Building and Revising Patient-Reported Outcomes Measures: Evaluating Item and Scale Functioning with the IRT Model

Bryce B. Reeve, Ph.D.

Outcomes Research Branch

Applied Research Program

Division of Cancer Control and Population Sciences

National Cancer Institute

Overview of Both Presentations

- **Illustrations of evaluating and adapting Outcomes measures based on IRT model results:**
  - **Example 1: Applying IRT modeling for evaluating and revising the Fear of Recurrence Scale. [Bryce]**
    - **Methods to evaluate item properties**
    - **Methods to evaluate scale properties**
    - **Revising questionnaires**
  - **Example 2: Creating a short screener for PTSD [Maria]**
  - **Example 3: Evaluating IRT model fit in the 16-item Substance Problems Index [Maria]**
- **Special Issues for using IRT modeling techniques.**
  - **Evaluating Model Assumptions [Maria]**
  - **Validity [Bryce]**
  - **Model Choice [Bryce]**
  - **Sample Size [Bryce]**
- **Final Comments and Q&A Session [Bryce, Maria, and Audience]**

Slide prepared by Bryce Reeve, Ph.D.

Fear of Recurrence (FOR) Scale

- **Fear of Recurrence: The degree of concern reported by patients about the chances of cancer returning at a future time.**
  - **Northouse, L.L. (1981). Mastectomy patients and the fear of cancer recurrence. *Cancer Nursing*, 4(3), 213-220.**
- **22 items**
- **5-Point Likert-type Response scale:**

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

- **Items were scored so that higher numbers reflect higher fear of disease recurrence.**

Thank you to Dr. Laurel Northouse, University of Michigan, for sharing her FOR scale and expertise for this project. Slide prepared by Bryce Reeve, Ph.D.

Psychometric Study is in Progress

- **Study Purpose: Evaluate questionnaire properties and develop a shortened scale.**
- **The purpose of this presentation is to illustrate the tools of IRT modeling and not a full psychometric evaluation of the FOR scale.**
- **123 Cancer survivors approximately 1-5 years after treatment ended.**
  - **Mellon, S., Northouse, L.L. (2001). Family survivorship and quality of life following a cancer diagnosis. *Research in Nursing & Health, 24*, 446-459.**
- **IRT model assumptions were evaluated.**

Thank you to Dr. Suzanne Mellon, University of Detroit – Mercy, for sharing her data and expertise for this project.

Slide prepared by Bryce Reeve, Ph.D.

Evaluating Item Properties

Slide prepared by Bryce Reeve, Ph.D.

Traditional (or Classical) Measures

- **Mean item scores provide an estimate of the difficulty or severity of the item.**
  - **Scoring items from 1-5, range 2.10 – 3.73:**
    - **Highest (3.73): "I would like to feel more certain about my health." Most people marked "agree" or "strongly agree" [reverse scored]**
    - **Lowest (2.10): "I feel optimistic as I focus on my future." Most indicated "agree" or "strongly agree"**
- **Item – total score correlations provide an estimate of the strength of relationship (discrimination ability) between the item and overall construct.**
  - **Ranged from: .27 - .71:**
    - **Highest: "I am bothered by the uncertainty of my health"**
    - **Lowest "I feel optimistic as I focus on my future."**
- **Coefficient alpha with item removed indicates how internal consistency changes when the item is removed from the scale.**
  - **Overall coefficient alpha (α) = .92.**
    - **Only 4 items decreased α to .91.**
- **Response category frequencies indicate over or under-utilized options used by respondents.**
  - **The "strongly agree" and "strongly disagree" responses received the fewest responses.**

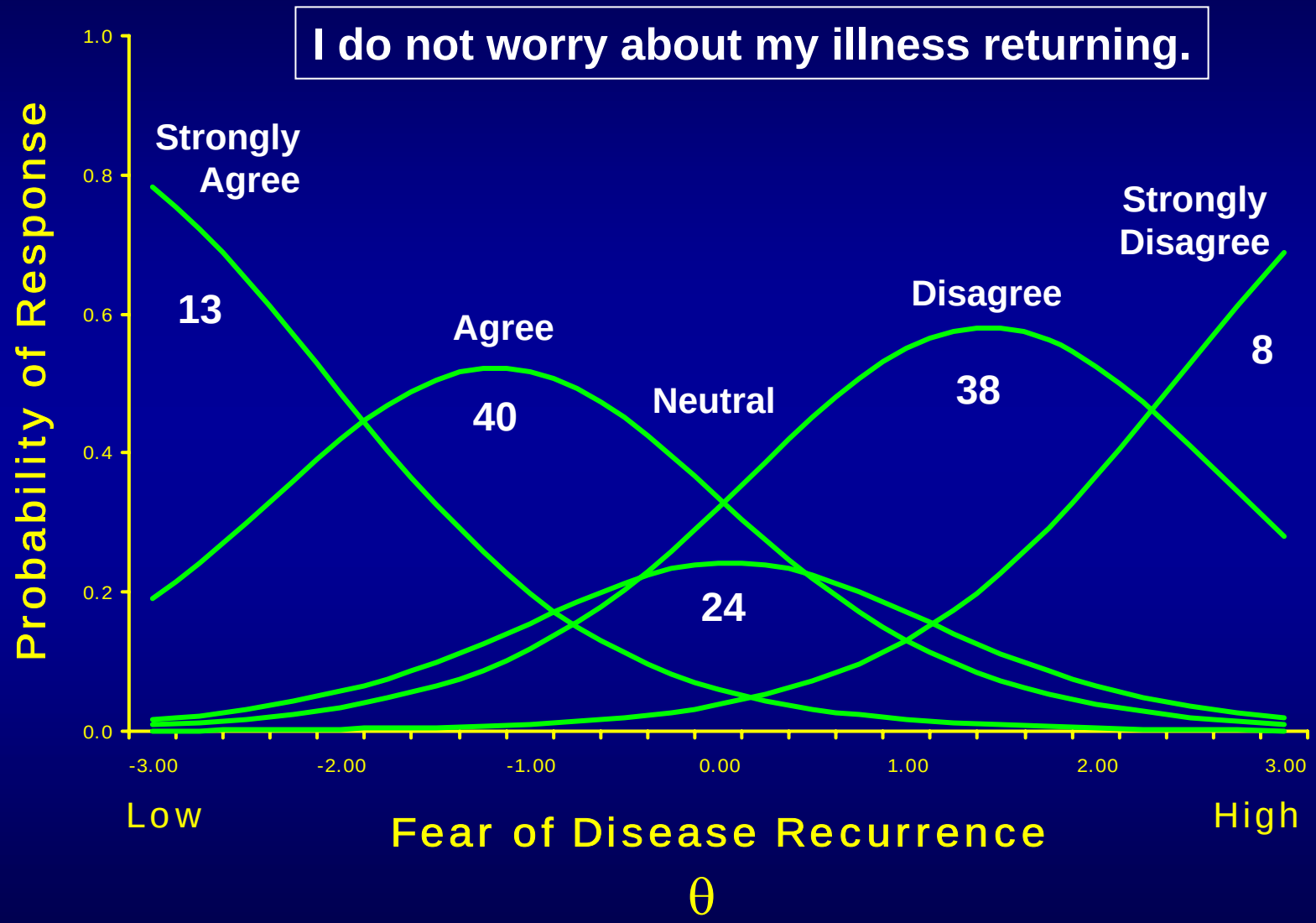
Slide prepared by Bryce Reeve, Ph.D.

Evaluating Item Properties with the IRT Model

- **Item Characteristic Curves (ICC)**
  - **ICCs model, in probabilistic terms, the relationship between a person's response to a question and his or her level on the construct (θ) being measured by the scale.**
    - **The steepness of the curves are defined by the discrimination power (*a* parameter) of the item.**
    - **Intersections between category curves are defined by the location/difficulty (*b* parameter) of the item.**
- **What can ICCs tell us about each item?**
  - **What is the appropriate number of response categories?**
  - **Which level of fear must a person have to endorse each response category?**

Slide prepared by Bryce Reeve, Ph.D.

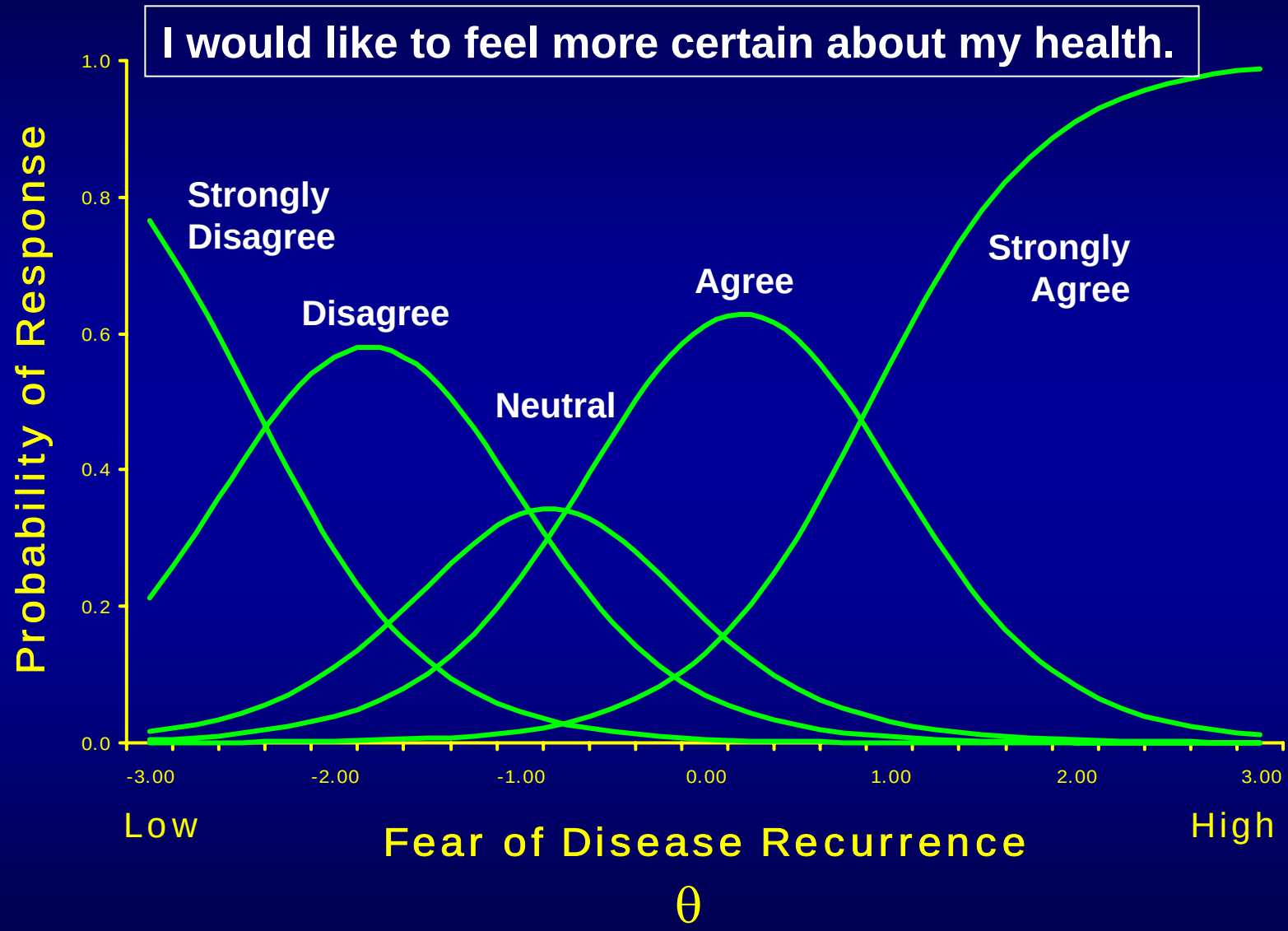
Item Characteristic Curves



***Parameter estimates from Samejima's Graded Response Model using MULTILOG software**

Slide prepared by Bryce Reeve, Ph.D.

Item Characteristic Curves



***Parameter estimates from Samejima's Graded Response Model using MULTILOG software**

Slide prepared by Bryce Reeve, Ph.D.

Item Characteristic Curves

Compared to other persons who have this illness, I feel that I worry less than they do about health concerns.

***Parameter estimates from Samejima's Graded Response Model using MULTILOG software**

Slide prepared by Bryce Reeve, Ph.D.

Evaluating Item Properties with the IRT Model

- **Item Information Curves (or Functions)**
  - **Information curves indicate the range over θ where an item is best at discriminating among individuals. Higher information denotes more precision (or reliability) for measuring a person's trait level.**
    - **The height of the curves (denoting more information) are defined by the discrimination power (*a* parameter) of the item.**
    - **Location of the information curves is determined by the threshold (*b*) parameter(s) of the item.**
- **What can item information curves tell us?**
  - **Which item(s) is most useful for measuring different fear levels?**
  - **What set of items are best for a short tailored survey?**
  - **When we have redundant items, which one is more informative?**

Slide prepared by Bryce Reeve, Ph.D.

Item Information Curves

***Parameter estimates from Samejima's Graded Response Model using MULTILOG software**

Slide prepared by Bryce Reeve, Ph.D.

Evaluating Scale Properties

Internal Assessment

Slide prepared by Bryce Reeve, Ph.D.

Traditional (or Classical) Measures

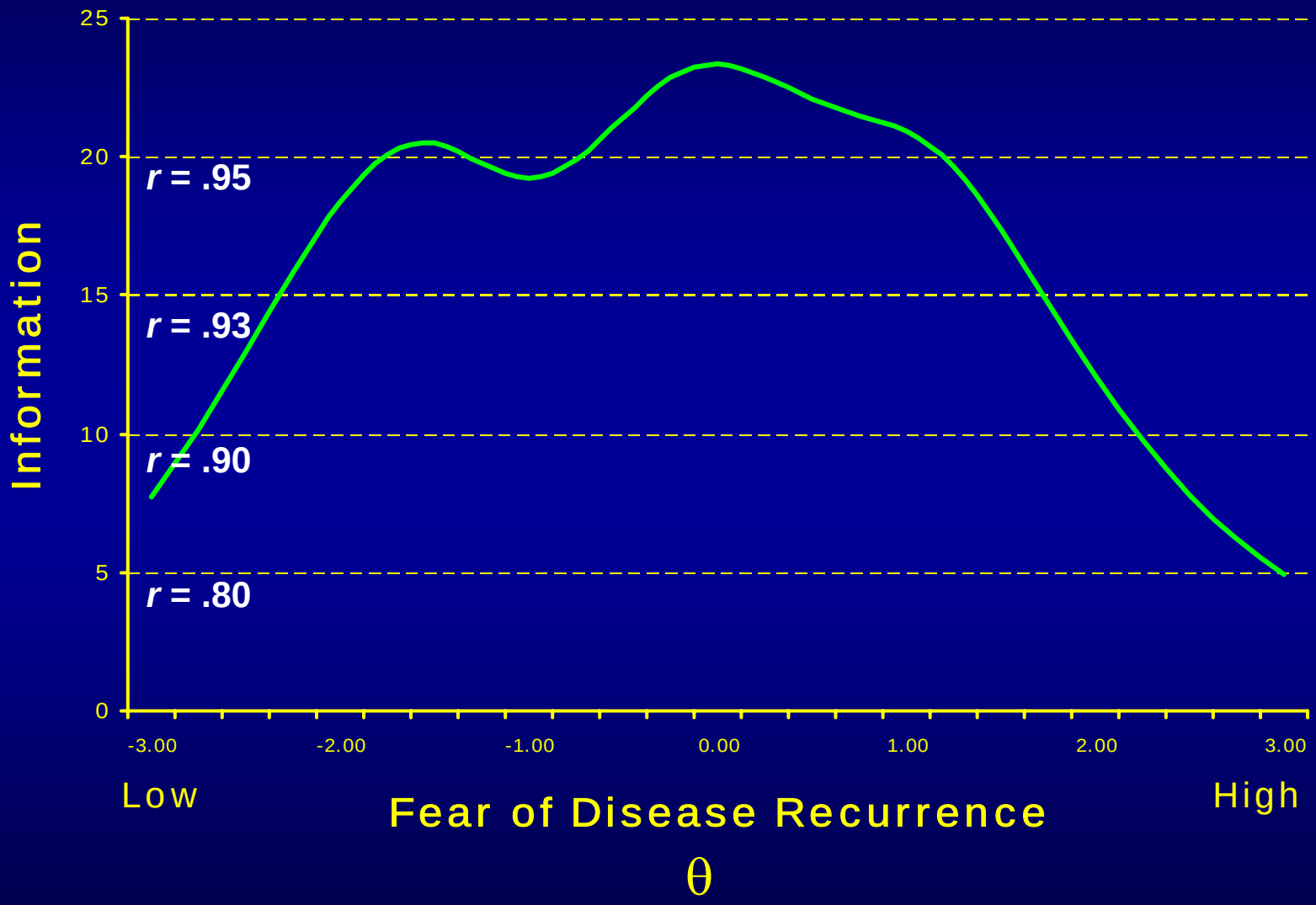
- **Reliability**
  - **Internal Consistency**
    - **Cronbach's Coefficient Alpha (α)**
- **FOR Scale: 22 items α = .92**
- **Question: Is the FOR scale reliable to measure an individual's fear score no matter what level of fear they may have?**

Evaluating Scale Properties with the IRT Model

- **Scale (or Test) Information Curve (or Function)**
  - **The scale information curve indicates the range over θ where a scale is best at discriminating among individuals. Higher information denotes more precision (or reliability) for measuring a person's trait level.**
    - **Sum of the item information curves.**
- **What can the scale information curve tell us?**
  - **How reliable is my scale for measuring different levels of a person's fear?**
  - **Where are there measurement gaps along my construct continuum?**

Slide prepared by Bryce Reeve, Ph.D.

FOR Scale Information Curve



***r* = approximate reliability**

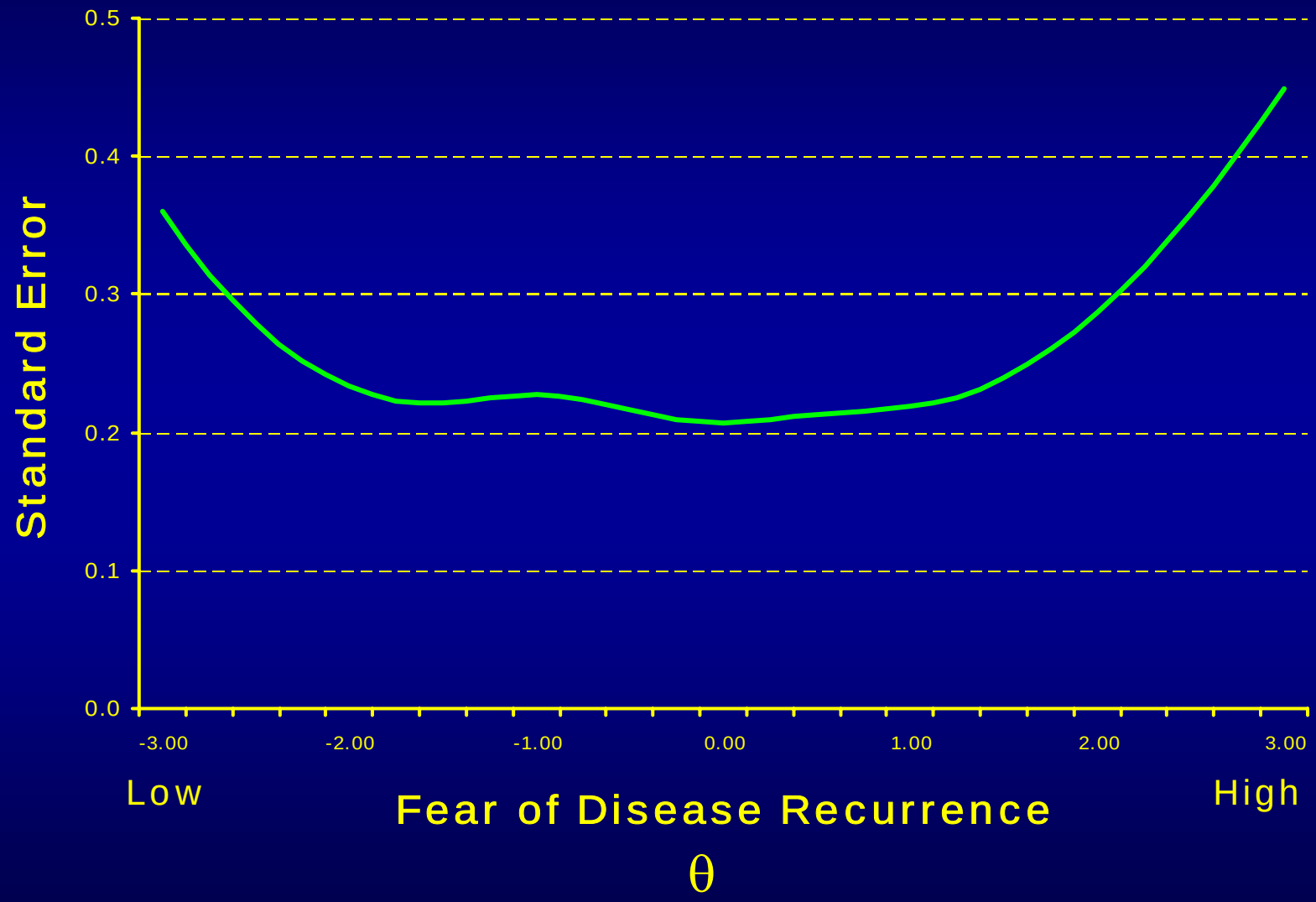
Slide prepared by Bryce Reeve, Ph.D.

Evaluating Scale Properties with the IRT Model

- **Standard Error of Measurement Curve (or Function)**
  - **The SEM curve describes an expected observed score fluctuation due to error in the measurement tool. Standard deviation of error about an estimated score**
  - **Inverse Square Root of Information.**
- **What can the SEM curve tell us?**
  - **How reliable is my scale for measuring different levels of a person's fear?**
  - **Where are there measurement gaps along my construct continuum?**

Slide prepared by Bryce Reeve, Ph.D.

FOR SEM Curve



Slide prepared by Bryce Reeve, Ph.D.

Revising Questionnaires

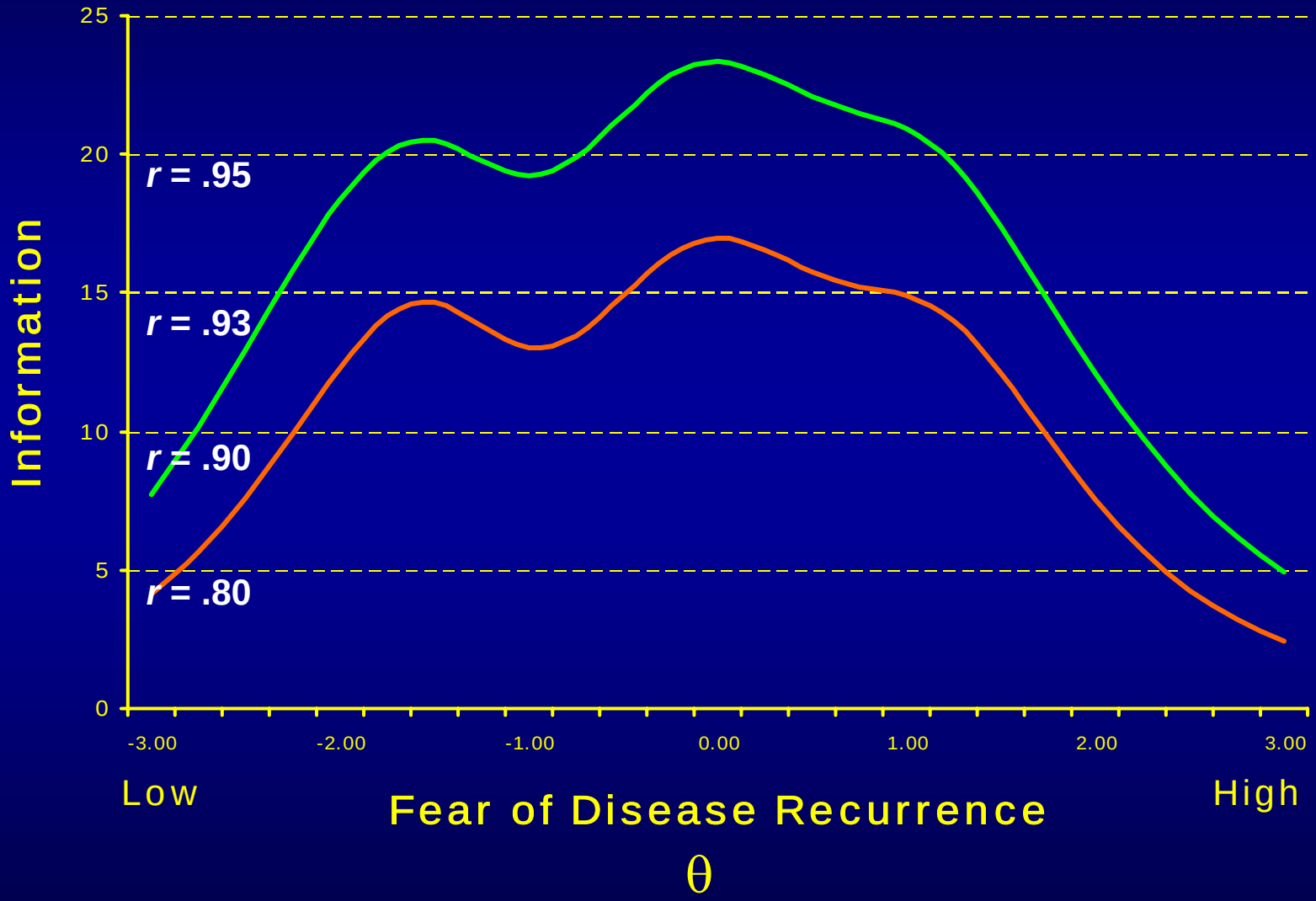
Slide prepared by Bryce Reeve, Ph.D.

How can IRT models inform questionnaire revisions?

- **Item Characteristic Curves can help us modify the response scale.**
- **Item information curves allow us to:**
  - **identify highly informative (reliable) items and weed out poor performing items.**
  - **Develop tailored instruments**
    - **Shortened version targeted to the study population**
    - **Screening or diagnostic tool**
      - **Maximize information at cut-off point(s).**
- **Scale Information Curves (and SEM curves) can help us determine the effect of removing an item or subset of items.**

Slide prepared by Bryce Reeve, Ph.D.

What is the reduction in information going from a 22 to 12 item scale?



***r* = approximate reliability**

Slide prepared by Bryce Reeve, Ph.D.

Building and Revising Outcomes Measures: Evaluating Item and Scale Functioning with IRT

Maria Orlando
RAND Corp.

Slide prepared by Maria Orlando, Ph.D.

Outline

- Example 1: Creating a short screener for PTSD
- Example 2: Evaluating IRT model fit in the 16-item Substance Problems Index
- Special issues for using IRT modeling techniques

Example 1 - Background

- School-based intervention program for victims of trauma
- Need to assess children for eligibility in program
  - Current assessment is too long
- Goal: Shorten 17-item measure of PTSD with minimum sacrifice to screening precision and scale integrity

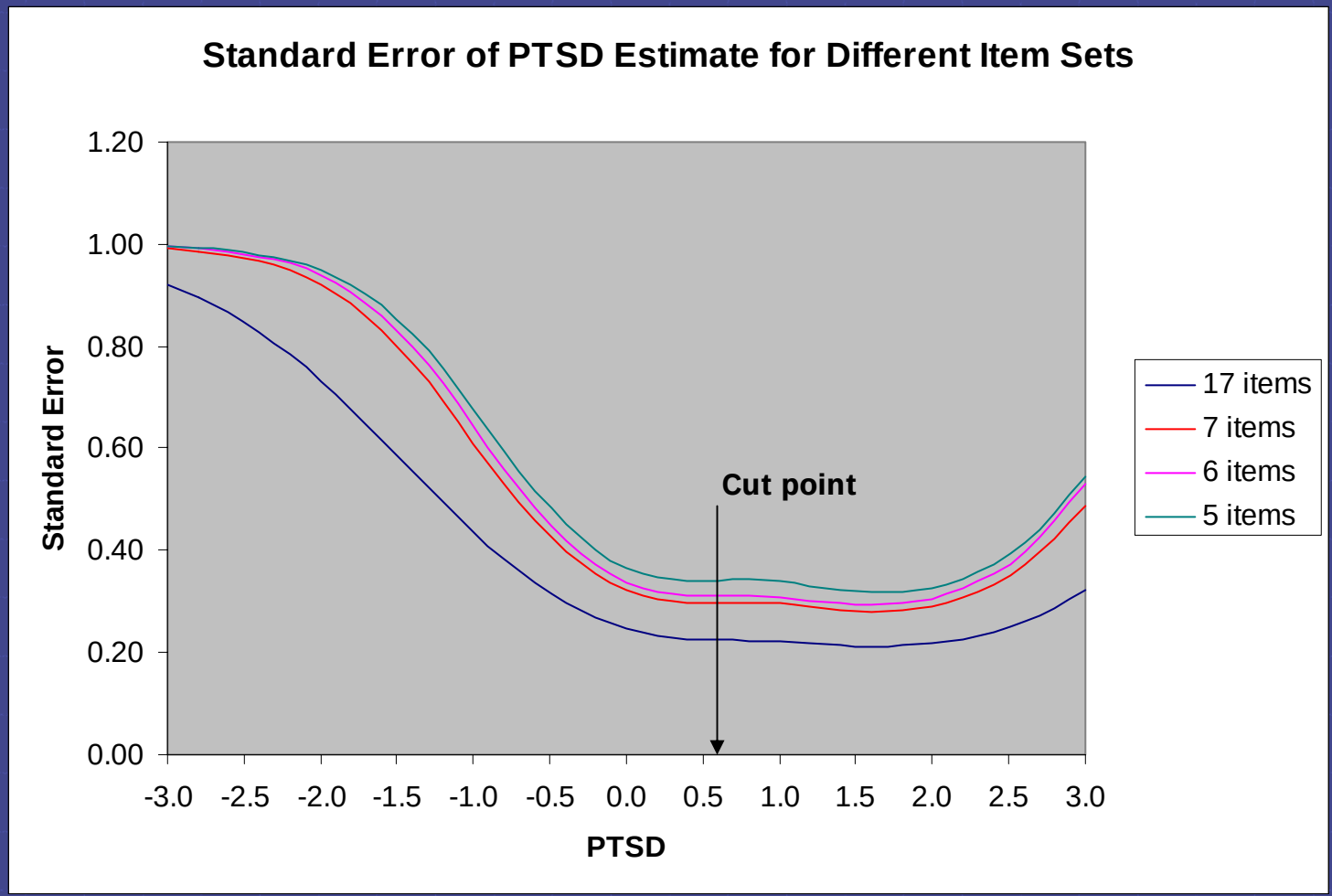
Example 1 – Method

- Sample
  - N=769 6th grade students from LAUSD
- Analytic approach
  - Calibrate 17 items with graded IRT model
  - Use results of calibration (item parameters, information) to choose candidate short scales
  - Evaluate performance (sensitivity, specificity) of candidate short scales

Example 1 – Item calibration

Item Content	a	b_1	b_2	b_3
The trauma came into head when you didn't want	2.53	0.16	1.50	1.93
Having bad dreams or nightmares	1.47	-0.82	0.99	1.76
Acting or feeling the trauma was happening again	2.52	0.38	1.48	2.03
Upset when think or hear about the trauma	2.47	0.06	1.34	1.94
Having feelings when think or hear about trauma	2.24	0.39	1.62	2.33
Trying not to think, talk and feel about trauma	2.44	0.30	1.24	1.63
Avoid activities, people,or place reminding of the trauma	2.23	0.66	1.58	2.06
Unable to remember an important part of the trauma	2.11	0.77	1.96	2.61
Less interest or not doing things used to do	1.81	0.63	1.90	2.62
Not feeling close to people around you	1.22	0.78	2.11	2.76
Unable to have strong feelings	1.92	0.51	1.67	2.26
Feeling your plans and hopes will not come true	1.57	0.47	1.55	2.16
Having trouble in sleep	1.90	0.23	1.36	1.94
Feeling irritable or having fits of anger	2.49	0.35	1.46	2.19
Having trouble concentrating	1.79	-0.02	1.46	2.25
Being overly careful	1.58	0.11	1.35	2.00
Being jumpy or easily startled	1.93	0.25	1.61	2.34

Example 1 – Choosing candidate item subsets



Slide prepared by Maria Orlando, Ph.D.

Example 1 – Evaluating candidate item subsets

	AUC	Sens.	Spec.
7 items	.985	97.0	88.0
6 items	.979	97.5	84.2
5 items	.974	95.5	88.7

Example 1 – Evaluating candidate item subsets

	Sample 1			Sample 2		
	AUC	Sens.	Spec.	AUC	Sens.	Spec.
7 items	.985	97.0	88.0	.897	95.1	84.4
6 items	.979	97.5	84.2	.890	92.1	82.1
5 items	.974	95.5	88.7	.874	88.1	86.7

Example 1 – Final short item set

Item Content	a	b_1	b_2	b_3
The trauma came into head when you didn't want	2.53	0.16	1.50	1.93
Acting or feeling the trauma was happening again	2.52	0.38	1.48	2.03
Upset when think or hear about the trauma	2.47	0.06	1.34	1.94
Trying not to think, talk and feel about trauma	2.44	0.30	1.24	1.63
Avoid activities, people,or place reminding the trauma	2.23	0.66	1.58	2.06
Feeling irritable or having fits of anger	2.49	0.35	1.46	2.19
Being jumpy or easily startled	1.93	0.25	1.61	2.34

Example 2 - Background

- 16-item Substance Problem Index (SPI) is very good indicator among adolescents
  - pre-treatment need
  - post-treatment success
- Want to examine properties of items and fit of IRT model

Example 2 - Method

- Sample
  - N=1,419 adolescents in residential and outpatient substance abuse treatment centers
- Analytic approach
  - Calibrate 16 items with 1PLM and 2PLM
  - Evaluate item fit with $S\text{-}X^2$
  - Create diagnostic plots to support fit index results

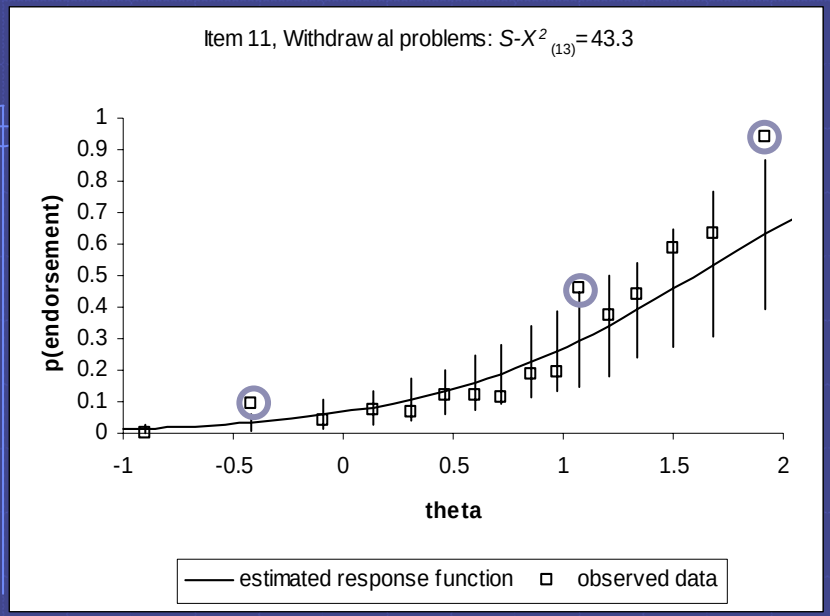
Example 2 – 1PLM v 2PLM

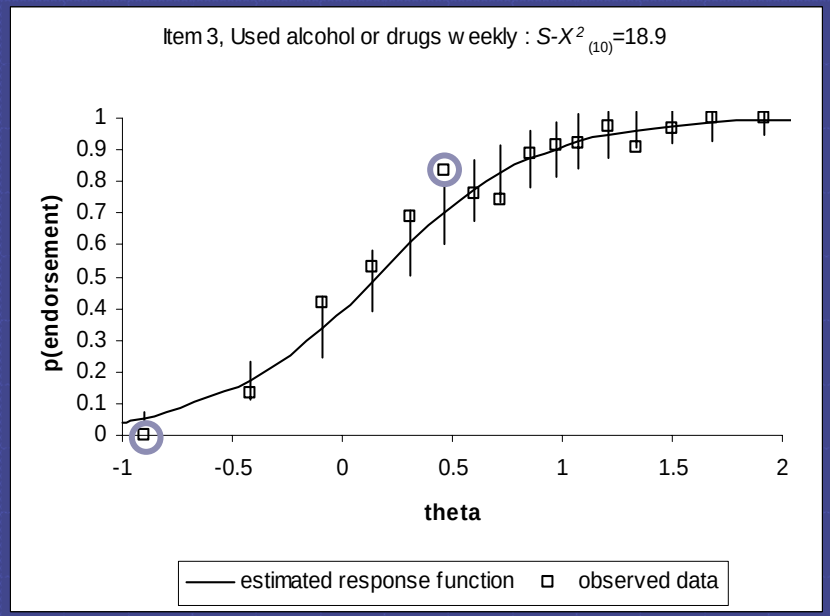
Model	–2*loglikelihood
1PLM	4380.4
2PLM	4080.5
Difference (15 df)	299.9

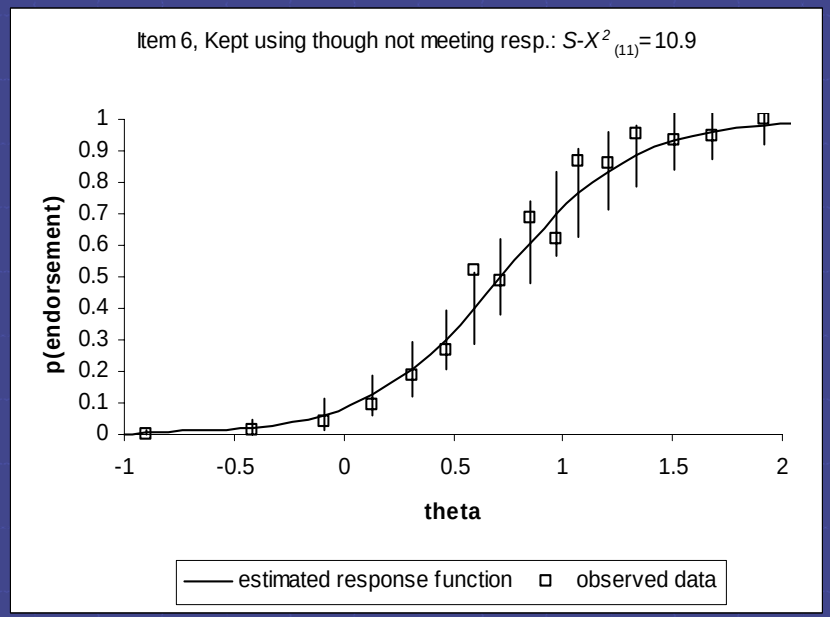
Example 2 – 2PLM Calibration and item fit of SPI items

Item content	a	b	$S\text{-}X^2$	df	*p*
Tried to hide when using	1.51	0.82	22.0	13	0.055
Others complained about your use	1.69	0.20	16.6	12	0.165
Used alcohol or drugs weekly	2.74	0.15	18.9	10	0.042
Use caused you to feel depressed, nervous etc.	2.69	1.01	11.5	12	0.487
Use caused numbness, tingling, blackouts etc.	2.09	1.52	21.4	13	0.065
Kept using even though not meeting responsibilities	3.33	0.72	10.9	11	0.452
Used in unsafe situations (e.g., driving a car)	2.15	1.21	19.6	13	0.106
Use caused you to have problems with the law	1.22	1.15	14.7	13	0.326
Kept using even though getting into fights, legal trouble	2.13	0.46	19.2	12	0.084
Needed more to get same high	2.71	1.02	15.3	12	0.225
Had withdrawal problems or used to avoid withdrawal	1.66	1.60	43.3	13	0.000
Used in larger amounts or more often than meant to	2.86	0.84	16.1	12	0.187
Unable to cut down or stop using	2.31	1.09	8.8	12	0.722
Spent a lot of time getting or feeling effects of drugs	2.92	0.51	17.9	11	0.084
Use caused you to give up, have problems with activities	3.54	1.02	18.2	11	0.077
Kept using even though adding to med/psych problems	3.13	1.02	7.2	12	0.847

Diagnostic item-fit plots of selected SPI items







Slide prepared by Maria Orlando, Ph.D.

Special issues for using IRT models to evaluate questionnaire properties

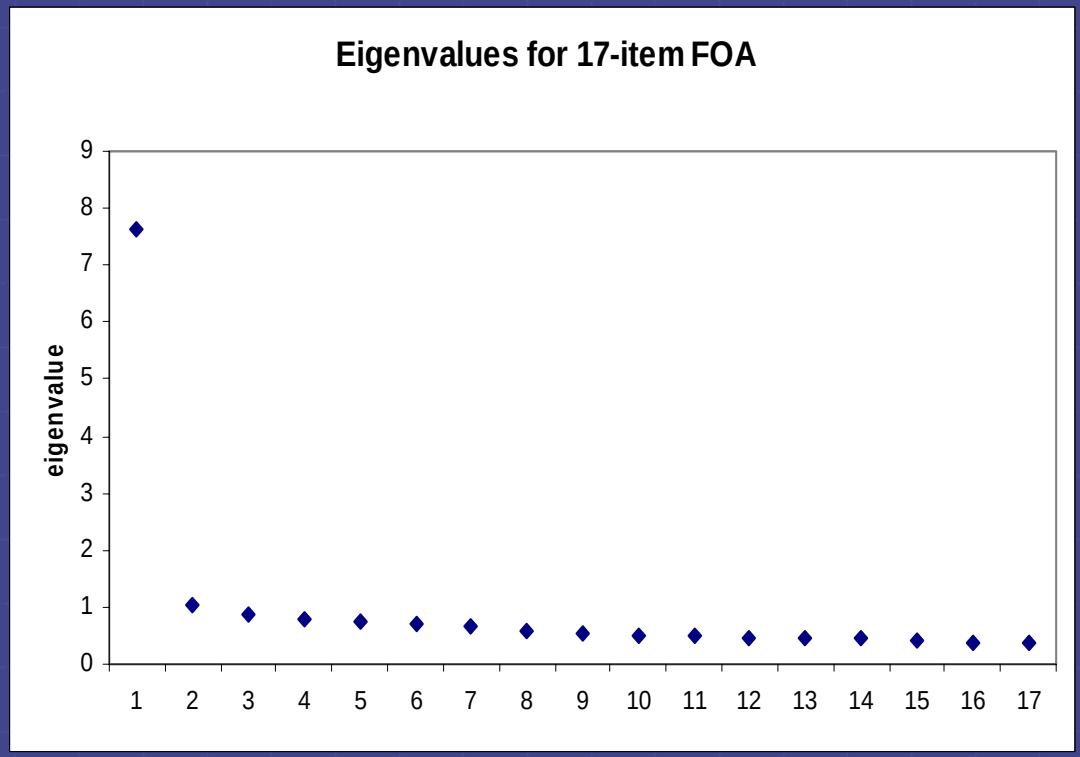
1. Evaluating model assumptions
2. Validity
3. Model choice
4. Sample size

Slide prepared by Maria Orlando, Ph.D.

Special issues: Evaluating model assumptions

- Model is appropriate for data
  - Model choice
  - Evaluating model fit
- Unidimensionality
  - covariance among the items can be explained by a single underlying factor
- Local independence
  - No excess covariation among subsets of items

Evaluating dimensionality



Loadings on first principal component

0.733
0.611
0.727
0.735
0.705
0.720
0.686
0.638
0.638
0.467
0.657
0.618
0.693
0.743
0.653
0.633
0.672

Slide prepared by Maria Orlando, Ph.D.

Local dependence

- Can arise among subsets of items with a common stem, or similar content, or items that are presented sequentially
- Detection
  - IRTNEW for dichotomous items
  - Examine residuals of CFA
  - Examine output from IRT calibration
- Solution
  - Omit one of the pair
  - Create a ‘testlet’

Special Issues: Validity

- **Validity Issues**
  - **Just because you used a sophisticated (statistical) modeling technique to evaluate and revise your questionnaire, it does not excuse you from evaluating the many important aspects of validity assessment**
  - **IRT modeling can help with validity issues.**
    - **Carefully addressing the assumptions of the IRT model and selecting items based on their content and properties will result in a carefully constructed and valid scale.**
    - **Evaluating measurement equivalence (i.e. DIF testing) when an instrument is adapted in a new language or a new population.**
    - **Building Block Design for scale construction.**

Special Issues: Model Choice

- **There are over a 100 models:**
  - **Parameteric and non-parametric models**
  - **Dichotomous and polytomous response categories**
  - **Unidimensional and multi-dimensional models**
  - **Differ by number of parameters they estimate**
- **Thus, choice of IRT model can be difficult!!**
- **There are a smaller subset of models that have found application in health outcomes research.**

Slide prepared by Bryce Reeve, Ph.D.

IRT Models You May See in Outcomes Research

Model	Item Response Format	Model Characteristics
Rasch / 1-Parameter Logistic	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
2-Parameter Logistic	Dichotomous	Discrimination and threshold parameters vary across items.
Graded Response	Polytomous	Ordered responses. Discrimination varies across items.
Nominal	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit (Rasch Model)	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale (Rasch Model)	Polytomous	Discrimination equal across items. Item threshold steps equal across items.
Generalized Partial Credit	Polytomous	Variation of Partial Credit Model with discrimination varying among items.

Slide prepared by Bryce Reeve, Ph.D.

Special Issues: Sample Size

- **The more the better!**
- **Rule of thumb?**
- **Well, it depends…..**
  - **Choice of model**
    - **More parameters to estimate require larger sample sizes**
  - **Number of response categories?**
    - **More categories means more parameters to estimate.**
  - **Purpose of study**
    - **Are you evaluating scale properties or calibrating items for an item bank?**
  - **Sampling distribution of the respondents**
    - **Prefer a spread of respondents over continuum.**
  - **How well does the data meet the assumptions of the model?**
  - **Relationship between the items and the underlying construct**
    - **Poor functioning items may require larger sample sizes.**

Slide prepared by Bryce Reeve, Ph.D.

Final Comments

- **IRT should be used in conjunction with CTT as a statistical tool to evaluate health outcomes data.**
- **It is important that a psychometrician work hand-in-hand with content experts throughout all phases of application from evaluating the assumption of unidimensionality to picking and choosing items.**

Slide prepared by Bryce Reeve, Ph.D.