Reliability and Error in Measurement

Instruments developed with Classical Test Theory and Item Response Theory

Allan J. Kozlowski, Ph.D., B.Sc. (PT)
Assistant Professor of Rehabilitation Medicine
Icahn School of Medicine at Mount Sinai
Disclosures

None.
Objectives

1. Review two frameworks for validity in rehabilitation measurement.

2. Review, compare, and contrast reliability from the Classical Test Theory (CTT) and Item-Response Theory (IRT) perspectives.

3. Discuss the implications for interpreting scores and conducting analyses using scores from CTT- and IRT-based measurement instruments.
Outline

1. Review two rehabilitation measurement validity frameworks
   1. Conventional (COSMIN)
   2. Contemporary (Messick)

2. Describe reliability
   1. From the Classical Test Theory (CTT) perspective
   2. From the Item-Response Theory perspective
   3. In regard to common sources of measurement error

3. Describe the implications for
   1. Interpreting scores at a single time point
   2. Interpreting change over time
   3. Modeling longitudinal data

4. Questions, Answers, and Discussion
Every measurement has error.
COSMIN: A Conventional Validity Framework
Conventional Validity Framework: COSMIN

- **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments

- [http://www.cosmin.nl/](http://www.cosmin.nl/)

- **Resources**
  - Taxonomy
  - Checklist
  - Systematic reviews of Measurement Properties

Conventional Validity Framework: COSMIN Definitions

Validity:

The degree to which an instrument measures the construct(s) it purports to measure.

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html
Conventional Validity Framework: COSMIN Definitions

- **Face:** … instrument *appears* to adequately reflect the construct

- **Content:** … instrument *content* adequately reflects the construct

- **Construct:** … scores of an instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct
  - Structural
  - Hypothesis Testing
  - Cross-Cultural

- **Criterion:** … degree to which the scores of an instrument adequately reflect a ‘gold standard’

- [http://www.cosmin.nl/cosmin-taxonomy_3_0.html](http://www.cosmin.nl/cosmin-taxonomy_3_0.html)
Conventional Validity Framework: COSMIN Definitions

Validity:

The degree to which an instrument measures the construct(s) it purports to measure.

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html
Messick: A Contemporary Validity Framework
Contemporary Validity Framework: Messick

**Construct Validity:** … is not simply a property of a measure but is a reflection of and resides in the conditions of its use.

- Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores

Contemporary Validity Framework: Messick

Six Aspects of Construct Validity:

- **Content:** relevance, representativeness, technical quality
- **Substantive:** empirical evidence fits theoretical basis
- **Structural:** scores ≡ instrument ≡ construct
- **Generalizability:** samples → population groups, settings, tasks
- **External:** convergent and discriminant evidence
- **Consequential:** implications of interpretation and action

Conventional ‘types’ are methods of establishing evidence.
Reliability from the Classical Test Theory Perspective
Classical Test Theory: Reliability definition

- **Reliability**: The degree to which the measurement is free from measurement error

\[ X = T + E \]

- \( X \) = Observed score
- \( T \) = True score
- \( E \) = Error

- Sample: assumes errors are normally distributed about a mean of zero

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html
Classical Test Theory: Reliability types

- **Internal consistency:** The degree of the interrelatedness among the items

- **Rater reliability:**
  - Inter-rater: consistency in scores across two or more raters
  - Intra-rater: consistency for a single rater

- **Test-retest reliability:** consistency over time (stable or unchanging sample)

- **Group average of summary scores**

- [http://www.cosmin.nl/cosmin-taxonomy_3_0.html](http://www.cosmin.nl/cosmin-taxonomy_3_0.html)
Reliability from the Item-Response Theory Perspective
Item Response Theory: Reliability

- **Item response theory (IRT):** attempts to explain the response of a person to an item based on the idea that the probability of a correct/keyed response is a function of
  - Person ability
  - Item difficulty

- IRT extends the CTT concept of reliability:
  - Measurement precision varies across ranges of item difficulty and person ability (i.e., observed scores).
  - Scores nearer the floor and ceiling tend to have larger errors than scores nearer the mid-range.

- Rasch model can be considered a 1-parameter IRT model

Item Response Theory: Reliability

- **Item response theory (IRT):** attempts to explain the response of a person to an item based on the idea that the probability of a correct/keyed response is a function of
  - Person ability
  - Item difficulty

- IRT extends the CTT concept of reliability:
  - Measurement precision varies across ranges of item difficulty and person ability (i.e., observed scores).
  - Scores nearer the floor and ceiling tend to have larger errors than scores nearer the mid-range.

- Rasch model can be considered a 1-parameter IRT model

Item Response Theory: Reliability

- **Misfit:**
  - Overfit: not enough variation in responses; may indicate redundant items
  - Underfit: unexpected patterns of responses; may indicate ‘noisy’ items
  - Infit: information-weighted fit statistic
  - Outfit: outlier-sensitive fit statistic

- **Differential item functioning (DIF):**
  - The extent to which item scores are influenced by confounding.
  - E.g., Male/female differences.

- **Category Disorder:**
  - Higher category is more likely at a lower point than a lower category

- Test effect of collapsing categories and removing items

# Rasch Measures and Standard Error

<table>
<thead>
<tr>
<th>Score</th>
<th>Measure</th>
<th>S. E.</th>
<th>Score</th>
<th>Measure</th>
<th>S. E.</th>
<th>Score</th>
<th>Measure</th>
<th>S. E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.00E</td>
<td>15.31</td>
<td>24</td>
<td>31.06</td>
<td>2.26</td>
<td>47</td>
<td>43.32</td>
<td>2.48</td>
</tr>
<tr>
<td>2</td>
<td>9.05</td>
<td>7.68</td>
<td>25</td>
<td>31.64</td>
<td>2.23</td>
<td>48</td>
<td>44.05</td>
<td>2.57</td>
</tr>
<tr>
<td>3</td>
<td>13.39</td>
<td>5.01</td>
<td>26</td>
<td>32.20</td>
<td>2.19</td>
<td>49</td>
<td>44.84</td>
<td>2.68</td>
</tr>
<tr>
<td>4</td>
<td>15.65</td>
<td>3.95</td>
<td>27</td>
<td>32.75</td>
<td>2.16</td>
<td>50</td>
<td>45.70</td>
<td>2.79</td>
</tr>
<tr>
<td>5</td>
<td>17.17</td>
<td>3.38</td>
<td>28</td>
<td>33.28</td>
<td>2.13</td>
<td>51</td>
<td>46.64</td>
<td>2.92</td>
</tr>
<tr>
<td>6</td>
<td>18.35</td>
<td>3.04</td>
<td>29</td>
<td>33.80</td>
<td>2.11</td>
<td>52</td>
<td>47.66</td>
<td>3.04</td>
</tr>
<tr>
<td>7</td>
<td>19.33</td>
<td>2.82</td>
<td>30</td>
<td>34.30</td>
<td>2.09</td>
<td>53</td>
<td>48.76</td>
<td>3.15</td>
</tr>
<tr>
<td>8</td>
<td>20.19</td>
<td>2.67</td>
<td>31</td>
<td>34.80</td>
<td>2.07</td>
<td>54</td>
<td>49.94</td>
<td>3.24</td>
</tr>
<tr>
<td>9</td>
<td>20.97</td>
<td>2.57</td>
<td>32</td>
<td>35.29</td>
<td>2.06</td>
<td>55</td>
<td>51.17</td>
<td>3.29</td>
</tr>
<tr>
<td>10</td>
<td>21.71</td>
<td>2.50</td>
<td>33</td>
<td>35.77</td>
<td>2.05</td>
<td>56</td>
<td>52.42</td>
<td>3.31</td>
</tr>
<tr>
<td>11</td>
<td>22.42</td>
<td>2.46</td>
<td>34</td>
<td>36.26</td>
<td>2.05</td>
<td>57</td>
<td>53.68</td>
<td>3.32</td>
</tr>
<tr>
<td>12</td>
<td>23.11</td>
<td>2.44</td>
<td>35</td>
<td>36.74</td>
<td>2.05</td>
<td>58</td>
<td>54.96</td>
<td>3.37</td>
</tr>
<tr>
<td>13</td>
<td>23.79</td>
<td>2.44</td>
<td>36</td>
<td>37.22</td>
<td>2.05</td>
<td>59</td>
<td>56.30</td>
<td>3.48</td>
</tr>
<tr>
<td>14</td>
<td>24.47</td>
<td>2.44</td>
<td>37</td>
<td>37.70</td>
<td>2.06</td>
<td>60</td>
<td>57.78</td>
<td>3.70</td>
</tr>
<tr>
<td>15</td>
<td>25.16</td>
<td>2.44</td>
<td>38</td>
<td>38.19</td>
<td>2.07</td>
<td>61</td>
<td>59.50</td>
<td>4.07</td>
</tr>
<tr>
<td>16</td>
<td>25.84</td>
<td>2.45</td>
<td>39</td>
<td>38.69</td>
<td>2.09</td>
<td>62</td>
<td>61.67</td>
<td>4.65</td>
</tr>
<tr>
<td>17</td>
<td>26.53</td>
<td>2.44</td>
<td>40</td>
<td>39.20</td>
<td>2.11</td>
<td>63</td>
<td>64.60</td>
<td>5.48</td>
</tr>
<tr>
<td>18</td>
<td>27.21</td>
<td>2.44</td>
<td>41</td>
<td>39.71</td>
<td>2.14</td>
<td>64</td>
<td>68.77</td>
<td>6.63</td>
</tr>
<tr>
<td>19</td>
<td>27.89</td>
<td>2.42</td>
<td>42</td>
<td>40.25</td>
<td>2.18</td>
<td>65</td>
<td>75.11</td>
<td>8.31</td>
</tr>
<tr>
<td>20</td>
<td>28.56</td>
<td>2.40</td>
<td>43</td>
<td>40.80</td>
<td>2.22</td>
<td>66</td>
<td>85.65</td>
<td>11.06</td>
</tr>
<tr>
<td>21</td>
<td>29.21</td>
<td>2.37</td>
<td>44</td>
<td>41.38</td>
<td>2.27</td>
<td>67</td>
<td>100.00E</td>
<td>17.31</td>
</tr>
<tr>
<td>22</td>
<td>29.85</td>
<td>2.33</td>
<td>45</td>
<td>41.99</td>
<td>2.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>30.46</td>
<td>2.30</td>
<td>46</td>
<td>42.63</td>
<td>2.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Item Response Theory: Other Benefits

- **Computer-Adaptive Test (CAT):** Subsequent questions depend on response to previous questions
  - Only include relevant items
  - Reduce burden to patient/subject: fewer items with similar reliability
  - Reduce burden to clinician or researcher: computer administration

- **Short Forms:**
  - Fixed
  - Tailored
Common Sources of Measurement Error
Common Sources of Error: CTT

- **Patient or Subject:**
  - Normal variability of construct
  - Comprehension

- **Instrument:**
  - Irrelevant items included
  - Important items excluded

- **Rater:**
  - Variability in skill
  - Variability in observation
  - Potential for bias?

- **Environment:**
  - Variations in equipment, space, etc.
  - Distractions to patient/subject
  - Distractions to rater

- Potential for bias?
Common Sources of Error: IRT

- **Patient or Subject:**
  - Normal variability of construct
  - Comprehension

- **Instrument:**
  - Irrelevant items included
  - Important items excluded

- **Rater:**
  - Variability in skill
  - Variability in observation
  - Potential for bias?

- **Environment:**
  - Variations in equipment, space, etc.
  - Distractions to patient/subject
  - Distractions to rater

- **Dimensionality**

- **Item Misfit**

- **DIF**
Implications for Interpreting Scores at a Single Time Point
Interpretation of Single Point Scores: CTT

- **Point Estimate:**
  - Observed score
  - Often ordinal

- **Margin of error:**
  - Standard Error of Measurement (SEM)
  - ±1 SEM ≡ 67% Confidence interval (CI)
  - ±1.96 SEM ≡ 95% CI

- **Conditional SEM**
  - Margin of error varies across scale range
Interpretation of Single Point Scores: IRT

- **Point Estimate:**
  - Rasch measure (logit or transformed)
  - Interval level
  - Summary or item score

- **Margin of error:**
  - Standard Error (SE)
  - ±1 SE ≡ 67% Confidence interval (CI)
  - ±1.96 SE ≡ 95% CI
  - Varies across scale range
Implications for Interpreting Change Over Time
Interpretation of Change over Time: CTT

- **Minimal Detectable Change (MDC):**
  - Margin of error for change score
  - Derived from stable (unchanged) sample
  - Two-point change: which time points?
  - Conditional MDC: varies for range of baseline score

- **Minimal Clinical Important Difference (MCID):**
  - Index for important change
  - Derived from sample who have changed an important amount
  - Importance anchored to
    - Patient
    - Clinician
    - Both
    - Other
Interpretation of Change over Time: IRT

- No MDC or MCID
- Change on summary score

Keyform Maps
- Pattern of change (Bode 2014)
- Item category thresholds (Velozo 2011)

Implications for Modeling Longitudinal Data
Implications for Modeling Longitudinal Data

- Do errors vary over time?
- Does timing of data collection matter?
- Ordinal or Interval?
Implications for Modeling Longitudinal Data

▶ Do errors vary over time?
  – Change in score distributions: random sampling (Bode 2014)
  – Response shift

▶ Does timing of data collection matter?

▶ Ordinal or Interval?
Implications for Modeling Longitudinal Data

- **Individual Growth Curve Models**
  - Hierarchical linear model
  - Explicitly model time
  - Account for within-person correlations of scores over time
  - Simultaneously model individual and group trajectories
  - Accommodate missing at random outcome scores and variable time point intervals
  - Linear, curvilinear, non-linear models
  - Covariate associations explain variance
  - More suited to large data sets


Summary

- Validity can be thought of as an attribute of scores, and the social consequences of interpreting and using scores for decision-making.
  - No such thing as a ‘valid and reliable instrument’

- CTT considers reliability and validity at the level of the instrument (scale or subscale summary scores only).

- IRT considers score reliability to be a probability function of person-ability and item-difficulty, and provides
  - Additional item-level diagnostic tests.
  - Item- or score-level reliability estimates (SEs)
  - Additional tools for interpretation (e.g., key-form maps)
  - Alternate modes of administration (CAT, tailored short forms)
Summary

- IRT offers advantageous for modeling change over time:
  - May reduce error at single time points
  - Strategies to account for variability in errors over time
  - Does not preclude developing CTT-based indices to aid interpretation
  - Enhances interpretation of change for longitudinal models
Questions, Answers, and Discussion