RTI HEALTH SOLUTIONS®

Quantitative Challenges Facing Patient-Centered Outcomes Research

ISPOR 19th Annual International Meeting
May 31-June 4, 2014

RTI Health Solutions

Research Triangle Park, NC, US
Ann Arbor, MI, US
Barcelona, Spain
Ljungskile, Sweden
Manchester, UK
Waltham, MA, US

www.rtihs.org
e-mail: rtihealthsolutions@rti.org

LEADING RESEARCH…
MEASURES THAT COUNT

QUANTITATIVE CHALLENGES FACING PATIENT-CENTERED OUTCOMES RESEARCH

PURPOSE

Patient-centered outcomes researchers collect data from patients and caregivers that can be used to guide healthcare decisions and improve healthcare delivery and outcomes. This workshop presents challenges associated with conducting quantitative data analysis of patient-centered data. At the end of the workshop, participants will have a better understanding of these analytical challenges and available approaches to successfully overcome them.

DESCRIPTION

Topic 1: Heterogeneity in patient-centered outcomes often translates into multidimensionality in data analysis. Different languages and cultures also contribute to heterogeneity of patient-centered outcomes that may lead to bias results. Methods to explore dimensionality and differential item functioning (DIF) are presented, including available software and programs.

Topic 2: Insufficient sample size may lead to large measurement errors or nonconvergent models. Too large of a sample size overpowers tests of significance. Recommendations for sample size when evaluating patient-centered measures are discussed. Rules of thumb for commonly used psychometric analyses, to ensure appropriate statistical inferences, are presented.

Topic 3: Missing data are inevitable, but non-random missing or skip-pattern questions by design may lead to bias and incorrect results. Patterns of missing data should be investigated with respect to demographics, disease severity, and study arms. A case study of a pattern-mixture model is used to identify groups of subjects with similar missing data patterns.

Topic 4: Low response rates are typically non-random and can threaten generalizability of results. Options for maximizing response rate and minimizing respondent burden, including use of IRT-based tools (short forms and computer adaptive tests) and multiple assessment platforms (hand-held devices such as tablets and smart phones) will be discussed.

Discussion Leaders

Wen-Hung Chen, PhD
Director of Psychometrics
RTI Health Solutions
3040 Cornwallis Road
Research Triangle Park, NC
27709
1-919-248-8522
wchen@rti.org

Lauren Nelson, PhD
Director of Psychometrics
RTI Health Solutions
3040 Cornwallis Road
Research Triangle Park, NC
27709
1-919-541-6590
lnelson@rti.org

Lori D. McLeod, PhD
Head, Psychometrics
RTI Health Solutions
3040 Cornwallis Road
Research Triangle Park, NC
27709
1-919-541-6741
lmcleod@rti.org

Maria Orlando Edelen, PhD
Senior Behavioral Scientist
RAND Corporation
20 Park Plaza, Suite 920
Boston, MA 02116
1-617-338-2059 ext. 8634
Maria_Edelen@rand.org

RTI HEALTH SOLUTIONS®

Quantitative Challenges Facing Patient-Centered Outcomes Research

Dimensionality and Differential Item Functioning

Wen-Hung Chen
ISPOR 19th Annual International Meeting
May 31-June 4, 2014

RTI Health Solutions

Research Triangle Park, NC, US
Ann Arbor, MI, US
Barcelona, Spain
Ljungskile, Sweden
Manchester, UK
Waltham, MA, US

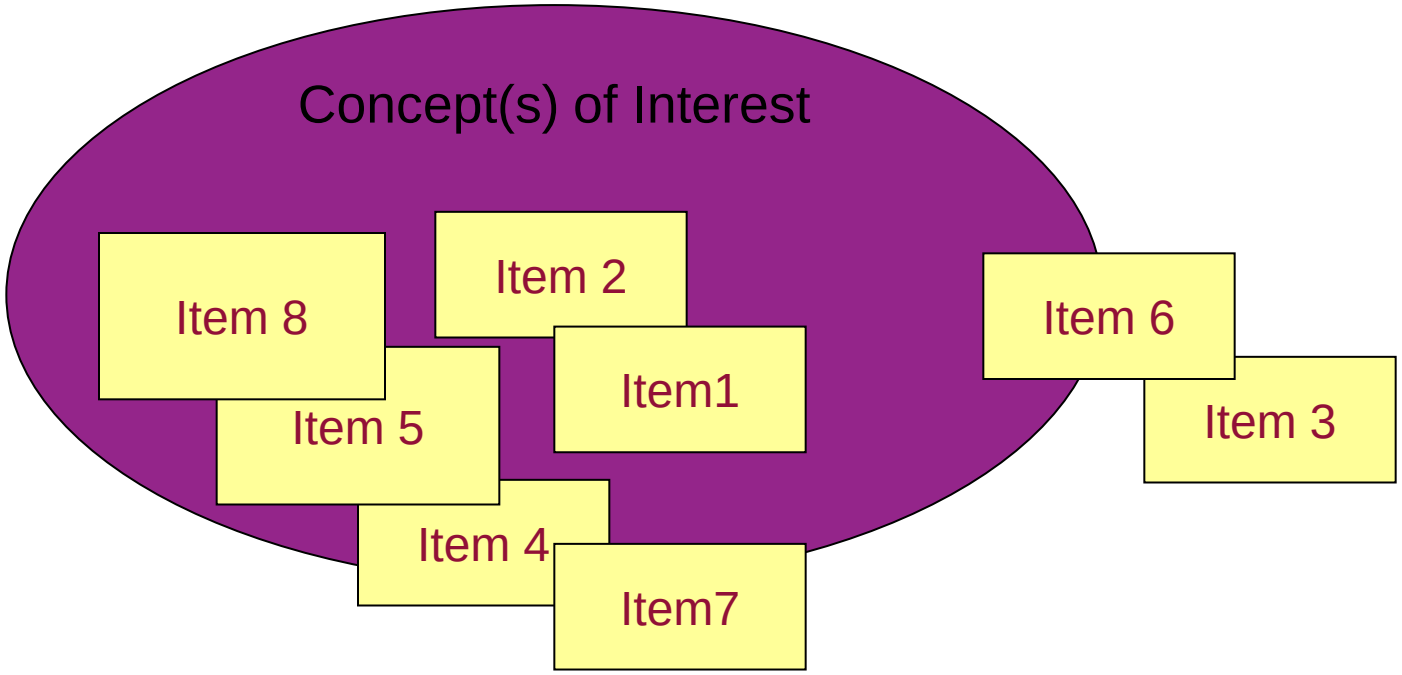
www.rtihs.org
e-mail: rtihealthsolutions@rti.org

Multidimensionality

- Many patient-reported outcome (PRO) measures are multidimensional in nature
  - The EuroQoL 5 Dimensions Questionnaire (EQ-5D)
  - SF-36

Concept of Interest Measured by Multiple Items

- Multiple variables (items) allow us to tap into more of the concept of interest (COI) in terms of content and to increase reliability, but…
- You have to make sense of all variables
  - How do they relate to one another and the COI?

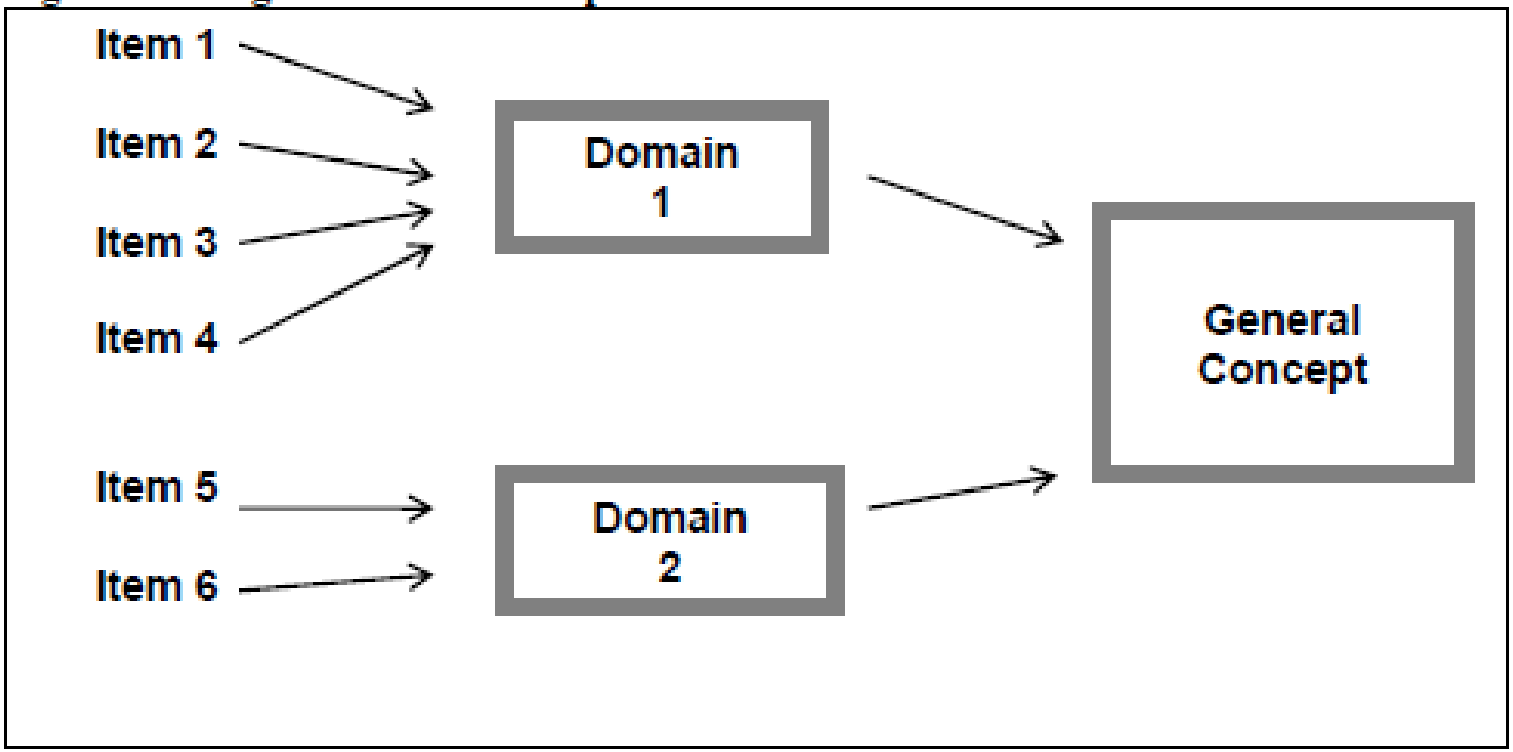


Conceptual Framework of a PRO Measure

- The adequacy of a proposed instrument to support a claim depends on the conceptual framework of the PRO instrument
- The conceptual framework explicitly defines the concepts measured by the instrument in a diagram that presents a description of the relationships between items, domains (subconcepts), and concepts measured and the *scores* produced by a PRO instrument (FDA PRO Guidance, 2009)

A Multidimensional Conceptual Framework

Figure 4. Diagram of the Conceptual Framework of a PRO Instrument



Common Approaches for Assessing Dimensionality

- Exploratory Factor Analysis (EFA): Attempts to discover how many factors (concepts/domains) are present and which variables they explain without making prior assumptions about the number of factors and how the items related to the factors

- Confirmatory Factor Analysis (CFA): Tests specific, theory-based hypotheses or to confirm EFA results about how many dimensions are present and which variables they explain

Factor Analysis

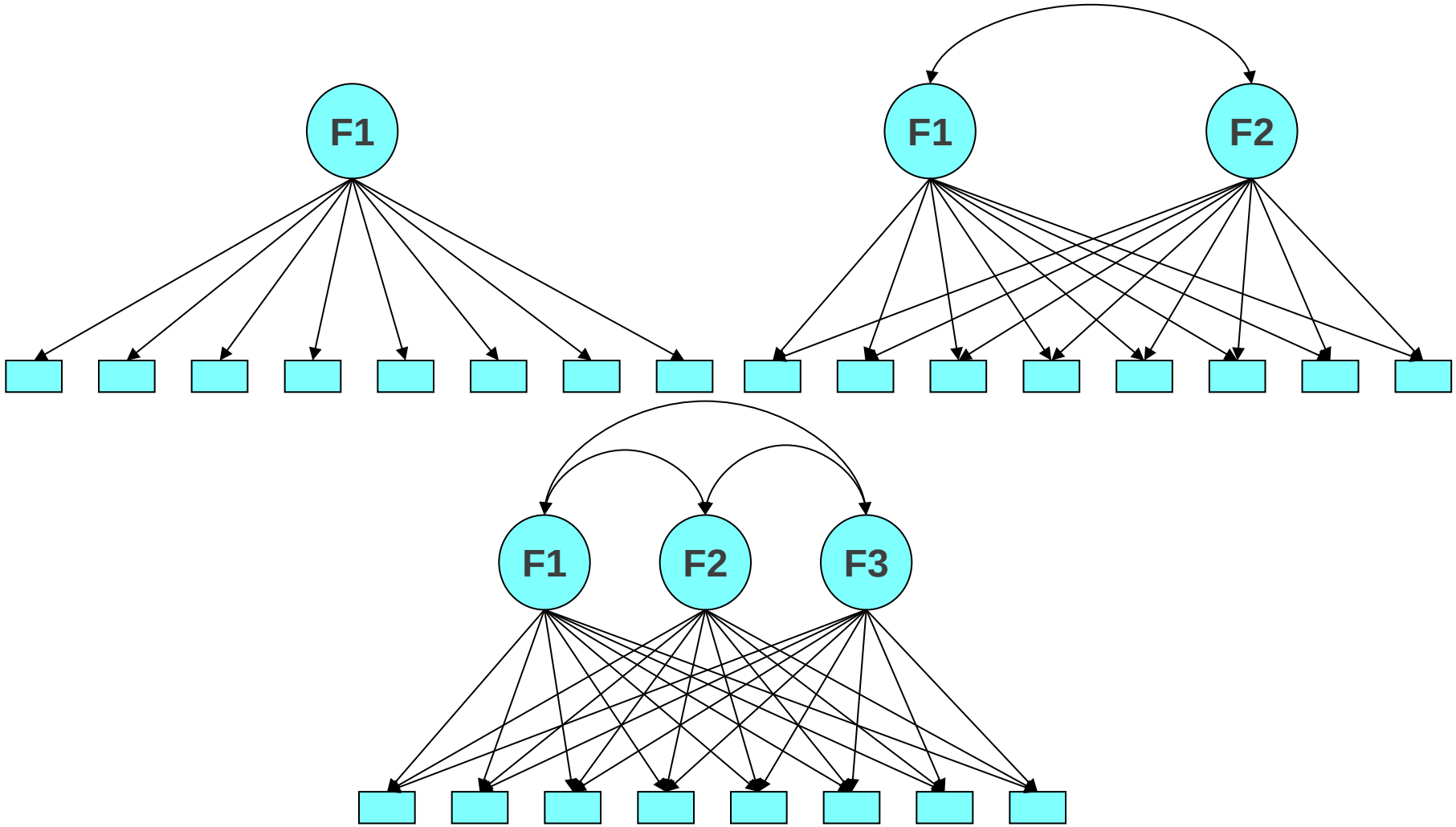
- Summarizes the pattern of correlations among a set of observed variables
- Variables correlated with one another but largely independent of others are combined into factors
- What accounts for the pattern of correlations among these variables? What do they have in common? That is, what is the "thing" the correlated variables are measuring?

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	1.00									
X2	0.40	1.00								
X3	0.50	0.40	1.00							
X4	0.35	0.50	0.01	1.00						
X5	0.07	0.05	0.50	0.30	1.00					
X6	0.11	0.09	0.60	0.40	0.03	1.00				
X7	0.04	0.12	0.40	0.35	0.02	0.40	1.00			
X8	0.12	0.04	0.03	0.07	0.40	0.30	0.07	1.00		
X9	0.09	0.11	0.06	0.05	0.50	0.40	0.50	0.40	1.00	
X10	0.05	0.07	0.08	0.02	0.05	0.06	0.60	0.35	0.50	1.00

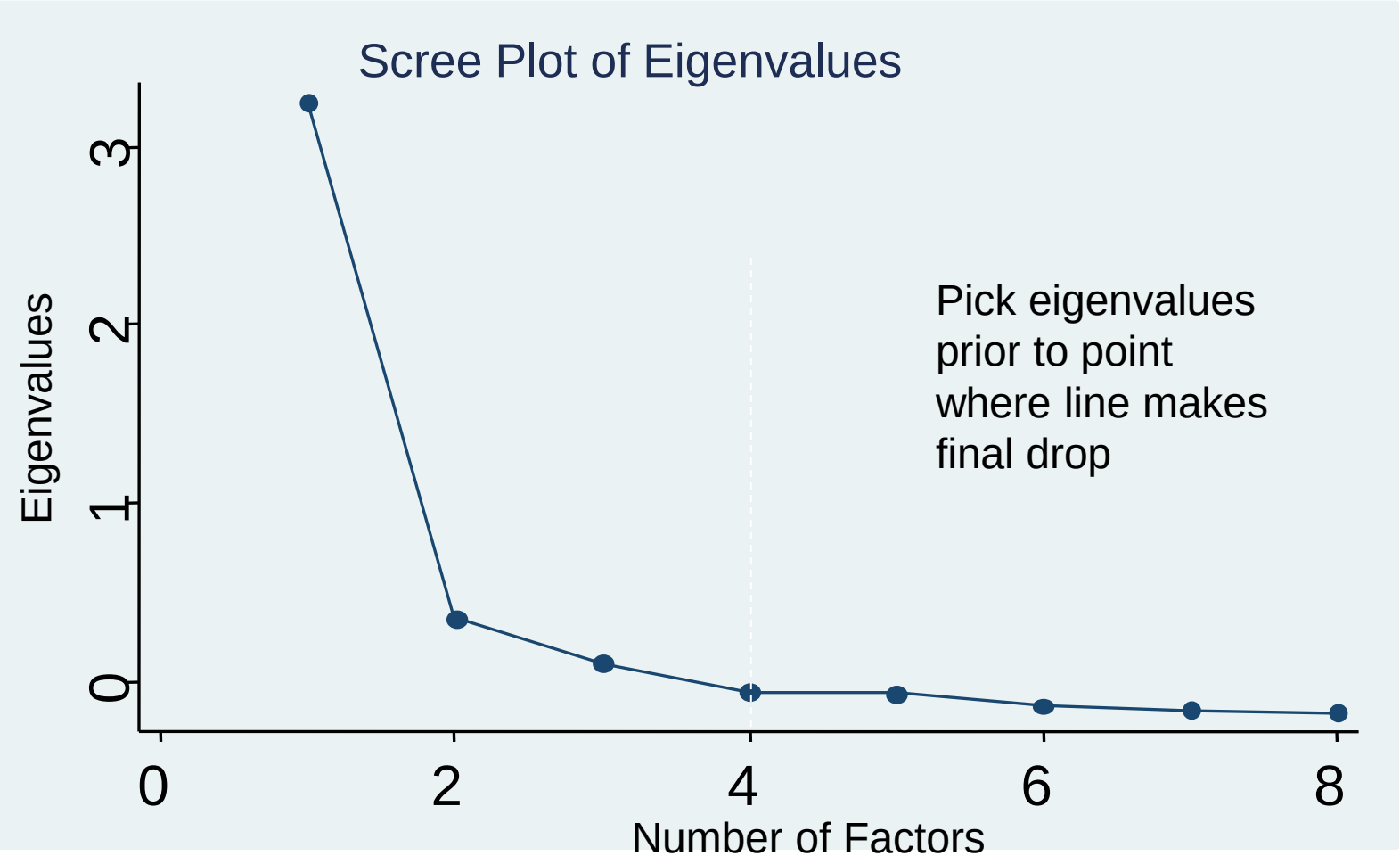
Example: EFA of the Center for Epidemiologic Studies Depression Scale 8-Items Version (CESD-8)

- Measures of depression (CESD-8):
  - Feel depressed
  - Everything is an effort
  - Sleep is restless
  - Feel happy
  - Feel lonely
  - Enjoys life
  - Feels sad
  - Unable to get going

CESD Possible Factor Structure



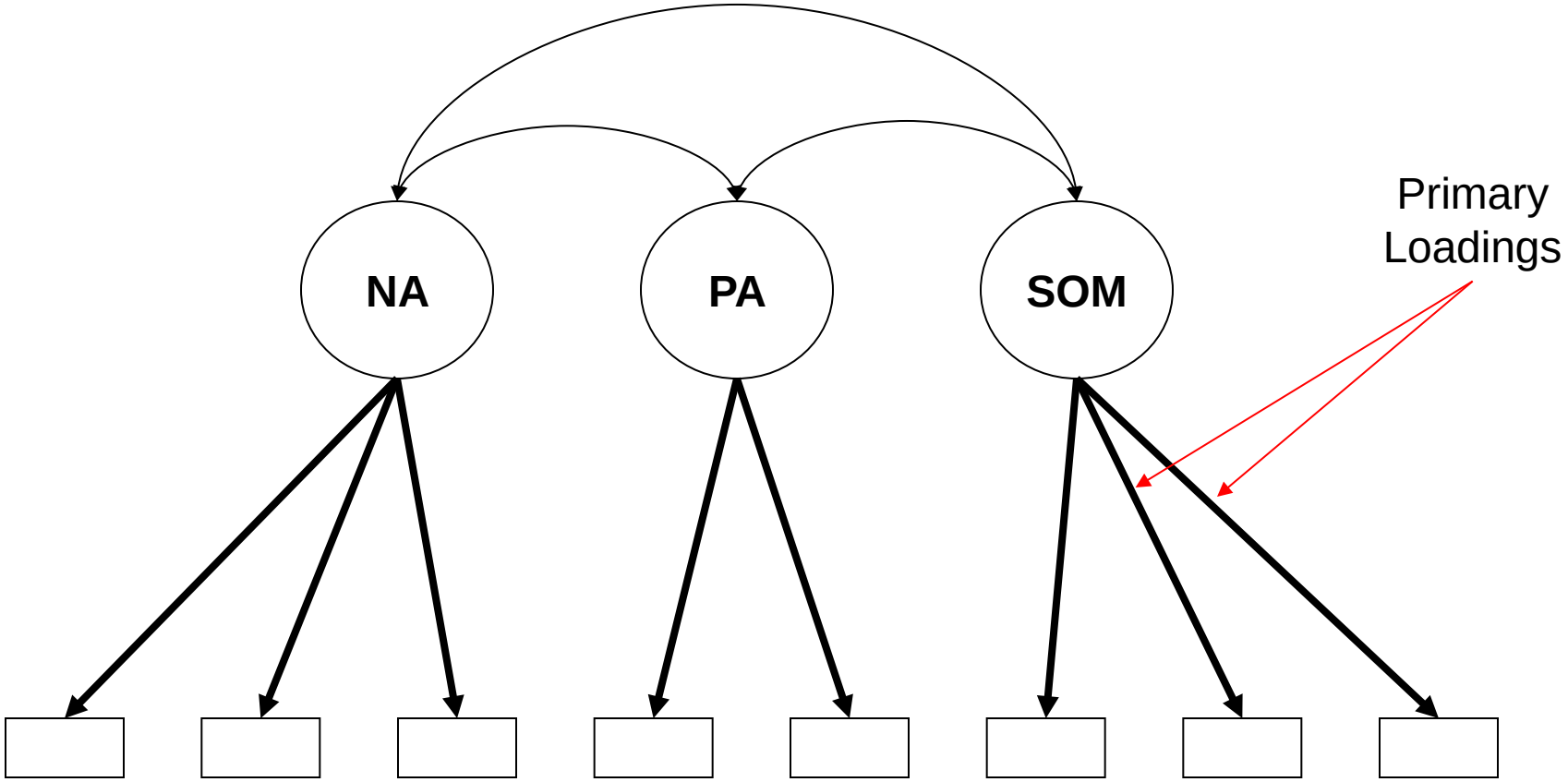
Eigenvalues and Scree Plot



CESD-8 Factor Structure Identified by EFA

- Negative affect/depression:
  - Depressed
  - Lonely
  - Sad
- Somatic:
  - Everything an effort
  - Sleep is restless
  - Unable to get going
- Positive affect:
  - Happy
  - Enjoy life

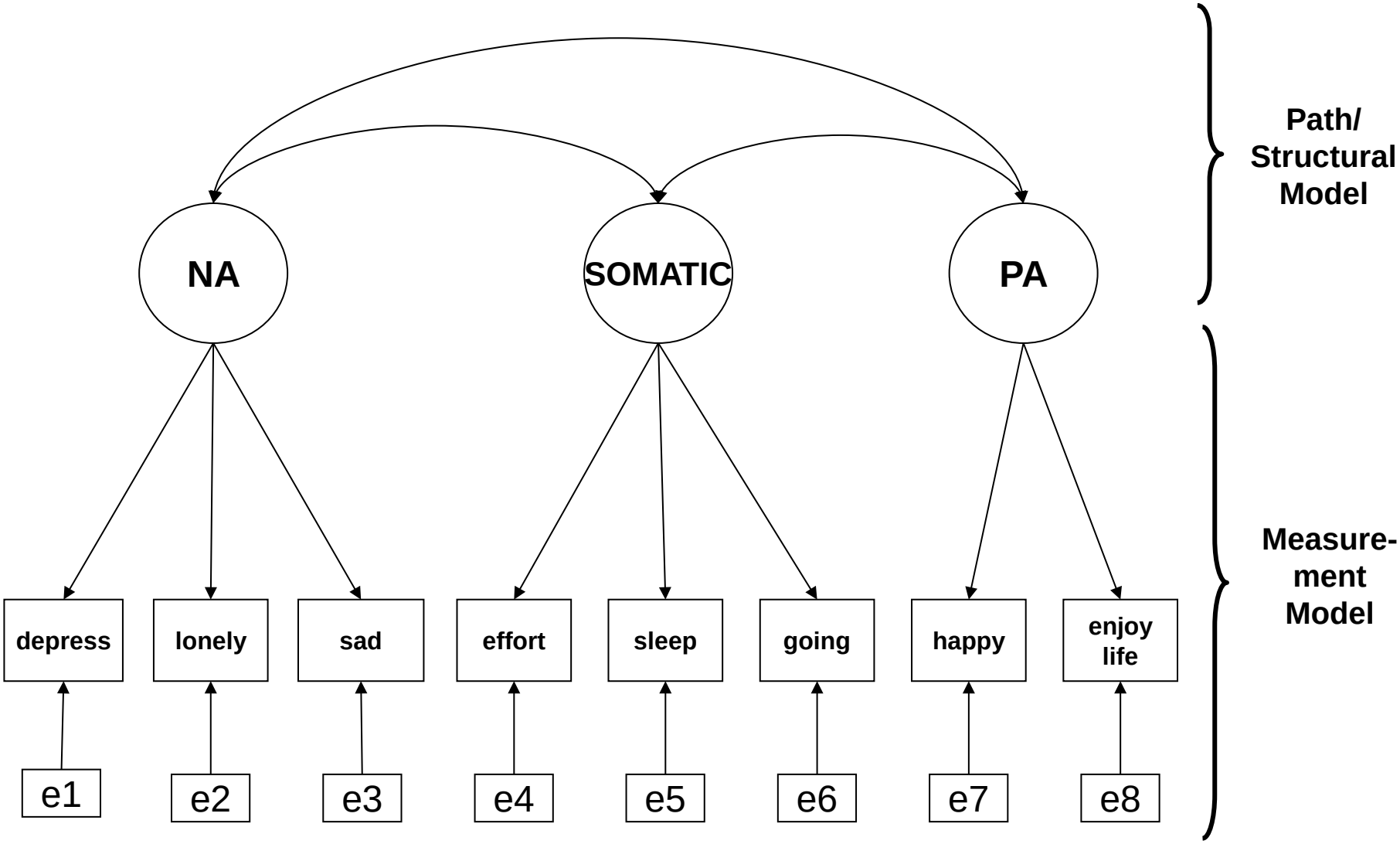
CESD-8 Factor Structure Identified by EFA



Confirmatory Factor Analysis (CFA)

- Confirmation of hypothesized factor structure of PRO measures
  - Do the results of the EFA hold up under CFA?
- Validation of PRO measures
  - Does a PRO relate to antecedents/consequences of interest?
  - Can assess multiple domains of PRO measures simultaneously
- Cross-validation of PRO measures
  - Is the factor structure the same for key subgroups (gender, race/ethnicity, age, country, treatment/control)?

CESD-8 Factor Structure Confirmed by CFA



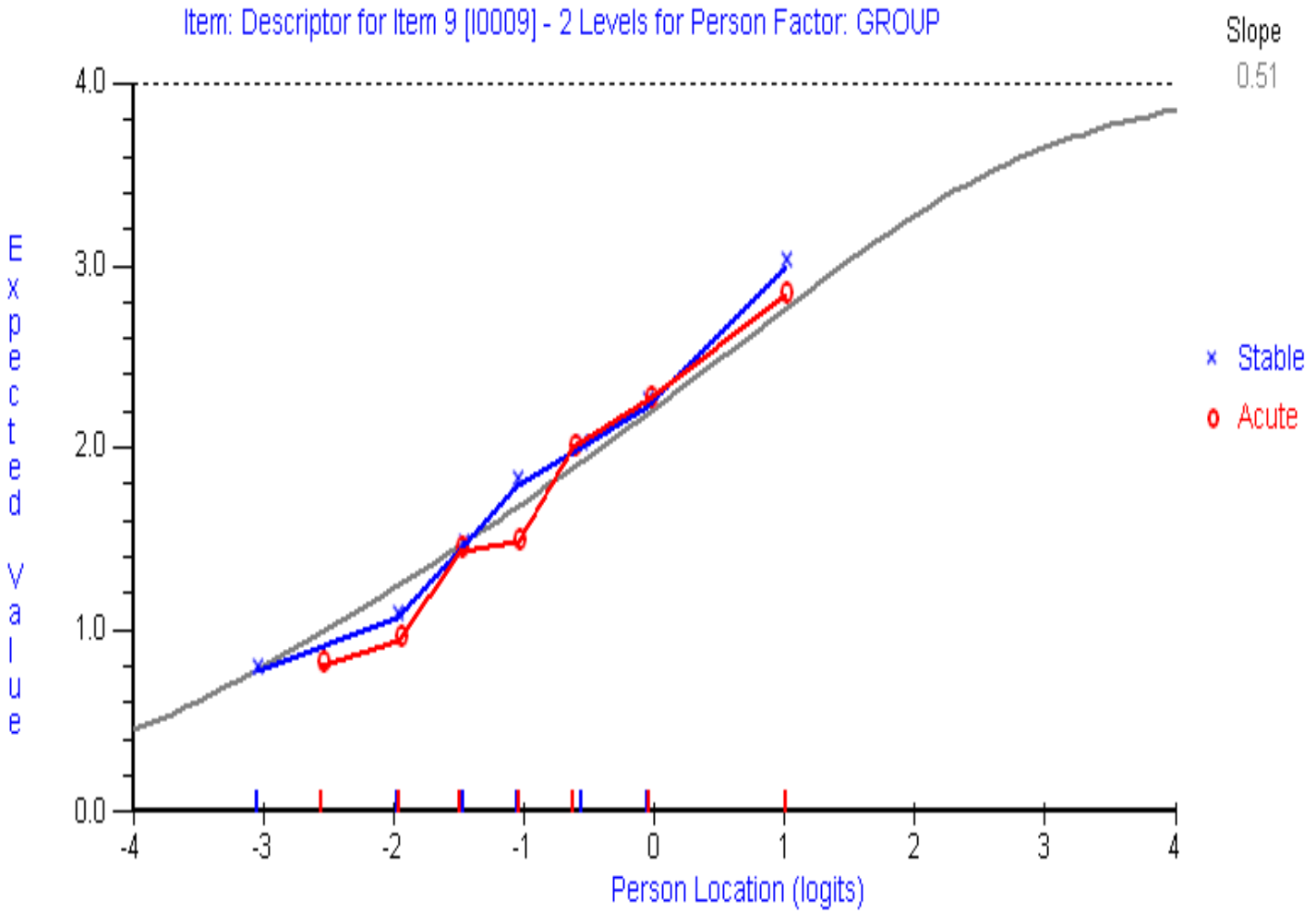
Leading Factor Analysis Software

- Mplus (most all-encompassing and rapidly becoming the most popular)
- AMOS (most user-friendly)
- EQS (one of the oldest programs; good combination of power and ease of use)
- LISREL (first structural equation modeling software)
- SAS PROC FACTOR and PROC CALIS (much improved from earlier releases and catching on)

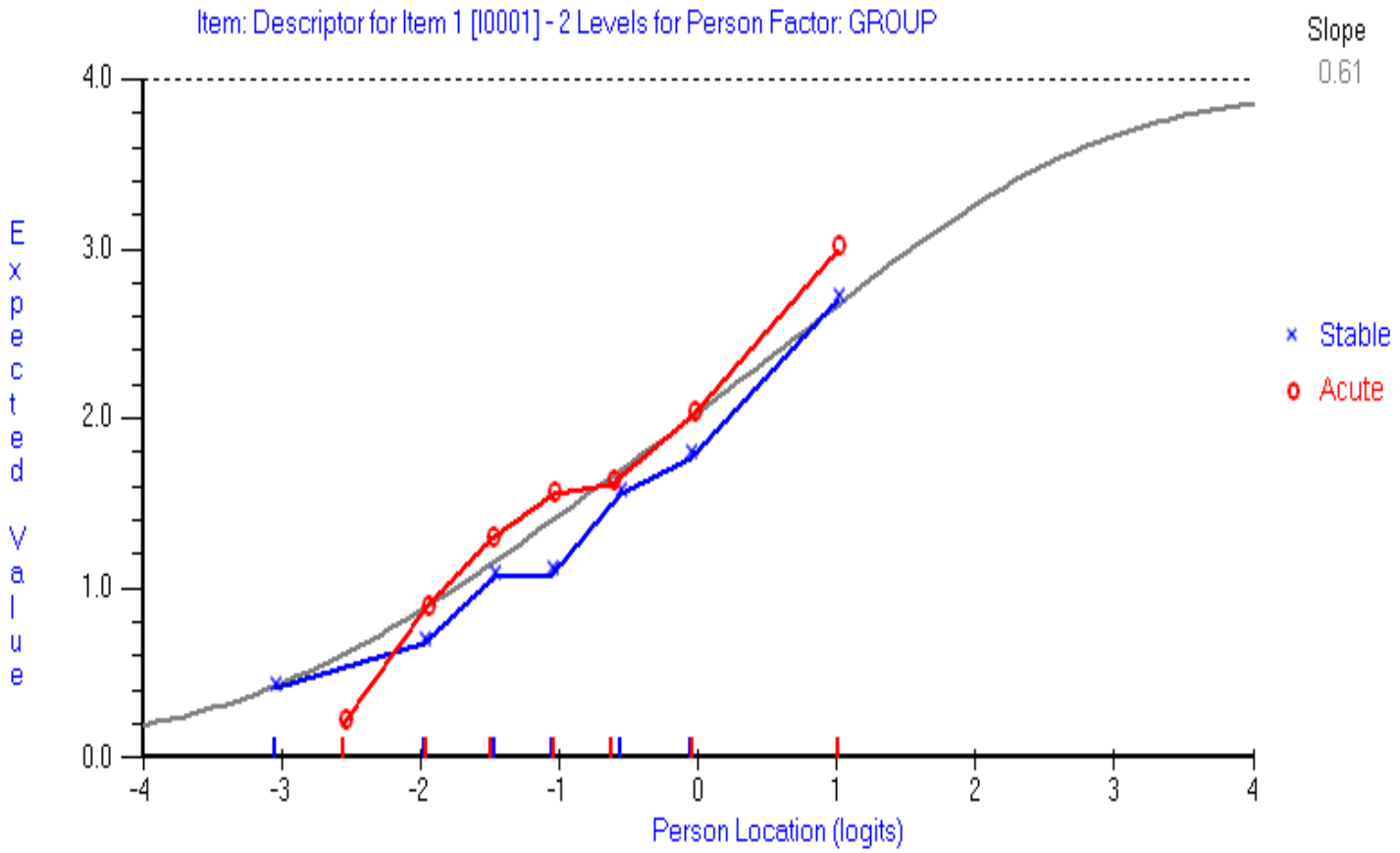
Differential Item Functioning (DIF)

- Heterogeneity among patients—such as age, gender, education level, language, culture, etc.—is likely to cause differential item function (DIF)
- DIF: Subjects from different groups (e.g., male vs. female) with the same latent trait (e.g., degree of depression) have a different probability of giving a certain response to an item (e.g., crying)
- For the score to be equivalent among groups, the items with DIF should be removed, or the DIF items should be scored differently according to group membership

Item Exhibits No DIF



Item Exhibits DIF



Example from PROMIS Pain Behavior Draft Item Bank

An item that exhibits DIF by gender

PAINBE27	**In the past 7 days**	**I had pain so bad it made me cry**	**1 = Had no pain 2 = Never 3 = Rarely 4 = Sometimes 5 = Often 6 = Always**

Items With DIF

- Assess the impact at the scale level
- Assess the content of the item
- Reword the item
- Remove the item
- Different scoring algorithms according to group memberships

Commonly Used Methods to Detect DIF

- Mentel-Haenszel Chi-square
  - Construct 3-fold contingency tables to compare proportion of each response categories between two (or more) groups
  - SAS, SPSS, R, etc.
- Item Response Theory (IRT)-Based Methods
  - Comparing the IRT item parameter estimates between two (or more) groups
  - IRTLRDIF, IRTPRO
- Rasch Model-Based Method
  - Analysis of variance on standardized residuals
  - RUMM2030
- Logistic Regression
  - Logistic regression model to compare the odds of observing each response categories between two (or more) groups
  - SAS, SPSS, R, etc.

RTI HEALTH SOLUTIONS®

Topic 2: Sample Size for Post Hoc PRO Analyses

Lori McLeod
lmcleod@rti.org

ISPOR 19th Annual International Meeting
May 31-June 4, 2014

RTI Health Solutions

- Research Triangle Park, NC, USA
- Ann Arbor, MI, USA
- Barcelona, Spain
- Ljungskile, Sweden
- Manchester, UK
- Waltham, MA, USA

www.rtihs.org
e-mail: rtihealthsolutions@rti.org

Sample Size Matters

- Too large
  - Overpowered for the PRO evaluation, and conclusions are compromised
    - "Statistical significance does not equate to meaningful difference"

- Too small
  - Models may not converge
  - Underpowered, and intended evaluations are not possible

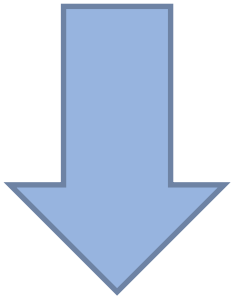
cosmologybus.typepad.com

Sample Size Matters

- Just right
  - Nonprimary clinical trial analysis: Justification after the fact (e.g., using confidence intervals) and then a responder analysis
  - Psychometric evaluation: Justification and planning for the sample size included in the study design discussions and documented

Sample Size: Reality

For most clinical studies, the PRO sample size is determined by the analytic demands for the primary endpoint	For psychometric evaluations of PRO measures, the sample size is often a compromise due to timeline, resources, and cost restrictions



No matter the circumstances, the size of the sample used in the post hoc PRO analyses should be considered when reporting results

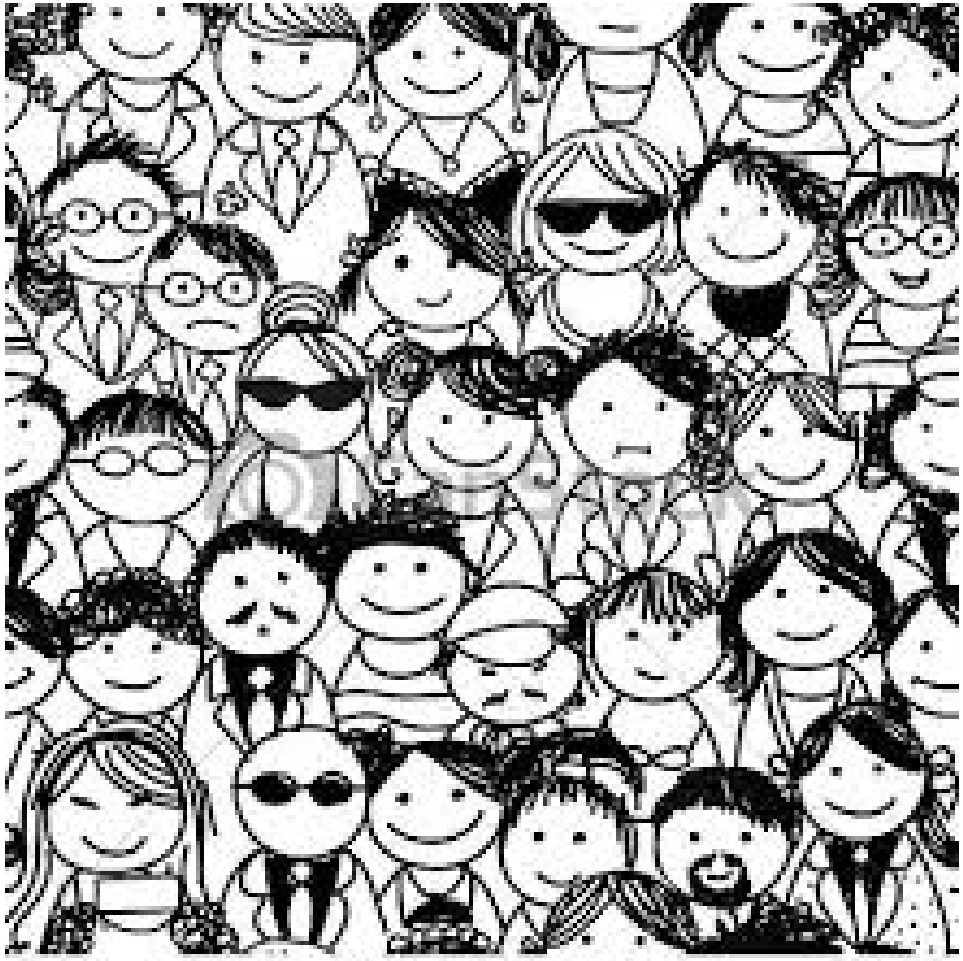
Note: Sample size is the number of patients completing the PRO measure

Recommendation: For "Just Right" Sample Sizes

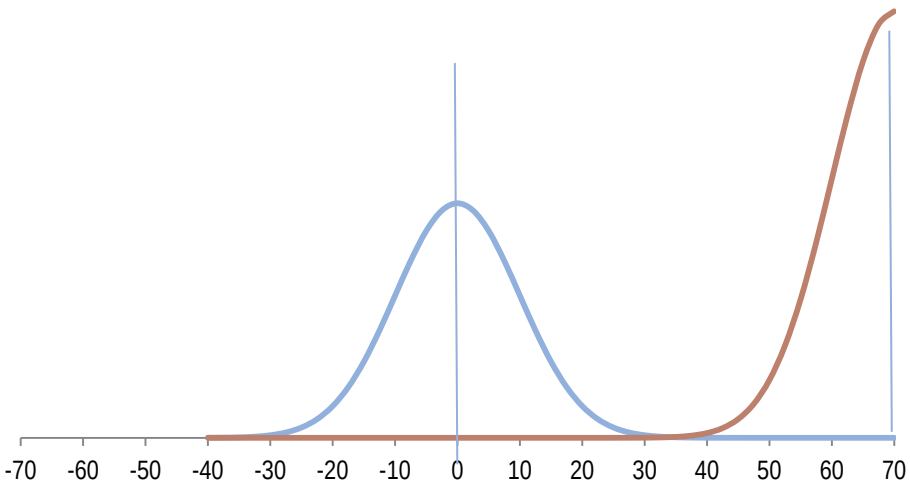
- Example: Group mean differences
  - If the test for the group means is statistically significant, further support the PRO score differences by reporting the proportion of responders by group (if available)
    - Use a distribution-based estimate if not available

Recommendations: If Sample Size Is Too Large

- Conduct the test on the full sample
- Provide information on the proportion of responders by group (if available or use distribution-based threshold)
- Conduct the test on multiple smaller subsamples (randomly select multiple subsamples using the clinical study stratum to sample and conduct the test)
- Provide confidence bands or effect size estimates



Example:
Clinical Trial Post Hoc PRO Analysis



$$n = \frac{2\sigma^2(Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

$$n = \frac{2\,(100)(0.84+1.96)^2}{5^2} = 63$$

- Sample size needed for 80% power and 0.05 level of significance is approximately 63 per arm
- Sample size of the clinical trial is 250 per arm, and resulting power is ~100% to detect 5 points difference in means!

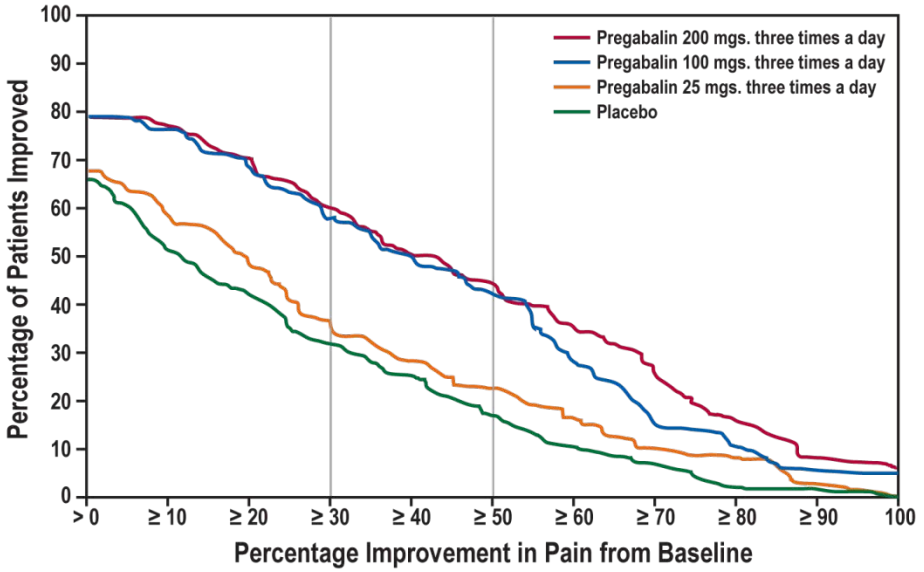
Furthermore, group mean tests ignore other aspects related to PRO scores:

- Score ranges
- Skewness

Example From Lyrica

Postherpetic Neuralgia (n=368)

- Treatment with LYRICA 100 mg and 200 mg three times a day statistically significantly improved the endpoint mean pain score and increased the proportion of patients with at least a 50% reduction in pain score from baseline
- Evaluated proposed cut scores of 30% and 50% improvement on an 11-point numerical pain rating scale to determine if three treatment groups differ from placebo



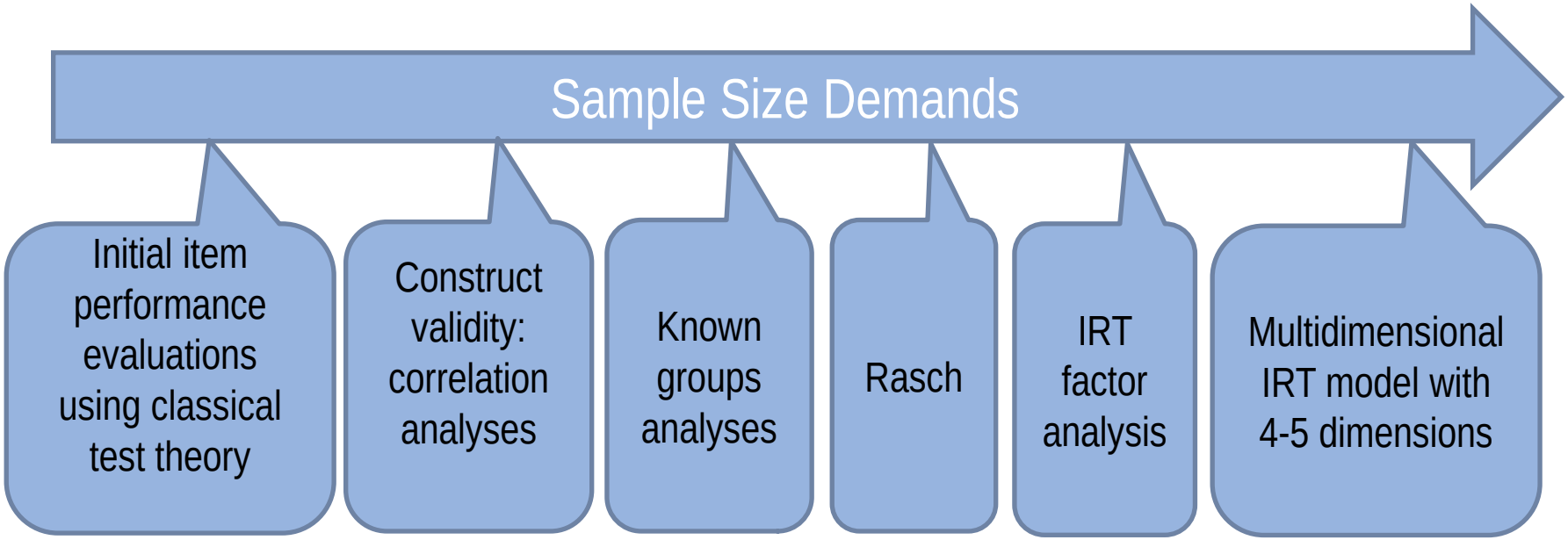
Lyrica, 2006 *Note: Positive change indicates improvement*

Recommendations: If Sample Size is Too Small

- Report descriptive information on the PRO (including effect size)
- Pool across studies (if same design, especially for cases with small populations such as orphan disease)
- Provide support for results based on published information from similar studies (e.g., compare effect sizes or clinically meaningful differences)
- Provide information on the proportion of responders by group (if available or use distribution-based threshold)

Psychometric Evaluations: Recommendations for Sample Size

- A priori state the key property for the psychometric evaluation
- Based on the key property, justify the sample size
- Incorporate confidence intervals for results for the secondary properties evaluated



Psychometric Evaluation: Sample Size Should Be Based on Key Property

-Additional Development post-ISPOR is underway for this content. Please contact lmcleod@rti.org for the current status.

“Real Life” Example

Property	n = 30	n = 50	n = 100	n = 150+
Cronbach’s alpha	0.88	0.89	0.87	0.87
Construct validity correlation	-0.42 (Less than 20)	-0.21 (Based on 35)	-0.35	-0.35
Known groups effect size	Not computed (Less than 10 in each group)	Apprx. 0 (Less than 20 in each group)	0.47	0.37

Summary and Fine Print

- Consider sample size when evaluating PRO measures or evaluating treatments using PRO measures
- One size does not fit all
  - We have provided recommendations/ rules of thumb
  - Justify the sample size after considering the complexity of the PRO measure, its intended use, the purpose of the evaluation, etc.

www.dailyfinance.com

RTI HEALTH SOLUTIONS®

June 4, 2014

Quantitative Challenges Facing Patient-Centered Outcomes

Missing Data

Lauren Nelson
lnelson@rti.org

Presented to: ISPOR 2014

RTI Health Solutions

- Research Triangle Park, NC, USA
- Ann Arbor, MI, USA
- Barcelona, Spain
- Ljungskile, Sweden
- Manchester, UK
- Waltham, MA, USA

www.rtihs.org
e-mail: rtihealthsolutions@rti.org

Is Missing PRO Data Inevitable?

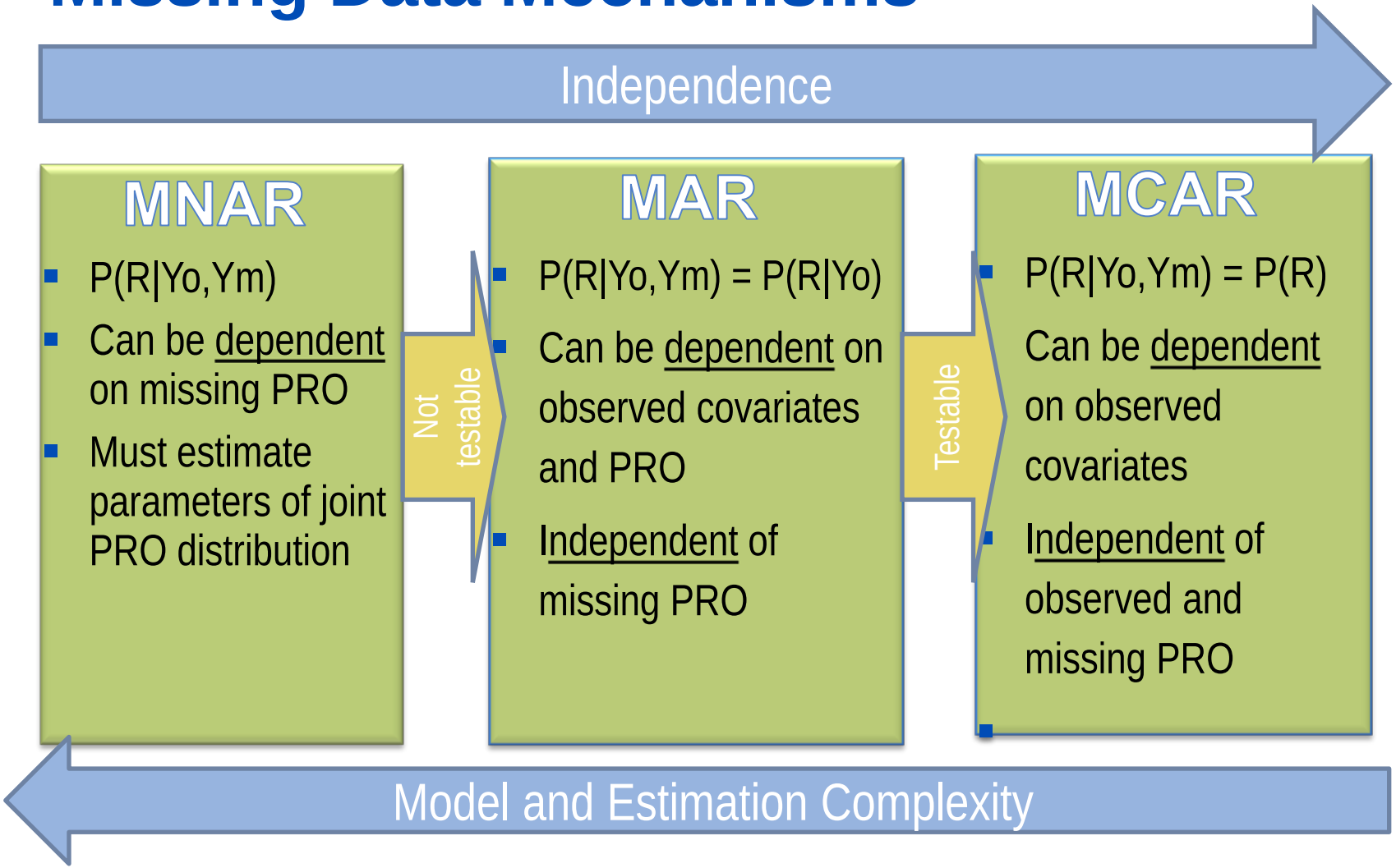
- Missing (incomplete) data commonly occurs in longitudinal studies despite well-planned and carefully executed studies
  - Declaration of Helsinki
- Magnitude of missingness varies
  - Trial length
  - Disease
  - Treatment
- Types of missing
  - Scale nonresponse
  - Item nonresponse

How Much Data Can Be Missing: "Rules of Thumb"

- Scale nonresponse
  - < 5% ignorable
  - 5%-20% may or may not impact conclusions
  - 30%-50% restrict conclusions
- Item nonresponse
  - < 5% ignorable
  - > 10% concern
- Seriousness of missing depends on the reasons or missing, study objective, and intended use
- Data should always be inspected for missing patterns
- Rules of thumb vary across the literature and types of missing data mechanisms
  - Missing Not at Random (MNAR)
  - Missing at random (MAR)
  - Missing Completely at Random (MCAR)

Fairclough (2010), Capparelli (2013)

Review:
Missing Data Mechanisms



R = distribution of missingness (0 = not missing, 1 = missing); Yo = observed PRO; Ym = missing PRO

Review:
Impact of Missing Data Mechanisms

- Important: Reduces power
  - A concern for rare conditions (orphan diseases) but not for overpowered, large-scale trials
- Most important: Can produce biased estimates and erroneous conclusions
  - Irrespective of sample size
  - Impacts variance estimates
  - Impacts random assignment (selection bias: subjects self-select), hence estimate

Prevention and Treatment of Missing Data in Clinical Trials

- First line of defense: Prevention
  - PRO scale data is precious!
  - Train sites and staff extensively on data collection
  - Consider patient burden
  - Complete evaluation of enrolled patients, irrespective of their adherence to study therapy or protocol
  - Collect auxiliary information on reasons for missing/dropout
- Last line of defense: Statistical

National Research Council (2010)

NRC Guidelines:
Identify Patterns of Scale Nonresponse

Pattern	Baseline	Week 4	Week 8	Week 12	Week 16
A	X	X			
B	X	X	X		
C	X	X		X	
D	X		X	X	
E	X	X	X	X	
F	X	X	X	X	X

NRC Guidelines: Statistical Approaches

- Scale nonresponse
  - Primary analysis: Select an analysis method with assumptions that are appropriate for MAR
    - Maximum likelihood estimation techniques
    - Bayesian multiple imputation
  - Secondary analysis: Select an alternative method (assuming MNAR) and conduct a sensitivity analysis
    - Pattern mixture models
    - Semi-parametric selection models
    - Use of auxiliary information may help MNAR approximate MAR
  - Compare results and conclusions
    - Currently no consensus on how to optimally synthesize results from the primary and secondary analyses

National Research Council (2010)

Guidelines for Item Nonresponse in Clinical Trials

Use the Developers' Algorithm

FDA (2009), Fairclough (2010), Capparelli, 2013

Developers' Algorithms

- Objective: Preserve reliability (classical test theory concept)
- Common approach: Single imputation
- "Half-rule" (person mean) single imputation
  - Subjects must respond to at least half of the items (otherwise missing)
  - Imputed response: Mean of nonmissing (within subject)
  - Score = sum of nonmissing items and imputed responses
- 'Maximum response possible' single imputation
  - Imputed response: Rescale the "worst" response option; multiply it by the ratio of the sum of nonmissing items to the possible total scale score
  - Score = sum of nonmissing responses and imputed responses

Fairclough (2010), Capparelli, 2013

Recommendations

-Additional Development post-ISPOR is underway for this content. Please contact lnelson@rti.org for the current status.

PRO Missing Data Frontier

- No data/information left behind (include missing data)
- Unified modern (i.e., model-based and computer intensive) scale and item nonresponse methods for clinical trial research
  - Measurement precision of the outcome measure will automatically be included. Trial models continue to treat an individual's score as if it was perfectly measured
  - Potentially will reduce bias and erroneous conclusions for "Treatment/No Treatment" high-stakes decisions
- Together, statisticians and psychometricians have the computing and brain power to accomplish this

Quantitative Challenges Facing Patient-Centered Outcomes Research

Maximizing Response Rates by Minimizing Burden

Maria Orlando Edelen

RAND Corporation

June 4, 2014, Montreal, Canada.

Overview

- The problem
- Leveraging modern resources to alleviate the problem
- The role of ‘modern’ measurement theory
  - IRT-based assessment
    - (short forms and computer adaptive tests)
- The role of technology and computer-based data collection

Low response rates are a threat to patient-centered outcomes research

- **Low response rates are typically non-random**
  - **Respondent fatigue or burnout**
  - **Irrelevant questions**
  - **Inconvenience**
  - **Lag time between recruitment and survey completion**

Most approaches to handling missing data require assumption of MCAR or MAR

- **MCAR = Missing Completely at Random**
- **MAR = Missing at Random**
- **Non-random missing data is**
  - **Difficult to manage analytically**
  - **A threat to generalizability of results**
- **Missing data is best avoided if at all possible!**

Modern resources can help to maximize response rate and minimize burden

- **Use of 'modern' measurement tools based on item response theory (IRT) can facilitate brief yet reliable assessment**
  - **short forms (SFs) and computer adaptive tests (CATs)**

- **Use of novel computer-based assessment platforms can increase convenience, reduce lag time, sustain attention**
  - **hand held devices, notepads, smart phones**

IRT has several features that facilitate creation of short precise instruments

- **Items and scores are placed on the same continuum**
- **The reliability of a score is a function of its location on the underlying measurement continuum**
- **Score reliability can be calculated based on responses to any given item or set of items**

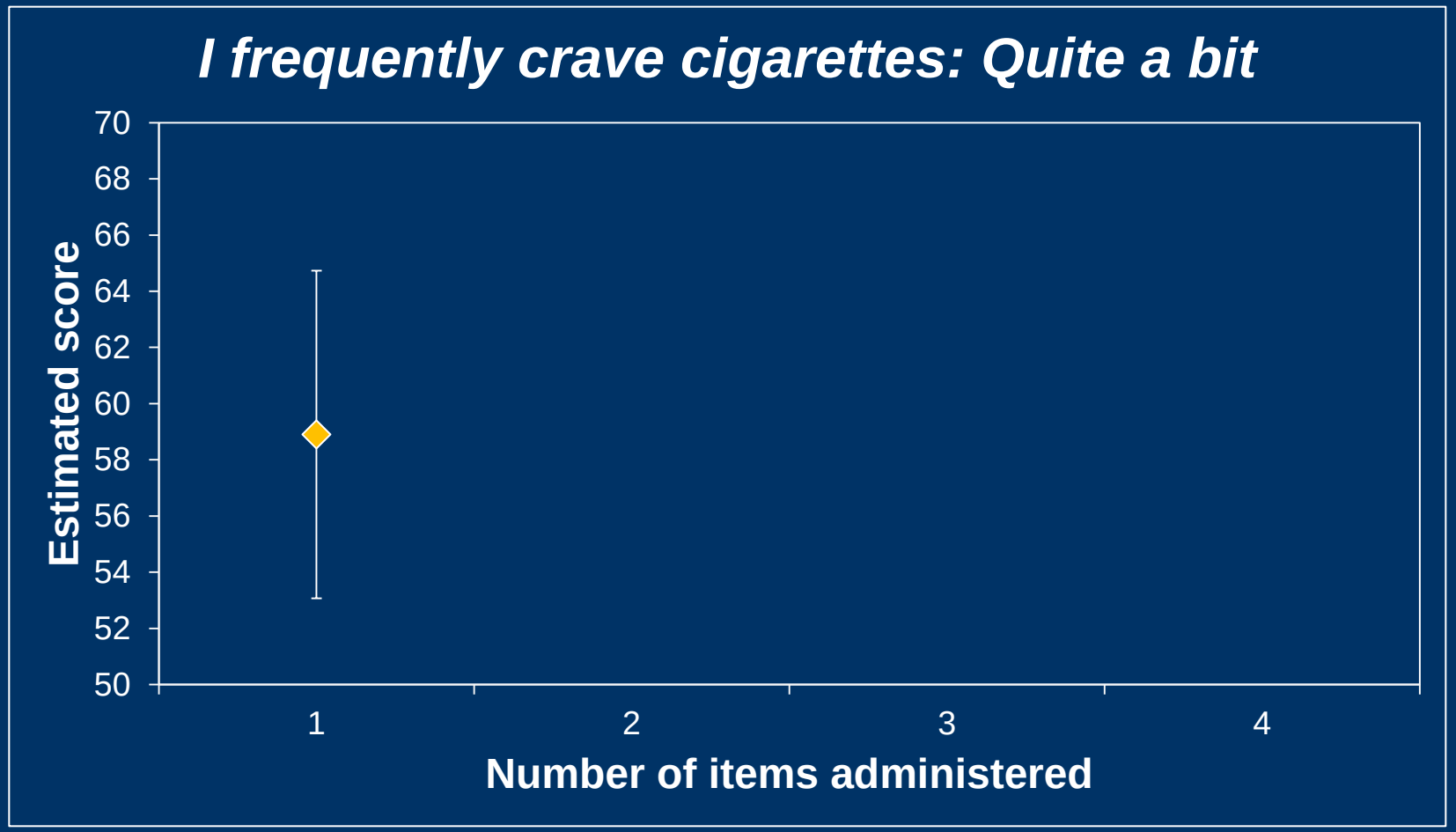
IRT-based SFs and CATs leverage these features

- **SFs are typically constructed to maximize precision**
  - **At a given point on the measurement continuum (e.g., cut-score for diagnosis)**
  - **Across the entire continuum**
- **Depending on measurement goals, different items will be selected for SF**
- **CAT takes this one step further by tailoring the item administration to the individual in real time**

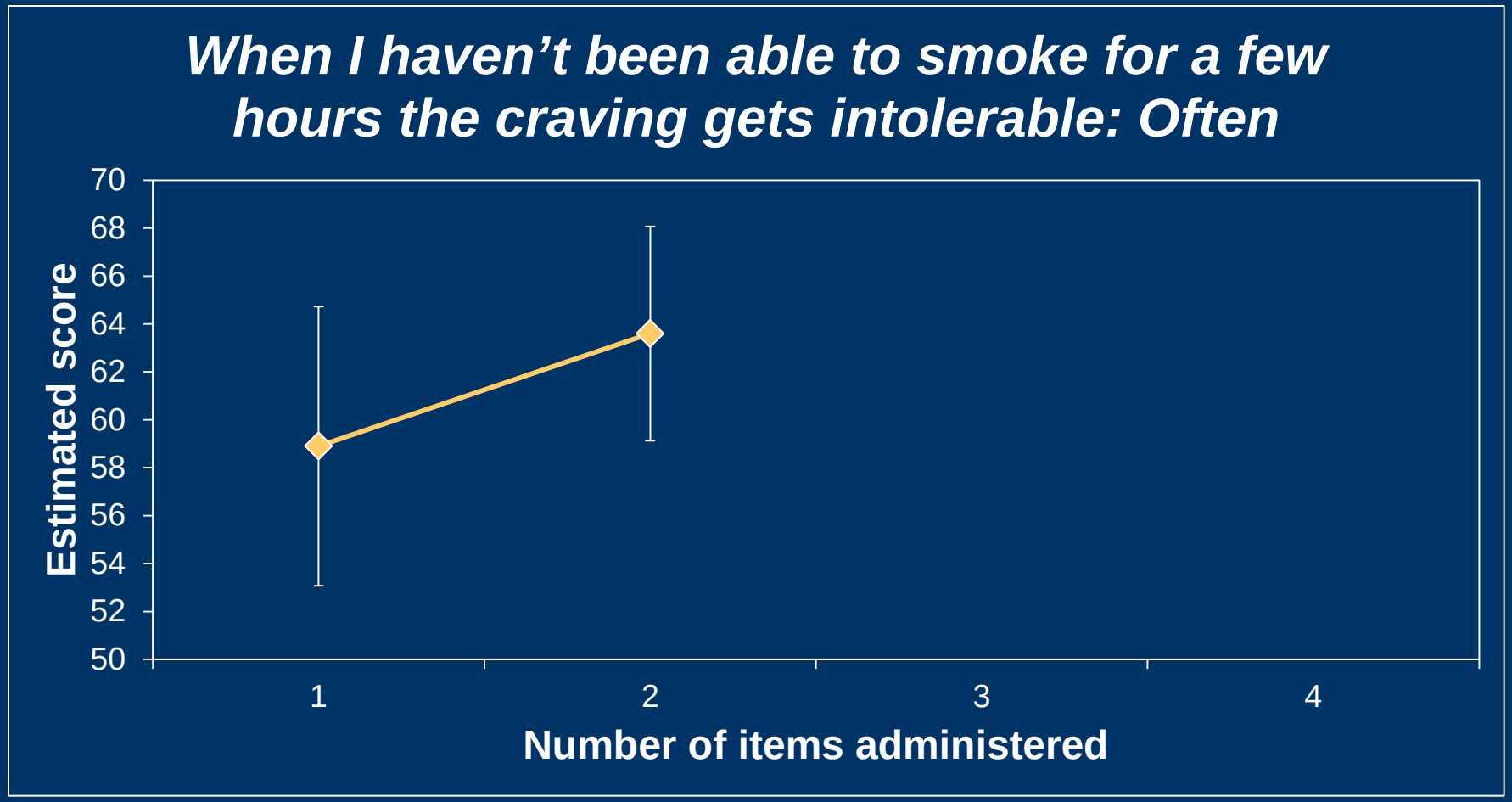
CAT DEMO

Using the PROMIS Smoking Assessment Toolkit *Nicotine Dependence* item bank

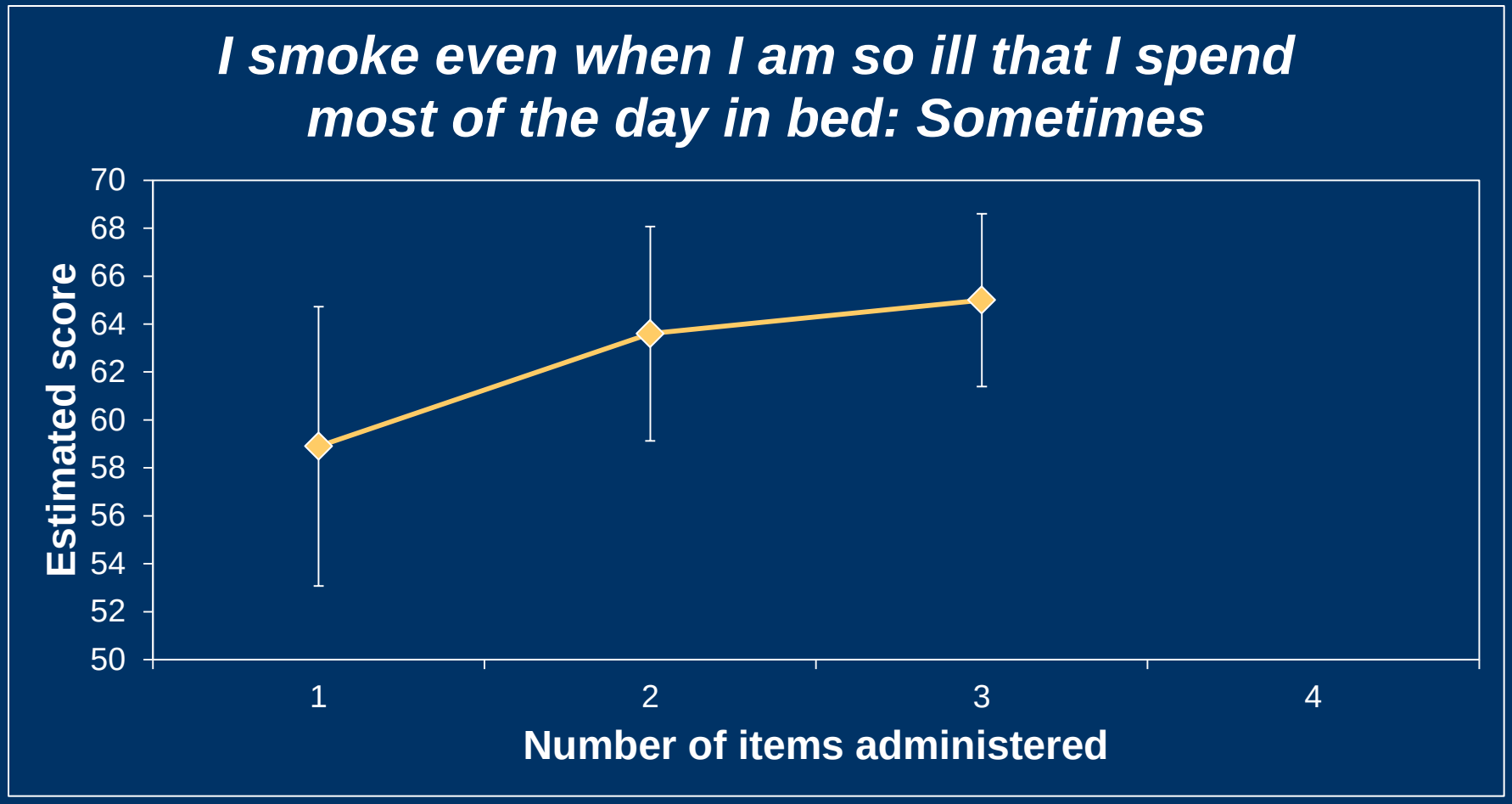
First item: Score = 58.9, SE = 5.8, rel=.66



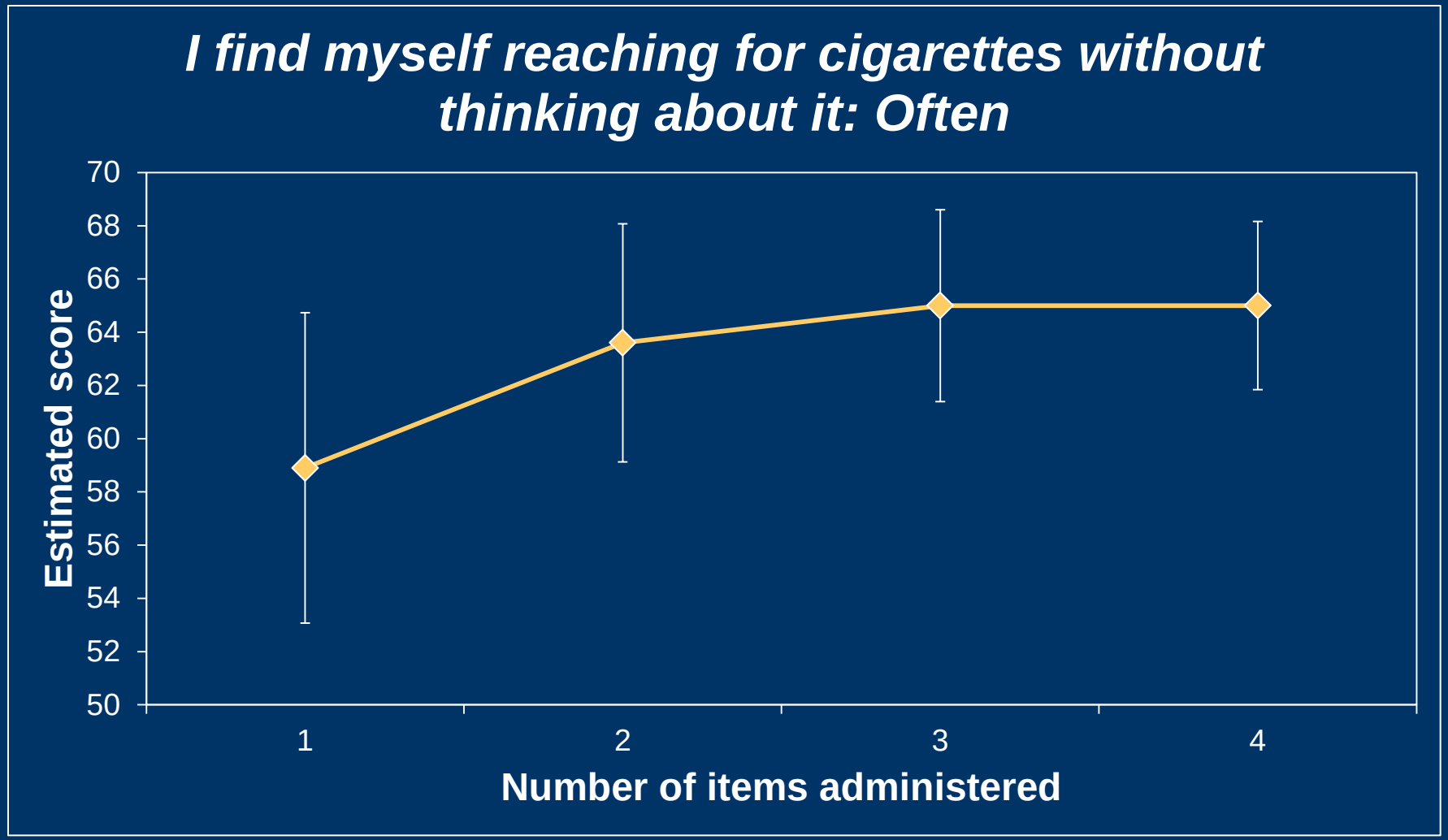
Second item: score = 63.6, SE = 4.5, rel=.80



Third item: score = 65.0, SE = 3.6, rel=.87



Fourth item: Score = 65.0, SE = 3.2, rel=.90



CAT administration produced highly reliable score estimate with just four items

Item	Response	Score	Reliability
I frequently crave cigarettes.	Quite a bit	58.9	0.66
When I haven't been able to smoke for a few hours, the craving gets intolerable.	Often	63.6	0.80
I smoke even when I am so ill that I spend most of the day in bed.	Sometimes	65.0	0.87
I find myself reaching for cigarettes without thinking about it.	Often	65.0	0.90

Use of computer-based assessment opens up a wide variety of assessment platforms

- Laptop “kiosk” for screening or survey completion in waiting room area
- Ipad or tablet that respondent can carry throughout visit
- Smartphone that respondent can take home
  - Scheduled prompts with links to survey
- Email reminders with links to survey
- Social media (e.g., Facebook and Twitter) for reminders, links and study updates to keep respondents involved

The platform can be chosen to maximize response rates and improve data quality

- **Make it easy and convenient for participants to get screened or respond to survey items**
- **Can incorporate voice support to ensure comprehension**
- **Can also collect additional useful data**
  - **Time to complete**
  - **GPS data**

Thank you

QUESTIONS?