QUALITATIVE AND QUANTITATIVE METHODS FOR ASSESSING MEASUREMENT EQUIVALENCE IN DIFFERENT SUBGROUPS

Ron D. Hays, Ph.D. and Leo S. Morales, M.D., Ph.D.

ISOQOL 2002 Annual Meeting

Workshop #3

Orlando, Florida, Caribe Royale
October 30, 2002 (9:00am-12:00 noon)

Course Syllabus

Example Focus Group protocol and report

Example Cognitive Interview protocol

Selected Readings

Web URLs

Session Outline

A Framework for Designing and Adapting Surveys.

Focus Groups and Cognitive Interviews.

Classical Test Theory to Assess Equivalence.

Item Response Theory (IRT) to Assess Equivalence.

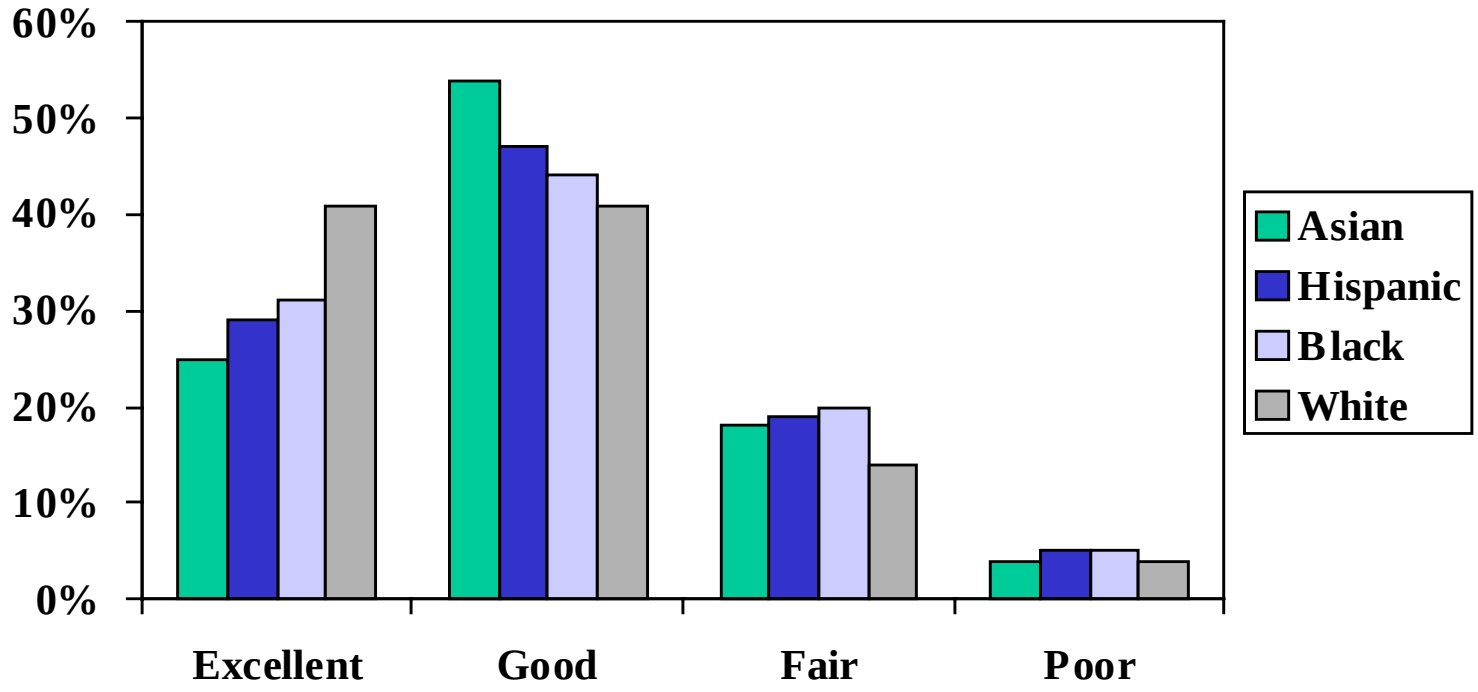
Introduction

Disparities in self-report measures are commonly reported (HRQOL, patient satisfaction)

Group comparisons require reliable, valid and equivalent measures

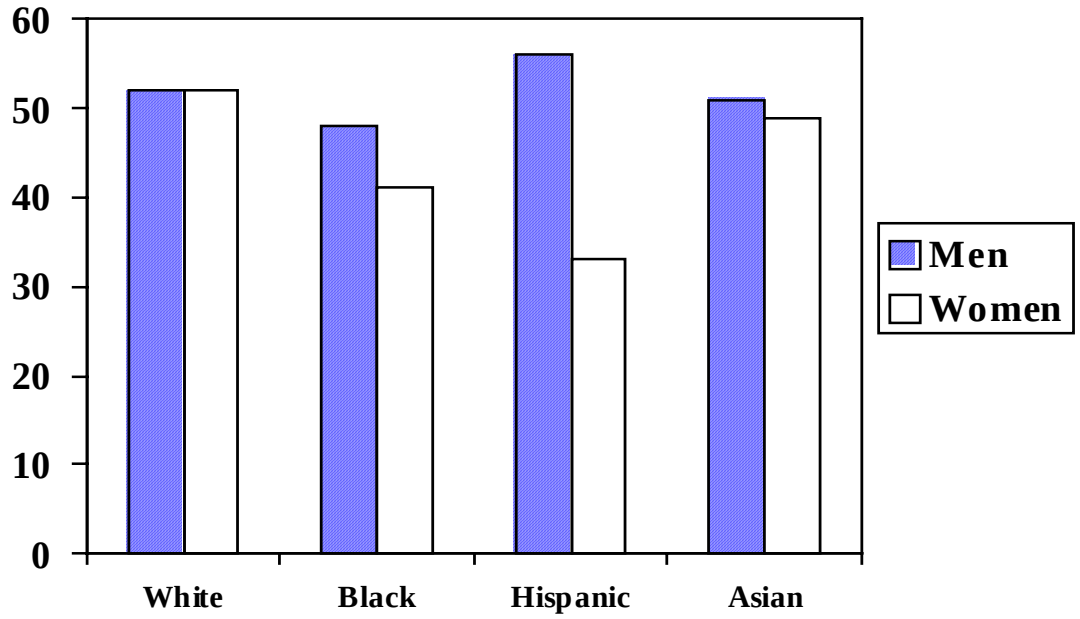
Framework to assure the performance of self-report measures in cross-cultural settings

General Health Rating, by Race and Ethnicity



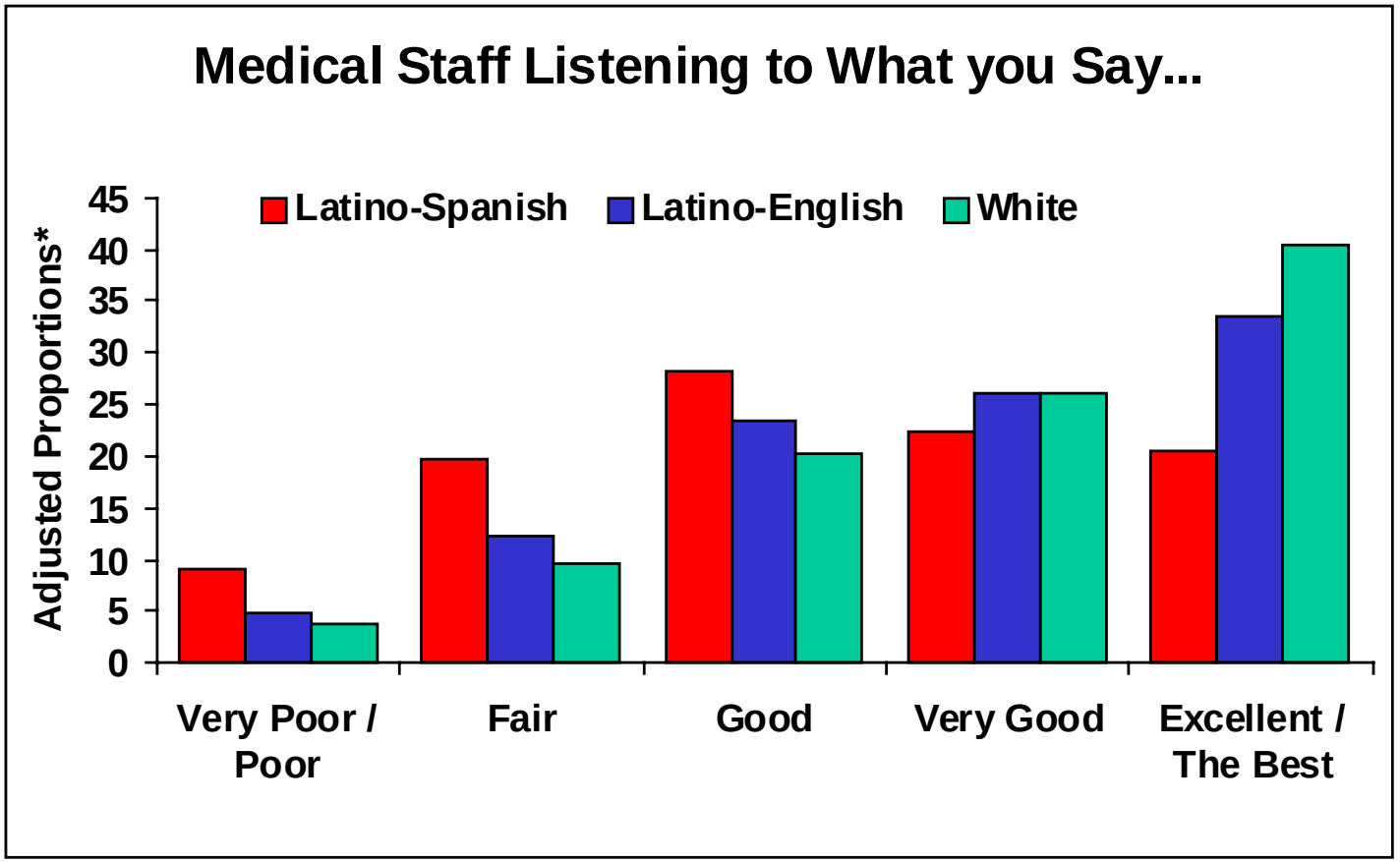
Source: Data from The Commonwealth Fund Minority Health Survey, 1994.

SF-20 Current Health Perceptions by Race/Ethnicity and Gender in the Medical Outcomes Study



Cunningham et al. JHCPU, 2000

Satisfaction by Ethnicity and Language



Morales et al. JGIM, 1999

Health Measures May Operate Differently in Various Groups

General health: excellent, very good, good, fair, poor response options

Spanish-speaking Hispanics were more likely to use "good" (buena) or "fair" (regular) responses than English-speakers

This may reflect differential meaning of the categories, or non-equivalence.

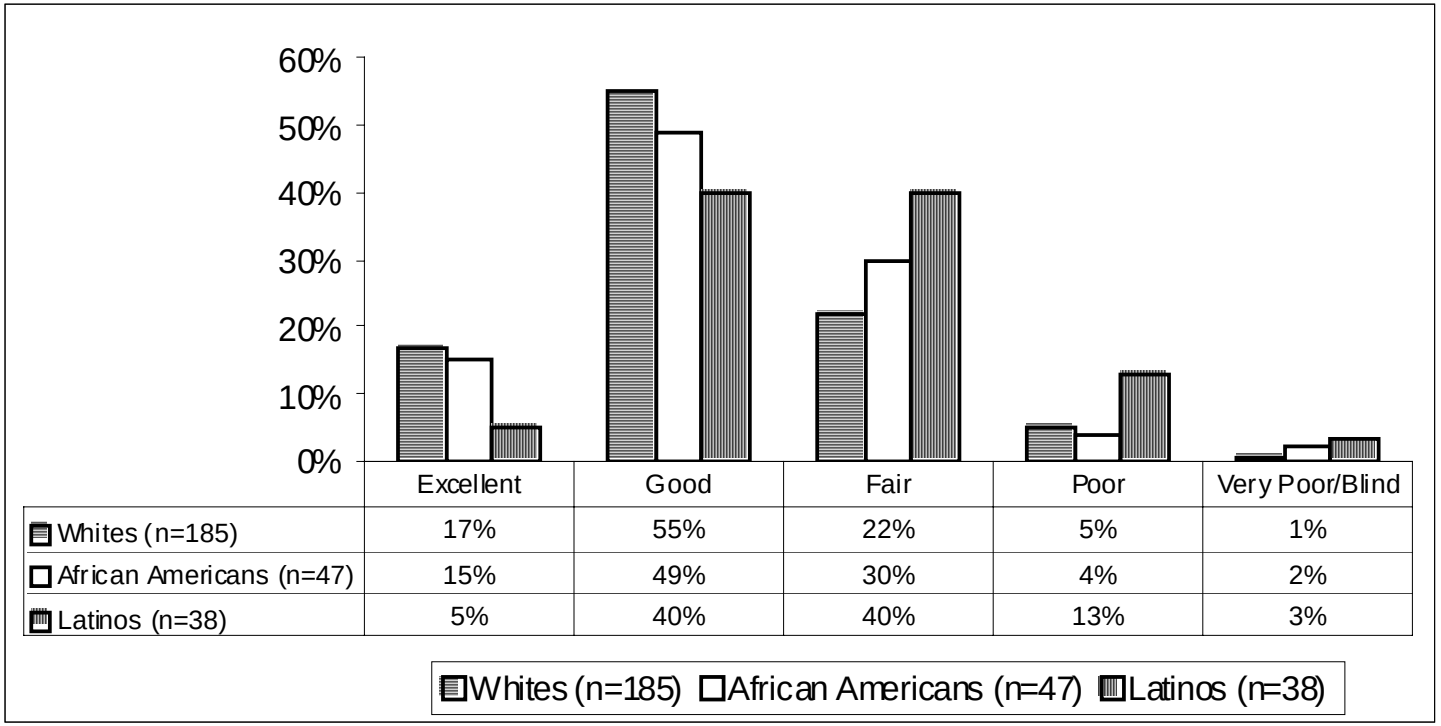
Angel and Guarnaccia, SSM 1989

A Measure Can Show A Difference Between Groups for One of Two Reasons:

The measure accurately reflects the underlying attribute for both groups when the two groups actually differ on the attribute.

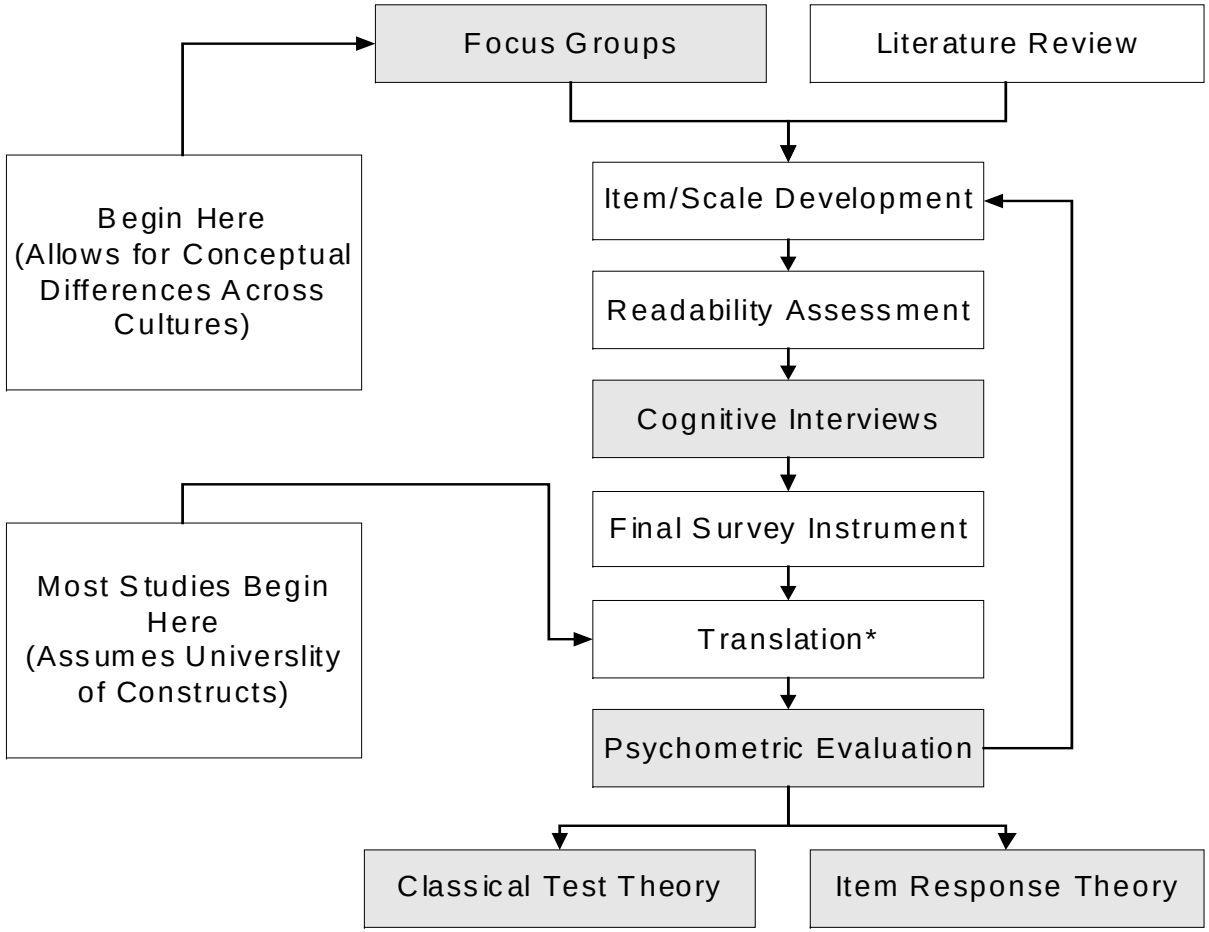
The measure inaccurately reflects the underlying attribute for one or both groups when the groups do or do not actually differ on the attribute.

Self-report of General Vision by Persons with 20/25 or Better Visual Acuity



	Excellent	Good	Fair	Poor	Very Poor/Blind
Whites (n=185)	17%	55%	22%	5%	1%
African Americans (n=47)	15%	49%	30%	4%	2%
Latinos (n=38)	5%	40%	40%	13%	3%

Survey Development Framework



Common Reasons for Adopting an Existing Measure

Replicate a previous study

So normative comparisons can be made

Existing measure is state-of-the-art

Save time/expense of developing new measure

FOCUS GROUPS

What are Focus Groups?

- Guided discussion held in a group
- Led by a moderator or team of moderators
- Participants chosen for relevance to research
- Discussion topics selected by researchers
- Recorded for later analysis

Advantages

Efficiency—can collect data from a group of people much more quickly and cheaply than individual interviews or large survey;

Allows researcher to interact directly with respondents;

Richness of data;

Allows participants to react and build upon other participants' responses;

Disadvantages

Convenience sample—cannot generalize to larger population;

Potentially biased results (biased by moderator or by overly influential participant);

Tendency to attach more value/credibility to findings than is warranted;

Open-ended nature of data may be difficult to summarize and interpret;

Stewart and Shamdasani (1991)

Uses of Focus Groups in Cultural Adaptation of Surveys

Assess whether domains currently covered in the survey adequately address needs and expectations of target population

Determine the need for developing new domains or expanding current domains

Assess whether concepts or specific terms used in the survey make sense to target population

Steps in Designing and Conducting Focus Groups

- Formulate the research question
- Identify the sampling frame
- Identify moderator
- Design the discussion guide
- Recruit the sample
- Conduct the focus group
- Analyze and interpret data
- Write final report

Source: Stewart and Shamdasani (1990)

Adapting Focus Group Methods to Cross-Cultural Needs (Strickland, 1999)

	Recommended	**Coastal Salish Tribe**
Focus group goal	Interactive dialogue	Difficult, but possible with extended time after meals
Participants	Should not be related	Not possible, most are related
	Should not participate more than once	Elders many not talk in the first meeting so must be invited more than once
Length	1-2 hours	Need 2-4 hours for rituals
Incentives	Gifts, food, money	Food, money, "gift away", presents made by the community
Seating	"U"	Circle

Special Issues in Conducting Focus Groups in Minority Communities

Recruitment

- Overcoming mistrust
- Personalistic forms of recruitment may work better

Group composition

- Group of strangers versus people who know each other;

Moderator

- From the community or familiar with racial/ethnic group;

Other Issues

Group dynamics

- Socially desirable/acceptable responses
- Fear/distrust
- Group norms/cultural constraints and patterns of communication

Community participation in the design of focus groups

Focus Group Logistics

HSPC/IRB

Recruitment and payment

Moderator and note-taker

Visitors

Audio-record (backup)

Background questionnaire

Focus Group Analysis

- Types of analysis strategies
  - Transcript-based
  - Tape-based
  - Note-based
  - Memory-based

    (Krueger, 1998)
- Computer software
  - Examples: Nudist, Atlas, WINMax
  - Indexing and cross-referencing tool
  - Not a substitute for analysis

Focus Group Example

Hepatitis C

- **rhays:**
- Example of unique issue that would be missed if survey only asked whether Jenny cut down on work instead of asking whether she adjusts her work behavior to ensure she doesn't deplete her energy reserve.
- Question 14a

Conservation of Energy

Because of Hepatitis C, Jenny has had to set "boundaries" with people. For example, she has had to close her office door and say "no" at times.

She has had to figure out how much energy she has to get to the finishing line. "If I don't have enough energy to get to the finishing line, it means I won't be available for everyone else."

- **rhays:**
- Question 15d, 16e, 25g

Longer Recovery Period

Candy has a full-time job (fixing earthquake damage) and "needs the rest of the day to be able to do that."

She engages in heavy work, but it takes her several hours to recover from it.

rhays:

Quest. 4a-d may not capture this very well.

Worklife accommodation and physical limitations

Manny is a self-employed builder who has owned his business for 25 years. He goes to bed at about 9:00 and gets up about 5:30 am. He wakes up several times during the night.

Manny has changed from participating actively on the job site with his employees to only setting up the job and supervision.

- **rhays:**
- Questions 12a, b, n

Symptoms of disease

Since he went off interferon, Manny has had a lot of muscle fatigue. His legs cramp when he surfs.

He can't arm wrestle his 12-year old son, etc. Forcing himself to do even 10 pushups now is difficult.

Manny has also been getting little pimples on his legs ("cryo globulins") due to inadequate blood circulation.

- **rhays:**
- Questions 14c, 14f, 27a

Medicine side effects

When taking interferon Manny said that he tended to get angry and yell and scream at his wife.

Manny also indicated that he gets a stomach ache daily somewhere between noon and 2 pm whether or not he has eaten. "All of the sudden I feel nauseous."

- **rhays:**
- Question 14e, 16d, 26e

Daily unpredictability leads to relationship changes

Tammy lost her best friend as a result of hepatitis C, because the friend didn't like her last minute cancellations caused by the disease.

However, it has brought her very close to her spouse—he is very supportive and even "babies" her now.

- rhays:
- Questions 26e, i, m

Keeping disease secret from others

Fred only recently told other people that he had hepatitis C (other than his wife). He told a few people and recommended that they get checked for it themselves.

He is more comfortable telling people now because he has more information about the disease and can explain what it is.

Cognitive Interviews

Cognitive Interviewing

To understand the thought processes employed in answering survey questions and to use this knowledge to construct, formulate, and ask better questions.

DeMaio and Rothbeb, 1996

Uses of Cognitive Testing for Cultural Adaptation of Surveys

Assess meaning of survey questions in target language.

Assess appropriateness of response options in target language.

Identify poor language usage.

Assess readability of survey questions in target language.

Model of Response Processes

Comprehension
- Understanding the questions

Retrieval
- Recalling information

Judgment
- Deciding relevance

Response
- Formulating answers

Tourangeau, 1984

Cognitive Interviewing Techniques

Concurrent Think Aloud

Respondent Debriefing

Confidence Ratings

Paraphrasing

Concurrent Think Aloud

Respondent verbalizes thoughts while going through the survey.

Interviewers encourages subject to think aloud:

- "Tell me what you are thinking"
- "Say more about that"

Specific probes

- "How did you decide to chose that answer?"
- "What does 'downhearted and blue' mean to you?"

Respondent Debriefing

Retrospective approach

- Interviewer asks respondent about completing the survey after completing the entire survey or a section of the survey.

Retrospective think aloud

- After completing survey, certain questions re-asked using think aloud.

Interviewer observation

- Identifies problematic skip patterns or questions in advance of debriefing

Confidence Ratings

Respondents rate their level of confidence in the answer they provide.

- Low confidence indicate lack of knowledge (especially proxies) or a difficult recall task

How old were you when you first started smoking?

- How confident are you in the answer you gave? Very confident, somewhat confident, not confident at all.

Paraphrasing

Respondents asked to paraphrase a question (repeat the question in your own words)

Example: Cognitive Testing from CAHPS®

Survey question:

- In the last 6 months, have you had problems getting to see a specialist? (yes/no)

Probes:

- Tell me more about that…
- What kind of doctor do you think of as a specialist?
- What is this question getting at? (IF YES) What kinds of problems have you had? Tell me about that.

Results of Cognitive Tests

Problems:

Infrequent events not well captured by yes/no format

Double negatives

Lost variability

Solution:

In the last six months, how often did doctors or other health professionals explain things in a way that you could understand?

Never, Sometimes, Usually, Always

Show Videotape

Summary

Cognitive testing identifies problems

- Question wording
- Response formats
- Instrument flow

Cognitive testing of alternate language survey versions can identify translation and equivalence problems

Break

Quantitative Analyses

Scoring

Average or sum all items in the same scale.

Transform raw average or sum to 0-100 possible range (linear transformation)

- **(raw score – minimum)* 100/(max – min)**

Field Test and Analyses

- Field test of translated survey instrument
- Psychometric analysis
  - Reliability estimates
    - Cronbach's alpha
  - Factor analysis
    - Test measurement invariance across groups
  - Item Response Theory (IRT) methods
    - Assess differential item functioning (DIF)

What's a Good Measure?

- Same person gets same score (reliability)
- Different people get different scores (validity)
- People get scores you expect (validity)
- It is practical to use (feasibility)

ANOVA Sources of Variance

Source	DF	Label
Between people (8)	7	BMS
Within people	48	WMS
Items (7)	*6*	JMS
Items x people	*42*	EMS

Intraclass Correlation and Reliability

Model	Reliability	Intraclass Correlation
One-Way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (K-1)MS_{WMS}}$
Two-Way Fixed	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{EMS} + (K-1)MS_{EMS}}$
Two-Way Random	$\frac{N(MS_{BMS} - MS)_{EMS}}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)MS_{EMS} + K(MS_{JMS} - MS_{EMS})/N}$

Reliability Minimum Standards

- For Group Comparisons
  - 0.70+ (Nunnally, 1978)
- For Individual Assessment
  - 0.90+ (Nunnally, 1978)

Multitrait Scaling Analysis

- Internal consistency reliability
  - Item convergence
- Item discrimination

Fake Multitrait/Multi-Item Correlation Matrix

	Trait #1	Trait #2	Trait #3
Item #1	0.80*	0.20	0.20
Item #2	0.80*	0.20	0.20
Item #3	0.80*	0.20	0.20
Item #4	0.20	0.80*	0.20
Item #5	0.20	0.80*	0.20
Item #6	0.20	0.80*	0.20
Item #7	0.20	0.20	0.80*
Item #8	0.20	0.20	0.80*
Item #9	0.20	0.20	0.80*

*Item-scale correlation, corrected for overlap.

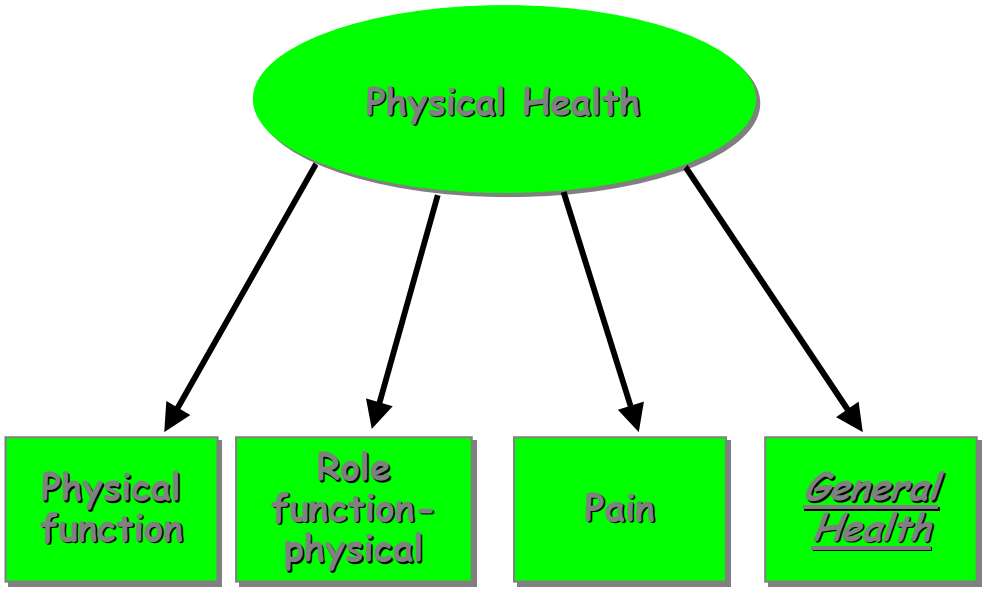
Two Steps in Exploratory Factor Analysis

Identify number of dimensions or factors

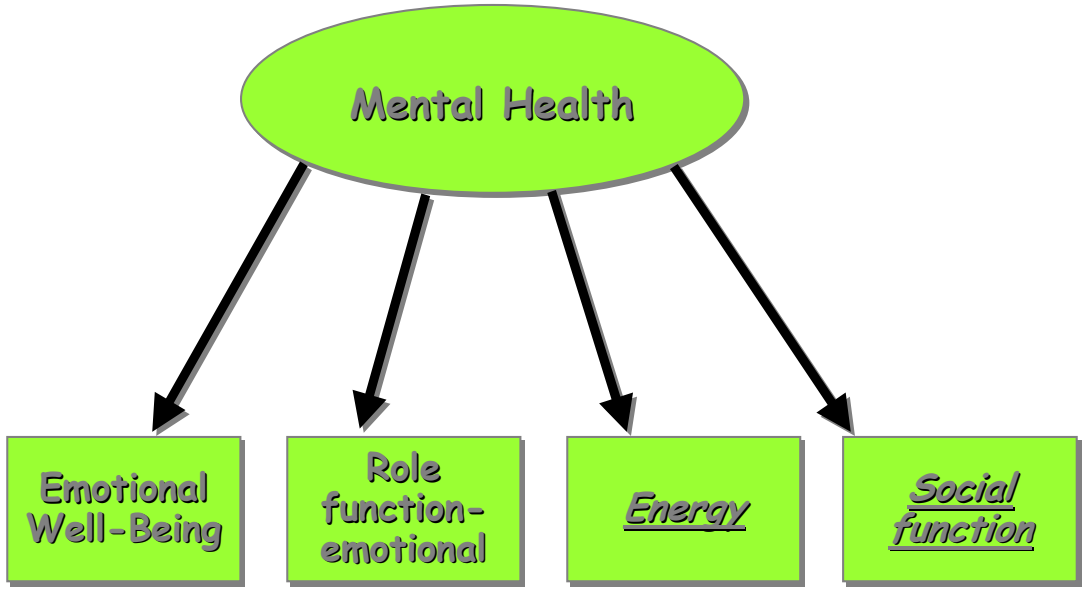
Rotate to simple structure

Fayers & Machin (1998) chapter

SF-36 Physical Health



SF-36 Mental Health



SF-36 Factor Analysis in Singapore

	English		Chinese		United States	
	Physical	Mental	Physical	Mental	Physical	Mental
PF	0.60	0.14	0.75	0.03	0.85	0.12
RP	0.85	0.12	0.78	0.25	0.81	0.27
BP	0.46	0.53	0.53	0.51	0.76	0.28
GH	0.14	0.74	0.32	0.66	0.69	0.37
VT	0.15	0.84	0.16	0.83	0.47	0.64
SF	0.49	0.56	0.48	0.56	0.42	0.67
RE	0.77	0.18	0.62	0.36	0.17	0.78
MH	0.12	0.83	0.10	0.86	0.17	0.87

Confirmatory Factor Analysis

- Compares observed covariances with covariances generated by hypothesized model
- Statistical and practical tests of fit
- Factor loadings
- Correlations between factors
- Regression coefficients

Fit Indices

- Normed fit index: $\frac{\chi^2_{null} - \chi^2_{model}}{\chi^2_{null}}$
- Non-normed fit index: $\frac{\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{model}}{df_{model}}}{\left(\frac{\chi^2_{null}}{df_{null}} - 1 \right)}$
- Comparative fit index: $1 - \left(\frac{\chi^2_{model} - df_{model}}{\chi^2_{null} - df_{null}} \right)$

Types of CFA Models

Configurally invariant model

- no cross-group equality constraints

Weak factorial invariance

- cross-group constraints on loadings

Partial weak factorial invariance

- released some constrains

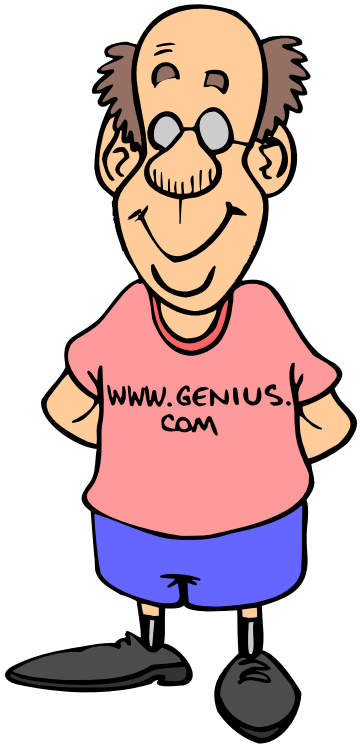
Factor correlation invariance

Partial factor correlation invariance

- constraints on some factor correlations

Higher order models

IRT



Overview

Brief Review of IRT Models

Examples of multiple-group analysis

What are IRT Models?

Mathematical equations that relate observed survey responses to a persons location on an unobservable latent trait (i.e., intelligence, patient satisfaction).

Types of IRT Models

- Unidimensional and multidimensional
- Dichotomous and polytomous
- Parameterization
  - One parameter: difficulty (location)
  - Two Parameter: difficulty and slope (discrimination)
  - Three Parameters: difficulty, slope, and guessing

IRT Model Assumptions

- Unidimensionality
  - One construct measured by items in a scale.
- Local Independence
  - Items uncorrelated when latent trait(s) have been controlled for.

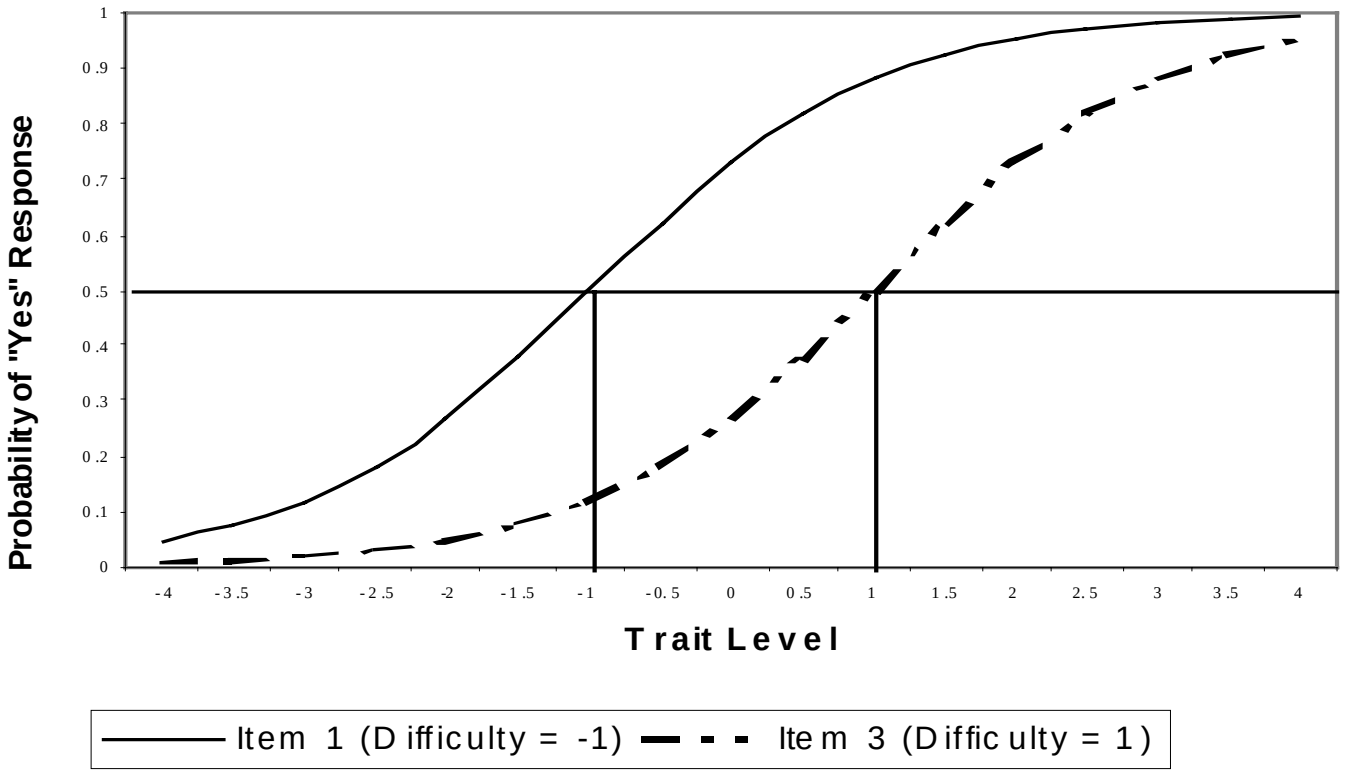
1-Parameter Logistic Model for (Dichotomous Outcomes)

$$P_i(\Theta) = \frac{e^{(\Theta - b_i)}}{1 + e^{(\Theta - b_i)}}$$

$P_i(\Theta)$ Probability that a randomly selected respondent with ability Θ (trait level) answers item i correctly.

b_i Item i difficulty.

Item Characteristic Curves
(1-Parameter Model)



2-Parameter Logistic Model for (Dichotomous Outcomes)

$$P_i(\Theta) = \frac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i(\Theta - b_i)}}$$

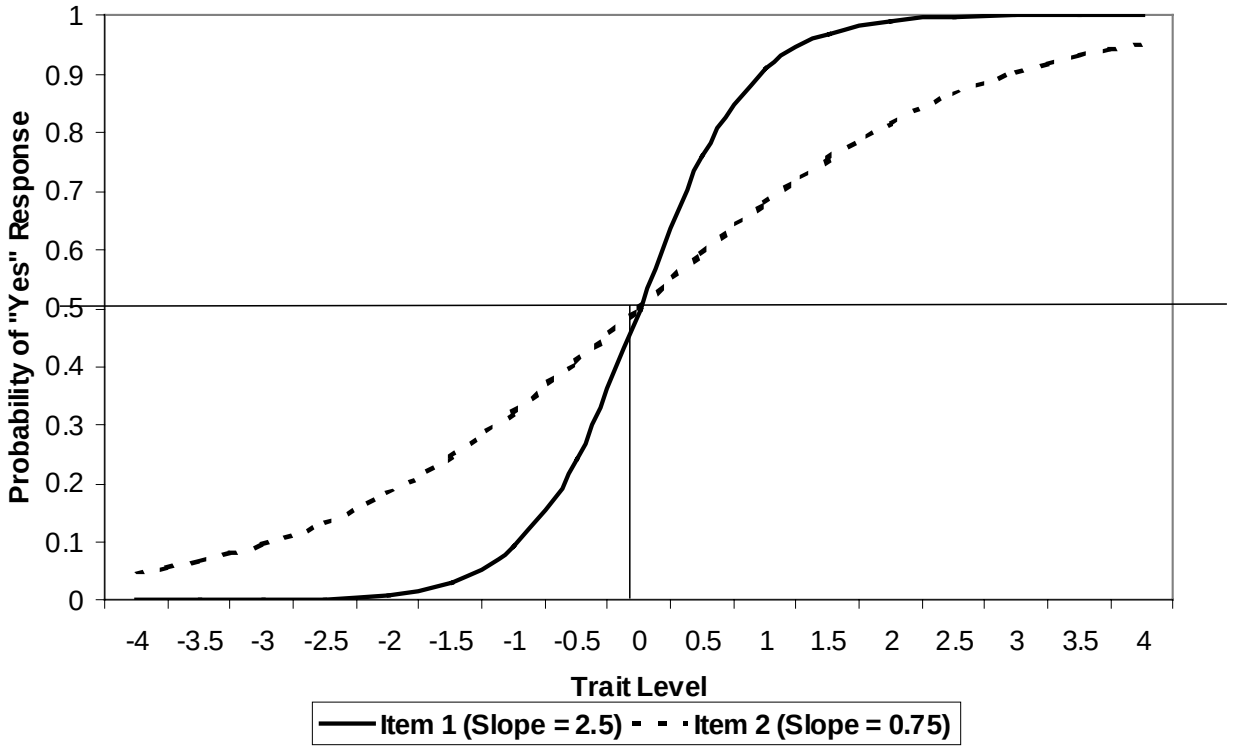
$P_i(\Theta)$ Probability that a randomly selected respondent with ability Θ (trait level) answers item i correctly.

b_i Item i difficulty.

a_i Item i slope.

D Scaling constant.

Item Characteristic Curves
(2-Parameter Model)



Features of IRT Models

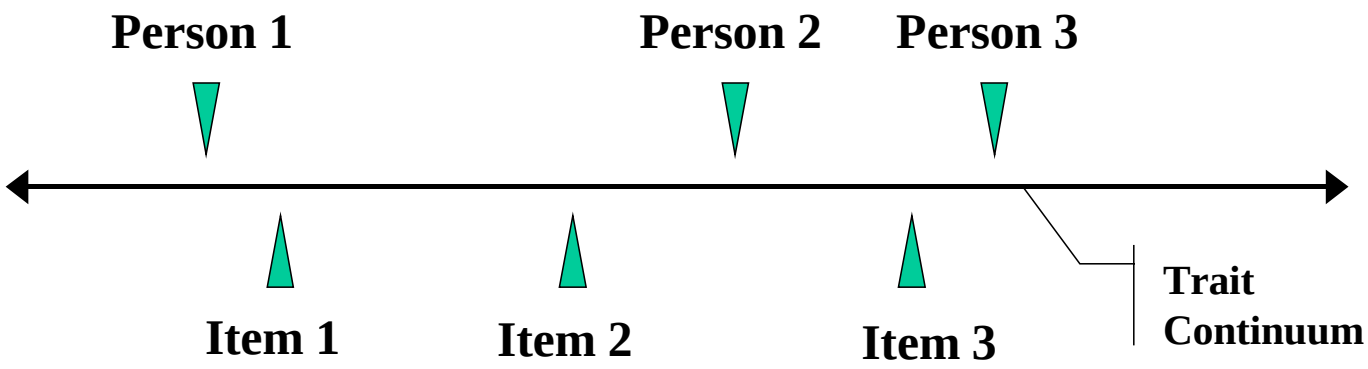
Question (item) and person characteristics that are sample independent.

Standard errors conditional on trait level.

Items can be selected to minimize standard error of ability estimate.

Equivalence studies.

Item Responses and Trait Levels



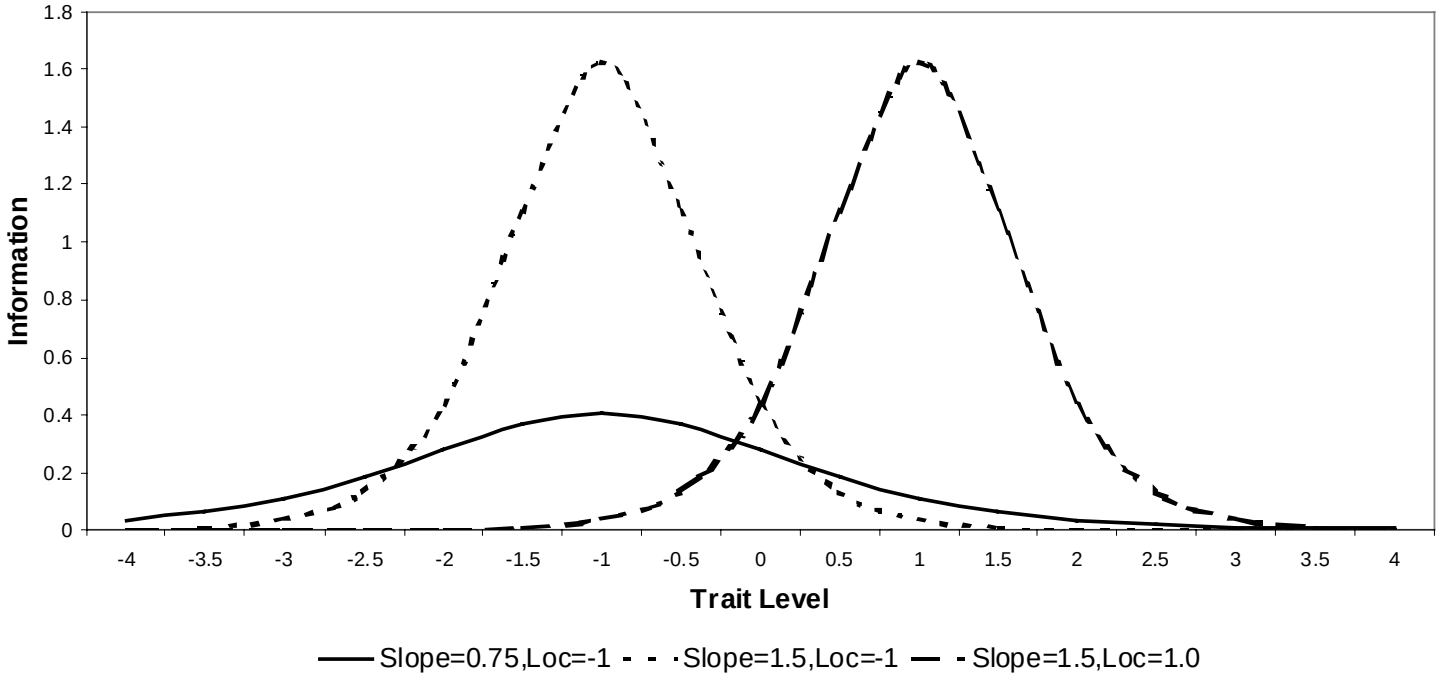
Item Information

$$I_i(\Theta) = \frac{[P_i'(\Theta)]^2}{P_i(\Theta)Q_i(\Theta)}$$

Where:
$P(\Theta)$ is item characteristic curve function
$Q(\Theta)=1-P(\Theta)$

Item Information

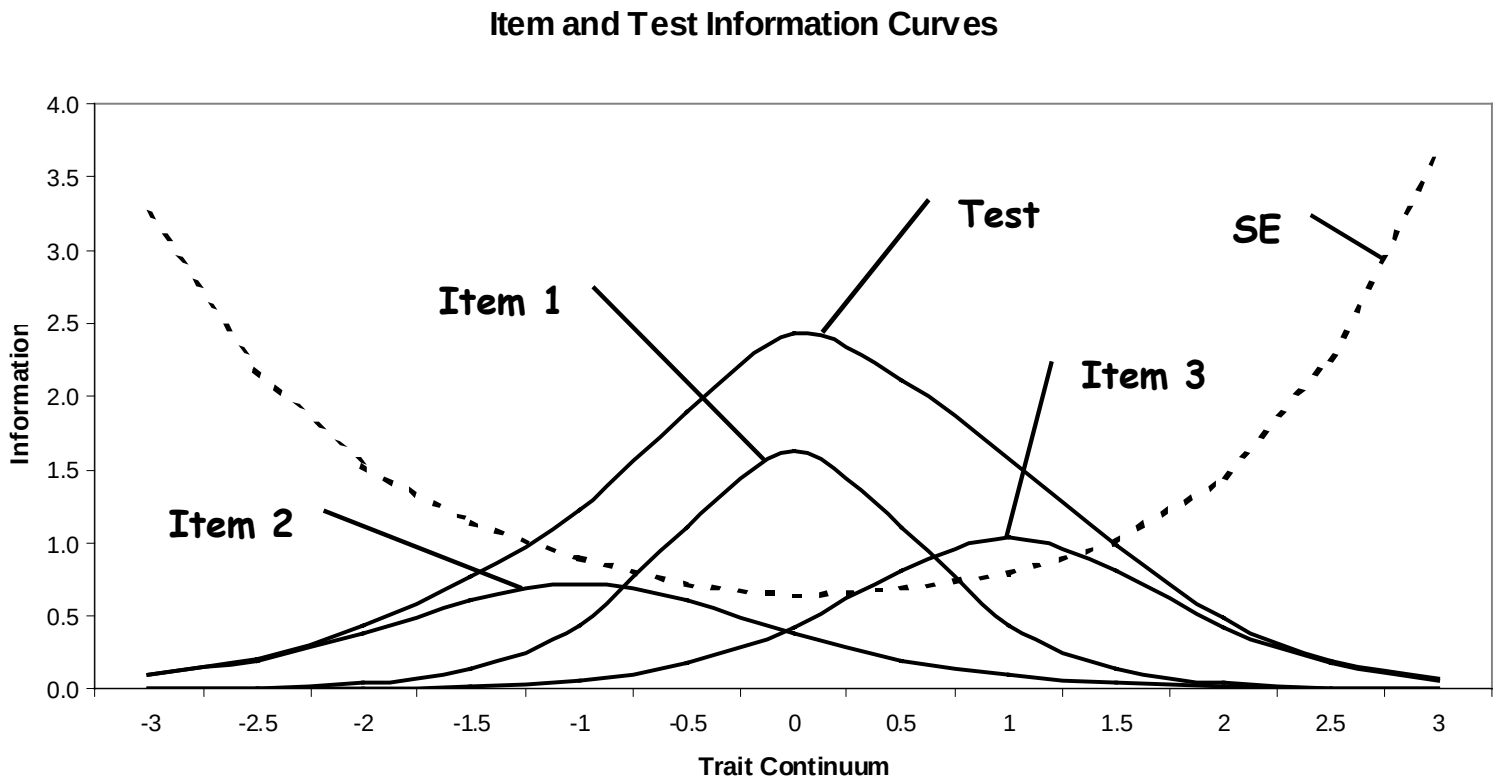
(2-parameter model)



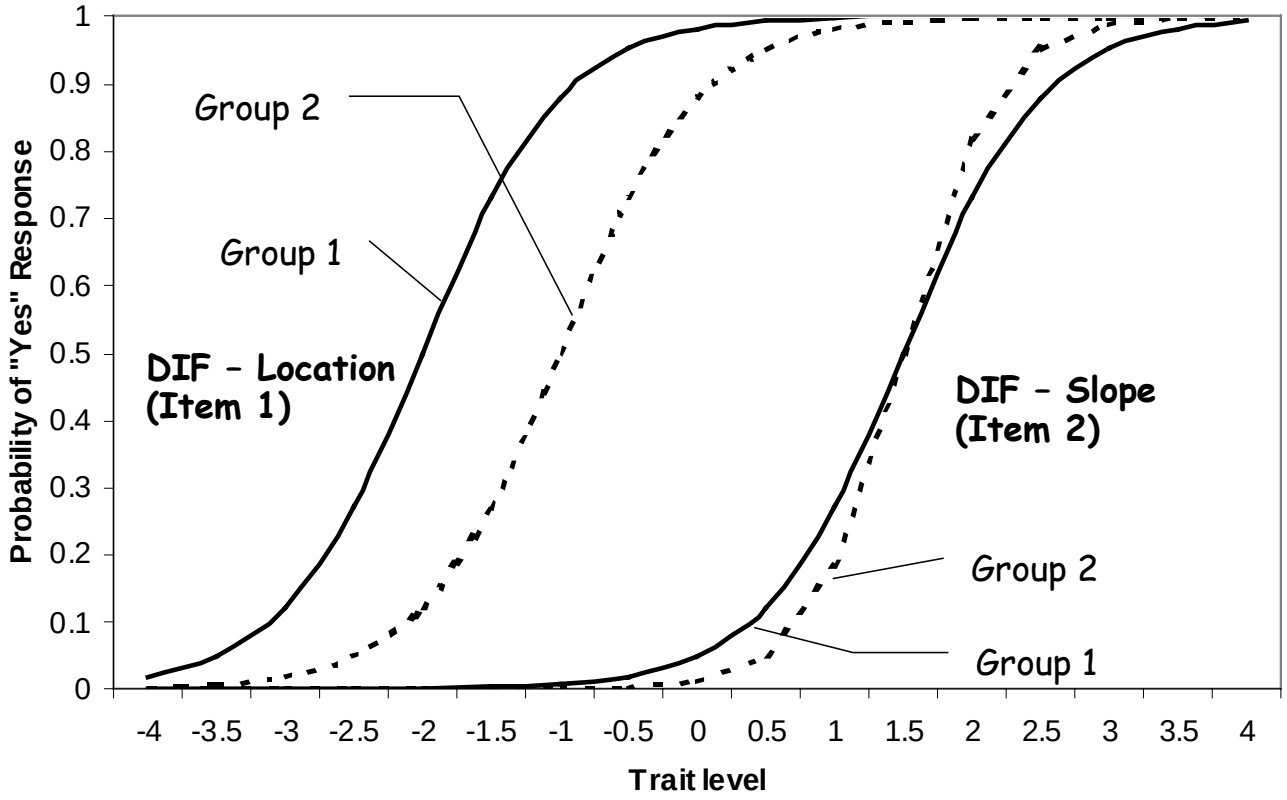
Test Information

$$I(\Theta) = \sum_{i=1}^{n} I_i(\Theta)$$

$$SE(\hat{\Theta}) = \frac{1}{\sqrt{I(\Theta)}}$$


Item and Test Information Curves
Information
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Test
SE
Item 1
Item 3
Item 2
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
Trait Continuum

"If…an item has a different item response function for one group than for another, it is clear that the item is biased."



Example 1: Data Source

California State Health Families Program 2000 and 2001 administrations of the CAHPS survey

28 participating health plans

Administered in 5 languages (n=26,671)

- English (n=11,231)
- Spanish (n=12,458)
- Cantonese (n=1,374)
- Korean (n=979)
- Vietnamese (n=629)

Response rate 2000=60%

Response rate 2001=58%

3-Category Items

How much of a problem:

- Getting a personal doctor or nurse?
- Getting referrals to see specialists?
- Getting care you or your doctor thought necessary?
- Getting approvals for care from plan?

A big problem, A small problem, Not a problem

4-Category Items

How often did doctors:

- Explain things in a way you could understand?
- Listen to you carefully?
- Spend enough time with you?
- Show respect for what you had to say?

Always, Usually, Sometimes, Never

Sample Characteristics-1

	N (unweighted)	% (weighted)
Respondent Age		
<25 years	1,926	7
25-34 years	11,469	44
35-44 years	10,680	41
?45 year	2,199	9
Respondent Education		
0-8 Grade	5,206	20
9-11 Grade	3,941	15
12 Grade	7,811	30
>12 Grade	9,103	34
Respondent Race/Ethnicity		
White	2,623	9
African American	473	2
Asian/Pacific Islander	4,639	18
Hispanic	16,729	63
Other	2,207	8

Sample Characteristics-2

	N (unweighted)	% (weighted)
Survey Language		
English	11,231	41
Spanish	12,458	47
Cantonese	1,374	6
Korean	979	4
Vietnamese	629	3
Child Age		
0-5 Years	9,006	34
6-9 Years	10,921	41
10-13 Years	6,744	25
Child Health Status		
Excellent	9,697	37
Very Good	8,412	32
Good	6,268	24
Fair/Poor	1,800	7

Testing for Differences in Item Functioning

- Identify anchor items using logistic discriminant analysis.
- Identify items with DIF using IRTLRTEST

Identifying Anchor Items Using Logistic Regression *Location Parameter*

Model 1: Group=Scale

Model 2: Group=Scale+Item

LLR Test: Model 1 versus Model 2

See: Miller, T. R., Spray, J. A. Summer 1993. Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scores Items. Journal of Educational Measurement 30:107-122.

Identifying Anchor Items
Slope Parameter

Model 2: Group=Scale+Item

Model 3: Group=Scale+Item+Scale*Item

LR Test: Model 2 versus Model 3

Identifying Anchor Items
Omnibus Test

Model 1: Group=Scale

Model 3: Group=Scale+Item+Scale*Item

LR Test: Model 1 versus Model 3

Example Logistic Regression Results

English Versus Korean	Model LL Values			Overall DIF	Slope DIF	Loc DIF
Items	Model 1 Scale	Model 2 Scale+Item	Model 3 Scale+Item +Scale*Item	Model 3-1 2 d.f.	Model 3-2 1 d.f.	Model 2-1 1 d.f.
pbnescr	20.62	30.56	35.15	14.53	4.59	9.94
pbcrdly	20.62	42.14	44.15	23.53	2.01	21.52
wtmor15	***20.62***	***22.13***	***23.52***	***2.90***		
drexpln	***20.62***	***22.91***	***24.58***	***3.96***		
drlistn	***20.62***	***22.09***	***22.69***	***2.07***		
drrespu	***20.62***	***21.31***	***23.42***	***2.80***		
drtmenf	20.62	41.2	44.21	23.59	3.01	20.58
stfresp	***20.62***	***20.74***	***23.64***	***3.02***		
stfhelp	***20.62***	***20.98***	***22.21***	***1.59***		
rtallcr	***20.62***	***20.66***	***21.53***	***0.91***		

IRTLRTEST

Graded Response Models estimated in Multilog

Models estimated when testing item *I (in all models anchor location and slope constrained)*

- Model 1: Test item parameters are constrained **(Test Constrained)**
- Model 2: Test item parameters not constrained **(Test Free)**
- Model 3: Test item location parameter constrained and slope unconstrained **(Test Location Constrained)**

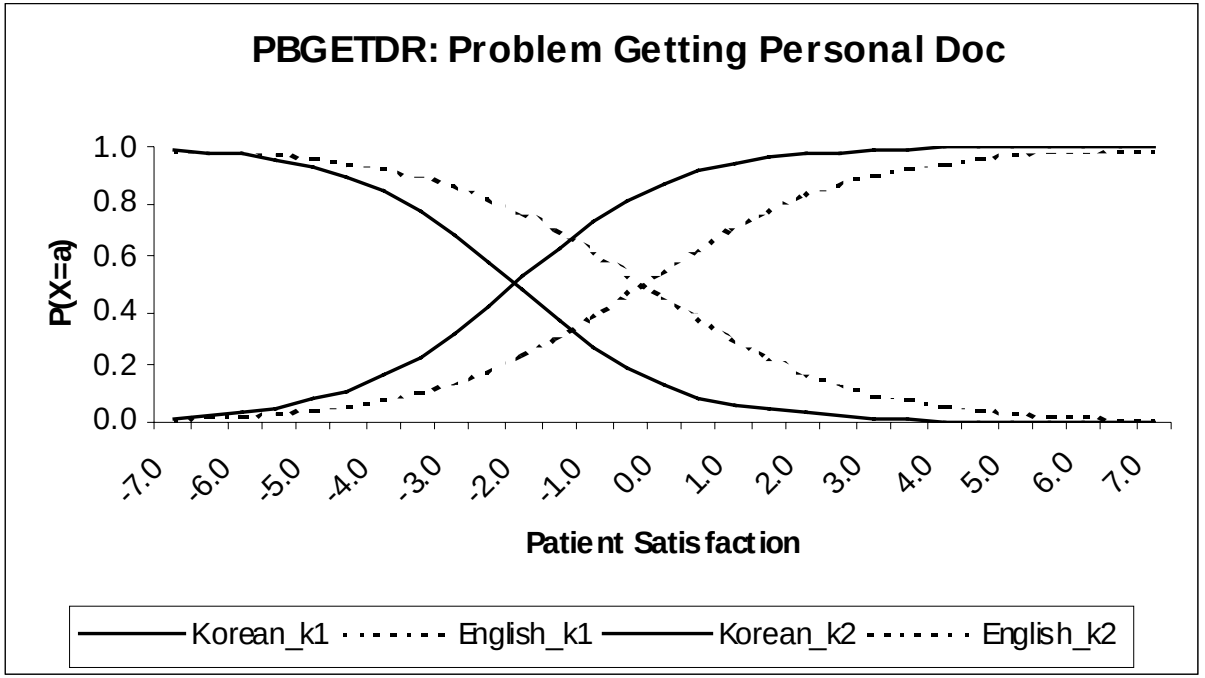
Hypothesis Tests

- Omnibus test for DIF: $-2(LL_{model1}-LL_{model2})$ [**Test Constrained versus Free**]
- Location parameter DIF: $-2(LL_{model3}-LL_{model2})$ [**Test Location Constrained versus Free**]
- Slope parameter DIF: $-2(LL_{model1}-LL_{model3})$ [**Test Constrained versus Location Constrained**]

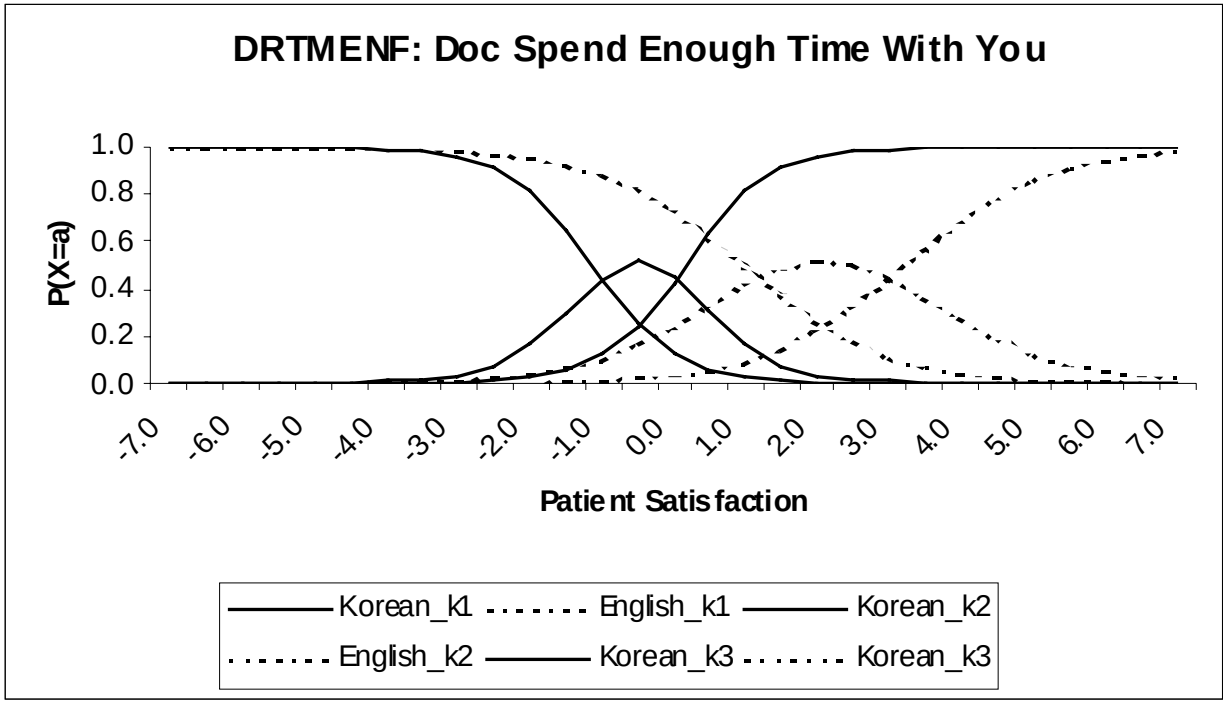
Example Results From IRTLRTEST

Item	Test	G^2	*d.f.*	R-Slope	R-Loc	F-Slope	F-Loc	Mean	SD		
1	Omnibus	38.4	2	1.01	-1.75	0.96	-1.09	-0.51	0.8		
1	Slope	0.1	1	1.01	-1.75	1.01	-1.07	-0.51	0.8		
1	Loc	38.3	1	1.03	-1.68	1.03	-1.68	-0.52	0.8		
2	Omnibus	120.2	2	0.82	-2.59	0.86	-1.08	-0.51	0.8		
2	Slope	0.1	1	0.82	-2.58	0.82	-1.10	-0.51	0.8		
2	Loc	120.1	1	0.85	-2.38	0.85	-2.38	-0.52	0.8		
3	Omnibus	43.3	2	1.12	-1.72	0.74	-1.32	-0.50	0.8		
3	Slope	4.4	1	1.08	-1.76	1.08	-1.10	-0.50	0.8		
3	Loc	38.9	1	1.10	-1.66	1.10	-1.66	-0.52	0.8		
4	Omnibus	1.9	2	0.98	-0.62	1.33	-0.45	-0.52	0.8		
Item	Test	G^2	*d.f.*	R-Slope	R-Loc1	R-Loc2	F-Slope	F-Loc1	F-Loc2	Mean	SD
5	Omnibus	10.2	3	1.89	-1.27	-0.37	1.41	-1.63	-0.55	-0.53	0.8
5	Slope	4.3	1	1.87	-1.28	-0.37	1.87	-1.44	-0.56	-0.53	0.8
5	Loc	5.9	2	1.85	-1.29	-0.38	1.85	-1.29	-0.38	-0.52	0.8

2-Category Response Item



3-Category Response Item



Example 2: Data Source

United Medical Group Association Study of patient experiences with care

Random sample of adult patients with $\geq$1 visit to their doctor during the prior year.

Response rate of 59%

Survey fielded October 1994 to June 1995

Example 7-Category Items

How do you rate:

- Medical staff listening to what you have to say
- Answers to your questions
- Explanations about prescribed medications

The best, excellent, very good, good, fair, poor, and very poor and not applicable

Total of 9 items

Testing for Differential Item Functioning

Estimates location and slope parameters for each group using Parscale 3.5 (Partial Credit Model)

Test equality of parameters between groups directly

Assess the impact of finding statistically significant DIF on test scores.

SAMPLE CHARACTERISTICS

	White	Hispanic
N	5,508	713
Age (mean, [SD])	52 (18)	42 (15)*
Male (%)	35	38
Married (%)	74	78*
≤12 Years Schooling (%)	69	47*
Health Status (mean, [SD]; 0-10, 10=Best)	7.3 (1.7)	7.2 (1.8)

*$p < 0.05$

DIF RESULTS - DIFFICULTY

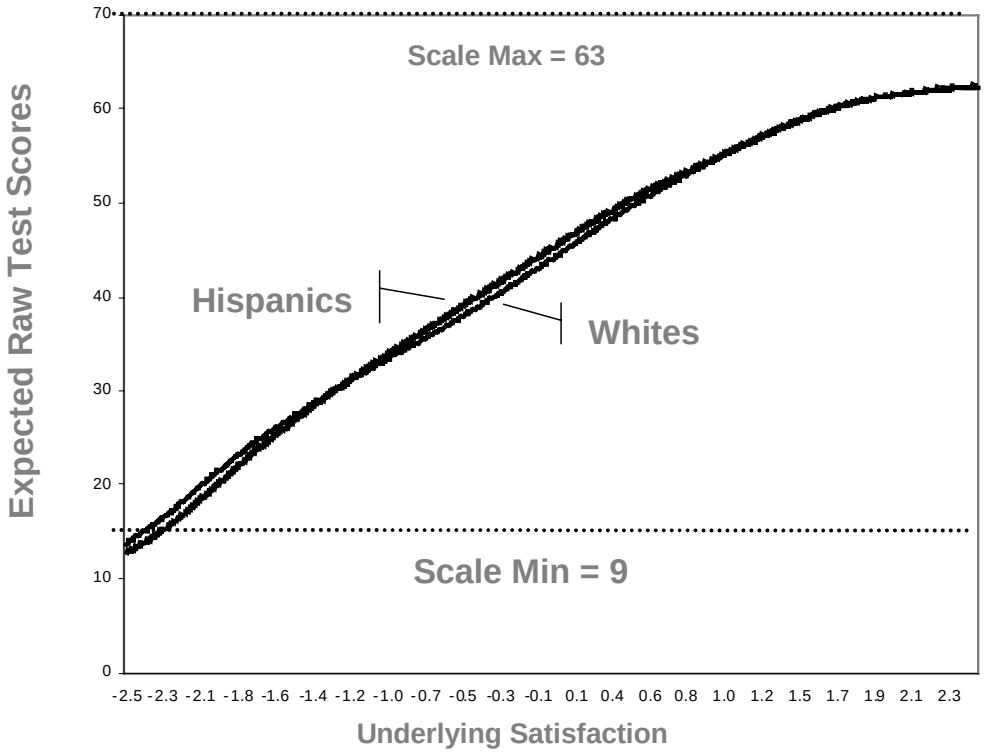
Item	Whites	Hispanics
1 (Difficulty (SE))	-0.71 (0.01)	-0.76 (0.04)
2	-0.74 (0.01)	-0.77 (0.03)
3	-0.72 (0.01)	-0.72 (0.04)
4	-0.58 (0.01)	-0.61 (0.04)
5	-0.55 (0.01)	-0.54 (0.04)
6	-0.83 (0.01)	-0.76 (0.03)
7	-0.77 (0.01)	-0.76 (0.03)
8	-0.61 (0.01)	-0.58 (0.03)
9	-0.43 (0.01)	-0.44 (0.04)

DIF RESULTS - SLOPES

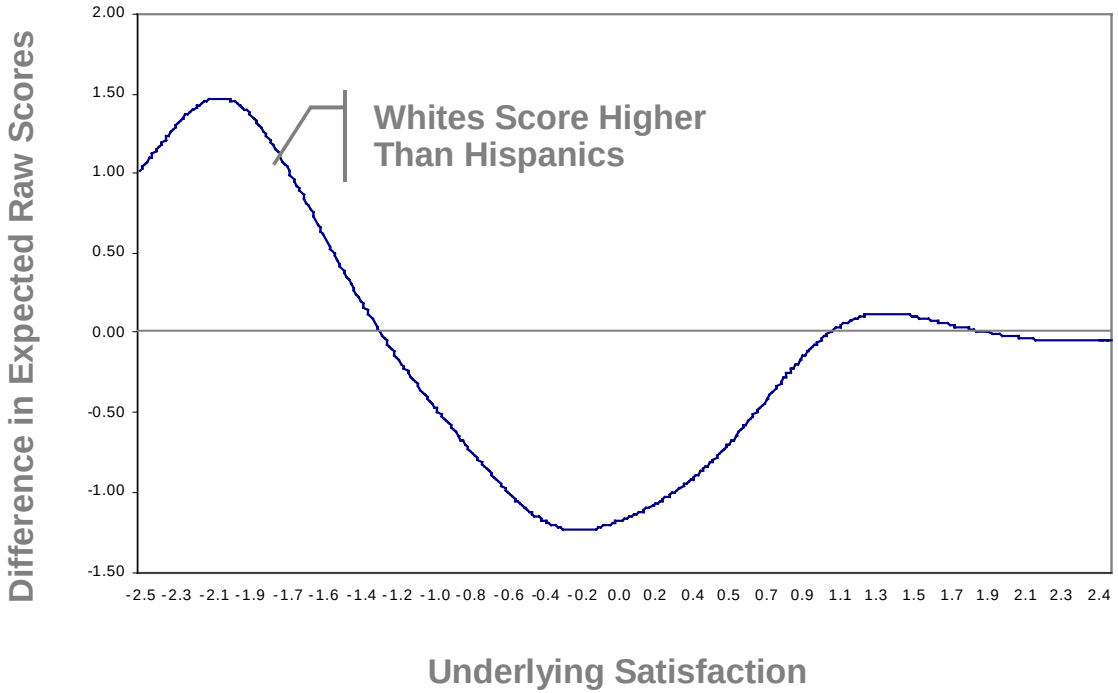
Item	Whites	Hispanics	
1 (Slope (SE))	2.99 (0.15)	2.90 (0.05)	
2	3.52 (0.15)	3.32 (0.06)	
3	2.09 (0.10)	2.00 (0.03)	
4	2.39 (0.09)	2.34 (0.04)	
5	2.84 (0.14)	2.53 (0.04)*	
6	3.09 (0.16)	3.70 (0.07)*	
7	3.97 (0.23)	4.08 (0.08)	
8	3.11 (0.15)	3.28 (0.05)	
9	1.77 (0.08)	1.78 (0.03)	

*P<0.05

EXPECTED RAW SCORES WITH DIF ITEMS



BIASING EFFECT OF DIF



CONCLUSIONS

- Two of nine items demonstrated statistically significant DIF.
- Group comparisons remained robust whether or not these items were included in the scale.
- IRT analysis of item and scale bias provides useful insights about impact of DIF.

THE END