DIFFERENTIAL ITEM FUNCTIONING AND HEALTH ASSESSMENT

Jeanne Teresi

- “Racial, ethnic and socioeconomic disparities are national problems that affect health care….
- Disparities in the health care system are pervasive”

(DHHS, Agency for Healthcare Research and Quality, National Healthcare Disparities Report, 2003)

Slide prepared by Jeanne Teresi, Ph.D.

USES OF DIF:

- **EVALUATE EXISTING MEASURES**

- **DEVELOP NEW MEASURES THAT ARE:**

  - Culture Fair,
  - Gender Equivalent,
  - Age invariant

Slide prepared by Jeanne Teresi, Ph.D.

BIAS IN HEALTH-RELATED MEASURES WILL ALWAYS EXIST:

- **Too many factors at play**
- **Too many cultural background variables exist**
- **Direction and level of bias unpredictable**

Slide prepared by Jeanne Teresi, Ph.D.

WHAT IS THE CASE FOR DIF ANALYSES ?

- Need to attempt identification of presence, magnitude, and impact
- Need to attempt adjustment

Slide prepared by Jeanne Teresi, Ph.D.

DEFINITIONS

- **DIF involves three factors:**
  - **Response to an item**
  - **Conditioning/matching health status variable**
  - **Background (grouping) variable(s)**
- **DIF can be defined as conditional probabilities or conditional expected item scores that vary across groups.**

Slide prepared by Jeanne Teresi, Ph.D.

- **Controlling for level of health status, is the response to an item related to group membership?**

- A randomly-selected person of average physical function interviewed in Spanish should have the same chance of responding in the unimpaired direction to a health status item as would a randomly selected person of average function interviewed in English.

Slide prepared by Jeanne Teresi, Ph.D.

EXAMPLE

- contingency table that examines the cross-tabulation of item response by group membership for every level (or grouped levels) of the attribute estimate

Slide prepared by Jeanne Teresi, Ph.D.

Two by two contingency table for item ‘Trouble with a Long Walk’ by race groups, conditioning on the sum score for the Physical Scale (score levels 26-32)

	Item Score		
Group	Trouble (0)	No Trouble (1)	Total
Focal (African American)	26 (68.4%)	12 (31.6%)	38 (100%)
Reference group (Whites)	149 (82.8%)	31 (17.2%)	180 (100%)
Total	175 (80.3%)	43 (19.7%)	218

Slide prepared by Jeanne Teresi, Ph.D.

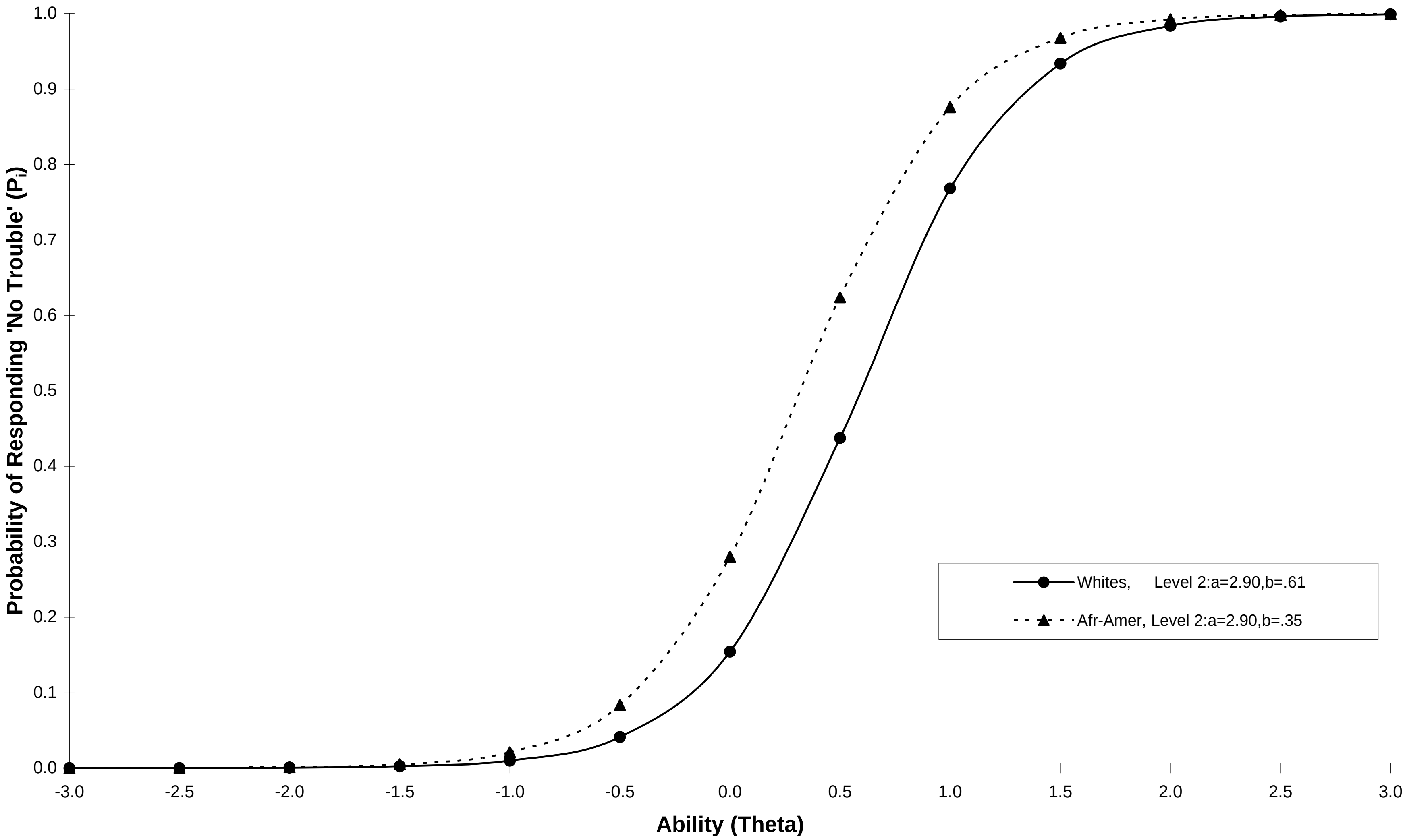
Uniform DIF Definitions

- DIF is in the same direction across the entire spectrum of disability (item response curves for two groups do not cross)
- DIF involves the location (b) parameters

- DIF is a significant main (group) effect in regression analyses predicting item response

Slide prepared by Jeanne Teresi, Ph.D.

Physical Functioning Scale
Plot of Boundary Response Functions
Item 5 - Trouble with a Long Walk



Slide prepared by Jeanne Teresi, Ph.D.

- The probability of a randomly selected black person of above average physical function (theta = .5) responding in a non-disordered direction to the item “trouble with a long walk” is higher (.62) than for a randomly selected white person (.44) at the same ability level. (Given equal ability, black respondents are more likely than white respondents to say “no trouble”.)

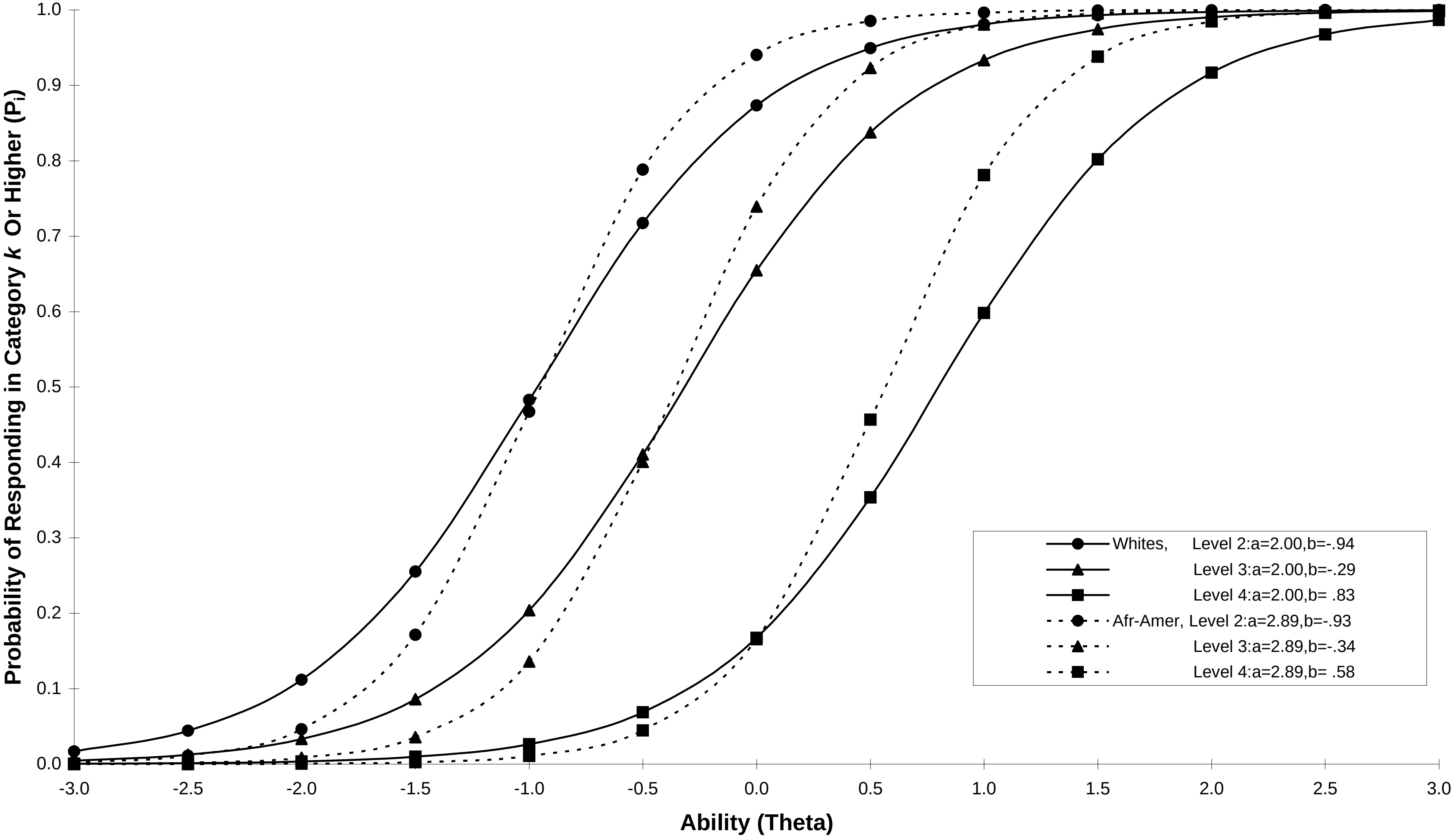
Slide prepared by Jeanne Teresi, Ph.D.

Non-Uniform DIF

- An item favors one group at certain disability levels, and other groups at other levels (or the probability of item endorsement is higher for group 1 at lower ability and higher for group 2 at higher ability)
- DIF involves the discrimination (a) parameters
- DIF is a significant group by ability interaction in regressions predicting item response
- DIF is assessed by examination of nested models comparing differences in log-likelihoods

Slide prepared by Jeanne Teresi, Ph.D.

Physical Functioning Scale
Plot of Boundary Response Functions
Item 9 - Limited in Work or Other Daily Activities



Slide prepared by Jeanne Teresi, Ph.D.

MAGNITUDE

- Magnitude of DIF

  Item level characteristic; e.g.,

  odds ratio,

  area statistic,

  beta coefficient or R square increment,

  expected item scores

Slide prepared by Jeanne Teresi, Ph.D.

Magnitude of non-uniform DIF

- Odds ratios for the item: “Walk one block” for different race groups and levels of ability (theta) (Gibbons and Crane, 2004)

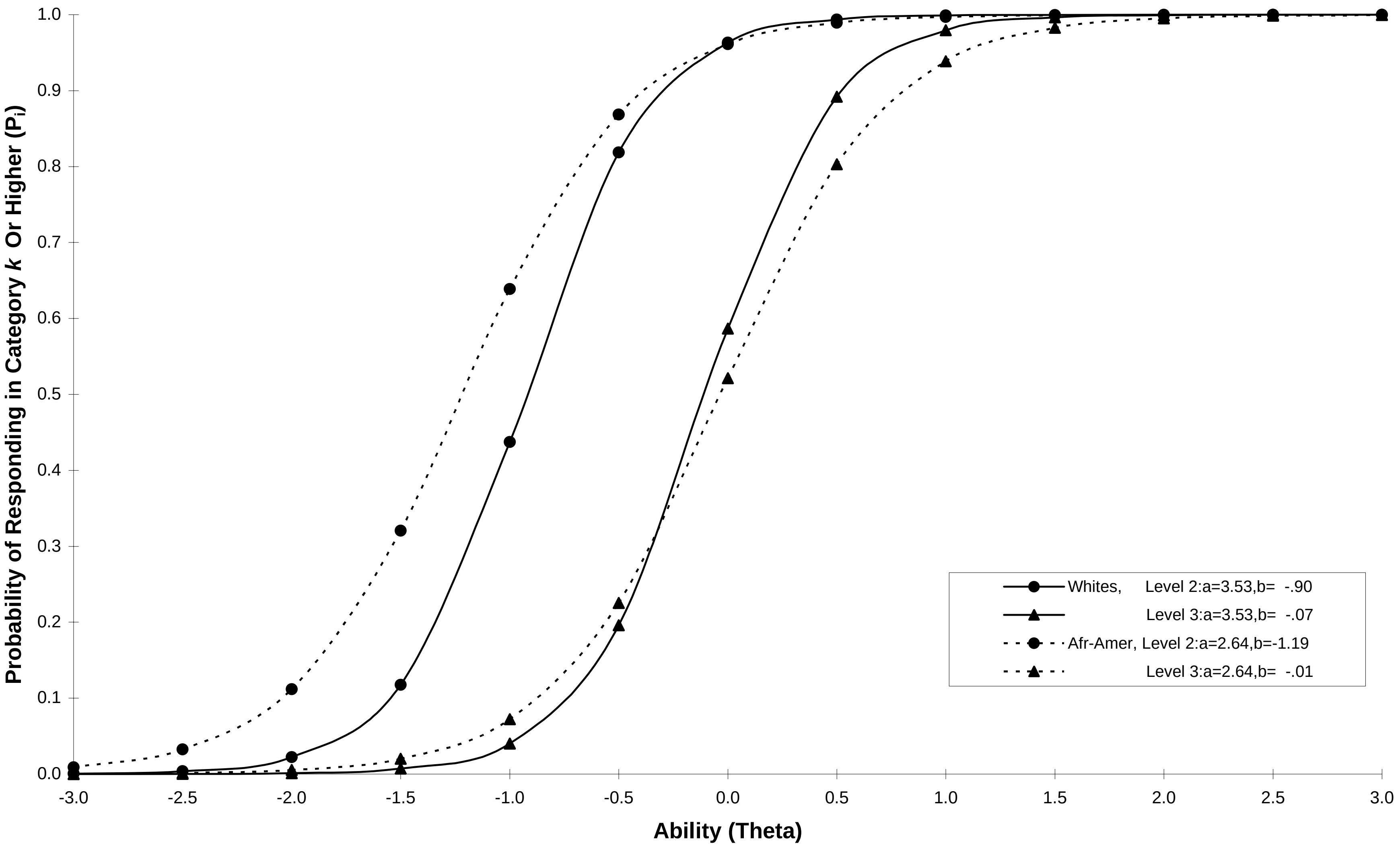
	Theta		
	-0.5	0	0.5
Black	.21	.85	3.4
White	.17	1.0	5.9

Slide prepared by Jeanne Teresi, Ph.D.

- Non-uniform DIF was found for 'Walk one block'
- Among people with low overall physical function, blacks had higher scores on 'Walk one block'. At higher levels of function, whites had higher scores on the item.
- This can be shown graphically:

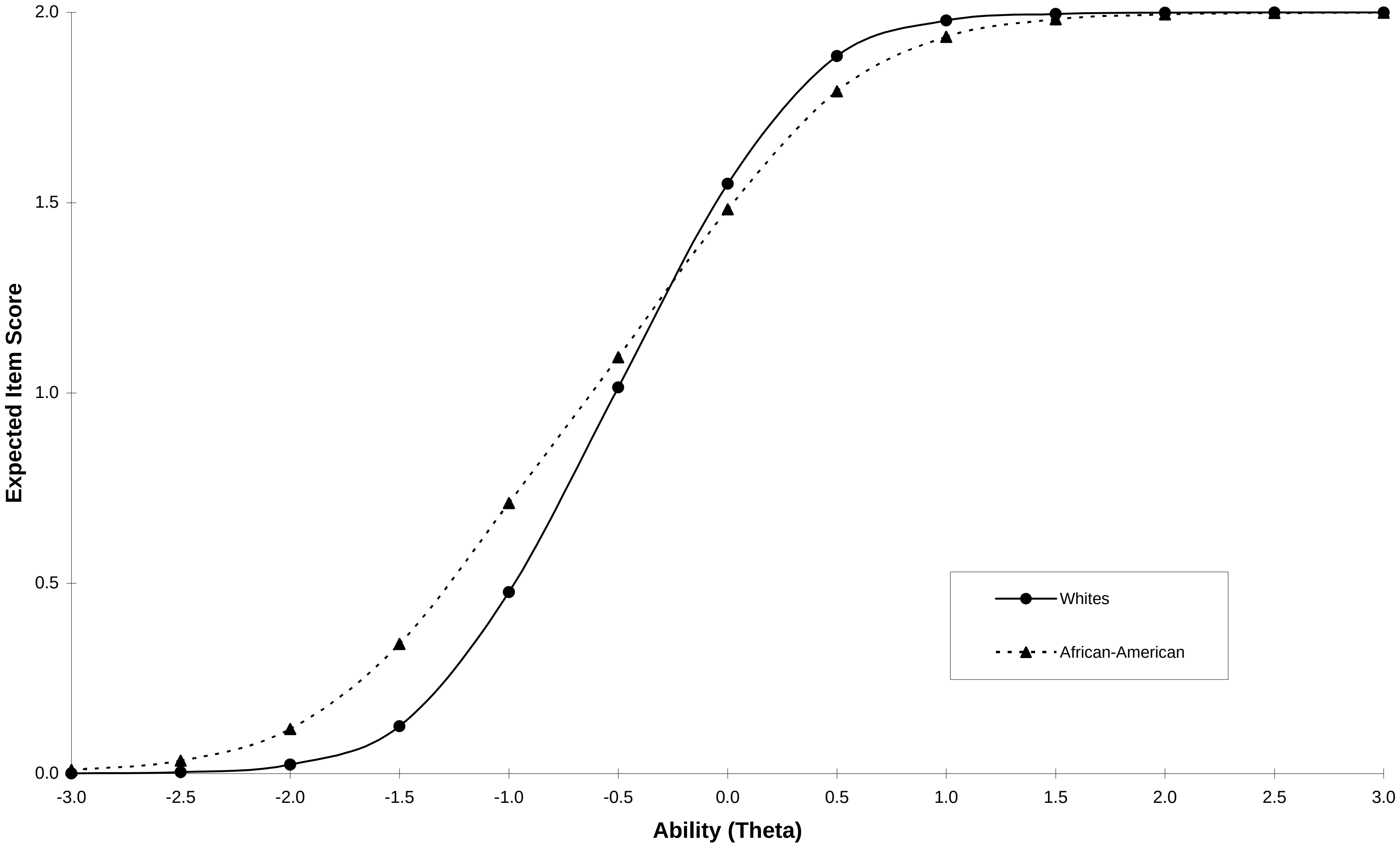
Slide prepared by Jeanne Teresi, Ph.D.

Physical Functioning Scale
Plot of Boundary Response Functions
Item 22 - Walking One Block



Slide prepared by Jeanne Teresi, Ph.D.

Physical Functioning Scale
Expected Item Score Functions by Race Groups
Item 22 - Walking One Block



Slide prepared by Jeanne Teresi, Ph.D.

- Item bias

  implies a substantive review

  the cumulative body of evidence suggests that the item may have different meaning or may be measuring an unwanted nuisance factor for one group as contrasted with another

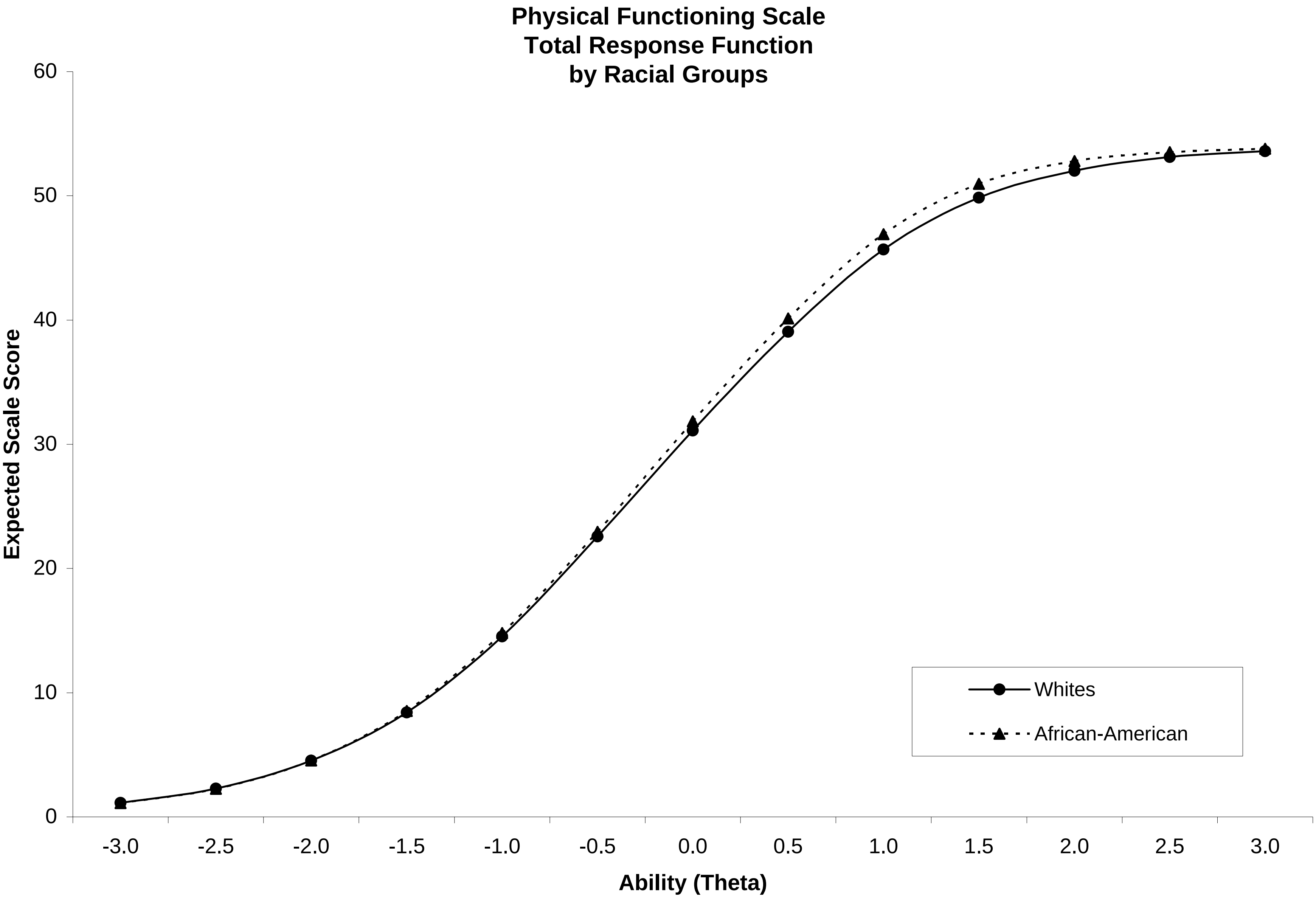
Slide prepared by Jeanne Teresi, Ph.D.

IMPACT

Impact in the context of health measures:

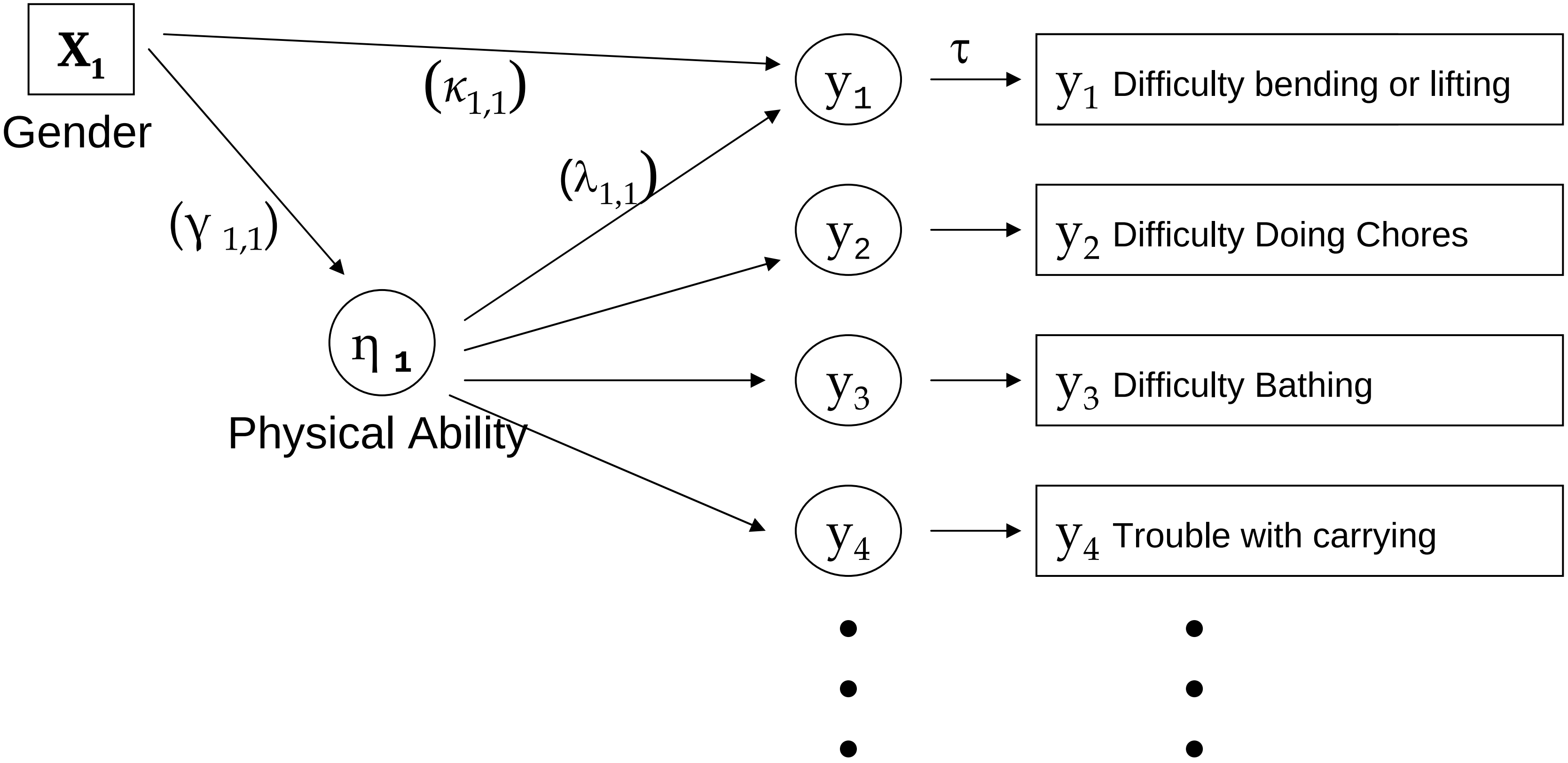
- differences in the health status distributions between or among studied groups;
- group differences in the total (test) response function;
- group differences in relationship of demographic variables to health status variables with and without adjustment for DIF.

Slide prepared by Jeanne Teresi, Ph.D.



Slide prepared by Jeanne Teresi, Ph.D.

Simplified Single Group MIMIC Model



The direct effect $(\kappa_{1,1})$ is an estimate of uniform DIF and $(\gamma_{1,1})$ is an estimate of impact.

Slide prepared by Jeanne Teresi, Ph.D.

ISSUES FOR CONSIDERATION

- What is a group?

Homogenous meaningful entities or

proxies for other variables and, as such, should be "deconstructed"

- If groups are to be used, there are numerous interactions that might be considered.

Slide prepared by Jeanne Teresi, Ph.D.

CONDITIONING VARIABLE

- **Observed matching variables**

  total or weighted raw score.

- **Latent variable**

  estimated using marginal maximum likelihood or other procedures

- **"Valid" target dimension**

  as distinct from secondary "nuisance" factors.

- **External "gold standard" diagnostic variable**

- **Internal "silver standard" "anchor" such as a vignette**

Slide prepared by Jeanne Teresi, Ph.D.

DIF METHODS

There are numerous review articles and books related to DIF. A few are:

- Camilli and Shepard, 1994;
- Holland and Wainer; 1993;
- Millsap and Everson, 1993;
- Potenza and Dorans, 1995;
- Thissen, Steinberg and Wainer, 1993.

Slide prepared by Jeanne Teresi, Ph.D.

Differences among DIF methods can be characterized according to whether they:

- are parametric or non-parametric;
- are based on latent or observed variables;
- treat the disability dimension as continuous;
- can model multiple traits;
- can detect both uniform and non-uniform DIF;
- can examine polytomous responses;
- can include covariates in the model;
- must use a categorical studied (group variable).

Slide prepared by Jeanne Teresi, Ph.D.

COMMON METHODS

Mantel-Haenszel (Holland and Thayer, 1988) and standardization (Dorans and Kulick, 1986)

based on contingency table and observed conditioning variable

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of MH/standardization

- Few model assumptions;
- Performs favorably in simulations (see Potenza and Dorans, 1995);
- Standardization provides empirical item-measure regressions;
- Provides magnitude measures;
- Is not labor intensive or complex.

Slide prepared by Jeanne Teresi, Ph.D.

SIBTEST

SIBTEST (Stout and colleagues; Shealy and Stout, 1993;

- Poly-SIBTEST (Chang, Mazzeo and Roussos, 1996),
- Crossing SIBTEST (Li and Stout, 1996)
- CATSIB (Nandakumar and Roussos, 2002)

(based on contingency table, theoretically anchored in the notion of a latent conditioning variable, but estimation is usually based on total continuous, observed disability score)

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of SIBTEST

- Non-parametric; model fit is not an issue in DIF detection;
- Allows modeling of multidimensional abilities;
- Provides DIF significance tests and magnitude estimates;
- Can detect crossing DIF with crossing SIB;
- Simulations show superior performance of Poly-SIB (in comparison to IRTLR and DFIT under several IRT models) in terms of false positives when groups have different ability distributions and the correct model is not known (Bolt, 2002)

Slide prepared by Jeanne Teresi, Ph.D.

LOGISTIC REGRESSION

Logistic regression (Swaminathan and Rogers, 1990);

Ordinal Logistic Regression (Zumbo, 1999; Crane van Belle and Larson, 2004)

(based on examination of regression predicting item response from main effects of group and ability and their interaction)

Usually uses observed conditioning variable, but IRT estimates can be used

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of LR

- Covariates can be included;
- Studied variable can be continuous;
- Can model multiple abilities;
- Can model non-uniform DIF;
- Performs well (in terms of detection rates) in simulations in the presence of non-uniform DIF;
- Provides magnitude measure;
- Easy to perform (unless IRT ability estimates are used)

Slide prepared by Jeanne Teresi, Ph.D.

IRT-Based Measures

IRT-based methods:

- Likelihood ratio test based on IRT (Thissen, 1991, 2001)

(based on examination of differences in fit between compact and augmented models that include additional free parameters representing non-uniform and uniform DIF)

Latent conditioning variable

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of IRT

- Well-developed theoretical models;
- Can examine uniform and non-uniform DIF;
- No equating required because of simultaneous estimation of group parameters;
- Can model missing data;
- Simulations show superior performance (in terms of power in comparison with non-parametric methods)

Slide prepared by Jeanne Teresi, Ph.D.

Area and DFIT Methods

- **Area and DFIT methods based on IRT** (Raju and colleagues, 1995; Flowers and colleagues, 1999)

  (based on IRT model with latent conditioning variable)

Non-compensatory DIF (NCDIF) indices:

average squared differences in item "true" or expected raw scores for individuals as members of the focal group and as members of the reference group.

(expected score is the sum of the (weighted) probabilities of category endorsement, conditional on disability).

Differential test functioning (DTF) :

based on the compensatory DIF (CDIF) index, and reflects group differences summed across items

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of DFIT

- Can detect both uniform and non-uniform DIF and shares the advantages of IRT models upon which it is based;
- Magnitude measures used for DIF detection;
- Impact of item DIF on the total score is examined;
- One simulation study (in comparison with IRTLR) showed favorable performance in terms of false DIF detection (Bolt, 2002)

Slide prepared by Jeanne Teresi, Ph.D.

MIMIC models

- MIMIC model based on structural equation approach to IRT (Muthén, 1984)

(latent continuous ability variable)

Uniform DIF examined using direct effects from measurement/SEM model)

Slide prepared by Jeanne Teresi, Ph.D.

Some Advantages of MIMIC

- Simultaneous modeling of group differences in the item response and underlying ability;
- Good impact measures;
- Can model multidimensional data;
- Can include covariates;
- Can adjust for impact of DIF

Slide prepared by Jeanne Teresi, Ph.D.

Future Directions: DIF, CAT, and IMPACT

- Use of DIF methods in the context of CAT.
- Need for better guidelines for DIF detection in the context of health-related measures
  - – Optimal magnitude indices,
  - – Optimal cut scores,
  - – Integration of significance testing and magnitude measures
  - – Further development of impact measures

Slide prepared by Jeanne Teresi, Ph.D.

Conclusions

- DIF cancellation at the aggregate level may still have an impact on an individual
- DIF assessment of measures remains a critical component of health disparities research, and of efforts to achieve cultural equivalence in an increasingly, culturally diverse society.

Slide prepared by Jeanne Teresi, Ph.D.