., lack of invariance at the item level
- Link item parameter estimates across groups reducing bias in cross-group assessment

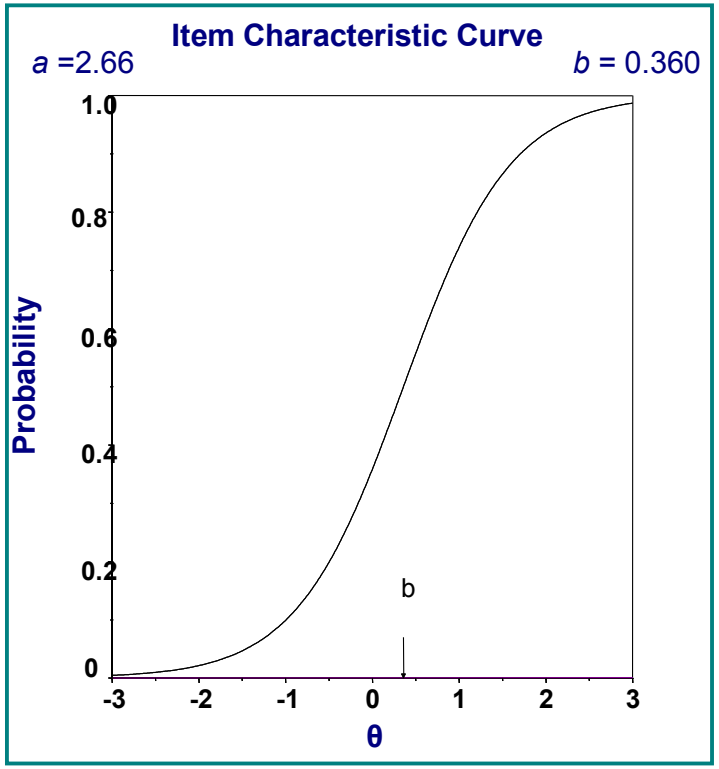
IRT Models

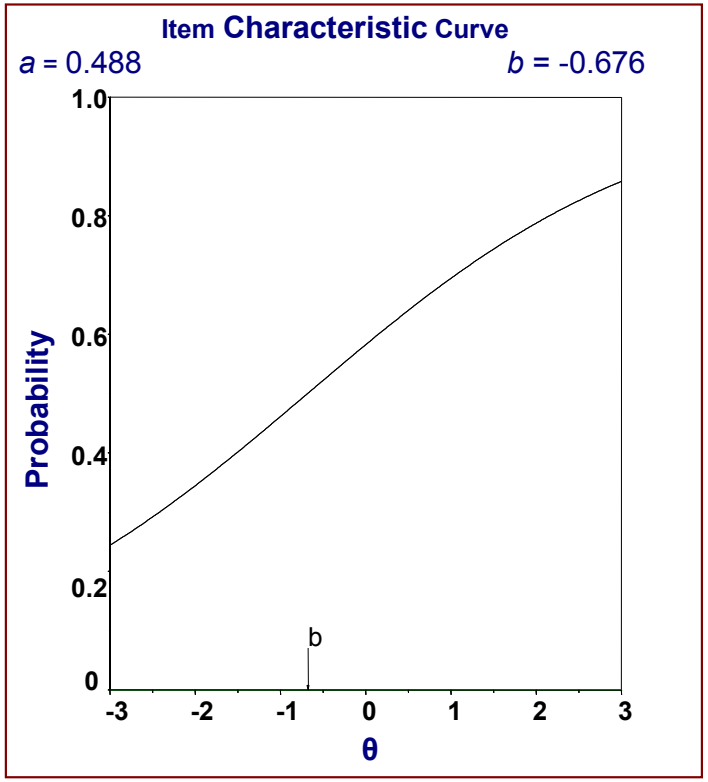
- One parameter (1PL) model (similar to Rasch) assumes identical discrimination parameter *(a)* and only location parameters *(b)* reflecting the level of function items measure are estimated
- 2PL model permits both *a* and *b* parameters to vary (Samejima's graded model, a variant of 2PL model applied to Likert scale items used in our studies)
- 3 PL model most often used in educational/ achievement testing and includes a *c* (guessing) parameter estimate where the probability of individuals responding correctly to an item that measures higher than trait levels they possess

Note on *a* and *b* Parameter Estimates

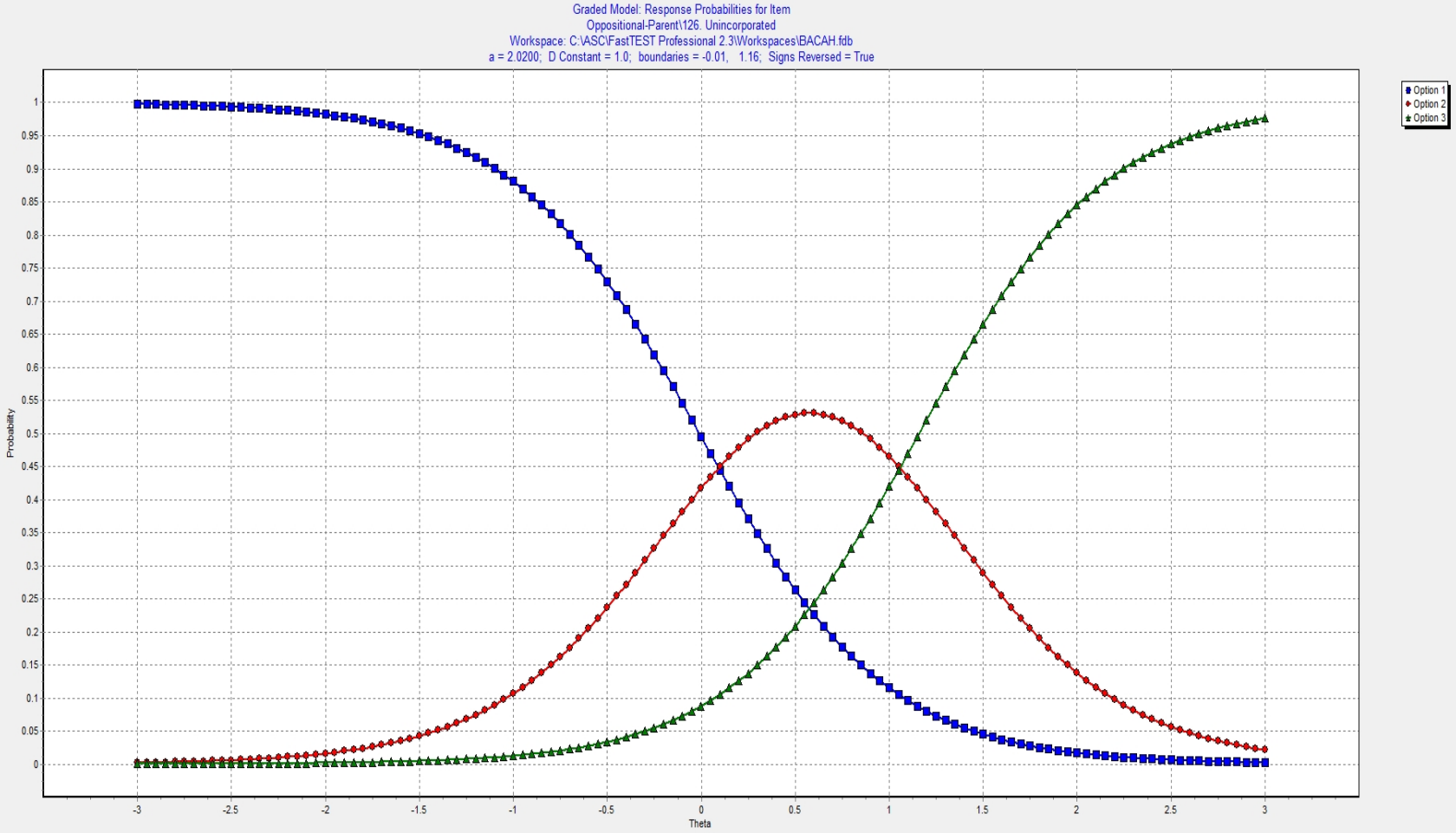
- *a* parameter estimate similar to a factor loading where higher a parameter estimate reflects greater discrimination (usually ranges form 0-3 and higher a parameter)
- *b* parameter estimate are standardized scores, similar to the intercept in CFA and ranges from negative to positive infinity (typically ranges from – 3 to 3)
  - Also known as boundary parameter estimates, where one less than the number points on a scale for items are estimated
- Note that IRT can accommodate mixed response formats and several different types of models can be estimated for different groups of items in a single run

Examples of ICC with Good Vs. Poor Discrimination





Item response Function for BACAH ODD Scale item “126. Uncooperative”



IRT Differential Item Functioning (DIF)

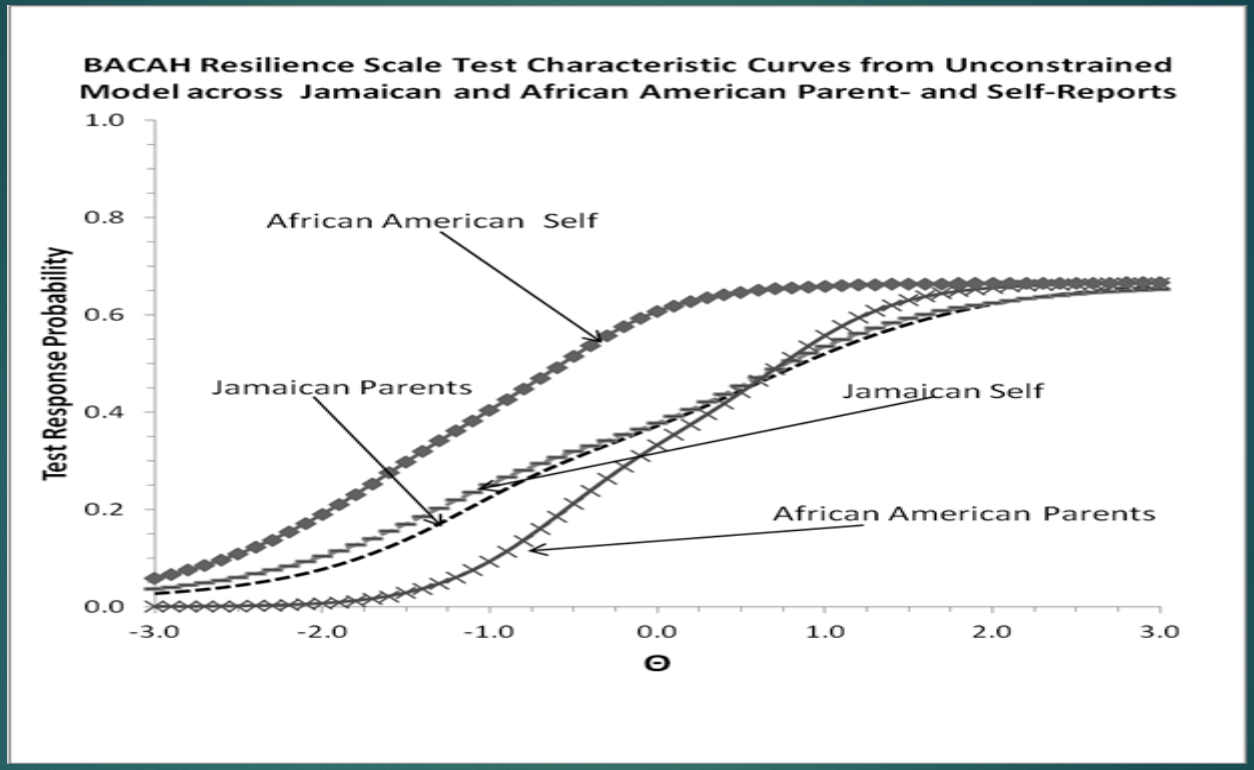
- IRT DIF is especially applicable to testing invariance at the item level, where significant DIF reflects absence of invariance
- Testing for DIF in the *a* parameter estimate tests for each item is a test of metric invariance (significant DIF called nonuniform DIF since response to an item for a group is higher at one end of the continuum and lower at the other end)
- Significant DIF for *b* parameter estimates is reflective of scalar invariance also called location DIF—indicating that scores for one group are uniformly higher than that of the other group
- Note that identifying items without significant DIF are essential in placing scales from different groups on an identical metric

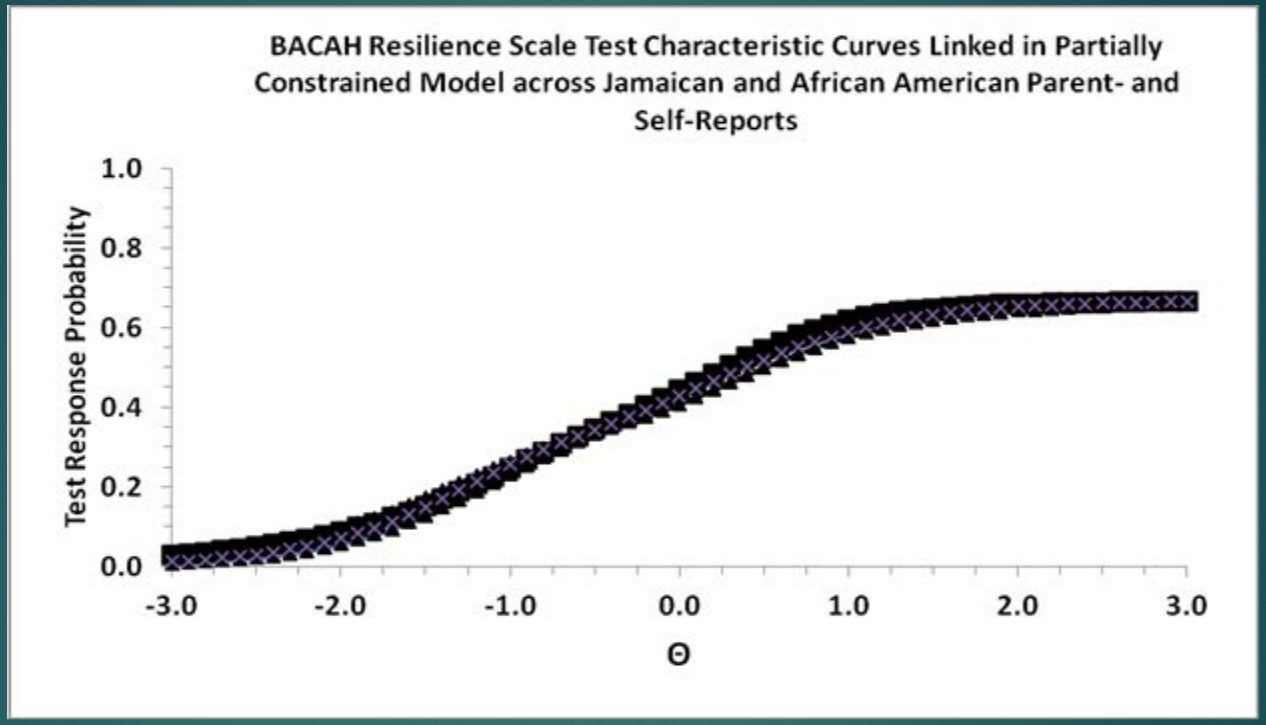
IRT Linking

- Derived from educational testing, where one might update or extend test without having to collect data on all old and all new items
- Many forms of linking exists but the test characteristic curve method is our choice (finds linking items by identifying items without significant DIF)
- Linking steps for BACAH Resilience scale across African American and Jamaican youth parent and self-reports
  - Identify identical items across ethnic and informant groups
  - Obtain data from each nationality X informant groups
  - Conduct confirmatory full information IRT factor analyses for each ethnic X informant group
  - Identify common items across each and use IRT analyses to identify items without significant DIF across various pairs of four groups
  - Constrain items without significant DIF and freely estimate item parameters across groups whose data are placed adjacent to each other.

Theoretical Cross-Group Linking of Items Without Significant DIF

First Item Group	Second Item Group	Third Item Group	Fourth Item Group	Fifth Item Group
A			A	A
B		B	B	B
C				
				D
E	E	E		
F	F			F
G	G			
	H	H		H
				I
	J	J		
	K	K		
	L		L	L
		M		
			N	
		O	O	


BACAH Resilience Scale Test Characteristic Curves from Unconstrained Model across Jamaican and African American Parent- and Self-Reports
Test Response Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Θ
African American Self
Jamaican Parents
Jamaican Self
African American Parents


BACAH Resilience Scale Test Characteristic Curves Linked in Partially Constrained Model across Jamaican and African American Parent- and Self-Reports
Test Response Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Θ

Discussion

- IRT is a group of modern measurement modeling procedures that are considered to outperform traditional methods
- We have shown that IRT is capable of permitting culture and informant specific measurement yet permit cross-group comparisons with reduced measurement bias

Why is IRT Infrequently Applied

- IRT very complex since most analyses occur at the item level and at least two and most often three or more parameters are estimated for each item
- IRT software applications complex to learn and to institute often requiring use of at least five different software programs to arrive at results such as those presented here
- Apart from M-Plus, I know of no other existing widely used software packages that include IRT analyses
- So why go through all of this trouble?
  - Easily lends itself to computerized adaptive testing that can shorten administration time by at least ½
  - CAT proven to provide identical or even more accurate test results than traditional testing
  - Permits easy use different sets of items on any dimension, as screening forms

Conclusion

- Researchers who conduct cross-ethnic group work bear the burden of proof that findings from their studies have minimal measurement bias across groups
  - Multi-informant measurement for children and adolescents makes job more complex
- Further research needed for the multiple behavioral and emotional strengths and problems scales
- Procedures demonstrated in this presentation can provide a foundation for scaffolding further research