What’s the Score? Moving from Items to Scores - Methods, Considerations, and Case Examples

Eighth Annual Patient-Reported Outcome Consortium Workshop

April 26 – 27, 2017 ■ Bethesda, MD
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 or the Critical Path Institute.

• 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
  – *Steve Blum, MBA, MA* – Director, Patient-Reported Outcomes, GlaxoSmithKline

Presenters
  – *Kathy Wyrwich, PhD* – Senior Research Advisor, Eli Lilly and Company
  – *Bryce Reeve, PhD* – Professor, University of North Carolina
  – *Cheryl D. Coon, PhD* – Principal, Outcometrix

Panelists
  – *Laura Lee Johnson, PhD* – Deputy Director, Division of Biometrics III, Office of Biostatistics, Office of Translational Sciences, CDER, FDA
Why a Session on Scoring?

- We have spent a lot of time discussing the development of instruments and the interpretation of scores (i.e. meaningful change), but have spent little time considering the underlying approaches for how instruments can and should be scored.
- As instrument developers, we often don’t thinking about the scoring of the measure until later in development.
- Yet, it is critical to ensure that we have a robust, “validated” scoring algorithm
- There is a lack of consensus on best approaches
- Today we will start to discuss some important considerations around scoring of PRO and other COA measures.
Session Outline and Objectives

• Provide an overview of the different types of scoring algorithms, including real-world examples of new and existing PRO measures

• To understand how psychometric methods can inform the handling of missing data and the impact of missing data on instrument scoring

• To understand the statistical implications for interpretation and handling of different types of data
What is a Score?

• A numerical “value” assigned to a patient’s responses on a questionnaire
  • Some scores are inherent to the response scale (i.e. visual analog scale or
    numerical rating scale), where as others need a value to be applied

• Intended to “quantify” aspects of an individual’s health state/status
  • How much/how often does an individual experience the concept of interest

• Needs to be based on a pre-specified “validated” algorithm
  • Number of different approaches that can be taken when generating a score
  • Algorithm should also provide instructions for handling of missing data
Scoring Terminology

- **Raw Score** – a “simple” score obtained by summing responses for items (raw score may still involve some transformation as some items may need to be reversed scored)

- **Weighted Score** – a score obtained by proportionally combining scores of individual items. Weights of items can either be proportionate (equal) or disproportionate (vary) by item or group of items

- **Transformed Score** – a score obtained by converting the scoring range for the scale

- **Norm-Based Score** – a transformed score that aligns the scale to normative values for a given population
Scores can be generated for **individual items**, for **groups of items**, for the **instrument/scale** as a whole or how it is implemented as an **endpoint**.

**Individual Item Level**
- Value for a single individual item

**Group/Domain Level**
- Value for a set of multiple items (various groups)

**Scale/Instrument Level**
- Total value of all of the items in the scale/instrument
- May not always have an overall total score

**Endpoint Level**
- How the score is implemented as a study endpoint
- Endpoint could be based on individual items or groups of items
What Does the PRO Guidance Say?

• The PRO Guidance touches on several scoring topics:
  • Item distributions and intervals
  • Weighting of items or domains
  • Appropriateness of domain and summary scores
  • Ability to detect change
Challenges with Scoring COA Measures

- What type of data do we have?
- How do we assign appropriate values to different response options?
- How do we combine scores for multiple items?
- When in the development of the measure is the scoring algorithm developed/validated?
Speakers and Panelists

• Kathleen W. Wyrwich, PhD
  • Senior Research Advisor, Eli Lilly and Company
    • Score Transformations
    • Ordinal versus Continuous Scores

• Bryce B. Reeve, PhD
  • Professor, Health Policy and Management, UNC Gillings School of Public Health
  • Member, Lineberger Comprehensive Cancer Center
  • Simple Scoring versus Complex (IRT) Scoring
Speakers and Panelists

• Cheryl D. Coon, PhD
  • Principal, Outcometrix
  • Practical Example – Data from Asthma Daily Symptom Diary
  • Handling Missing Data

• Laura Lee Johnson, PhD
  • Deputy Director, Division of Biometrics III, Office of Translational Sciences, CDER, FDA
  • FDA Response
Moving from Items to Scores: Issues of Perceived Score Continuity

Kathy Wyrwich, PhD
Sr. Research Advisor, Eli Lilly and Company
How Ordinal Items Become Scale Scores?

• With the widespread use of the SF-36 over the past 25 years, 0-100 point scales have become ubiquitous in health status measurement.
  • Essentially, we take the possible raw SUM score range (with all item facing the same direction) and conduct a linear transform yielding the final 0-100 point scale.

• Another popular scoring rubric is a the scale AVERAGE score
  • For example, the Asthma Quality of Life Questionnaire (AQLQ; 1992) has four domains:
    • Symptoms (11 items), Activity Limitation (12 items), Emotional Function (5 items), Environmental Exposure (4 items)
    • Responses are on a 7-point Likert scale (1 = severely impaired to 7 = not impaired at all)
    • Domain scale scores are the average items scores within the domain (e.g., 3.14)

• Hence, the average scale score is within the range of all response options and remains comparable to the items
Summed vs. Average Scale Scores

Advantages to **Summed** Scoring

1. Reflects the total of the item responses

1. When transformed to 0-100, gives an initial feel on where the scale score falls within the range

Advantages to **Average** Scoring

1. Contains the range of the response option scores (gives a feel for where the score falls within the range

2. Can be scored with missing item(s) without imputation of the missing item(s)

What happens when we create these scale scores that deserves awareness?
• For example, the 36-item Short Form (SF-36) Physical Functioning Scale: 10 items, each with 3 response options:

1 = Yes, limited a lot
2 = Yes, limited a little
3 = No, not limited at all

• Raw scores move in increments of 1 unit, and range from 10-30 points.

• The linear transform for the original SF-36 scoring stretches the raw score of 10 down to 0 and the highest raw score (30) to 100.

• With this new 0-100 point scale, we have become accustomed to treating resulting scores as though we are working with a 101 point scale and always treat the score scores as continuous data.
Picturing the process: Original SF-36 Scoring

10 - 30 Raw Score

0 - 100 Transformed Score

Formula and example for transformation of raw scale scores to 0-100 scale scores:

\[
\text{Transformed Scale} = \frac{(\text{Actual raw score} - \text{lowest possible raw score})}{\text{Possible raw score range}} \times 100
\]

Why 0-100 Transformation on SF-36 Original?

• All items are scored so that a high score defines a more favorable health state.
• Each item is scored on a 0 to 100 range so that the lowest and highest possible scores are 0 and 100, respectively.
• Scores represent the percentage of total possible score (100) achieved.
**Important to Keep in Mind: State Change**

*State Change*: Smallest amount that an individual patient/respondent’s score can change is they improve or worsen by only one response level on only one item on the scale.

In other words, at the individual level, we need to see these original SF-36 0-100 scale scores as a series of *lily pads that skip across the pond...and not a continuous line*

For example, on the SF-36 Physical Functioning Scale:

- If a patient changed only 1 item to the next highest response [e.g., “Yes, limited a little” (at baseline) to “No, not limited at all” (at end of study)] but stayed at the same response on all other 9 items, their score would change by 5 points over time.
The Role of State Change in Determining a Responder Definition

- Using this example from the SF-36 Physical Functioning Scale, any reasonable estimates for the responder definition should be a multiple of 5 points.
- That is, if improvement in Physical Functioning is an important change over time, the responder definition “choices” that align with patient change scores are: 5, 10, 15, 20, etc.
- Please also note that for a condition where treatment seeks to delay or reduce the rate of decline, a responder definition of 0, -5, -10, etc. may also be reasonable, and should be examined through anchor-based methods.

Examining the state change can help to best use the information empirically derived via anchor-based methods.
The advantage of the standardization and norm-based scoring of the 8 SF-36v2 scales is that results for one scale can be meaningfully compared with the other scales and their scores have a direct interpretation in relation to the distribution of scores in the 1998 general U.S. population. Specifically, all scores above or below 50 are above or below the average, respectively, in the 1998 general U.S. population. Because the standard deviation is 10 for all 8 scales, each one point difference or change in scores also has a direct interpretation. A one point difference or change is one-tenth of a standard deviation unit or an effect size of 0.10. Lastly, norm-based scoring provides the basis for comparing scale scores across Version 1.0 and Version 2.0 standard forms.
Examining the SF-36V2 Normed-Based Scoring for the Physical Functioning Scale Scores

**Standard Form (past four weeks recall)**

**Step 1.**
Formulas for z-score standardization of SF-36v2 scales (Standard Form):

\[
PF_Z = \frac{PF - 83.29094}{23.75883}
\]

where PF = 0-100 original transformed score

**Step 2.**
Norm-based transformation of SF-36v2 z-scores (Standard Form):

Norm-Based PF: \( PF = 50 + (PF_Z \times 10) \)

**Acute Form (past week recall)**

**Step 1.**
Formulas for z-score standardization of SF-36v2 scales (Acute Form):

\[
PF_Z = \frac{PF - 82.62455}{24.43176}
\]

where PF = 0-100 original transformed score

**Step 2.**
Norm-based transformation of SF-36v2 z-scores (Acute Form):

Norm-Based PF: \( PF = 50 + (PF_Z \times 10) \)

Picturing the process: Normed-Based SF-36V2 Scoring

10 - 30 Raw Score

0 - 100 Original Transformed Score

14.9  \( State \ Change = \sim 2.10 \)
for Standard Form

16.2  \( State \ Change = \sim 2.05 \)
for Acute Form
Examining the SF-36V2 Normed-Based Scoring for the Physical Functioning Scale Scores

**Standard Form (past four weeks recall)**

**Step 1.** Formulas for z-score standardization of SF-36v2 scales (Standard Form):

\[ PF_Z = \left( PF - 83.29094 \right) / 23.75883 \]

where PF = 0-100 original transformed score

**Step 2.** Norm-based transformation of SF-36v2 z-scores (Standard Form):

Norm-Based PF: \( PF = 50 + (PF_Z \times 10) \)

**Acute Form (past week recall)**

**Step 1.** Formulas for z-score standardization of SF-36v2 scales (Acute Form):

\[ PF_Z = \left( PF - 82.62455 \right) / 24.43176 \]

where PF = 0-100 original transformed score

**Step 2.** Norm-based transformation of SF-36v2 z-scores (Acute Form):

Norm-Based PF: \( PF = 50 + (PF_Z \times 10) \)

Example: EORTC QLQ-C30 Scales

- All of the scales and single-item measures range in score from 0 to 100.
- A high scale score represents a higher response level.
- Thus, a high score for a functional scale represents a high/healthy level of functioning, but a high score for a symptom scale / item represents a high level of symptomatology/problems.

What is happening with the EORTC Single Items? Example: Appetite Loss

Picturing the process:
0-100 Appetite Loss Scoring

1-4 Raw Score

0 - 100 Transformed Score

State Change = 33.3 points
• **Be Mindful of the Scale Score State Change Level**
  • Very useful in the process of determining responder definitions and other issues in scale score interpretation

• **Be Respectful of Ordinal Data**
  • Just because ordinal data is put through a linear transformation and/or a multi-step score norming process, at the individual level, scores are still changing/moving over time across distinct lily pads
Consideration of Simple Sum Scores and Item Response Theory (IRT) Scores for PRO Measurement

Bryce B. Reeve, PhD
Professor, Health Policy and Management
Member, Lineberger Comprehensive Cancer Center
bbreeve@email.unc.edu

Mian Wang, PhD
Psychometrician,
Lineberger Comprehensive Cancer Center
Sum (or Average) Scores

• Positives:
  • Easy to calculate
  • Can be easy to interpret

• Negatives
  • For small # of items, the aggregated score is more ordinal than interval level
  • Treats each item as equal in terms of discrimination and level of severity/difficulty
  • Distance between adjacent response categories treated equally.
  • Missing values presents challenges (scores are either not estimated or requires imputation)
  • Sum score does not partial out measurement error

“T’m going to need tech support.”
What is Item Response Theory (IRT)?

1. Theory for Scale Construction
   • For this presentation, I am treating Rasch models as a subset of IRT models

2. Methodology for:
   • Evaluating the properties of items within a scale and the overall scale
   • Refining a scale
   • Scoring an individual
   • Linking multiple scales on to a common metric
   • Tailoring a measure to an individual or group.
     • Computerized Adaptive Testing (CAT)

3. IRT is designed for:
   • Modeling latent “unobservable” variables (traits, domains, \( \theta \))
   • Multi-item Scales
PROMIS Anxiety 4-Item Short Form (custom)

• In the past 7 days,
  - EDANX30: I felt worried.
  - EDANX33: I felt terrified.
  - EDANX46: I felt nervous.
  - EDANX53: I felt uneasy.

• Response options:
  - 1 = Never
  - 2 = Rarely
  - 3 = Sometimes
  - 4 = Often
  - 5 = Always
Samejima’s Graded Response IRT Model: Item Characteristic Curves

I felt worried.

\[ a = 3.03 \quad b_1 = -0.52 \quad b_2 = 0.32 \quad b_3 = 1.35 \quad b_4 = 2.30 \]

I felt terrified.

\[ a = 2.58 \quad b_1 = 1.15 \quad b_2 = 1.82 \quad b_3 = 2.72 \quad b_4 = 3.59 \]

\[
P(X_i = k | \theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{ik+1})}}
\]
IRT Model Approach to Scoring an Individual

For polytomous response data, the Samejima’s graded response IRT Model for 1 item:

\[
P(X_i = k|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{ik+1})}}
\]

Because of the property of local independence, the likelihood of any response pattern can be defined as the serial product of the individual item response probabilities.

\[
L(\text{response pattern}|\theta) = \prod_{i=1}^{n\text{items}} P(X_i|\theta)
\]

\[
L(2,3|\theta) = P(X_1 = 2|\theta) \times P(X_2 = 3|\theta)
\]

If we have knowledge of the distribution of this trait in the population, we can add the population distribution function.

\[
L(\text{response pattern}|\theta, \phi(\theta)) = \prod_{i=1}^{n\text{items}} P(X_i|\theta) \phi(\theta)
\]

We maximize this Likelihood to get a person’s level on the construct.
**PROMIS Anxiety Sum Scores vs IRT EAP & T-Scores**

<table>
<thead>
<tr>
<th>Response Pattern</th>
<th>Worried</th>
<th>Terrified</th>
<th>Nervous</th>
<th>Uneasy</th>
<th>Sum Score</th>
<th>EAP*</th>
<th>SE&lt;sub&gt;EAP&lt;/sub&gt;</th>
<th>PROMIS T-score</th>
<th>PROMIS SE&lt;sub&gt;T&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP1</td>
<td>5-Always</td>
<td>1-Never</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>-0.30</td>
<td>0.43</td>
<td>47</td>
<td>4.3</td>
</tr>
<tr>
<td>RP2</td>
<td>4-Often</td>
<td>2-Rarely</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>0.13</td>
<td>0.42</td>
<td>51.3</td>
<td>4.2</td>
</tr>
<tr>
<td>RP3</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>8</td>
<td>0.31</td>
<td>0.36</td>
<td>53.1</td>
<td>3.6</td>
</tr>
<tr>
<td>RP4</td>
<td>1-Never</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>3-Sometimes</td>
<td>8</td>
<td>0.49</td>
<td>0.36</td>
<td>54.9</td>
<td>3.6</td>
</tr>
</tbody>
</table>

*EAPs (expected a posteriori) are estimated using a Bayesian approach which finds the expected value of the posterior distribution. EAPs are in the metric of $\theta$ which follows a standard normal distribution $N(0, 1)$. PROMIS T-scores are normed with a mean of 50 and SD of 10 in the general US population.
### PROMIS Anxiety Sum Scores vs IRT EAP & T-Scores

<table>
<thead>
<tr>
<th>Response Pattern</th>
<th>Worried</th>
<th>Terrified</th>
<th>Nervous</th>
<th>Uneasy</th>
<th>Sum Score</th>
<th>EAP*</th>
<th>SE_{EAP}</th>
<th>PROMIS T-score</th>
<th>PROMIS SE_{T}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP1</td>
<td>5-Always</td>
<td>1-Never</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>-0.30</td>
<td>0.43</td>
<td>47</td>
<td>4.3</td>
</tr>
<tr>
<td>RP2</td>
<td>4-Often</td>
<td>2-Rarely</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>0.13</td>
<td>0.42</td>
<td>51.3</td>
<td>4.2</td>
</tr>
<tr>
<td>RP3</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>8</td>
<td>0.31</td>
<td>0.36</td>
<td>53.1</td>
<td>3.6</td>
</tr>
<tr>
<td>RP4</td>
<td>1-Never</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>3-Sometimes</td>
<td>8</td>
<td>0.49</td>
<td>0.36</td>
<td>54.9</td>
<td>3.6</td>
</tr>
</tbody>
</table>

*EAPs (expected a posteriori) are estimated using a Bayesian approach which finds the expected value of the posterior distribution. EAPs are in the metric of $\theta$ which follows a standard normal distribution $N(0, 1)$. PROMIS T-scores are normed with a mean of 50 and SD of 10 in the general US population.
### PROMIS Anxiety Sum Scores vs IRT EAP & T-Scores

<table>
<thead>
<tr>
<th>Response Pattern</th>
<th>Worried</th>
<th>Terrified</th>
<th>Nervous</th>
<th>Uneasy</th>
<th>Sum Score</th>
<th>EAP*</th>
<th>SE$_{EAP}$</th>
<th>PROMIS T-score</th>
<th>PROMIS SE$_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP1</td>
<td>5-Always</td>
<td>1-Never</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>-0.30</td>
<td>0.43</td>
<td>47</td>
<td>4.3</td>
</tr>
<tr>
<td>RP2</td>
<td>4-Often</td>
<td>2-Rarely</td>
<td>1-Never</td>
<td>1-Never</td>
<td>8</td>
<td>0.13</td>
<td>0.42</td>
<td>51.3</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP3</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>1-Never</td>
<td>8</td>
<td>0.31</td>
<td>0.36</td>
<td>53.1</td>
<td>3.6</td>
</tr>
<tr>
<td>RP4</td>
<td>1-Never</td>
<td>1-Never</td>
<td>3-Sometimes</td>
<td>3-Sometimes</td>
<td>8</td>
<td>0.49</td>
<td>0.36</td>
<td>54.9</td>
<td>3.6</td>
</tr>
</tbody>
</table>

*EAPs (expected a posteriori) are estimated using a Bayesian approach which finds the expected value of the posterior distribution. EAPs are in the metric of $\theta$ which follows a standard normal distribution $N(0, 1)$. PROMIS T-scores are normed with a mean of 50 and SD of 10 in the general US population.
Association between Sum Score & IRT Score

- Pearson $r = 0.97$
Association between Sum Score & IRT Score (a Counterexample)

• Pearson $r = 0.63$

• The four made-up items differ from the original four items only in terms of their discrimination parameters.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Original $a$</th>
<th>Made-up $a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worried</td>
<td>3.03</td>
<td>0.1</td>
</tr>
<tr>
<td>Terrified</td>
<td>2.58</td>
<td>1</td>
</tr>
<tr>
<td>Nervous</td>
<td>3.4</td>
<td>5</td>
</tr>
<tr>
<td>Uneasy</td>
<td>3.66</td>
<td>10</td>
</tr>
</tbody>
</table>
IRT Scoring

• Positives
  • IRT has the advantage that it extracts all information available in the item responses (except error).
  • In contrast to CTT, IRT models each response pattern is usually associated with a unique estimate of theta (score) except the Rasch model
  • Interval level scores
  • Scores with missing data can still be estimated (however, standard error is higher relative to same response pattern with no missing).
  • Scores can be compared or combined across PRO scales with different questions/items (basis for CAT)

• Negatives
  • IRT is a complex methodology
  • IRT scoring is complex to apply (automated technology makes this simple)
  • Strict assumptions of the model
Conclusions

• We have focused on scoring, but need to emphasize the great need to design PRO measures using IRT methods/theory to result in measures that are reliable and responsive to change.

• The quality of a PRO measure is related to the attention the developer(s) took to use qualitative and quantitative methods integrating multiple perspectives throughout the development process.

• The scoring algorithm should be selected based on:
  • how the PRO measure was designed,
  • its intended use (e.g., PROMIS provides multiple short forms),
  • to be comparable with other studies that used the same PRO measure,
  • interpretability of the score.
Moving from Items to Scores: An Example from the Asthma Daily Symptom Diary

Cheryl D. Coon, PhD
Principal, Outcometrix
Acknowledgments

- PRO Consortium Asthma Working Group
- FDA Qualification Review Team
- Expert Panelists
- Adelphi Values
• The PRO Consortium Asthma Working Group is developing an asthma symptom diary for use in deriving primary or secondary endpoints in future clinical trials.

• A stand-alone quantitative pilot study generated observational data from 219 subjects for the purposes of assessing reliability and validity and determining a provisional scoring algorithm.

• A robust set of analyses were used to assess the structure of the items and to assess the tolerance of missing data across items and across days within a week.

➔ Intentional analyses were undertaken to produce interpretable scores.
In qualitative interviews, eight core symptoms of asthma were identified.

These eight symptoms are assessed via eight items that are administered in a morning diary and in an evening diary on an 11-point numeric rating scale:

1. Difficulty breathing
2. Wheezing
3. Shortness of breath
4. Pressure (heavy feeling)
5. Chest tightness
6. Chest pain
7. Cough
8. Mucus (phlegm)

None As bad as you can imagine
0 1 2 3 4 5 6 7 8 9 10
Methods

• How should item responses be combined into scores?
  • Principal components analysis (PCA)
  • Confirmatory factor analysis (CFA) (unidimensional and second-order)
  • Item response theory (IRT)
  • Inter-item correlations

• How should missing data be handled across items?
  • Cronbach’s alpha-if-item-deleted
  • Standard deviation simulation

• How should missing data be handled across days?
  • Spearman-Brown prophecy formula
  • Standard deviation simulation
• Scree plots for the morning and evening diaries showed one dominant eigenvalue (6.42-6.62), explaining 80-83% of the variance among the items.

→ There is evidence for a unidimensional structure.
Instrument Structure:
Unidimensional CFA

• All items had strong factor loadings (>0.70).
• Some fit statistics indicated adequate model fit (i.e., CFI, NNFI, WRMR), while RMSEA values were much larger than expected (0.17-0.23).
• Modification indices, which indicate excess relationship not explained by the model, were high for two item pairs: Item 4 (chest pressure) and item 5 (chest tightness) (54-101), and item 7 (cough) and item 8 (mucus/phlegm) (25-80).

→ An alternative model should be considered to see if different relationships between the ADSD items might better explain the data.

CFI = comparative fit index; NNFI = non-normed fit index; RMSEA = root mean square error of approximation; WRMR = weighted root mean square residual.
A second-order CFA was conducted that assumes the ADSD items are clustered into breathing symptoms, chest symptoms, and cough symptoms, which combine together to measure overall asthma symptom severity.

- All items had strong factor loadings (>0.80).
- Some fit statistics indicated good model fit (i.e., CFI, NNFI, WRMR), while RMSEA values improved (0.08-0.10).
- Compared to the unidimensional model, the modification indices were considerably lower (all < 15).

→ The ADSD items do measure one overarching domain, but the symptom groupings may need to be considered in the scoring algorithm.

CFI = comparative fit index; NNFI = non-normed fit index; RMSEA = root mean square error of approximation; WRMR = weighted root mean square residual.
• Slopes for many items were above thresholds for what would typically be considered unlikely (i.e., >4)
  • Item 4 (chest pressure) had the highest slopes (5.9-6.0), with item 3 (shortness of breath) and item 5 (chest tightness) also exceeding expected thresholds.

→ There is evidence for local dependence, with item 4 (chest pressure) being particularly problematic.
Instrument Structure: Inter-item Correlations

• The inter-item correlation between item 4 (chest pressure) and item 5 (chest tightness) was very high (0.94-0.95)
  • Consideration of the scatterplots showed that participants were providing nearly identical responses to the two items, thus indicating potential redundancy.

→ One of these items can be deleted without losing distinct information.
• How should the eight ADSD items be grouped together to form scores?
  • While the second-order CFA revealed that sets of items are nested within subdomains (i.e., breathing symptoms, chest symptoms, cough symptoms), the items are well-related and appropriate to combine to form one overall asthma total symptom score.

• Should all eight ADSD items be retained?
  • No, participant data demonstrate that item 4 (chest pressure) and item 5 (chest tightness) are redundant among respondents.
  • **Delete chest pressure**, as it showed some evidence of race and ethnic differential item functioning, and clinical experts associate “pressure” with cardiovascular rather than respiratory issues.
• What weighting scheme should be used to construct the total score?
  • Two alternatives were considered: (1) a total average score across all seven items (i.e., equal weighting) and (2) a total average score based on the worst symptom score within each of the three subdomains.
  • Reliability and validity analyses showed little difference between these two scoring algorithms.
  • In the absence of superiority of one of the scoring algorithms over the other, the total symptom score based on equal item weighting was selected because **simple scoring is preferred where possible**, especially for scoring feasibility and interpretation within clinical practice.
• ADSD scores computed based on only two observed items can be considered reliable (alpha>0.7).

• Reliability increases as more items are observed and exceeds 0.9 with four observed items.

• Beyond four items, the incremental benefit of additional observed items on reliability is less pronounced.

→ Highly reliable ADSD scores can be computed using observed data from four or more items.
The pattern of standard deviations of ADSD scores was examined as one item per person was randomly set to missing in an incrementally increasing fashion.

The standard deviation with no missing data (i.e., all seven items observed) did not differ noticeably from the standard deviation when one or two items were missing.

The standard deviation of ADSD scores gradually increased as four or more items were missing.

The stability of ADSD scores begins to suffer when four items are missing.
• How should missing data be handled at the item-level?
  • While sufficiently reliable scores could be produced with only two items, when more than 3 items were missing, there was a noticeable increase in the variability of scores.
  • Thus, 4 items out of 7 must be non-missing to be able to compute a total score.
Number of Days Required: Spearman-Brown Prophecy Formula

- ADSD data from a single day provide reliable scores (ICC>0.7).
- Reliability increases markedly when data from two days are considered (ICC>0.8).
- Reliability is even stronger when four or more days are considered (ICC>0.9).
- Beyond four days the incremental benefit of additional days’ data on reliability is less pronounced.

⇒ Highly reliable weekly scores can be computed using four or more days of ADSD data.

ICC = intra-class correlation coefficient.
The pattern of standard deviations of ADSD weekly scores was examined as one day per person was randomly set to missing in an incrementally increasing fashion.

The standard deviation with no missing data (i.e., all seven days observed) did not differ noticeably from the standard deviation when one or two days of data were missing.

The standard deviation of ADSD weekly scores gradually increased as four or more days of data were missing.

The stability of weekly scores begins to suffer when four days of data are missing.
• How should missing data be handled at the day-level?
  • While sufficiently reliable scores could be produced with only one day of data, when more than 3 days were missing, there was a noticeable increase in the variability of scores.
  • Thus, 4 days out of 7 must be non-missing to be able to compute a weekly score.
Summary

• **Be Intentional**
  • Decisions were based on analyses carefully designed to inform those decisions.
  • Decisions were made with the clinical context in mind.

• **Be Interpretable**
  • There were multiple scoring algorithms that made sense psychometrically, but the final decision was based on parsimony and the interpretability of resulting scores.
Panel Discussion and Q & A

Moderator
– Steve Blum, MBA, MA – Director, Patient-Reported Outcomes, GlaxoSmithKline

Presenters
– Kathy Wyrwich, PhD – Senior Research Advisor, Eli Lilly and Company
– Bryce Reeve, PhD – Professor, University of North Carolina
– Cheryl D. Coon, PhD – Principal, Outcometrix

Panelists
– Laura Lee Johnson, PhD – Deputy Director, Division of Biometrics III, Office of Biostatistics, Office of Translational Sciences, CDER, FDA