Optimizing Qualitative and Quantitative Research: How to Make the Process More Efficient

Sixth Annual

Patient-Reported Outcome Consortium Workshop

April 29 - 30, 2015 ■ Silver Spring, MD

CRITICAL PATH INSTITUTE
a decade of excellence

10 YEARS

Disclaimer

The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies, the U.S. Food and Drug Administration, the Critical Path Institute, the PRO Consortium, or the ePRO Consortium.

These slides are the intellectual property of the individual presenters and are protected under the copyright laws of the United States of America and other countries. Used by permission. All rights reserved. All trademarks are the property of their respective owners.

Session Participants

Moderator

- J. Jason Lundy, PhD

Presenters and Panelists

- J. Jason Lundy, PhD – Principal, Outcometrix
- Stacie Hudgens, MA – Managing Partner, Clinical Outcomes Solutions
- R.J. Wirth, PhD – Managing Partner, Vector Psychometric Group, LLC
- Wen-Hung Chen, PhD – Reviewer, Study Endpoints, SEALD, OND, CDER, FDA

Session Outline

- Discuss opportunities to optimize the COA instrument development process
- Present approaches for maximizing the information available in a sample
  - Analyzing cognitive interviewing data quantitatively
  - Longitudinal IRT in small samples
- Comments and questions

Sample Optimization in Scale Development

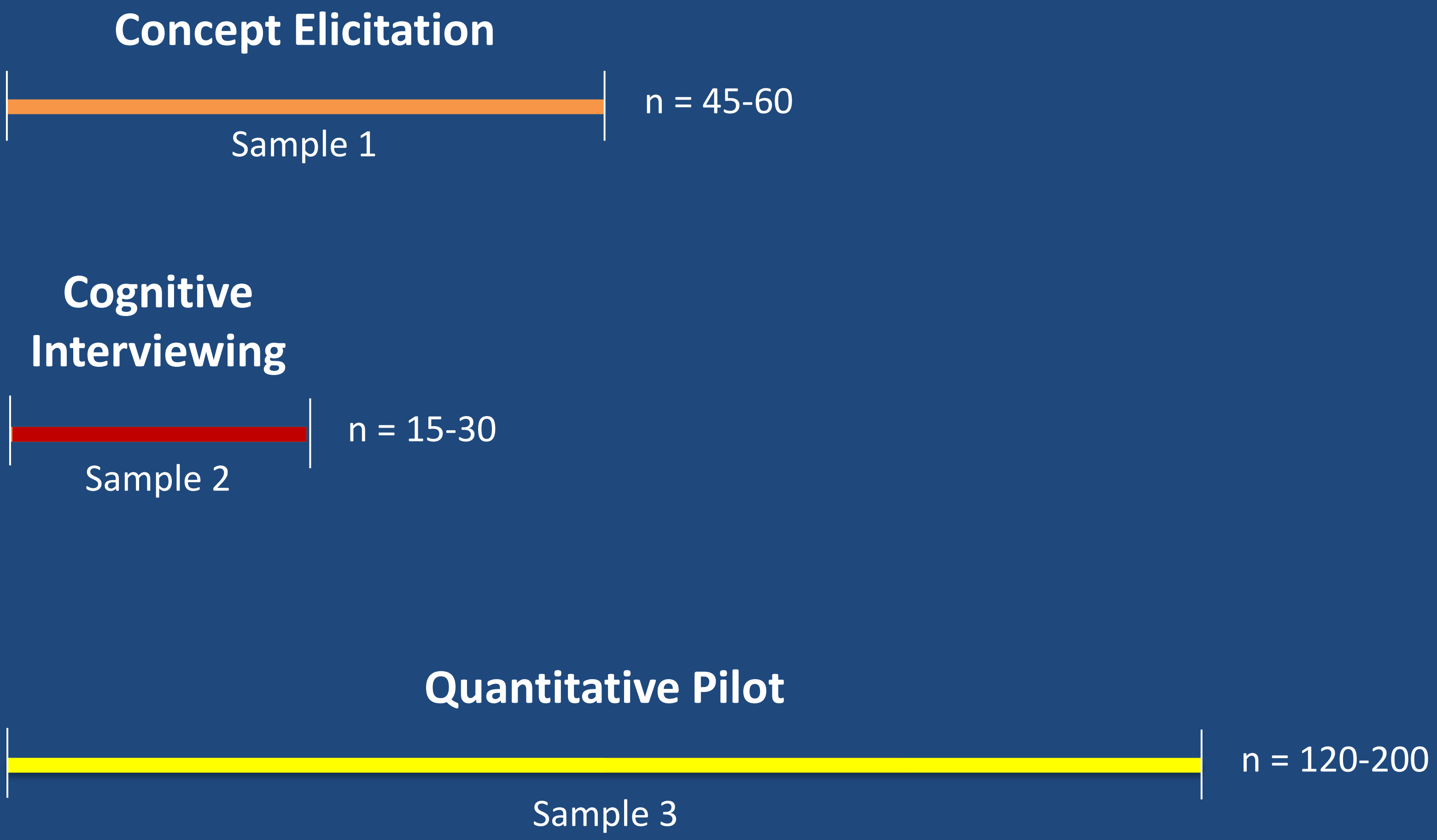
J. Jason Lundy, PhD
Outcometrix

Sixth Annual
Patient-Reported Outcome Consortium
Workshop

April 29 - 30, 2015 ■ Silver Spring, MD

Why change what works?

Traditional Instrument Development



Concept Elicitation

n = 45-60

Sample 1

Cognitive Interviewing

n = 15-30

Sample 2

Quantitative Pilot

n = 120-200

Sample 3

Why should we try to maximize information from a sample?

- Recruitment is expensive
  - Time and money
- If we collect additional information, we can:
  - Reduce overall sample size;
  - Reduce rounds of revision;
  - Have more confidence in the scale moving into the later rounds of testing;
  - Get to confirmatory testing faster.
    - Gain experience with the instrument in clinical trials

The problem with small sample qualitative interviews

- Goal of qualitative interviews are to elicit concepts and reduce sources of measurement error Qualitative interviews have their own sources of measurement error
  - Differences in how interviews are conducted may lead to variation in their quality and;
  - The way the interview data is interpreted may vary, so that a particular report may be taken to have different meanings
  - Threats to validity include confirmation bias and context bias
    - Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*, *73*(1), 32-55.

Qualitative Sample Size Determination

- Concept Elicitation - “No rule can be provided to determine either the sample size or number of iterations needed to reach saturation in PRO instrument development.”
  - Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. *Value in Health*, *14*(8), 967-977.

  *Post-hoc Bayesian approaches are available to predict the probability that saturation has been achieved – See Williams LA, Berger D, and Johnson VE. in the special issue on “Advances in Clinical Outcome Assessments” in *Therapeutic Innovation & Regulatory Science*, November 2015.

- Cognitive Interviews - “Although Willis has suggested that seven to 10 interviews are sufficient to confirm patient understandability of the item, the number of interviews needed is a function of the complexity of the instrument, the diversity of the population of interest, and the number of questionnaire iterations necessary to fully explore patient understanding of items.”
  - Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value in Health*, *14*(8), 978-988.

Cognitive Interview Sample Size

- There is a general lack of agreement on sample size for qualitative research, particularly for cognitive interviews
  - – The literature provides ranges from 5 to 75 subjects
    - Perneger, T. V., Courvoisier, D. S., Hudelson, P. M., & Gayet-Ageron, A. (2015). Sample size for pre-tests of questionnaires. *Quality of Life Research*, *24*(1), 147-151.
  - – However, the ability of an interview to detect a problem is bound by the laws of basic probability
    - If the sample size is too small, the probability that no participant will report any given problem can be large.
  - – Hence, sample sizes can be estimated for various values of problem prevalence, probability of detection, and power

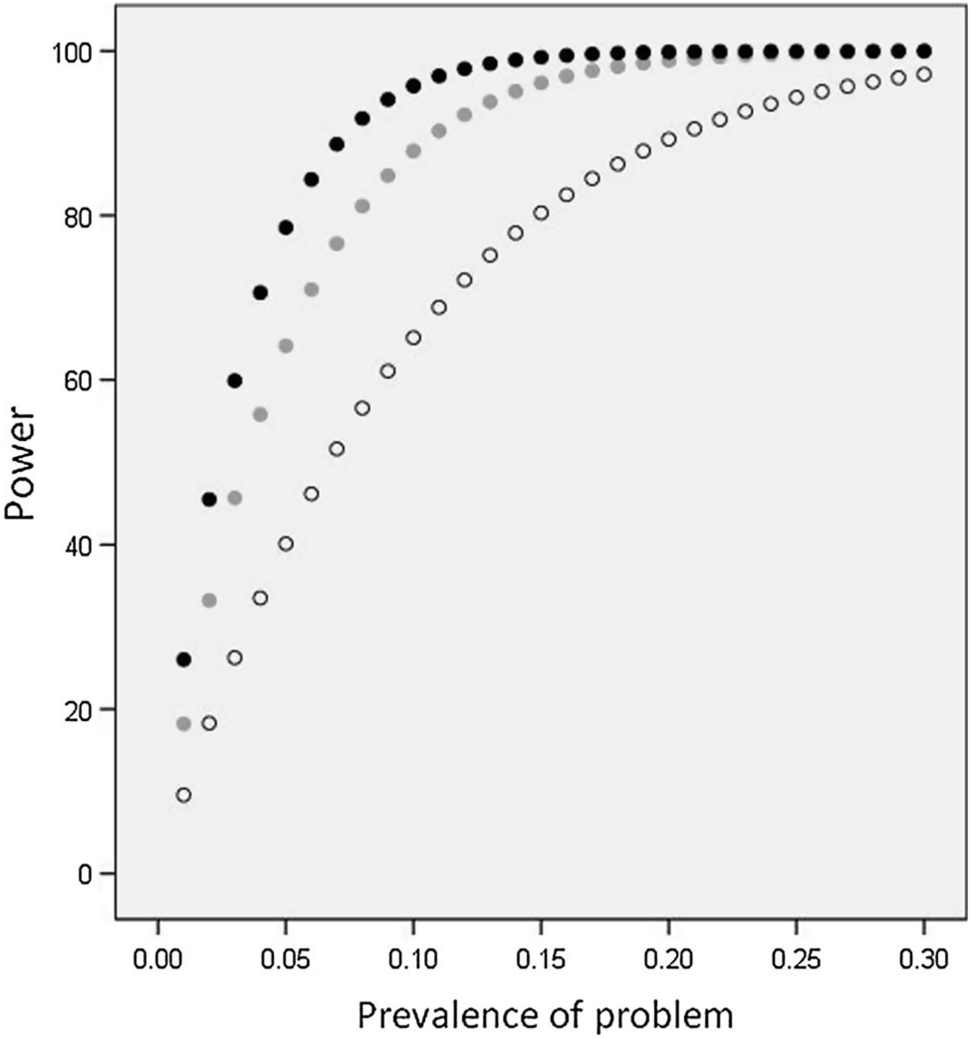
Cognitive Interview Sample Size

Sample size to detect at least one problem in an interview can be computed as:

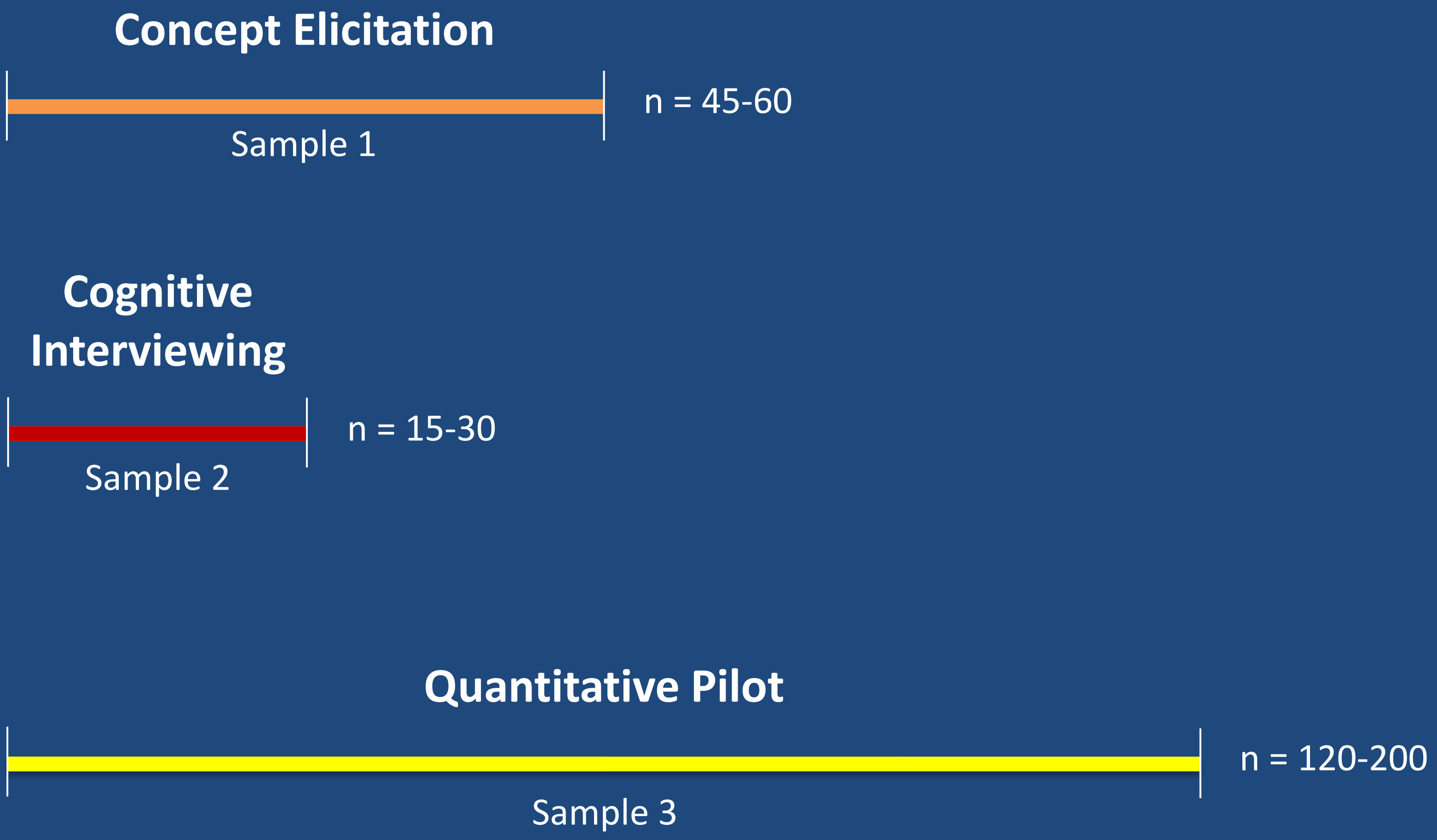
$$n = \ln(1\text{-power})/\ln(1\text{-}p)$$

“…sample sizes of 30 or more should be preferred for pre-tests whenever possible to achieve a reasonable power to detect fairly prevalent problems.”

Perneger, T. V., Courvoisier, D. S., Hudelson, P. M., & Gayet-Ageron, A. (2015). Sample size for pre-tests of questionnaires. *Quality of Life Research*, *24*(1), 147-151.



Traditional Approach



Concept Elicitation

n = 45-60

Sample 1

Cognitive Interviewing

n = 15-30

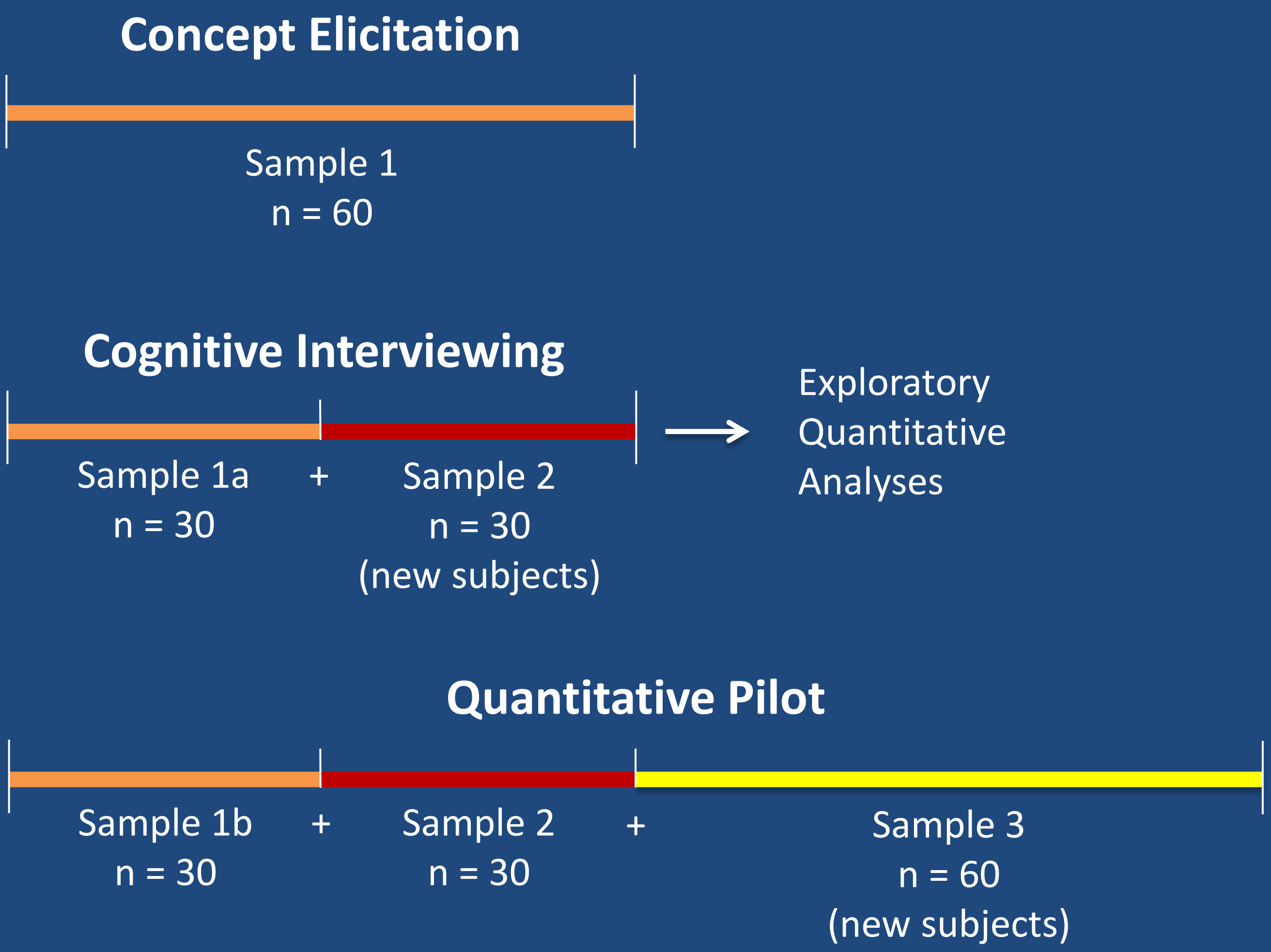
Sample 2

Quantitative Pilot

n = 120-200

Sample 3

Reduce, Reuse, Recycle Example



Concept Elicitation

Sample 1
n = 60

Cognitive Interviewing

Sample 1a
n = 30

+

Sample 2
n = 30
(new subjects)

→ Exploratory
Quantitative
Analyses

Quantitative Pilot

Sample 1b
n = 30

+

Sample 2
n = 30

+

Sample 3
n = 60
(new subjects)

For Illustration Purposes Only

Key Considerations

- Your mileage may vary
  - Sample sizes will need to be determined for each study independently
  - The number of new subjects should equal the number of reused subjects, at a minimum
- For cognitive interviews:
  - Initial round can use prior concept elicitation sample to confirm item conceptualization
  - Subsequent interviews in a new sample to reconfirm concept and subject understanding
  - Descriptive statistics and CTT to describe item performance

Conclusions

- Sample sizes for qualitative research can be computed
- Increasing sample sizes during cognitive interviews, *while lowering overall sample size through reuse*, provides better problem detection and the opportunity for a *real* mixed methods approach
- Reusing the sample(s) can provide additional information, and enhance subject engagement

Collecting and Analyzing Quantitative Data from Cognitive Interviews

Stacie Hudgens, MA
Clinical Outcomes Solutions

Sixth Annual
Patient-Reported Outcome Consortium
Workshop

April 29 - 30, 2015 ■ Silver Spring, MD

Why consider mixed methods approach?

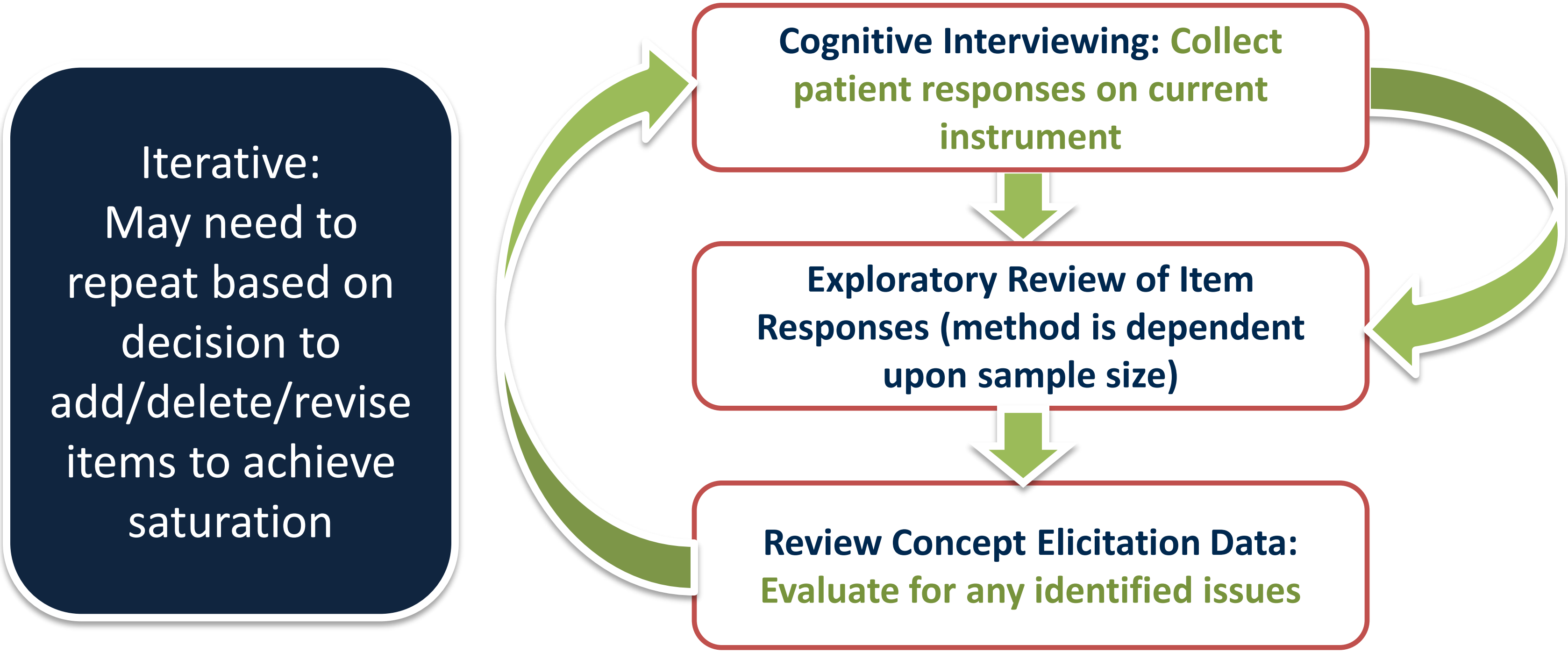
- Useful for optimizing the data collected from a limited sample
  - Recruitment is expensive
  - Opportunity to collect additional information
  - Useful in hard to recruit samples – rare/orphan indications
- Exploratory quantitative analyses as secondary objective to conceptual understanding in cognitive interviews
  - Early indication of rating scale utilization (overall and by strata)
  - Distributional characteristics (early discrimination across severity levels)
- Potential to reuse cognitive interviewing sample for quantitative pilot study

Activities Supporting this Topic

- 2011
  - PRO Consortium Mixed Methods Panel convened to discuss psychometric methods for understanding early item performance
- 2012
  - Utilizing Rasch Measurement for Mixed Methods: 2011 Webinar and 2012 ISPOR Workshop
  - IAC Symposium Perspectives on Mixed Methods 2012
  - Mixed Methods Meeting at the FDA 2012
- 2015
  - ISOQOL SIG on Mixed Methods formed

Conceptualization of the Approach

- In addition to concept elicitation and cognitive debriefing with patients, along with the assessment of saturation, exploratory, descriptive analysis of responses may be implemented iteratively during instrument development.



Iterative: May need to repeat based on decision to add/delete/revise items to achieve saturation

Cognitive Interviewing: Collect patient responses on current instrument

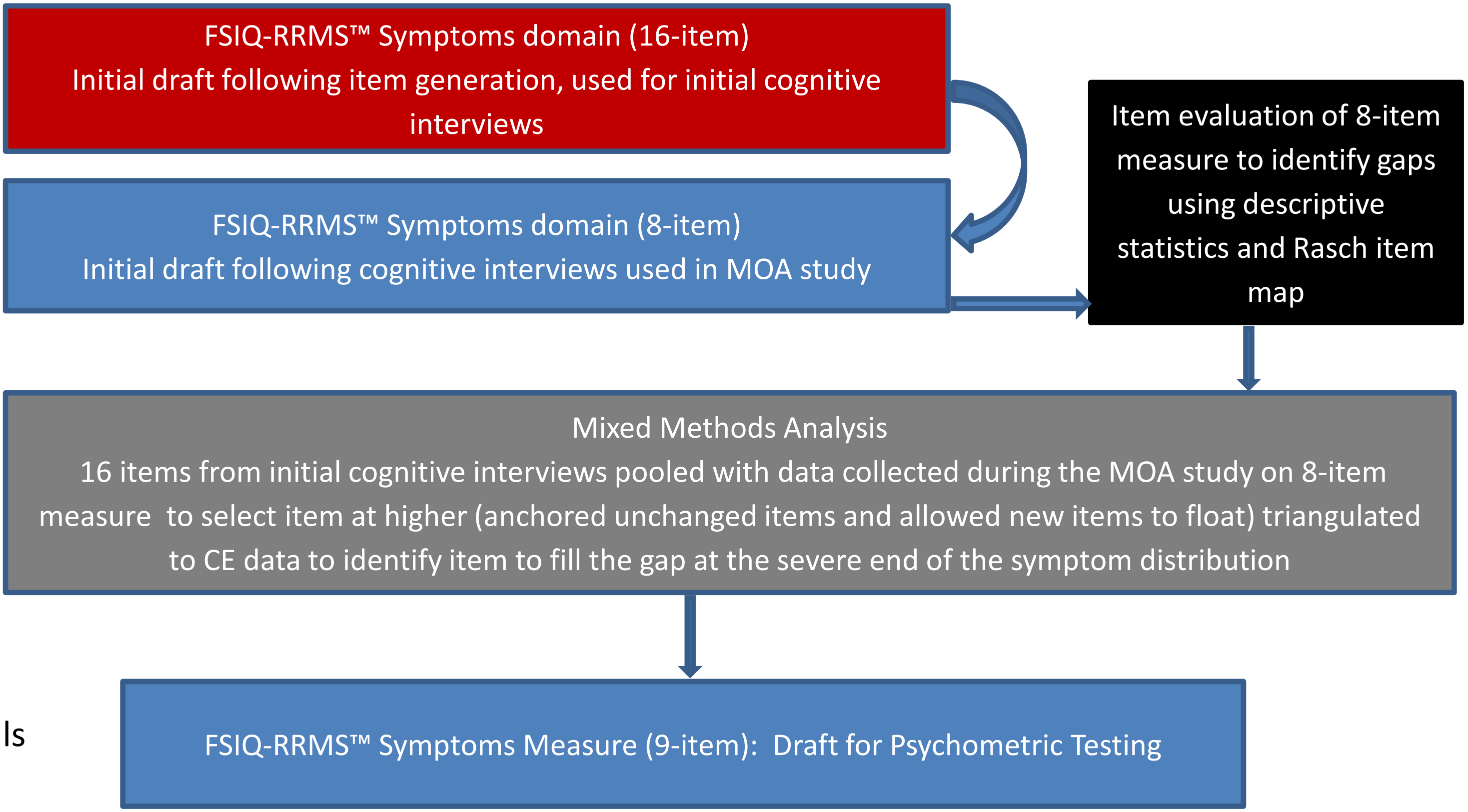
Exploratory Review of Item Responses (method is dependent upon sample size)

Review Concept Elicitation Data: Evaluate for any identified issues

Example: Fatigue in Multiple Sclerosis

The utilization of mixed methods in this example resulted in the following:

- Inclusion of a key item at the severe tail of the severity distribution prior to full psychometric testing
- Ability to understand the early distribution characteristics of the rating scale and item severity continuum



FSIQ-RRMS™ Symptoms domain (16-item)
Initial draft following item generation, used for initial cognitive interviews

FSIQ-RRMS™ Symptoms domain (8-item)
Initial draft following cognitive interviews used in MOA study

Item evaluation of 8-item measure to identify gaps using descriptive statistics and Rasch item map

Mixed Methods Analysis
16 items from initial cognitive interviews pooled with data collected during the MOA study on 8-item measure to select item at higher (anchored unchanged items and allowed new items to float) triangulated to CE data to identify item to fill the gap at the severe end of the symptom distribution

FSIQ-RRMS™ Symptoms Measure (9-item): Draft for Psychometric Testing

Courtesy of Actelion Pharmaceuticals Ltd.

Conclusions

- Cognitive interviewing is a critical step in pilot testing of an instrument. Optimizing this data collection for multiple uses only assists in the improvement of measures prior to full scale psychometric testing (e.g., observational or clinical trial endpoint evaluation)
- While there has been concern about "contamination" from using data collected during qualitative research, conceptually, it is less of a concern in qualitative testing as we have the ability to improve the measure as we move through instrument development
  - [In qualitative research] "data collection and analysis is often progressive, in that a second or subsequent interview in a series should be 'better' than the previous one" (van Teijilingen and Vanora, 2002)
- Data collected within cognitive interviewing may be used as an early pilot where data are evaluated in a purely exploratory manner and not used for hypothesis testing as is the case in psychometric evaluation
- Additional benefit:
  - Easy to implement within the interview
  - Data entry and analysis is quick
  - Interpretation may lead to the addition of items, early understanding of scale endorsement, and understanding of the item hierarchy (e.g., frequency, intensity)

van Teijlingen, Edwin, and Vanora Hundley. "The importance of pilot studies." *Nursing Standard* 16, no. 40 (2002): 33-36.
Peat, J., Mellis, C., Williams, K. and Xuan W. (2002), Health Science Research: A Handbook of Quantitative Methods, London: Sage.

Scale Development using Longitudinal IRT

R.J. Wirth, PhD
Vector Psychometric Group, LLC

SIXTH ANNUAL
PATIENT-REPORTED OUTCOME CONSORTIUM WORKSHOP

April 29 - 30, 2015 ■ Silver Spring, MD

Thank You

C-Path and the **FDA** for this opportunity and to the session team for comments on previous drafts of this talk

Linda Deal and **Shire** for allowing me to use their data

Carrie Houts for running the analyses and for comments on earlier drafts

Rob Morlock for insisting that we explore this methodology

Mike Edwards & **Li Cai** for comments on earlier drafts

Outline

- Goal of scale development
- PROs and sample size
- Longitudinal IRT is one possible solution
- Real-world example
- Conclusion

Goals of Psychometric Analysis

- To understand the measurement properties of a set of items. This allows us to develop a scoring rubric that reliably and accurately reflects individuals' standing on a construct of interest

- Our job as psychometricians is to maximize the confidence people have in the scores and, by extension, the results of analyses based on those scores

PROs and Sample Size

- Sample size is one issue particularly problematic in rare or hard to assess populations

- We can make several choices when we have less than ideal sample sizes, one of which is to use simpler models (e.g., classical test theory). That's a reasonable thing to do, but it has predictable results (like large confidence intervals)
  - Example: we have 20 items, N = 50, and we find a coefficient alpha of 0.8

$$\text{95\% CI: } 0.40 \ \leq \ \alpha \leq 1.00$$

- So we’re left with a sub-optimal method with almost as many issues

PROs and Sample Size

- Sample size is one issue particularly problematic in rare or hard to assess populations

- We can make several choices when we have less than ideal sample sizes, one of which is to use simpler models (e.g., classical test theory). That's a reasonable thing to do, but it has predictable results (like large confidence intervals)
  - Example: we have 20 items, N = 125, and we find a coefficient alpha of 0.8

$$95\%\ \text{CI: } 0.55 \leq \alpha \leq 1.00$$

- So we're left with a sub-optimal method with almost as many issues

One Solution

- How do we move towards optimal methods when we cannot get larger samples?
  - We need maximize the sample we have
- Using longitudinal (multilevel) IRT models allow us to incorporate multiple assessments from each subject into our scale development process
- While this doesn’t mean we can simply collect 10 assessments from 100 people in one weekend, it does gives us options when multiple assessments are available

An Example: Traditional IRT

- Deal et al.’s (in press) psychometric examination of the Yale-Brown Obsessive Compulsive Scale Modified for Binge Eating (YBOCS-BE)

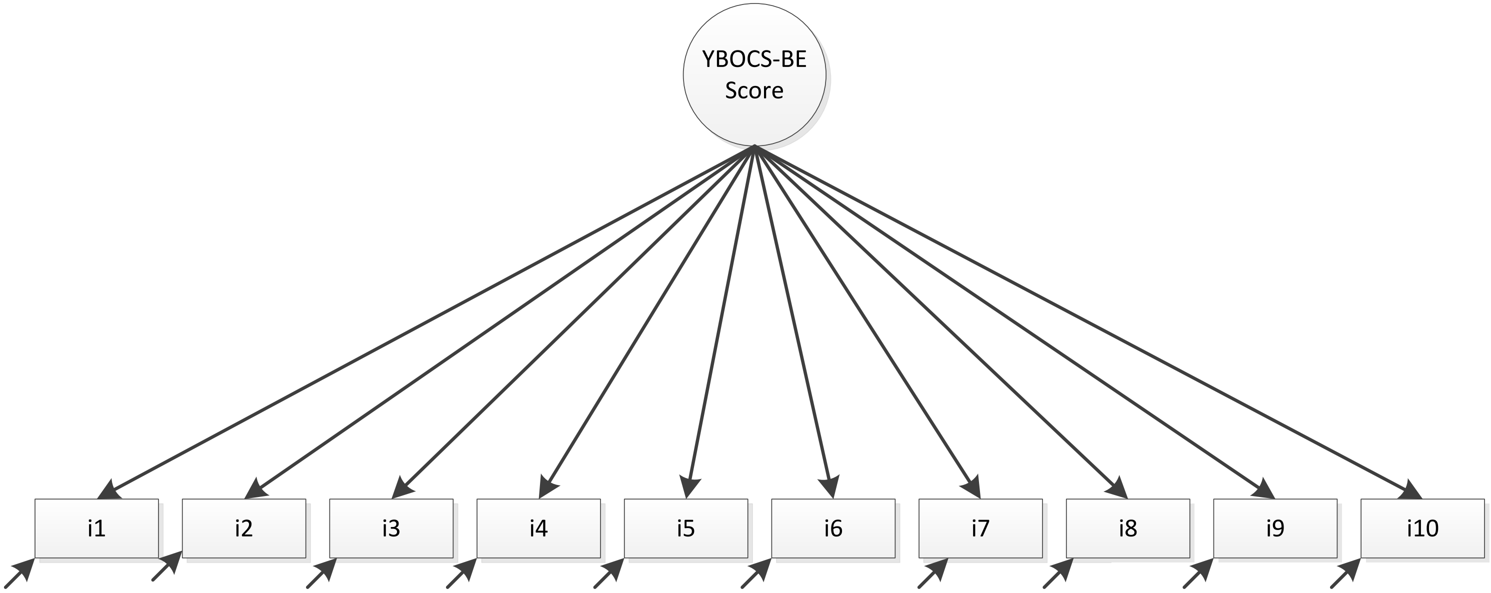
- 270 (255 used here) men & women 18-55 years of age, $25 \leq \text{BMI} \leq 45$ kg/m^2, met DSM-IV-TR criteria for a diagnosis of an eating disorder & had a confirmed diagnosis of binge eating disorder

- IRT analyses conducted using flexMIRT® v. 3.0

- Other analyses conducted using SAS® 9.2

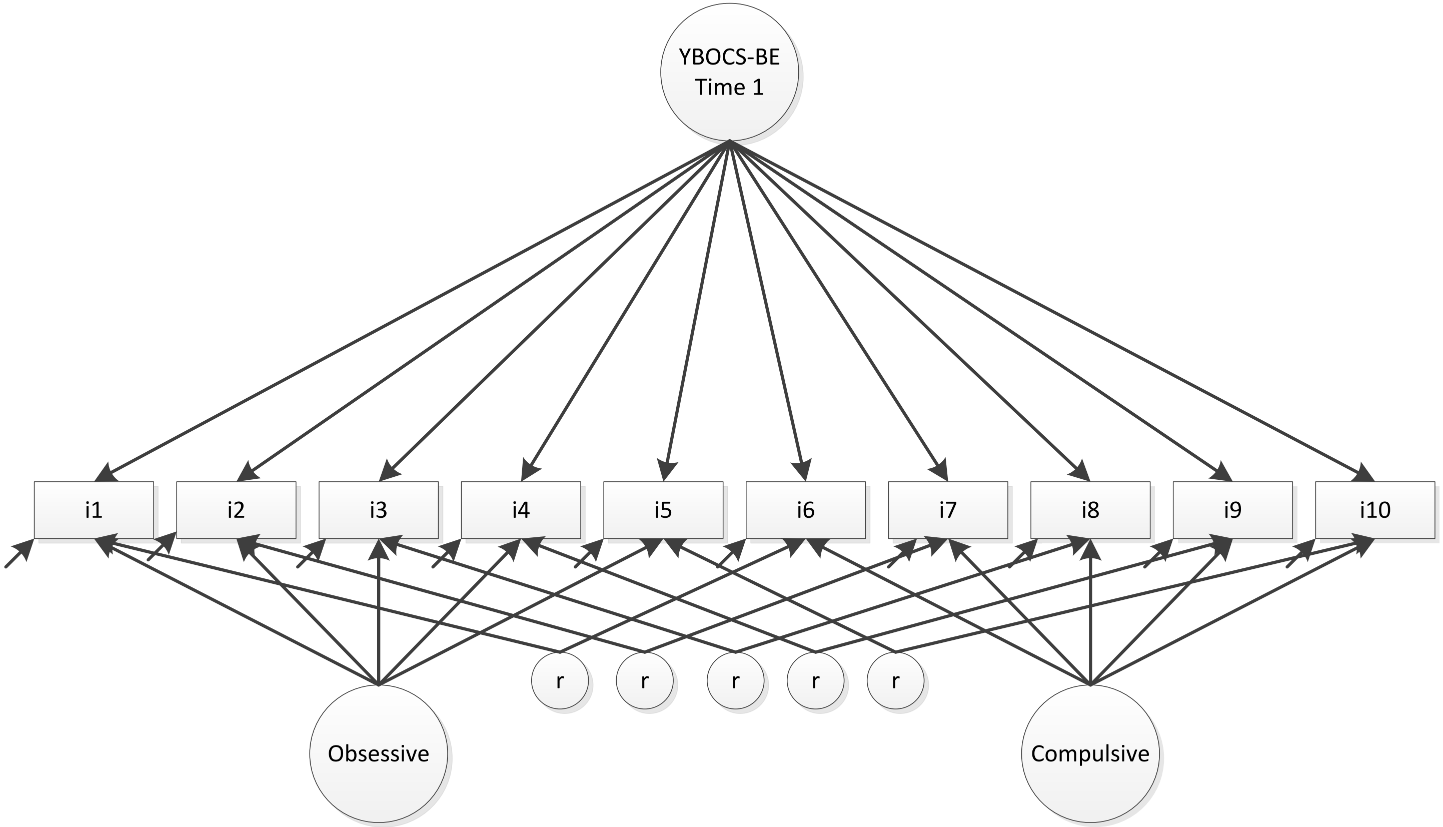
An Example: Traditional IRT

- YBOCS-BE has 10 items, 5 response options
- Measures obsessive and compulsive behavior
- Interested in a single overall score – consistent with original scale



An Example: Traditional IRT

For the stats geeks - the IRT model for these items was actually a little more complicated



An Example: Traditional IRT

Taken from Table 5: Mean Change in the YBOCS-BE Total Score over Time

	Week 3	Week 7	Week 11	Cohen's Effect Size
Low Dose				
# patients	62	57	53	0.42
Mean Score	-1.4	-1.7	-1.9	
Mean Change from Baseline	-1.4	-1.7	-1.9	
Medium Dose				
# patients	63	57	55	0.44
Mean Score	-1.6	-1.8	-1.9	
Mean Change from Baseline	-1.5	-1.7	-1.7	
High Dose				
# patients	63	57	55	0.62
Mean Score	-1.8	-1.9	-2.1	
Mean Change from Baseline	-1.7	-1.9	-2.0	

An Example: Longitudinal IRT

- Let’s instead assume we only had N = 125

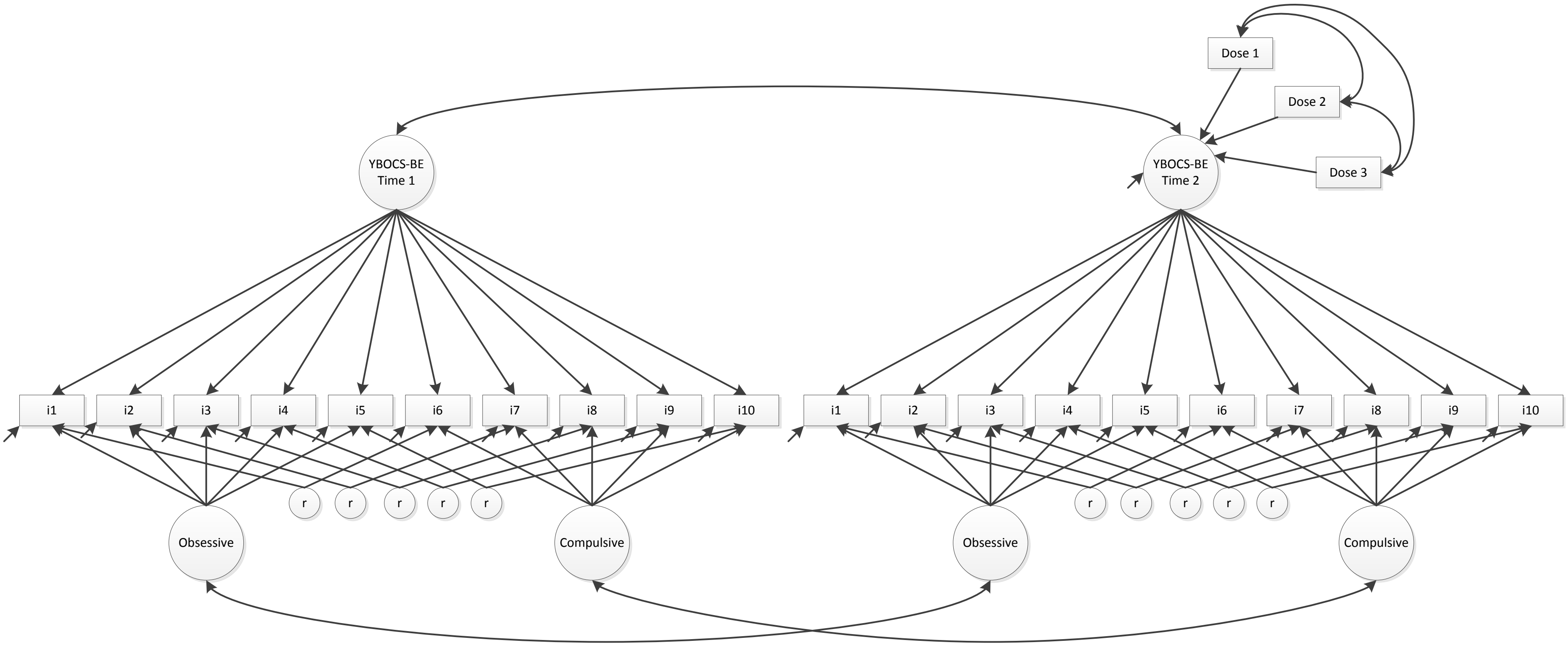
- This is a very small N (especially given the complexity of the statistical model)

- Trying to examine these items using only N=125 and single time point just doesn’t work

- What if we try using multiple time points?
  - If we make some invariance assumptions, we can actually get results consistent with the published results using only N = 125 and two time times of assessment

An Example: Longitudinal IRT

For the stats geeks:



An Example: Comparing N = 225 & 125

Mean Change in the YBOCS-BE Total Score over Time

	Week 3	Week 7	Week 11	Cohen's Effect Size
Low Dose				
# patients	62	57	53	0.42/0.37
Mean Score	-1.4/-1.0	-1.7/-1.6	-1.9/-1.8	
Mean Change from Baseline	-1.4/-1.3	-1.7/-1.8	-1.9/-2.1	
Medium Dose				
# patients	63	57	55	0.44/0.44
Mean Score	-1.6/-1.2	-1.8/-1.6	-1.9/-1.9	
Mean Change from Baseline	-1.5/-1.5	-1.7/-1.9	-1.7/-2.2	
High Dose				
# patients	63	57	55	0.62/0.62
Mean Score	-1.8/-1.4	-1.9/-1.9	-2.1/-2.2	
Mean Change from Baseline	-1.7/-1.8	-1.9/-2.3	-2.0/-2.6	

An Example: Comparing N = 225 & 125

Mean Change in the YBOCS-BE Total Score over Time

	Week 3	Week 7	Week 11	Cohen's Effect Size
Low Dose				
# patients	62	57	53	0.42/0.37
Mean Score	-1.4/-1.0	-1.7/-1.6	-1.9/-1.8	
Mean Change from Baseline	-1.4/-1.3	-1.7/-1.8	-1.9/-2.1	
Medium Dose				
# patients	63	57	55	0.44/0.44
Mean Score	-1.6/-1.2	-1.8/-1.6	-1.9/-1.9	
Mean Change from Baseline	-1.5/-1.5	-1.7/-1.9	-1.7/-2.2	
High Dose				
# patients	63	57	55	0.62/0.62
Mean Score	-1.8/-1.4	-1.9/-1.9	-2.1/-2.2	
Mean Change from Baseline	-1.7/-1.8	-1.9/-2.3	-2.0/-2.6	

Summary

- The goal of psychometric analyses is to provide reliable scores we can have confidence in using

- Study designs can limit our ability to collect the number of subjects we may want

- Using longitudinal (conditional) IRT models provides one way to use all the information we have in a set of data to maximize the quality of our estimates and inferences

- This can be especially beneficial in rare or hard to assess populations as it provides a way to use more desirable and informative measurement methods to understand your subjects (and treatments) with confidence

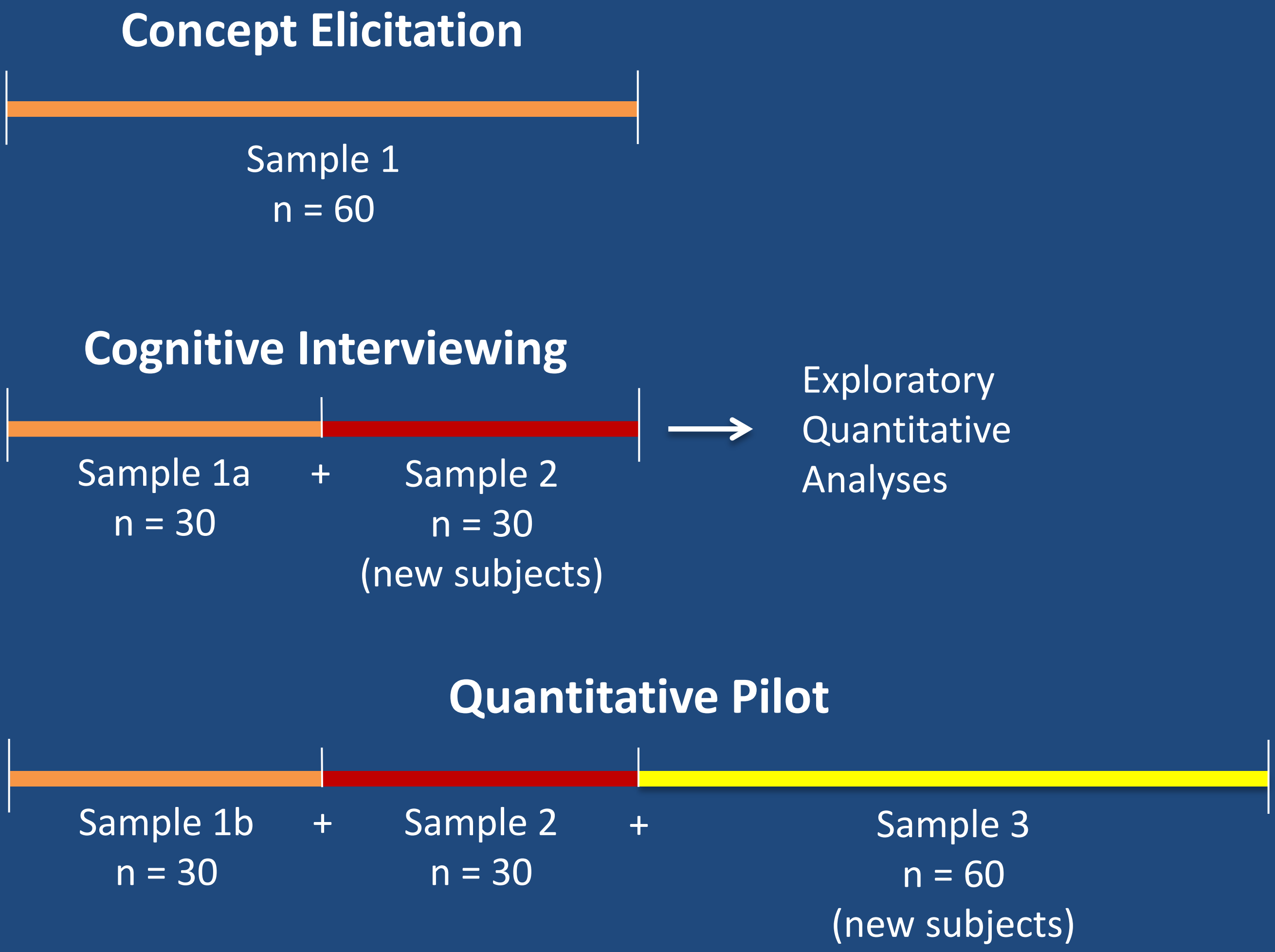
Thank You

RJWirth@VPGcentral.com

Disclaimer

The views expressed in this presentation are those of the speaker, and do not necessarily represent an official FDA position.

Reduce, Reuse, Recycle Example



Concept Elicitation

Sample 1
n = 60

Cognitive Interviewing

Sample 1a
n = 30

+

Sample 2
n = 30
(new subjects)

→ Exploratory
Quantitative
Analyses

Quantitative Pilot

Sample 1b
n = 30

+

Sample 2
n = 30

+

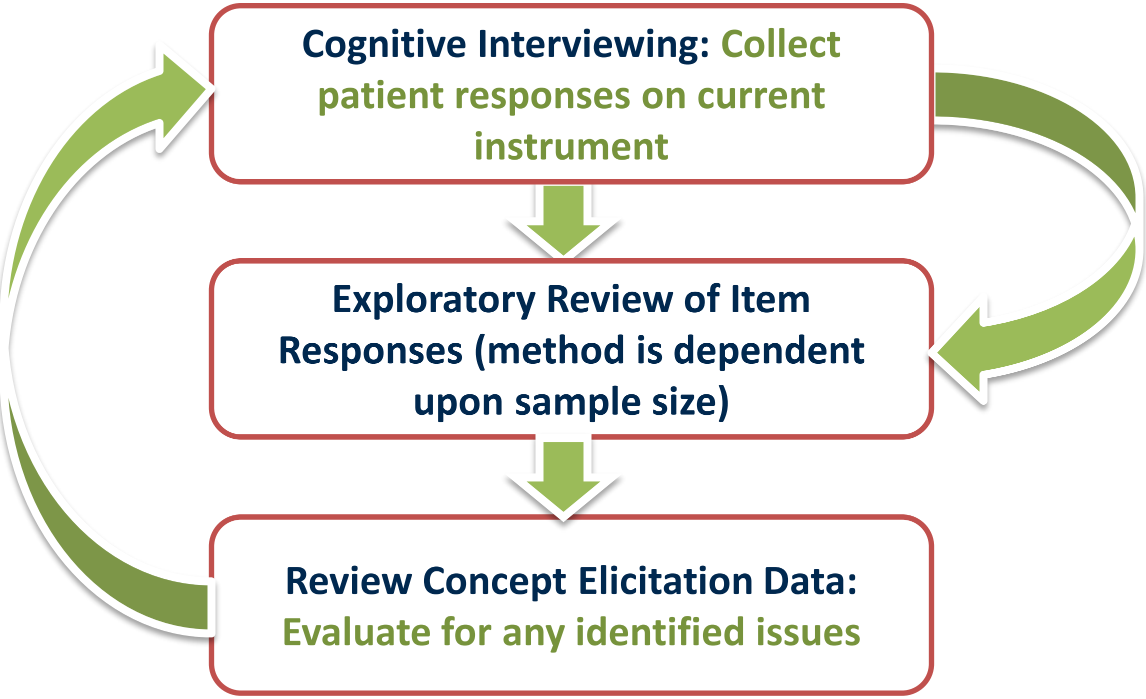
Sample 3
n = 60
(new subjects)

For Illustration Purposes Only

Conceptualization of the Approach

- In addition to concept elicitation and cognitive debriefing with patients, along with the assessment of saturation, exploratory, descriptive analysis of responses may be implemented iteratively during instrument development.

Iterative:
May need to repeat based on decision to add/delete/revise items to achieve saturation



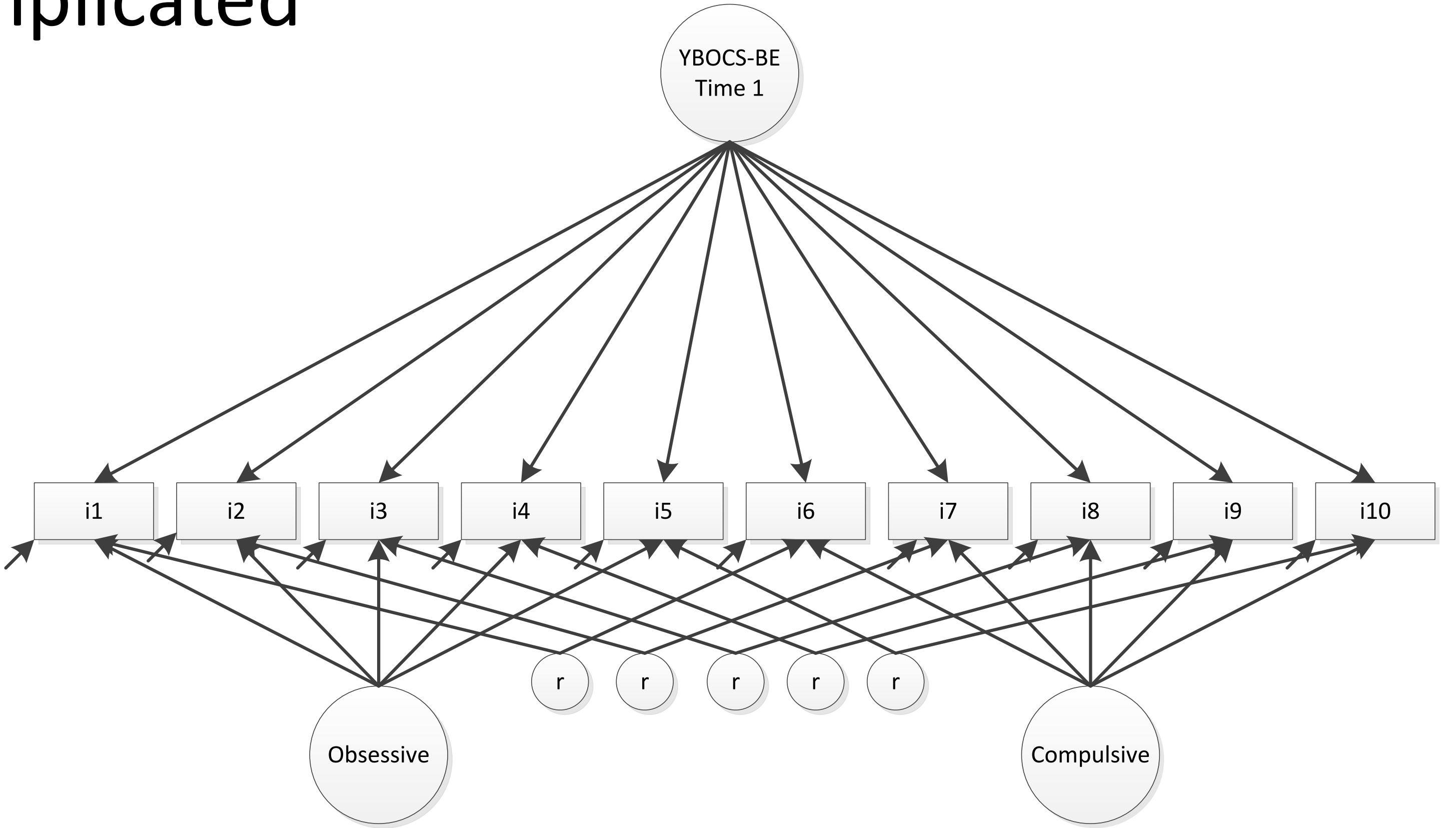
Cognitive Interviewing: Collect patient responses on current instrument

Exploratory Review of Item Responses (method is dependent upon sample size)

Review Concept Elicitation Data: Evaluate for any identified issues

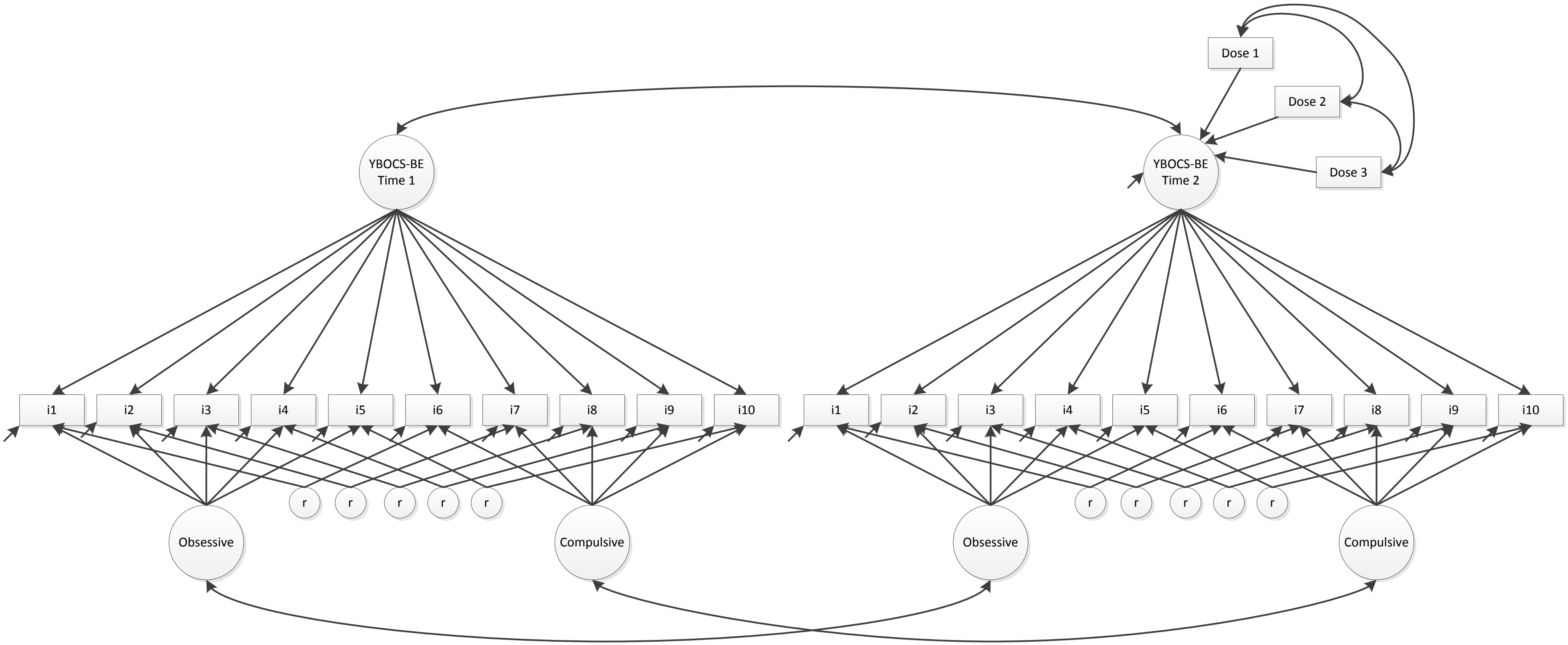
An Example: Traditional IRT

For the stats geeks - the statistical model for these items was actually a little more complicated



An Example: Longitudinal IRT

For the stats geeks:



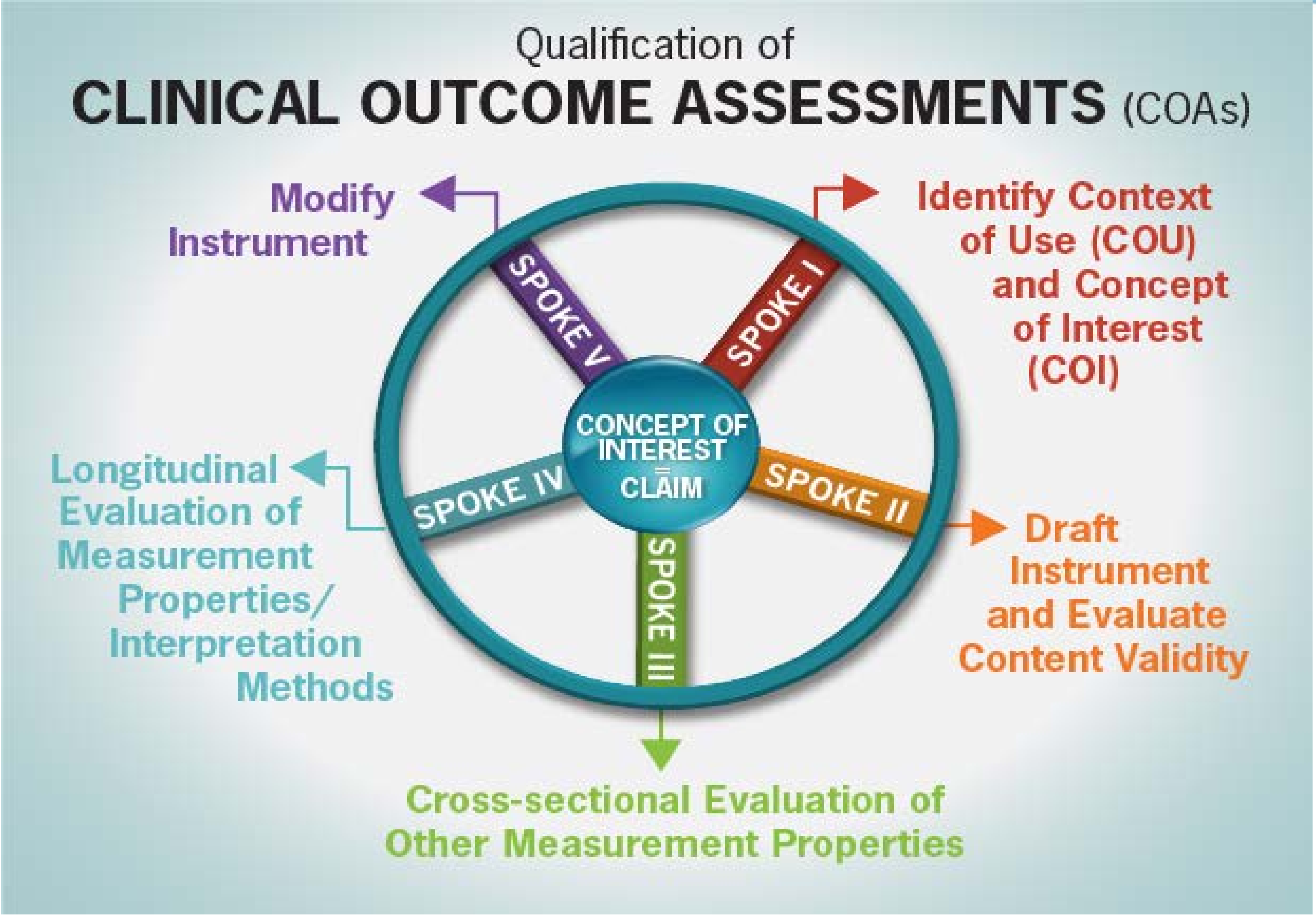
An Example: Comparing N = 225 & 125

Mean Change in the YBOCS-BE Total Score over Time

	Week 3	Week 7	Week 11	Cohen's Effect Size
Low Dose				
# patients	62	57	53	0.42/0.37
Mean Score	-1.4/-1.0	-1.7/-1.6	-1.9/-1.8	
Mean Change from Baseline	-1.4/-1.3	-1.7/-1.8	-1.9/-2.1	
Medium Dose				
# patients	63	57	55	0.44/0.44
Mean Score	-1.6/-1.2	-1.8/-1.6	-1.9/-1.9	
Mean Change from Baseline	-1.5/-1.5	-1.7/-1.9	-1.7/-2.2	
High Dose				
# patients	63	57	55	0.62/0.62
Mean Score	-1.8/-1.4	-1.9/-1.9	-2.1/-2.2	
Mean Change from Baseline	-1.7/-1.8	-1.9/-2.3	-2.0/-2.6	

Simulation Study

- **in vitro**
- **in vivo**
- **in silico**
  - There is still a lot to learn of different simulation models and methods in different context
  - It may be useful to add to the research agenda: better understanding of the role and appropriate use of simulation study in clinical outcome assessment development
  - Research questions such as could simulation studies help inform study design? sample size? study duration? frequency of assessment? relationship among items and between items and sample,? and generating hypotheses?



Qualification of

CLINICAL OUTCOME ASSESSMENTS (COAs)

Modify Instrument
SPOKE V
SPOKE I
Identify Context of Use (COU) and Concept of Interest (COI)
CONCEPT OF INTEREST = CLAIM
Longitudinal Evaluation of Measurement Properties/ Interpretation Methods
SPOKE IV
SPOKE II
Draft Instrument and Evaluate Content Validity
SPOKE III
Cross-sectional Evaluation of Other Measurement Properties

Panel Discussion

- Comments from Wen-Hung Chen, PhD
  - Reviewer, Study Endpoints, SEALD, OND, CDER, FDA
- Questions from the Audience

Session Participants

Moderator

- J. Jason Lundy, PhD

Presenters and Panelists

- J. Jason Lundy, PhD – Principal, Outcometrix
- Stacie Hudgens, MA – Managing Partner, Clinical Outcomes Solutions
- R.J. Wirth, PhD – Managing Partner, Vector Psychometric Group, LLC
- Wen-Hung Chen, PhD – Reviewer, Study Endpoints, SEALD, OND, CDER, FDA