Methods for testing psychometric equivalence: CFA and IRT

Steve Gregorich, Ph.D.

University of California, San Francisco
RCMAR Center for Aging in Diverse Communities

GSA Preconference Workshop

November 21, 2003

Concepts of psychometric equivalence: Basic questions

First basic question

- Does the instrument measure the same construct (or latent variable or factor) across groups?

  This concerns whether the instrument is an equally valid measure of the targeted construct in each population group

Concepts of psychometric equivalence: Basic questions

Second basic question

- Can the item/scale scores be directly and meaningfully compared across population groups?

  This concerns whether the same scale of measurement obtains in each population group, or whether response bias is present

  All else being equal, members of some groups may respond differently to items, systematically obtaining higher or lower scores than members of other groups

  If this happens, group measurements will be contaminated and will not reflect true differences on the construct of interest

  Example: Fahrenheit and Celsius scales both measure temperature, but the two measurements cannot be directly compared.

Concepts of psychometric equivalence: Basic questions

- Another example of measurement bias

Two primary care practices

In practice #1 patients are weighed while wearing their street clothes

In practice #2 patients are weighed while wearing an examination gown

All else being equal, patients in practice #2 will appear heavier

Observed differences in patient weights across practices
will *not* reflect true differences in patient weights

The differing protocols will bias cross-practice comparison of weights

Bias can occur for reasons other than procedural/protocol differences

Concepts of psychometric equivalence: Example data

- Example IPC data

  4 items of the IPC Empowerment scale

  Measured in the Latino-English and Latino-Spanish language samples

praise: how often did doctors praise you for how you were taking care of yourself?

control: how often did doctors give you a sense of control over your health?

diet: how often did doctors help you feel that sticking to your treatment would make a difference?

prevent: How often did doctors help you feel that you can prevent some health problems?

Ordered responses: 1=*never*, 2=*rarely*, 3=*sometimes*, 4=*usually*, 5=*always*

Concepts of psychometric equivalence: EFA

- The so-called exploratory factor analysis (EFA) framework

  EFA uses observed variables (items) to 'identify' latent variables (or factors or constructs)

  Latent variables are not directly observed

  Latent variables are thought to be responsible for item response

  Items are imperfect measures of the latent variable

  When a set of items shares one latent variable in common,
  they are said to be unidimensional. (i.e., measure one latent)

Concepts of psychometric equivalence: EFA

- Exploratory factor analysis model

I fit a one-factor model separately to the data from Latinos completing interviews in English and Spanish

The model fit well in both groups: $\chi^2_2 = 5.77$ and $\chi^2_2 = 1.70$, *n.s.*

Subjectively compare the factor loadings across groups....

items	English-language interviews		Spanish-language interviews
praise	.72	⟷	.78
control	.77	⟷	.90
diet	.66	⟷	.62
prevent	.79	⟷	.82

Concepts of psychometric equivalence: EFA

- Exploratory factor analysis model

  What do the results suggest?

  The model fit well in both groups

  Suggesting that the items were unidimensional in both groups

  Corresponding factor loadings were similar across groups

  Similarity of factor loadings across groups suggests that the meaning of the latent variable is similar across groups

  But were they similar enough? Subjective assessment

  Finally, the analysis does not address potential response bias

Concepts of psychometric equivalence: CFA

- The so-called confirmatory factor analysis (CFA) framework

  Simultaneously fit a factor model to data from two or more groups

  Assess how well the model 'fits' the data in each group

  Formally compare the model parameters across groups
  e.g., are the factor loadings equivalent across groups?

  Test for response bias

Concepts of psychometric equivalence: CFA

- Factor models are really sets of linear regression equations

- Quick review of linear regression models

In bivariate linear regression, individual outcomes are expressed as...

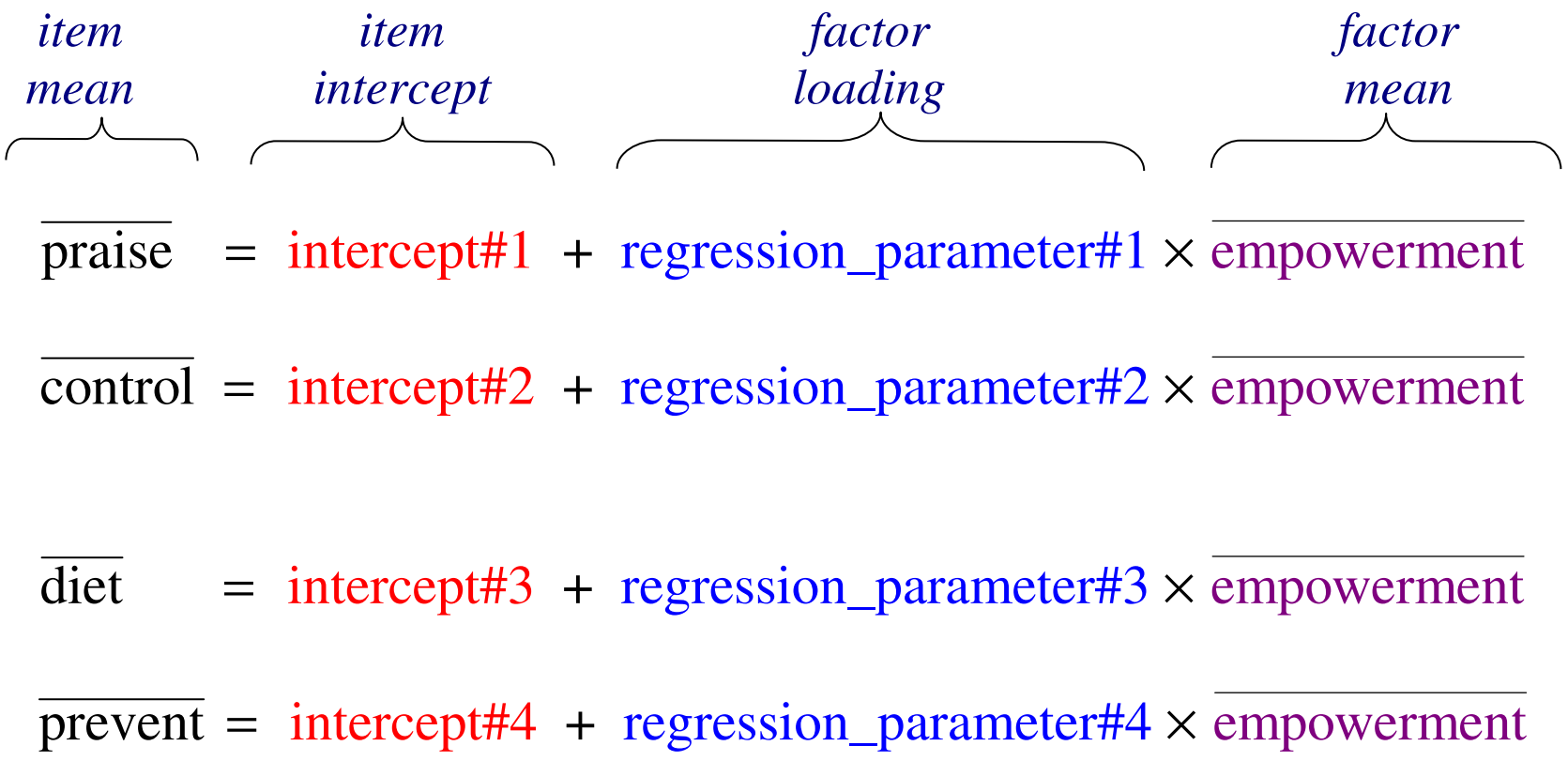
$$\text{outcome} = \text{intercept} + \text{regression\_parameter} \times \text{explanatory\_variable} + \text{residual}$$

In bivariate linear regression, mean outcomes are expressed as...

$$\overline{\text{outcome}} = \text{intercept} + \text{regression\_parameter} \times \overline{\text{explanatory\_variable}}$$

Concepts of psychometric equivalence: CFA

- Generically, in one group, the factor model would be a set of 4 bivariate linear regression equations...one equation for each item...



- Note that the intercepts and regression parameters are specific to each item

Concepts of psychometric equivalence: IRT

- IRT models are really sets of (usually) logistic regression equations (2PL)
- Assume that the IPC items used a binary yes/no response format.
- Quick review of logistic regression models

In bivariate logistic regression, individual predicted outcomes equal...

$$logit\ Pr(Y=1|X) = \text{intercept} + \text{regression\_parameter} \times \text{explanatory\_variable}$$

In bivariate logistic regression, overall probability of the outcome equals...

$$logit\ \overline{Pr(Y=1|X)} = \text{intercept} + \text{regression\_parameter} \times \overline{\text{explanatory\_variable}}$$

Concepts of psychometric equivalence: IRT

- Generically, in one group, the IRT model is a set of 4 bivariate logistic regression equations...one equation for each item...

probability of response | *difficulty* (β) | *discrimination* (α) | *trait mean* ($\bar{\theta}$)

$$logit\ \overline{Pr(\text{praise}=1|\theta)} = \text{intercept\#1} + \text{regression\_parameter\#1} \times \overline{\text{empowerment}}$$

$$logit\ \overline{Pr(\text{control}=1|\theta)} = \text{intercept\#2} + \text{regression\_parameter\#2} \times \overline{\text{empowerment}}$$

$$logit\ \overline{Pr(\text{diet}=1|\theta)} = \text{intercept\#3} + \text{regression\_parameter\#3} \times \overline{\text{empowerment}}$$

$$logit\ \overline{Pr(\text{prevent}=1|\theta)} = \text{intercept\#4} + \text{regression\_parameter\#4} \times \overline{\text{empowerment}}$$

- Note that the intercepts and regression parameters are specific to each item

Concepts of psychometric equivalence: CFA and IRT

- The CFA and IRT models allow comparison of corresponding parameter estimates across samples

  Are corresponding parameter estimates equivalent across groups?

  Equivalence of corresponding *factor loading* or *item discrimination* parameters suggests that the same construct is being measured in all groups (i.e., the instrument is equally valid in all groups)

  Equivalence of corresponding *item intercept* or *item difficulty* parameters suggests that response bias will not contaminate group comparisons

CFA two-group, one-factor partial invariance model: Latinos

English interview	Spanish interview
item mean / *item intercept* / *factor loading* / *factor mean*	*item mean* / *item intercept* / *factor loading* / *factor mean*
$\overline{\text{praise}} = -0.86 + 1.22 \times \overline{E}_{ENG}$	$\overline{\text{praise}} = -0.86 + 1.22 \times \overline{E}_{SPN}$
$\overline{\text{control}} = -0.63 + 1.27 \times \overline{E}_{ENG}$	$\overline{\text{control}} = -0.63 + 1.27 \times \overline{E}_{SPN}$
$\overline{\text{diet}} = 0.50 + 1.00 \times \overline{E}_{ENG}$	$\overline{\text{diet}} = 0 + 1.00 \times \overline{E}_{SPN}$
$\overline{\text{prevent}} = -0.35 + 1.22 \times \overline{E}_{ENG}$	$\overline{\text{prevent}} = -0.63 + 1.22 \times \overline{E}_{SPN}$

A two-group, one-factor partial invariance model: Latinos

- Corresponding factor loadings were equivalent across groups

  Suggests that the items represented the same underlying construct across Latinos who completed the interview in English and Spanish.

- Corresponding item intercepts for *praise* and *control* were equivalent across groups

  Suggests that these items allow for unbiased comparison of groups

- Corresponding item intercepts for *diet* and *prevent* were not equivalent across groups

  Suggests that group comparisons based upon these items are biased

Impact of partial invariance on group comparisons

- I created composite scores for each respondent by taking the mean of all 4 items

  Possible values ranged from 1 to 5 (i.e., never to always)

  The English mean, 3.56, was higher than the Spanish mean, 3.50.

  But this difference was not significant

- I next created composite scores for each respondent by taking the mean of the 2 unbiased items

  The English mean, 3.38, was lower than the Spanish mean, 3.51.

  Here the difference was significant, $p < .05$ (Kruskal-Wallis test)

Impact of partial invariance on group comparisons

- In this example, the IPC Empowerment scale appeared to measure the same construct across groups of Latinos who completed the interview in English and Spanish
- However, evidence of response bias was found
- When comparing groups using the 4-item composite response bias for *diet* and *prevent* made it appear that there were no group differences
- When comparing groups using the unbiased 2-item composite a group difference was observed
- Especially for *soft* measures, such as attitudes, opinions, motives, and self-reported behaviors, it is important to assess whether response bias will contaminate group comparisons.
- Practical implications for research into health disparities