Psychometrics – Rasch Item Response Theory vs. Classical Test Theory

Joseph C. Cappelleri
Pfizer Inc
joseph.c.cappelleri@pfizer.com

Kathleen W. Wyrwich
Evidera
Kathy.wyrwich@evidera.com

Patient Reported Outcomes and Clinical Outcome Assessment Summit
May 6-7, 2014
Philadelphia, Pennsylvania
Outline

- **Classical Test Theory**
  - Assumptions
  - Item difficulty, discrimination, curves, weighting
  - Reliability
  - Sample size
  - Example

- **Item Response Theory**
  - Assumptions
  - Item characteristic curves, information, fit
  - Differential item functioning
  - Sample size
  - Example
CLASSICAL TEST THEORY
Assumptions

• Assumes each person has true score on a concept of interest
  – Observed score = True score + Error
  – Obtained if there were no errors in measurement
  – Expected over an infinite number of independent administrations
  – True score not observed but estimated by observed score

• Key assumptions
  – Random errors are normally distributed
  – Random errors are uncorrelated with true score
  – Expected value of error is zero
Item Difficulty

• Consider a set of binary items (can be extended to ordinal items)

• Item difficulty is measured by the proportion of respondents who “endorse” an item (here “endorsing” implies a favorable response)

• Items with high proportions of endorsement are easy items while items with low proportions of endorsement are difficult items

• Total score for an individual is based on how many items endorsed

• Items with proportions of 0 or 1 are useless because they do not differentiate among individuals on the concept of interest

• Best to create items with varying difficulty with an average proportion of endorsement across items of 0.50
Item Discrimination

- Proportion of endorsement (item difficulty) and the “extreme group method” can be used to calculate an *item discrimination index*

- The more the item discriminates among subjects with different attributes, the higher its discrimination index

- The opportunity of an item to have the highest discrimination index occurs when its proportion of endorsement is 0.50
Item Discrimination Index – Extreme Group Method

• Step 1: Partition subjects who have the highest and lowest overall scores into upper and lower groups
  – E.g., upper group: top 25%, lower group: bottom 25%

• Step 2: Determine the proportion who endorsed each item in the upper and lower groups

• Step 3: Subtract this pair of proportions from the two groups to arrive at a discrimination index for each item
## Item Discrimination Index - Illustration

<table>
<thead>
<tr>
<th>Item</th>
<th>Proportion Endorsed for Upper Group</th>
<th>Proportion Endorsed for Lower Group</th>
<th>Item Discrimination Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.20</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>0.70</td>
<td>0.65</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>0.70</td>
<td>-0.60</td>
</tr>
</tbody>
</table>

Item 1 is the best at discriminating  
Item 2 is the second best  
Item 3 is poor at discriminating  
Item 4 is very poor
Item Curves

• Provides more fine-grained information on an item than the overall proportion endorsed or discrimination index

• Produced by plotting the percentage of subjects choosing each response option on the vertical axis by the total score on the horizontal axis (expressed as a percentile)

• A good item has its probability of endorsing increasing monotonically with increasing total score (e.g., by showing an S-shaped curve)
Item 1: Equally good at discriminating across the continuum of the attribute
Item 2: Discriminates better at the lower end than at the upper end of the attribute
Item 3: Discriminates better at upper end, especially between 70th and 80th percentiles
Corrected Item-to-Total Correlation

• Another assessment of item discrimination

• Measures how well an item correlates with the sum of the remaining items

• Best to have moderate-to-high correlations

• Items with low correlation indicate that they do not go with the rest of the items
Item Weighting

- Differential when item are given more weight or less weight when being combined into a total score
  - Three ways to assign differential weights: item reliability, factor loadings, corrected item-to-total coefficients

- This is in contrast to giving each item equal weight
  - Each item contributes equally
  - Generally preferred strategy when items are substantially inter-correlated in measuring a single concept

- Items can be averaged or summed to produce total (raw) scores
  - Scores can be linearly transformed to a Z-score so that the mean is 0 and standard deviation of 1, analogous to the ability parameter (“theta”) metric in item response theory
Reliability

• Internal consistency – Cronbach’s alpha
  – If items are measuring the same concept, they should elicited similar response
  – Function of average inter-item correlation and number of items

• Test-retest – intraclass correlation
  – Captures the stability or reproducibility of the measure
  – Correlation of measure on two occasions between which there is no change
Sample Size Considerations

• Samples as small as 30 individuals can provide useful descriptive information about the psychometric performance of measures
  – Based on empirical evidence and experience as well as knowledge of statistical theory

• Multivariate methods, such as exploratory factor analysis and confirmatory factor analysis, can be considered but require larger samples
Example

• Self-Esteem And Relationship (SEAR) questionnaire for men with erectile dysfunction

• 14 items with response options
  almost never/never
  a few times (much less than half the time)
  sometimes (about half the time)
  most times (much more than half the time)
  almost always/always

• Higher score signify a more favorable response (1, least favorable; 5, most favorable)
## Item-Level Discriminant Tests: SEAR Questionnaire

<table>
<thead>
<tr>
<th>SEAR Item</th>
<th>Item-to-Total Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain: Sexual Relationship Satisfaction</td>
</tr>
<tr>
<td>1. I felt relaxed about initiating sex with my partner</td>
<td>0.68</td>
</tr>
<tr>
<td>2. I felt confident that during sex my erection would last long enough</td>
<td>0.74</td>
</tr>
<tr>
<td>3. I was satisfied with my sexual performance</td>
<td>0.85</td>
</tr>
<tr>
<td>4. <em>I felt that sex could be spontaneous</em></td>
<td>0.62</td>
</tr>
<tr>
<td>5. I was likely to initiate sex</td>
<td>0.63</td>
</tr>
<tr>
<td>6. I felt confident about performing sexually</td>
<td>0.82</td>
</tr>
<tr>
<td>7. I was satisfied with our sex life</td>
<td>0.82</td>
</tr>
<tr>
<td>8. My partner was unhappy with the quality of our sexual relations</td>
<td>0.57</td>
</tr>
<tr>
<td>9. I had good self-esteem</td>
<td>0.48</td>
</tr>
<tr>
<td>10. I felt like a whole man</td>
<td>0.56</td>
</tr>
<tr>
<td>11. <em>I was inclined to feel that I am a failure</em></td>
<td>0.37</td>
</tr>
<tr>
<td>12. I felt confident</td>
<td>0.51</td>
</tr>
<tr>
<td>13. <em>My partner was satisfied with our relationship in general</em></td>
<td>0.55</td>
</tr>
<tr>
<td>14. I was satisfied with our relationship in general</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Classical Test Theory (CTT) vs. Item Response Theory (IRT)

- CTT: Focus most often is on the total score

- IRT: Focus is on entire pattern of responses to all test items by an individual

- CTT and IRT provide different yet useful and complementary ways to examine responses to a series of items
ITEM RESPONSE THEORY
What is Item Response Theory?

- A statistical theory consisting of mathematical models expressing the probability of a particular response to a scale item as a function of the (latent or unobserved) attribute of the person and of certain parameters or characteristics of the item
Assumptions

• Unidimensionality: items tap only one dimension

• Local independence: condition on the same attribute level, items are not correlated

• Model fit: fit between model estimates and observed data (residual for item fit and also for person fit)
Thick lines indicate, from left to right, the location of items of increasing difficulty.

Arrows indicate the location of subjects A, B, and C on the ability continuum.
Dichotomous Item Response Models  
(Attribute $\theta$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Mathematical Form</th>
<th>Item Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-parameter logistic</td>
<td>$P_{ij}(\theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$</td>
<td>Difficulty ($b$)</td>
</tr>
<tr>
<td></td>
<td>$\ln \left[ \frac{P_{ij}(\theta_j)}{1 - P_{ij}(\theta_j)} \right] = (\theta_j - b_i)$</td>
<td></td>
</tr>
<tr>
<td>Two-parameter logistic</td>
<td>$P_{ij}(\theta_j) = \frac{1}{1 + e^{-\alpha_i(\theta_j - b_i)}}$</td>
<td>Difficulty ($b$), Discrimination ($\alpha$)</td>
</tr>
<tr>
<td>Three-parameter logistic</td>
<td>$P_{ij}(\theta_j) = c + (1 - c) \left[ \frac{1}{1 + e^{-\alpha_i(\theta_j - b_i)}} \right]$</td>
<td>Difficulty ($b$), Discrimination ($\alpha$), Guessing ($c$)</td>
</tr>
</tbody>
</table>
Item Characteristic Curves for Two Items of Differing Difficulty: Rasch Model

The graph shows two item characteristic curves for different physical functioning levels. The x-axis represents physical functioning (θ) ranging from -3 to 3, while the y-axis represents the probability of a positive response. Two items are depicted:

- **Item A: Walk**
- **Item B: Run**

The curves demonstrate how the probability of a positive response changes with varying physical functioning levels. The solid line represents Item A: Walk, and the dashed line represents Item B: Run.
Item Characteristic Curves for Two Items of Differing Discrimination and Difficulty

![Graph showing item characteristic curves for two items of differing discrimination and difficulty. The x-axis represents physical functioning (θ) ranging from -3 to 3, and the y-axis represents the probability of a positive response. The graph includes two curves, one for Item A: Walk and one for Item B: Run. Key points include (θ=0.70, p=0.80) for Item A and (θ=1.93, p=0.80) for Item B.](image-url)
Item Characteristic Curves for Two Items of Differing Intercepts, Difficulty, and Discrimination

- **Item A: Walk**
- **Item B: Run**

Probability of Positive Response

Physical Functioning ($\theta$)

POOR

EXCELLENT
Polytomous Response Models

- Partial credit model – generalization of the one-parameter (Rasch) dichotomous IRT model
  - Category threshold parameters are akin to difficulty thresholds in binary IRT models
  - These parameters reflect the attribute level at which a particular response category of an item becomes as likely (50% chance) to be responded to as the previous category
  - Rating scale model is a special case

- Graded response model – extension of the two-parameter dichotomous IRT model

\[ P_{ik} (\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{i(k+1)})}} \]
Ashworth Scale: 
Clinical Assessment of Degrees of Muscle Tone
During the last seven days, how much of the time have you accomplished your daily activities as a result of your physical health?
## Common IRT Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Item Response Format</th>
<th>Model Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch / 1-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination power equal across all items. Threshold varies across items.</td>
</tr>
<tr>
<td>2-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination and threshold parameters vary across items.</td>
</tr>
<tr>
<td>Graded Response</td>
<td>Polytomous</td>
<td>Ordered responses. Discrimination varies across items.</td>
</tr>
<tr>
<td>Nominal</td>
<td>Polytomous</td>
<td>No pre-specified item order. Discrimination varies across items.</td>
</tr>
<tr>
<td>Partial Credit (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination power constrained to be equal across items.</td>
</tr>
<tr>
<td>Rating Scale (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination equal across items. Item threshold steps equal across items.</td>
</tr>
<tr>
<td>Generalized Partial Credit</td>
<td>Polytomous</td>
<td>Variation of Partial Credit Model with discrimination varying among items.</td>
</tr>
</tbody>
</table>
Item Information Function for a Good Item (Item 1) and a Poorer Item (Item 2)

\[ I(\theta)_i = a_i^2 P_i (1 - P_i) \]
Notes:
1. More easy items than difficult ones; items on moderate activity needed
2. Some items have the same difficulty
3. Patients tend to cluster at the higher end of the scale, indicating that they will endorse most items
Fit Indices

- More common to assess item fit than model fit
- Key concept is residual
- Monotonicity of average attribute estimates with increases of category level of item
- Rasch models have input and output mean square statistics for item fit (and also person fit)
- Rasch models offer person separation (reliability) index
Item and Person Fit Indices in Rasch Models

- Infit mean square statistic is more sensitive to pattern of responses to items targeted on the person and vice versa (influenced by response patterns).

- Output mean square statistic is more sensitive to responses to items with difficulty far from a person and vice versa (imputed responses, lucky guesses, careless mistakes).

- Values between 0.5 and 1.5 are in the acceptable range.

- When values are greater than 1.0, the observed variance is more than expected.
  - A value of 1.7 indicates 70% more variation in the observed data than the Rasch model predicted.

- When values are less than 1.0, the observed variance is less than expected.
  - A value of 0.4 indicates 60% [i.e., 100%\(1 – 0.4\)] less variation in the observed response pattern than was modeled.
Person Separation Index

• Reliability index that reflects how accurately or precisely the scores separate or discriminate among persons

• A summary measure based on genuine person separation (true score variance) relative to it and also measurement error (observed score variance)

• Counterpart is Cronbach’s alpha from classical test theory
Differential Item Functioning (DIF)

• Uniform DIF
  – Item shows the same amount of DIF between groups whatever the level of the total scale score

• Non-uniform DIF
  – Magnitude of group effect depends on the total scale score

• Approaches exist (such as logistic regression) for evaluation of DIF

• Does not necessarily imply item bias (it depends)
Factors Influencing Sample Size

• Choice of model

• Number of response categories

• Study purpose

• Sample distribution of respondents

• Number of items

• Generally 250 to 500 respondents (less for Rasch models)
Example

- National Eye Institute-Visual Function Questionnaire (NEI-VFQ)
- Consider its six-item near-vision subscale
- Each item has the same set of response options
  - 0 = stop doing because of eyesight
  - 25 = extreme difficulty
  - 50 = moderate difficulty
  - 75 = a little difficulty
  - 100 = no difficulty at all
- Domain score ranges from 0 (worst) to 100 (best)
Rasch Model Example: Near-vision Subscale of NEI-VFQ

<table>
<thead>
<tr>
<th>Item</th>
<th>Item (from most to least difficult)</th>
<th>Item Difficulty</th>
<th>Infit Mean Square</th>
<th>Outfit Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Small Print</td>
<td>1.45</td>
<td>1.03</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Seeing Close Up</td>
<td>0.93</td>
<td>0.88</td>
<td>1.13</td>
<td></td>
</tr>
<tr>
<td>Reading Newsprint</td>
<td>0.77</td>
<td>0.71</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Reading Bills/Mail</td>
<td>-0.34</td>
<td>0.92</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Findings objects on shelf</td>
<td>-0.78</td>
<td>1.28</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>Shaving/hair styling/makeup</td>
<td>-2.02</td>
<td>1.39</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td><strong>Mean across items</strong></td>
<td><strong>0.00</strong></td>
<td><strong>1.04</strong></td>
<td><strong>0.96</strong></td>
<td></td>
</tr>
</tbody>
</table>
Probability Curves of Three Items: Near-vision Subscale of NEI-VFQ


Do CTT Results = Rasch /IRT Findings?

Kathy Wyrwich, PhD

May 7, 2014
Two Case Examples

1. **PedsQL**
   - Pediatric Quality of Life Inventory (Version 4.0)

2. **SGRQ vs. SGRQ-C**
   - St. George Respiratory Questionnaire-COPD
PedsQL (Version 4.0) Generic Core Scales

- The Generic Core Scales include parallel child self-report and parent proxy-report formats

- Child self-report includes ages 5-7, 8-12, and 13-18 years

- Parent proxy-report includes ages 2-4, 5-7, 8-12, and 13-18 years

- Four domains of functioning:
  - Physical (8 items), Emotional (5 items), Social (5 items), School (5 items)

- The 4 domains can be summed = overall HRQoL (Total Score)

- Emotional + Social + School = Psychosocial Functioning

Example items:

*In the past ONE month, how much of a problem has your child had with ...*

<table>
<thead>
<tr>
<th>Physical Functioning (problems with...)</th>
<th>Never</th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Walking more than one block</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Running</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Participating in sports activity or exercise</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
PedsQL 4.0 Generic Core Scales

- Used in over 60 publications, often demonstrating validation in a specific pediatric population
- Used in over 1,000 research studies
CTT

- Classical test theory (CTT) has served as the main approach to the development and evaluation of rating scales.

- CTT focuses predominantly on person-level statistics, such as means and standard deviations, and on test-level statistics such as:
  - reliability
  - known-groups
  - construct validity
# CTT Results (Varni et al. *Medical Care* 2001)

## Ceiling/Floor Effects, Reliability

### Table 1. Scale Descriptives and Internal Consistency Reliability for PedsQL 4.0 Child Self-Report and Parent Proxy-Report

<table>
<thead>
<tr>
<th>Scale</th>
<th>Number of Items</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Percentage Floor Ill/Healthy</th>
<th>Percentage Ceiling Ill/Healthy</th>
<th>Total Sample</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-Report</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>23</td>
<td>960</td>
<td>79.62</td>
<td>15.26</td>
<td>0.0/0.0</td>
<td>1.9/7.2</td>
<td>0.88</td>
<td>0.85–0.90</td>
</tr>
<tr>
<td>Physical health</td>
<td>8</td>
<td>959</td>
<td>80.19</td>
<td>19.30</td>
<td>0.0/0.0</td>
<td>13.1/25.8</td>
<td>0.80</td>
<td>0.77–0.85</td>
</tr>
<tr>
<td>Psychosocial health</td>
<td>15</td>
<td>958</td>
<td>79.37</td>
<td>15.70</td>
<td>0.0/0.0</td>
<td>5.2/12.0</td>
<td>0.83</td>
<td>0.78–0.85</td>
</tr>
<tr>
<td>Emotional functioning</td>
<td>5</td>
<td>958</td>
<td>78.10</td>
<td>20.66</td>
<td>0.3/0.8</td>
<td>22.4/29.8</td>
<td>0.73</td>
<td>0.70–0.78</td>
</tr>
<tr>
<td>Social functioning</td>
<td>5</td>
<td>958</td>
<td>84.09</td>
<td>18.50</td>
<td>0.0/0.0</td>
<td>33.2/47.1</td>
<td>0.71</td>
<td>0.67–0.74</td>
</tr>
<tr>
<td>School functioning</td>
<td>5</td>
<td>933</td>
<td>75.87</td>
<td>19.71</td>
<td>0.3/0.5</td>
<td>13.0/23.1</td>
<td>0.68</td>
<td>0.59–0.75</td>
</tr>
<tr>
<td><strong>Proxy-Report</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>23</td>
<td>1622</td>
<td>80.87</td>
<td>16.73</td>
<td>0.2/0.0</td>
<td>4.1/10.3</td>
<td>0.90</td>
<td>0.89–0.90</td>
</tr>
<tr>
<td>Physical health</td>
<td>8</td>
<td>1613</td>
<td>81.38</td>
<td>23.18</td>
<td>2.3/0.0</td>
<td>18.5/39.6</td>
<td>0.88</td>
<td>0.85–0.89</td>
</tr>
<tr>
<td>Psychosocial health</td>
<td>15</td>
<td>1621</td>
<td>80.58</td>
<td>16.52</td>
<td>0.2/0.0</td>
<td>5.6/13.8</td>
<td>0.86</td>
<td>0.80–0.86</td>
</tr>
<tr>
<td>Emotional functioning</td>
<td>5</td>
<td>1622</td>
<td>77.95</td>
<td>20.67</td>
<td>1.4/0.1</td>
<td>19.5/29.5</td>
<td>0.77</td>
<td>0.69–0.80</td>
</tr>
<tr>
<td>Social functioning</td>
<td>5</td>
<td>1615</td>
<td>85.38</td>
<td>19.17</td>
<td>0.5/0.0</td>
<td>34.4/58.1</td>
<td>0.75</td>
<td>0.68–0.78</td>
</tr>
<tr>
<td>School functioning</td>
<td>5</td>
<td>1417</td>
<td>77.80</td>
<td>22.00</td>
<td>1.7/0.3</td>
<td>15.5/34.5</td>
<td>0.76</td>
<td>0.57–0.78</td>
</tr>
</tbody>
</table>

Note: Percentage Floor/Ceiling = the percentage of scores at the extremes of the scaling range.
<table>
<thead>
<tr>
<th>Scale</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>df</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>662</td>
<td>74.22</td>
<td>18.40</td>
<td>2, 1575</td>
<td>128.39</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>199</td>
<td>80.42</td>
<td>15.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>717</td>
<td>87.61</td>
<td>12.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>653</td>
<td>73.28</td>
<td>27.02</td>
<td>2, 1566</td>
<td>91.72</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>199</td>
<td>81.81</td>
<td>20.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>717</td>
<td>89.32</td>
<td>16.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychosocial health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>661</td>
<td>74.80</td>
<td>18.16</td>
<td>2, 1574</td>
<td>98.95</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>199</td>
<td>79.56</td>
<td>15.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>717</td>
<td>86.58</td>
<td>12.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>661</td>
<td>73.05</td>
<td>23.27</td>
<td>2, 1575</td>
<td>38.90</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>199</td>
<td>78.82</td>
<td>18.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>718</td>
<td>82.64</td>
<td>17.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>657</td>
<td>79.77</td>
<td>21.91</td>
<td>2, 1568</td>
<td>72.55</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>198</td>
<td>83.58</td>
<td>18.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>716</td>
<td>91.56</td>
<td>14.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronically ill</td>
<td>601</td>
<td>71.08</td>
<td>23.99</td>
<td>2, 1376</td>
<td>73.25</td>
<td>0.001</td>
</tr>
<tr>
<td>Acutely ill</td>
<td>167</td>
<td>74.74</td>
<td>20.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>611</td>
<td>85.47</td>
<td>17.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: For each scale, each subgroup significantly different from every other subgroup P < 0.05.
**TABLE 7**
One-Way ANOVA Values Comparing Healthy Children and Children with Cancer (On- or Off-Treatment): PedsQL 4.0 Generic Core (Acute Version) Child Report

<table>
<thead>
<tr>
<th>Scale</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Difference</th>
<th>df</th>
<th>f</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>105</td>
<td>68.92</td>
<td>15.97</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>41</td>
<td>70.88</td>
<td>17.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>73</td>
<td>77.66</td>
<td>15.25</td>
<td>a &lt; c**, a, b &lt; d***</td>
<td>3320</td>
<td>16.78</td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>105</td>
<td>83.41</td>
<td>14.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>105</td>
<td>65.54</td>
<td>23.14</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>41</td>
<td>73.17</td>
<td>18.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>73</td>
<td>80.01</td>
<td>18.66</td>
<td>a &lt; c, d***</td>
<td>3320</td>
<td>13.86</td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>105</td>
<td>82.60</td>
<td>19.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychosocial health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>105</td>
<td>71.04</td>
<td>15.17</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>41</td>
<td>69.74</td>
<td>19.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>73</td>
<td>76.51</td>
<td>16.03</td>
<td>a, b &lt; d***, c &lt; d*</td>
<td>3320</td>
<td>14.22</td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>105</td>
<td>84.03</td>
<td>15.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>105</td>
<td>68.81</td>
<td>21.24</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>41</td>
<td>69.15</td>
<td>22.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>73</td>
<td>77.67</td>
<td>20.19</td>
<td>a &lt; c*, a &lt; d***, b &lt; d**</td>
<td>3320</td>
<td>10.30</td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>105</td>
<td>83.12</td>
<td>18.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>105</td>
<td>77.19</td>
<td>18.29</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>41</td>
<td>70.98</td>
<td>24.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>73</td>
<td>79.64</td>
<td>19.78</td>
<td>a &lt; d**, b &lt; d***</td>
<td>3320</td>
<td>7.81</td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>105</td>
<td>86.74</td>
<td>18.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School functioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Txₐ</td>
<td>92</td>
<td>66.22</td>
<td>19.60</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Off Tx ≤ 12ₐ</td>
<td>37</td>
<td>69.32</td>
<td>22.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Tx &gt; 12ₐ</td>
<td>62</td>
<td>71.41</td>
<td>18.12</td>
<td>a &lt; d***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthyₐ</td>
<td>63</td>
<td>79.39</td>
<td>18.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA: analysis of variance; SD: standard deviation; On Tx, on-treatment sample; Off Tx ≤ 12: off-treatment ≤ 12 mos sample; Off Tx > 12: off-treatment > 12 mos/long-term survivor sample.
### Convergent/Discriminant Validity (Varni et al. 2001)

**Table 4. Pearson Correlations Between Indicators of Morbidity and Child Self-Report and Parent Proxy-Report**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Care Needed</th>
<th>School Missed</th>
<th>Work Missed</th>
<th>Work Routine Impact</th>
<th>Work Concentration Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
<td>( n )</td>
<td>( r )</td>
<td>( n )</td>
<td>( r )</td>
</tr>
<tr>
<td>Self-Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-0.24</td>
<td>865</td>
<td>-0.22</td>
<td>598</td>
<td>NS</td>
</tr>
<tr>
<td>Physical</td>
<td>-0.27</td>
<td>864</td>
<td>-0.21</td>
<td>596</td>
<td>-0.20</td>
</tr>
<tr>
<td>Psychosocial</td>
<td>-0.18</td>
<td>863</td>
<td>-0.20</td>
<td>598</td>
<td>NS</td>
</tr>
<tr>
<td>Emotional</td>
<td>-0.13</td>
<td>863</td>
<td>-0.13</td>
<td>597</td>
<td>NS</td>
</tr>
<tr>
<td>Social</td>
<td>-0.13</td>
<td>863</td>
<td>-0.11</td>
<td>598</td>
<td>NS</td>
</tr>
<tr>
<td>School</td>
<td>-0.17</td>
<td>841</td>
<td>-0.24</td>
<td>591</td>
<td>-0.21</td>
</tr>
<tr>
<td>Proxy-Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-0.38</td>
<td>1522</td>
<td>-0.29</td>
<td>816</td>
<td>-0.30</td>
</tr>
<tr>
<td>Physical</td>
<td>-0.38</td>
<td>1515</td>
<td>-0.26</td>
<td>811</td>
<td>-0.30</td>
</tr>
<tr>
<td>Psychosocial</td>
<td>-0.31</td>
<td>1521</td>
<td>-0.25</td>
<td>816</td>
<td>-0.23</td>
</tr>
<tr>
<td>Emotional</td>
<td>-0.19</td>
<td>1521</td>
<td>-0.12</td>
<td>817</td>
<td>-0.20</td>
</tr>
<tr>
<td>Social</td>
<td>-0.27</td>
<td>1518</td>
<td>-0.13</td>
<td>812</td>
<td>-0.24</td>
</tr>
<tr>
<td>School</td>
<td>-0.28</td>
<td>1324</td>
<td>-0.33</td>
<td>786</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

Note: All displayed correlation coefficients (\( r \)) are significant at \( P < 0.01 \).
Rasch and Item Response Theory

- Modern psychometric approaches, such as Item Response Theory, which includes Rasch analysis, bring increased attention to item-level statistics.

- The Rasch measurement model determines the relationship between a person’s performance on any item in a scale and the level of the trait that person possesses.

- Rasch measurement model provides empirical evidence to support how well the items on a scale measure the construct of interest—providing insights on the validity of a scale.
Rasch Analysis of the PedsQL 4.0

- Amin et al. Rasch analysis of the PedsQL: an increased understanding of the properties of a rating scale. *Journal of Clinical Epidemiology* 2012
  - Examined the PedsQL 4.0 Generic Core Scales (parent report) using Rasch analysis
  - Data from a Canadian study of 376 parents of children aged 2-17 years old on active cancer treatment
The 23-item PedsQL 4.0 scale did not fit the Rasch model (Amin et al. *Journal of Clinical Epidemiology* 2012)

- $\chi^2$ test yielded there was a significant difference between observed data and expected data based on the Rasch model for the 23-item construct of HRQoL

- Important limitations in the PedsQL including:
  - disordered item thresholds
  - local dependency

- Rasch item curves showed that respondents did not discriminate between the five-point response categories

- Person-Item targeting was also problematic
Item Thresholds: The Ideal Case
Problematic Item Thresholds

![Graph showing probability distributions for different response frequencies: Never, Sometimes, Almost Never, and Often, with corresponding thresholds for person location (logits).]
Disordered Thresholds: 23-Item PedsQL
Local Dependency: 23-Item PedsQL

**Emotional Functioning (problems with...)**

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Feeling afraid or scared</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Feeling sad or blue</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Feeling angry</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4. Trouble sleeping</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5. Worrying about what will happen to him or her</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Person-item threshold map of the PedsQL 4.0 (parent-report)
Additional PedsQL 4.0 Notes

- In the Digital Commons/Open Access thesis, Amin reported on Rasch analysis of the four functioning domains and found problematic items in all domains

http://digitalcommons.mcmaster.ca/cgi/viewcontent.cgi?article=7074&context=opendissertations
Figure 5b: Person-item threshold map for physical function summary score

Person-Item Threshold Distribution

(Grouping Set to Interval Length of 0.20 making 45 Groups)

- Ceiling effect (items do not discriminate where high functioning portion of sample lies)

- Floor effect (items do not discriminate where low functioning portion of sample lies)
This study revealed that, in all domains, the five response categories did not perform adequately.
Category Probability Curves Of Five Response Categories Of The PedsQL 4.0 (Jafari et al. HQLO 2012)
Two Other Rasch Analyses of PedsQL 4.0 Items with Same Conclusions

- **Korean Translation**
  - Kook S, Varni JW. Validation of the Korean version of the Pediatric Quality of Life Inventory (PedsQL™) 4.0 Generic Core Scales in school children and adolescents using the Rasch model. *Health Qual Life Outcomes* 2008;6:41.

- **Singapore Translation**
PedsQL: Do CTT Results = Rasch / IRT Findings?

- CTT was mostly affirmative, while the Rasch analyses consistently found the same problematic areas of concern

- These learning yielded better designed / conceptualized domains in PROMIS pediatric work

- The PROMIS work has demonstrated the CTT = IRT findings
Psychometric properties of the PROMIS® pediatric scales: precision, stability, and comparison of different scoring and administration options

James W. Varni · Brooke Magnus · Brian D. Stucky · Yang Liu · Hally Quinn · David Thissen · Heather E. Gross · I-Chan Huang · Darren A. DeWalt

Accepted: 23 September 2013 / Published online: 2 October 2013 © Springer Science+Business Media Dordrecht 2013

Abstract

Objectives The objectives of the present study are to investigate the precision of static (fixed-length) short forms versus computerized adaptive testing (CAT) administration, response pattern scoring versus summed score conversion, and test–retest reliability (stability) of the Patient-Reported Outcomes Measurement Information System (PROMIS®) pediatric self-report scales measuring the latent constructs of depressive symptoms, anxiety, anger, pain interference, peer relationships, fatigue, mobility, upper extremity functioning, and asthma impact with polytomous items.

Methods Participants (N = 331) between the ages of 8 and 17 were recruited from outpatient general pediatrics and subspecialty clinics. Of the 331 participants, 137 were diagnosed with asthma. Three scores based on item response theory (IRT) were computed for each respondent: CAT response pattern expected a posteriori estimates, short-form response pattern expected a posteriori estimates, and short-form summed score expected a posteriori estimates. Scores were also compared between participants with and without asthma. To examine test–retest reliability, 54 children were selected for retesting approximately 2 weeks after the first assessment. Results A short CAT (maximum 12 items with a standard error of 0.4) was found, on average, to be less precise than the static short forms. The CAT appears to have limited usefulness over and above what can be accomplished with the existing static short forms (8–10 items). Stability of the scale scores over a 2-week period was generally supported. Conclusion The study provides further information on the psychometric properties of the PROMIS pediatric scales and extends the previous IRT analyses to include precision estimates of dynamic versus static administration, test–retest reliability, and validity of administration across groups. Both the positive and negative aspects of using CAT versus short forms are highlighted.

J. W. Varni
Department of Pediatrics, College of Medicine, Texas A&M University, College Station, TX, USA

J. W. Varni (✉)
Department of Landscape Architecture and Urban Planning, College of Architecture, Texas A&M University, 3137 TAMU, College Station, TX 77843-3137, USA
e-mail: jvarni@arch.tamu.edu

B. Magnus · Y. Liu · H. Quinn · D. Thissen
Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
St. George Respiratory Questionnaire (SGRQ)  
Versus  
St. George Respiratory Questionnaire-Chronic Obstructive Pulmonary Disease (SGRQ-C)
SGRQ

- SGRQ was published in 1991 as a standardized self-completed instrument for measuring impaired health status and perceived well-being in airways disease, such as:
  - chronic obstructive pulmonary disease (COPD)
  - asthma
  - bronchiectasis
  - post tuberculosis (Jones et al. 1991)

- The SGRQ contains 50 items in three domains:
  - Symptoms
  - Activity
  - Impact

- Each provide a domain score, and the domains can be combined to produce an overall total score
## Jones et al. 1992 Validation Study
### Known-Groups Validity

| SGRQ Scores Compared with Responses to Items from the MRC Respiratory Symptoms Questionnaire |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
|                                  | Symptoms (mean ± SD)             | Activity (mean ± SD)             | Impact (mean ± SD)               | Total (mean ± SD)                |
|                                  | n                               |                                 |                                 |                                 |
| Cough for 3 months               |                                 |                                 |                                 |                                 |
| No                               | 60                              | 51.9 ± 23.7                     | 50.3 ± 23.7                     | 27.2 ± 18.8                     | 29.2 ± 17.0                     |
| Yes                              | 81                              | 70.4 ± 11.1                     | 59.4 ± 23.7                     | 43.5 ± 23.3                     | 43.5 ± 19.7                     |
| r² For model                     | 0.35                            | 0.03                            | 0.09                            | 0.11                            |
| p Value                          | < 0.0001                        | 0.026                           | < 0.0001                        | < 0.0001                        |
| Sputum for 3 months              |                                 |                                 |                                 |                                 |
| No                               | 64                              | 54.1 ± 15.2                     | 53.2 ± 25.4                     | 30.9 ± 19.9                     | 42.0 ± 18.5                     |
| Yes                              | 77                              | 69.5 ± 12.0                     | 57.4 ± 22.9                     | 42.8 ± 22.9                     | 52.1 ± 19.5                     |
| r² For model                     | 0.24                            | 0.0                             | 0.06                            | 0.06                            |
| p Value                          | < 0.0001                        | 0.3                             | 0.0015                          | 0.002                           |
| Wheeze                           |                                 |                                 |                                 |                                 |
| None                             | 14                              | 46.4 ± 12.5                     | 36.3 ± 19.3                     | 15.9 ± 13.0                     | 27.2 ± 13.2                     |
| < Daily                          | 40                              | 53.2 ± 13.2                     | 46.8 ± 22.2                     | 25.8 ± 13.4                     | 37.9 ± 13.4                     |
| Daily                            | 87                              | 69.4 ± 23.3                     | 61.7 ± 23.3                     | 46.2 ± 22.5                     | 55.2 ± 18.9                     |
| r² For model                     | 0.32                            | 0.12                            | 0.25                            | 0.25                            |
| p Value                          | < 0.0001                        | < 0.0001                        | < 0.0001                        | < 0.0001                        |

67
Construct Validity - Jones et al. 1992

<table>
<thead>
<tr>
<th></th>
<th>Symptom</th>
<th>Activity</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEV&lt;sub&gt;1&lt;/sub&gt;</td>
<td>(-)0.01†</td>
<td>(-)0.08‡</td>
<td>(-)0.07§</td>
</tr>
<tr>
<td>FVC</td>
<td>(-)0.06§</td>
<td>(-)0.10‖</td>
<td>(-)0.20‖</td>
</tr>
<tr>
<td>SaO&lt;sub&gt;2&lt;/sub&gt;, rest</td>
<td>(-)0.02†</td>
<td>(-)0.04‖</td>
<td>(-)0.01†</td>
</tr>
<tr>
<td>6-MWD</td>
<td>(-)0.07§</td>
<td>(-)0.35‖</td>
<td>(-)0.35‖</td>
</tr>
<tr>
<td>MRC dyspnea grade</td>
<td>0.13‖</td>
<td>0.50‖</td>
<td>0.44‖</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.12‖</td>
<td>0.20‖</td>
<td>0.38‖</td>
</tr>
<tr>
<td>Depression</td>
<td>0.08‡</td>
<td>0.23‖</td>
<td>0.39‖</td>
</tr>
<tr>
<td>SIP physical</td>
<td>0.12‖</td>
<td>0.38‖</td>
<td>0.48‖</td>
</tr>
<tr>
<td>SIP psychosocial</td>
<td>0.07§</td>
<td>0.28‖</td>
<td>0.42‖</td>
</tr>
<tr>
<td>SIP total</td>
<td>0.11‖</td>
<td>0.39‖</td>
<td>0.54‖</td>
</tr>
</tbody>
</table>

* Correlation expressed as proportion of the variance in SGRQ score ($r^2$) attributable to variance in the other variable. FEV<sub>1</sub> and FVC expressed as percentage of predicted normal. (-) Negative correlation between variables.
† NS, not significant, $p > 0.01$. 
Responsiveness of Total Score - Jones et al. 1992

Fig. 1. Change in total SGRQ score and change in MRC dyspnea grade in 133 patients with chronic airflow limitation studies over 1 yr.
Meguro et al. 2006

- Over a decade use, the developers of the SGRQ became aware that some of this instrument’s 50 items were:
  - confusing to COPD patients as demonstrated by much more missing data and/or
  - relatively weaker relationships when compared to other items in their domain

- An improvement effort commenced over a decade after the original instrument’s development, using SGRQ data from large studies in COPD

- The resulting SGRQ-C is a shorter 40-item version of the SGRQ that has:
  1. removed the items with the weakest measurement properties in the original instrument, and
  2. revised some of the best items

- Moreover, the SGRQ-C yields domain and total scores that are directly comparable with the original SGRQ
The SGRQ-C Adaptation Process Meguro et al. 2006

- The SGRQ developers were aware of consistently missing responses for the item measuring *impact of chest trouble on paid employment*, and this item was removed.

- Rasch analyses were conducted within each of the SGRQ domains of symptoms, activity and impact.

- The Rasch model examined each of the patient’s responses, and through probabilistic modeling determines the patient’s “ability” level in each domain compared to other patients in the dataset under the assumption that all of the domain’s items measure the same concept.

- With this simultaneous interaction between the domain items and the patient’s ability level, each item can be tested to determine how well it fit the domain’s structure.
Decision Process of SGRQ Item Removals for Creating the SGRQ-C

- **A priori** criteria for removing items were:
  1. Items with multiple response categories were removed if ordered thresholds could not be achieved through combining categories and rescoring.
  2. Dichotomous items were removed if the statistical tests showed that misfit was present. However, an item that fitted less well could be retained if its removal caused a previously ‘good’ item to misfit.
  3. An item could be removed if its item location was counter-intuitive; for example an item representing severe impairment that had more affirmative responses than an item representing mild impairment.
Decision Process of SGRQ Item Removals for Creating the SGRQ-C

- Where more than one item met the removal criteria to a similar extent, item content was used to inform removal decisions.

- The PSI (scale reliability index) for each component was calculated before and after removing each item.

- If removal worsened the PSI, the item was retained. This procedure was used to maintain the overall discriminatory function of each SGRQ component.

- Finally, the PSI was calculated for the whole set of items retained in the revised version and compared with that for the original 50 item set.
<table>
<thead>
<tr>
<th>Item</th>
<th>Threshold between response categories (logits)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Categories 1-2</td>
</tr>
<tr>
<td>1</td>
<td>-1.55</td>
</tr>
<tr>
<td>2</td>
<td>-0.69</td>
</tr>
<tr>
<td>3</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>-0.65</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
</tr>
<tr>
<td>6</td>
<td>0.66</td>
</tr>
<tr>
<td>7</td>
<td>-1.90</td>
</tr>
</tbody>
</table>

**Item 1:** Over the last year I have coughed: most days a week, several days a week, a few days a month, only with chest infections, not at all

**Item 2:** Over the last year I have brought up phlegm (sputum): responses as for Item 1

**Item 3:** Over the last year I have had shortness of breath: responses as for Item 1

**Item 4:** Over the last year I have had attacks of wheezing: responses as for Item 1

**Item 5:** During the last year how many severe or very unpleasant attacks of chest trouble have you had: more than 3 attacks, 3 attacks, 2 attacks, 1 attack, no attacks

**Item 6:** How long did the worst attack of chest trouble last: a week or more, 3 or more days, 1 or 2 days, less than a day

**Item 7:** Over the last year, in an average week, how many good days (with little chest trouble) have you had: none, 1 or 2, 3 or 4, nearly every day, every day.
Symptoms Domain Solutions in SGRQ-C

- **Combining response categories for Symptoms items**
  - The tests for logical ordering of response categories in the original SGRQ showed that not all were correctly ordered, but after combining some adjacent response categories, logical ordering of response categories was achieved for all items except item 6.
    - **SGRQ Item 1.** ‘Over the last year I have coughed’: combined ‘several days a week’ and ‘a few days a month’. In the SGRQ-C these items are represented as ‘several days a week’
    - **SGRQ Item 2.** Same as Item 1
    - **SGRQ Item 3.** combined ‘several days a week’ and ‘a few days a month’ and ‘only with chest infections’
    - **SGRQ Item 4.** ‘Over the last year I have had attacks of wheezing’: No categories had to be combined
Symptoms Domain Solutions in SGRQ-C

- Combining response categories for Symptoms items - Continued

  - **SGRQ Item 5.** ‘During the last year, how many severe or very unpleasant attacks have you had’: combined ‘more than 3 attacks’ with ‘3 attacks’. In the SGRQ-C these items are represented as ‘3 or more attacks’. Also, combined ‘2 attacks’ with ‘1 attack’. In the SGRQ-C these items are represented as ‘1 or 2 attacks’

  - **SGRQ Item 6.** ‘How long did the worst attack of chest trouble last’: Logical ordering of this item could not be achieved despite combining adjacent items. It was removed

  - **SGRQ Item 7.** ‘Over the last year in an average week, how many good days (with little chest trouble) have you had’: combined with ‘3 or 4’ with ‘nearly every day’. In the SGRQ-C these items are represented as ‘most days are good’. Note: the original response category ‘1 or 2’ was not combined with any other, but in the SGRQ-C it has been renamed ‘a few good days’
Symptoms Domain Solutions in SGRQ-C

- **Renaming items in the Symptom component of the SGRQ**
  - In the original version of the SGRQ a recall period was specified for some items. This recall period has been problematic for some users, so we have removed a specific recall period and renamed the corresponding items in the SGRQ-C accordingly:
    - **SGRQ Item 1.** ‘Over the last year I have coughed’ is now in SGRQ-C as: ‘I cough’
    - **SGRQ Item 2.** ‘Over the last year I have brought up phlegm (sputum)’ is now in SGRQ-C as: ‘I bring up phlegm (sputum)”
    - **SGRQ Item 3.** ‘Over the last year I have had shortness of breath’: is now in SGRQC as: ‘I have shortness of breath’
    - **SGRQ Item 4.** ‘Over the last year I have had attacks of wheezing’ is now in SGRQ-C as: ‘I have attacks of wheezing’
    - **SGRQ Item 5.** ‘During the last year, how many severe or very unpleasant attacks have you had’: is now in SGRQ-C as: ‘How many attacks of chest trouble did you have in the last year?”
    - **SGRQ Item 6.** ‘How long did the worst attack of chest trouble last’. Is not in the SGRQ-C
    - **SGRQ Item 7.** ‘Over the last year in an average week, how many good days (with little chest trouble) have you had’ is now in SGRQ-C as: ‘How often do you have good days (with little chest trouble)?’
Overview of SGRQ to SGRQ-C Changes

- Ten items were deleted—the worst fitting items in each domain.

- For most SGRQ symptom items with the response options that originally included “a few days a week”—this response was removed.

- The one-year recall period included in many SGRQ items’ stems was deleted and replaced with simplified wording (e.g., “I cough”).

- With these item deletions and revisions to the SGRQ, the SGRQ-C scoring algorithm for each domain and the total scores were also revised to yield domain and overall scores that were nearly equivalent to the SGRQ.

- As a result, the adapted instrument has demonstrated similar test-retest reliability and change scores over 16 weeks compared to the original SGRQ with similar relationships were also observed for the SGRQ with other COPD measures, like FEV\textsubscript{1}, FVC, and six-minute walk distance (Meguro et al. 2006).
Adjustment to Compare the SGRQ-C Score with the SGRQ Score

- Adjustment is required to compare SGRQ-C scores with the SGRQ
  - Symptoms:
    - SGRQ score = (SGRQ-C x 0.99) + 0.94 units
  - Activity:
    - SGRQ score = (SGRQ-C x 0.87) + 7.01 units
  - Impacts:
    - SGRQ score = (SGRQ-C x 0.88) + 2.18 units
  - Total score:
    - SGRQ score = (SGRQ-C x 0.90) + 3.10 units
Item Map for SGRQ-C Items Identified by Domain

SGRQ Domain

- Symptom
- Activity
- Impact

Location

Q16 Brth Walk up hill
Q3 Breathless
Q15 Brth walking up stairs
Q38 walk slower than others
Q33 housework takes time
Q13 Walking around at home
Q14 Walking out on the level
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth - makes me tired
Q9 Brth makes me tired
Q8 Cough
Q7 Good days
Q6 Can't move from chair / bed
Q5 Attack f
Q4 Wheeze
Q1 Cough
Q2 Phegm
Q3 Breathless bending over
Q21 Breathless bending over
Q40, one flight of stairs stop
Q39 housework takes time
Q38 walk slower than others
Q34 housework takes time
Q33 housework takes time
Q32 Cough Brth - makes me tired
Q31 Everything is so much effort
Q30 Exercise not safe
Q29 I become frail invalid
Q28 Cough Brth - sleep
Q27 Not in control of chest
Q26 Afraid panic when out of breath
Q25 My chest is nuisance
Q24 Cough brth - embarrassing
Q23 Easily Exhausted
Q22 Cough Brth - sleep
Q21 Cough Brth - sleep
Q20 Cough Brth - makes me tired
Q19 Cough Brth - makes me tired
Q18 Cough hurts
Q17 Daily life - can't go shopping
Q16 Daily life - can't go shopping
Q15 Brth makes me tired
Q14 Walking out on the level
Q13 Walking around at home
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth makes me tired
Q9 Important problem
Q8 Cough
Q7 Good days
Q6 Can't move from chair / bed
Q5 Attack f
Q4 Wheeze
Q3 Breathless bending over
Q21 Breathless bending over
Q40, one flight of stairs stop
Q39 housework takes time
Q38 walk slower than others
Q34 housework takes time
Q33 housework takes time
Q32 Cough Brth - makes me tired
Q31 Everything is so much effort
Q30 Exercise not safe
Q29 I become frail invalid
Q28 Cough Brth - sleep
Q27 Not in control of chest
Q26 Afraid panic when out of breath
Q25 My chest is nuisance
Q24 Cough brth - embarrassing
Q23 Easily Exhausted
Q22 Cough Brth - sleep
Q21 Cough Brth - sleep
Q20 Cough Brth - makes me tired
Q19 Cough Brth - makes me tired
Q18 Cough hurts
Q17 Daily life - can't go shopping
Q16 Daily life - can't go shopping
Q15 Brth makes me tired
Q14 Walking out on the level
Q13 Walking around at home
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth makes me tired
Q9 Important problem
Q8 Cough
Q7 Good days
Q6 Can't move from chair / bed
Q5 Attack f
Q4 Wheeze
Q3 Breathless bending over
Q21 Breathless bending over
Q40, one flight of stairs stop
Q39 housework takes time
Q38 walk slower than others
Q34 housework takes time
Q33 housework takes time
Q32 Cough Brth - makes me tired
Q31 Everything is so much effort
Q30 Exercise not safe
Q29 I become frail invalid
Q28 Cough Brth - sleep
Q27 Not in control of chest
Q26 Afraid panic when out of breath
Q25 My chest is nuisance
Q24 Cough brth - embarrassing
Q23 Easily Exhausted
Q22 Cough Brth - sleep
Q21 Cough Brth - sleep
Q20 Cough Brth - makes me tired
Q19 Cough Brth - makes me tired
Q18 Cough hurts
Q17 Daily life - can't go shopping
Q16 Daily life - can't go shopping
Q15 Brth makes me tired
Q14 Walking out on the level
Q13 Walking around at home
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth makes me tired
Q9 Important problem
Q8 Cough
Q7 Good days
Q6 Can't move from chair / bed
Q5 Attack f
Q4 Wheeze
Q3 Breathless bending over
Q21 Breathless bending over
Q40, one flight of stairs stop
Q39 housework takes time
Q38 walk slower than others
Q34 housework takes time
Q33 housework takes time
Q32 Cough Brth - makes me tired
Q31 Everything is so much effort
Q30 Exercise not safe
Q29 I become frail invalid
Q28 Cough Brth - sleep
Q27 Not in control of chest
Q26 Afraid panic when out of breath
Q25 My chest is nuisance
Q24 Cough brth - embarrassing
Q23 Easily Exhausted
Q22 Cough Brth - sleep
Q21 Cough Brth - sleep
Q20 Cough Brth - makes me tired
Q19 Cough Brth - makes me tired
Q18 Cough hurts
Q17 Daily life - can't go shopping
Q16 Daily life - can't go shopping
Q15 Brth makes me tired
Q14 Walking out on the level
Q13 Walking around at home
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth makes me tired
Q9 Important problem
Q8 Cough
Q7 Good days
Q6 Can't move from chair / bed
Q5 Attack f
Q4 Wheeze
Q3 Breathless bending over
Q21 Breathless bending over
Q40, one flight of stairs stop
Q39 housework takes time
Q38 walk slower than others
Q34 housework takes time
Q33 housework takes time
Q32 Cough Brth - makes me tired
Q31 Everything is so much effort
Q30 Exercise not safe
Q29 I become frail invalid
Q28 Cough Brth - sleep
Q27 Not in control of chest
Q26 Afraid panic when out of breath
Q25 My chest is nuisance
Q24 Cough brth - embarrassing
Q23 Easily Exhausted
Q22 Cough Brth - sleep
Q21 Cough Brth - sleep
Q20 Cough Brth - makes me tired
Q19 Cough Brth - makes me tired
Q18 Cough hurts
Q17 Daily life - can't go shopping
Q16 Daily life - can't go shopping
Q15 Brth makes me tired
Q14 Walking out on the level
Q13 Walking around at home
Q12 Brth washed dressed
Q11 Brth makes me tired
Q10 Brth makes me tired
Q9 Important problem
SGRQ: Do CTT Results = Rasch /IRT Findings?

- CTT was mostly affirmative, while the Rasch analysis found the same problematic areas of concern

- These learnings yielded fewer and improved items in the SGRQ-C
Conclusions

- CTT learnings and Rasch/IRT methods can complement each other, with the latter providing greater understanding of:
  - the items in a domain, and
  - the items’ relationships to the persons being measured

- Key drawbacks to the universal use of Rasch/IRT methods in all psychometric analyses are:
  1. Sample sizes needed to run the analyses
  2. Software and expertise needed to run the analyses and properly interpret the results

- Ideally, CTT and IRT results should both indicate that a measure is truly fit for purpose as we seek tools with the precision to demonstrate important treatment effects