Beyond Rulers and Scales: Developing Measurement Tools for Research, Education, and Clinical Care

Su-Ting T. Li, MD, MPH\textsuperscript{1} and Daniel C. West, MD\textsuperscript{2}

\textsuperscript{1}University of California, Davis
\textsuperscript{2}University of California, San Francisco

APPD Workshops, April 17, 2010
PAS Workshop, May 4, 2010
Su-Ting T. Li, MD, MPH
Daniel C. West, MD

Have documented that they have no relevant financial relationships to disclose or COIs to resolve.
Unapproved or Off Label Disclosures for

Su-Ting T. Li, MD, MPH
Daniel C. West, MD

Presenter: Drs. Li and West have documented that their presentation will not involve discussion of unapproved or off-label, experimental or investigational use.
Goals

1. Identify the characteristics of validity

2. Identify the characteristics of reliability

3. Develop a plan to design your own valid and reliable measurement tool
Workshop Outline

- 8:45-8:55  Introductions
- 8:55-9:15  Define validity
- 9:15-10:00  Small group exercise #1
- 10:00-10:20  Large group exercise
- 10:20-10:40  Define reliability
- 10:40-11:20  Small group exercise #2
- 11:20-11:40  Large group exercise
- 11:40-11:45  Summary
Ground Rules and Introductions

• First time you comment, please introduce yourself (name and where you are from)
• Who is interested in developing a tool for clinical care or research?
• Who is interested in developing a tool for medical education?
• What do you want to gain from attending this workshop?
## Validity vs. Reliability

<table>
<thead>
<tr>
<th></th>
<th>Validity</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>Tool measures what it is intended to measure</td>
<td>Tool is dependable and reproducible</td>
</tr>
<tr>
<td>Example</td>
<td>Washing machine that gets clothes clean</td>
<td>Washing machine that runs and goes through all the cycles dependably</td>
</tr>
<tr>
<td>Determination</td>
<td>Compare to “gold standard” or other measures of validity</td>
<td>Use measurement theory</td>
</tr>
</tbody>
</table>
Validity vs. Reliability

- Reliable
  - Not Valid
- Valid
  - Not Reliable
- Neither Reliable Nor Valid
- Both Reliable And Valid
Measurements and Outcomes

• What is the purpose of measurement?
  – Discriminate between different states, characteristics, behaviors, skills

• Measure something important
  – Should be linked to specific aim or goal
  – Not just something that is easy to measure
    • Outcome of interest needs to be measurable
Establishing Validity

• What are you trying to measure?
  – Basic math ability

• What are the theoretical constructs (component parts) of what you are trying to measure
  – Addition
  – Subtraction
  – Multiplication
  – Division

• What are the elements of each construct? (ie., specific items in your measurement)
  – Example: addition ability
    • 19 + 27 + 42 = ? (add multiple-digit numbers)
    • 1.5 + 2.6 = ? (add decimals)
Types of Validity

- Face validity
- Content validity
- Construct validity
  - Convergent validity
  - Divergent validity
- Criterion validity
  - Criterion-standard validity ("gold standard")
  - Predictive validity
  - Concurrent validity
- External validity
Face Validity

- General perception
- On first glance, does it measure the right thing
- Critically important from the point of view of those to be assessed (e.g. to get cooperation)
- Weakest evidence of validity
Content Validity

Agreement between construct and items in construct

• Provide **evidence** that items in assessment
  – MATCH construct
  – Include ALL important aspects of construct

Example: Respiratory distress
  – SaO2
  – Accessory muscle use
  – Mental status
  – Inspiratory breath sounds (air movement)
  – Expiratory breath sounds (wheeze)
Content Validity
How do you provide evidence?

• Medical literature
  – Guidelines
  – Studies

• Expert review
  – Who are the content experts?
  – How will you incorporate experts into development of your tool?
    • Focus groups to determine construct items
    • Develop consensus through Delphi method where expert panel iteratively rates validity of items
Construct Validity

Agreement between construct and measurement

Convergent validity

– Are different measures of same construct consistent?

  • *If measuring addition skills, people who answer an addition problem correctly should tend to answer other addition problems correctly*

Divergent validity

– Are measures of different constructs independent?

  • *Ability to solve math problems independent of ability to identify different animals*
Criterion Validity
Comparison to some other valid measure

- Criterion-standard ("gold standard")
- Predictive validity
- Concurrent validity
Is There a Gold Standard?

• If there is a gold standard, how will you compare your measure to the gold standard?
  – Correlation
  – Sensitivity/specificity/likelihood ratios
  – Regression analysis (allows you to estimate effect)

• But, there is usually no gold standard
  – In that case, compare with other criterion and construct measures to build a case for validity
Criterion Validity

• Predictive validity:
  – Does the score on the measurement tool predict performance in the future?
    • *Does a high score on the respiratory distress scale correlate with future intubation/ventilation?*

• Concurrent validity:
  – Is the score on the measurement tool consistent with current performance?
    • *Do NICU fellows score higher on a neonatal resuscitation tool than PGY1s?*
    • *Do residents who score higher on an assessment of interpersonal skills completed by nurses also score higher on an assessment of interpersonal skills by patients and families?*
External Validity

• How will you determine whether your tool is generalizable to other settings or outside the conditions of your study?
  – Collaborate with institutions different from your own to test this tool
    • Example: *Test respiratory distress tool developed in an academic children’s hospital at a community hospital.*
  – Are results similar?
Using Our Worksheet For Validity

Step 1: Determine constructs

<table>
<thead>
<tr>
<th>What do you want to measure?</th>
<th>Theoretical Constructs (i.e. component parts)</th>
<th>Instrument Items (i.e. measurable characteristic of constructs)</th>
</tr>
</thead>
</table>
| Resident professionalism and communication skills with patients/families | Professionalism | - Truthful, upfront and frank  
- Follows through with commitments/promises |
|                              | Communication | - Encourages families and patient to ask questions  
- Answers questions clear, understandable way |
# Step 2: Establish Content Validity

<table>
<thead>
<tr>
<th>Content Validity</th>
<th>Content Experts</th>
<th>Published Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources?</td>
<td>-Patients/Families</td>
<td>-Charter of Medical Professionalism</td>
</tr>
<tr>
<td></td>
<td>-Residents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Medical educators</td>
<td>-ABIM patient survey used for MOC</td>
</tr>
<tr>
<td>How will you use content experts?</td>
<td>-Focus groups for item development</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Delphi method: Content experts rate items iteratively for inclusiveness, validity, importance</td>
<td></td>
</tr>
</tbody>
</table>
Step 3: Establish construct validity

<table>
<thead>
<tr>
<th>Validity Type</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergent Validity</td>
<td>Confirmatory factor analysis to determine whether items within the same construct are rated similarly</td>
</tr>
<tr>
<td>Divergent Validity</td>
<td>Factor analysis to determine whether items in different constructs are independent</td>
</tr>
</tbody>
</table>
Step 4: Establish Criterion Validity

<table>
<thead>
<tr>
<th>Type of Validity</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion-standard validity</td>
<td>No gold standard exists</td>
</tr>
<tr>
<td>Predictive validity</td>
<td>Regression analysis to determine whether residents who score very low on communication skills are more likely to have more patient complaints</td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>Pearson’s correlation to determine whether patient and nurse evaluations of communication skills are similar (could also do by regression)</td>
</tr>
</tbody>
</table>
### Step 5: Establish external validity

<table>
<thead>
<tr>
<th>External Validity</th>
<th>Find collaborators at other residency programs (different size, geographic region, academic/community)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pilot tool with multi-institutional study to determine generalizability across institutions</td>
</tr>
</tbody>
</table>
Small group #1:

- Individually
  - Use worksheet to develop plan to establish validity
- Small group (2-3 people)
  - Share and critique plan
  - Then share plan with larger group
Importance of Reliability

• Goal is to measure intended characteristic, attribute, or feature
• But all measurements have error
• Error defined as *variability* in measured score for reasons other than differences in what you are trying to measure.
  - Goal is to minimize error
• Establish reliability using Measurement Theory
Sources of Error (i.e. Variability)

- **Raters**
  - Training, perspective, knowledge, ability, etc.

- **Occasion**
  - Performance varies from one day to next based on conditions unique to that occasion

- **Item**
  - Some questions harder or easier than others
  - Some attributes harder or easier to rate

- **Setting**
  - Inpatient, outpatient, critical care, etc.
Error Interactions

- Effect of source of error can vary depending on characteristics of an individual or other source of error

- Examples:
  - Rater might grade harder depending on mood or what they think of the person they are rating
  - Because of clinical experience, one rater might find it easier to rate use of accessory muscles for respiration than another rater
Minimizing Error

• Need to understand error in order to maximize reliability of an assessment tool

• Things you can do:
  – Rater training
  – Optimize assessment occasions (e.g. conditions)
  – Clarity of tool
  – Anchors for response scales
  – Enhance validity of tool
Measurement Theory

Definition: *Evaluates quality and dependability of measurements*

1. Classical Test Theory
2. Generalizability Theory
3. Item Response Theory
Classical Test Theory

Measured score = True score (expected) + error

• Produces reliability coefficient on scale of 0 to 1
• Familiar measurements of reliability
  – Inter-rater reliability
  – Test-retest reliability (i.e. parallel tests)
  – Internal consistency (Cronbach-alpha)
• Limitation is that it measures only one source of error at a time (error cannot be decomposed)
  – Cannot determine how to modify administration
Applications of Classical Test Theory

• Categorical measure
  – Checklist tools
  – Kappa statistic to measure agreement between two different raters

• Continuous measure
  – Likert scales (e.g. agree/disagree)
  – Could use Kappa, but not very robust
  – Pearson correlation between two raters or two different rating occasions
Generalizability Theory (G-Theory)

*Measured score = Universe score + error*

- Underlying assumption:
  - The measured attribute is in steady state
  - Score variability on different occasions, types of items, and by different raters are due to error

- Error can be decomposed

- Produces generalizability coefficient on scale of 0-1
  - Represents confidence with which one can generalize measurement across all acceptable (and important) conditions of measurement
Performing a G-Study

\[ \text{Measured score} = \text{Universe score} + \text{error} \]

- Uses ANOVA to estimate amount of variance attributable to different sources
- Universe score variance is indicator of how much test scores vary due to differences in object of measurement (want this to be high)
- Relative contributions of error to measured score can be determined
  - Source of error in analysis referred to as a “facet”
  - The more facets, the larger the sample size required
Sources of Variance in Measurement

- **Difference Among Persons**
  - $P$ = person (universe score)

- **Sources of Error**
  - $R$ = rater
  - $O$ = occasion
  - $PR$ = person x rater
  - $PO$ = person x occasion
  - $RO$ = rater x occasion
  - $PRO$ = person x rater x occasion
  - Residual = unmeasured error
Interpreting Reliability or Generalizability Coefficients

0.8-1.0  Good reliability or generalizability
0.6-0.8  Marginal reliability
< 0.6    Poor reliability
Decision-Study

- Possible only using G-theory
- Use variance estimates from ANOVA to perform “what if” scenario
  - Vary the most important sources of error to reach G-coefficient $\geq 0.8$
- Determine best conditions to administer the assessment
  - Number of raters required
  - Number of occasions required
# Using Our Worksheet For Reliability

## Step 1: Identify Potential Sources of Error

<table>
<thead>
<tr>
<th>Potential Sources of Error (i.e. variability)</th>
<th>Strategies to Minimize Error (i.e. ways to reduce undesirable score variance)</th>
<th>Limitations (e.g. practical constraints)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences in the way patients/families (raters) score items</td>
<td>Train raters</td>
<td>Could train nurses, not practical to train patients</td>
</tr>
<tr>
<td>Performance by occasion</td>
<td>Avoid post-call assessments</td>
<td>Schedule challenge and might want this info</td>
</tr>
<tr>
<td>Performance by item or question</td>
<td>Clarify items</td>
<td></td>
</tr>
<tr>
<td>Performance by location (e.g. PICU vs. clinic)</td>
<td>Confine assessment to one location</td>
<td>Lose valuable information</td>
</tr>
<tr>
<td>Resident performance by construct (e.g. prof vs. I/C)</td>
<td>Stratify analysis by construct</td>
<td></td>
</tr>
</tbody>
</table>
### Classical Test Theory Design

<table>
<thead>
<tr>
<th>Reliability Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test/Retest Reliability</td>
<td>Administer assessment, then repeat same assessment one week later (Pearson correlation)</td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>Have two different raters evaluate each resident in the study group on the same day (Kappa or Pearson)</td>
</tr>
<tr>
<td>Intra-rater reliability</td>
<td>Have same rater evaluate resident, then same rater evaluate same resident (ideally same resident encounter) one week later (Pearson correlation)</td>
</tr>
</tbody>
</table>

### Practical Constraints:
- Cannot separate rater from occasion
- Cannot have same rater score same encounter twice
Step 2: Establish Reliability
Generalizability Theory (G-Study)

<table>
<thead>
<tr>
<th>G-Study Design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Object of Measurement</strong></td>
</tr>
<tr>
<td><strong>Facet 1</strong></td>
</tr>
<tr>
<td><strong>Facet 2</strong></td>
</tr>
<tr>
<td><strong>Facet 3</strong></td>
</tr>
</tbody>
</table>

Practical Constraints:
- Cannot separate rater and occasion
- Not certain we would want to generalize over constructs
G-Study: Nurse Evaluation of Resident Interpersonal and Communication Skills

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>R (Resident)</td>
<td>0.06</td>
<td>11</td>
</tr>
<tr>
<td>O (Occasion)</td>
<td>0.004</td>
<td>0.6</td>
</tr>
<tr>
<td>I (Item)</td>
<td>0.02</td>
<td>2.6</td>
</tr>
<tr>
<td>RO</td>
<td>0.14</td>
<td>22.9</td>
</tr>
<tr>
<td>RI</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>OI</td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>ROI + unmeasured error</td>
<td>0.38</td>
<td>63.3</td>
</tr>
</tbody>
</table>
Step 3: Administration of Tool
Determine how to maximize reliability

• Decision Study (D-Study)
  – Only possible with G-Study (need ANOVA)
  – Person/Occasion and Person/Occasion/Item/Error is source of most variance
  – Therefore, D-Study should vary occasion and item
  – Goal to find practical combination that generates G-coefficient $\geq 0.8$

• Classical Test Theory
  – Make reasonable guess, then restudy
## D-Study: Nurse Evaluation of Resident Interpersonal and Communication Skills

<table>
<thead>
<tr>
<th>Occasions</th>
<th>G-Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 Item</td>
</tr>
<tr>
<td>10</td>
<td>0.55</td>
</tr>
<tr>
<td>12</td>
<td>0.59</td>
</tr>
<tr>
<td>15</td>
<td>0.65</td>
</tr>
<tr>
<td>18</td>
<td>0.69</td>
</tr>
<tr>
<td>20</td>
<td>0.71</td>
</tr>
</tbody>
</table>
D-Study: Nurse Evaluation of Resident Interpersonal and Communication Skills

<table>
<thead>
<tr>
<th>Occasions</th>
<th>1 Item</th>
<th>2 Items</th>
<th>3 Items</th>
<th>4 Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.55</td>
<td>0.66</td>
<td>0.70</td>
<td>0.73</td>
</tr>
<tr>
<td>12</td>
<td>0.59</td>
<td>0.70</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>15</td>
<td>0.65</td>
<td>0.74</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>18</td>
<td>0.69</td>
<td>0.78</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>20</td>
<td>0.71</td>
<td>0.79</td>
<td>0.83</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Small Group Exercise: Putting It Into Practice

• Use back page of worksheet

• Individually make own plan to test reliability of tool
  – Identify sources of error, methods to minimize error, and limitations to administration
  – Design study to test some aspect of reliability

• Split into groups of 2 or 3
  – Discuss and critique among the group
Key Points

1. All measurements have error
2. Understanding error is key to administration and interpretation of assessment
3. Sources of error can interact with each other
4. Should be able to identify sources of error (i.e. variance), identify ways to minimize error, and design basic study to establish reliability
5. Generalizability theory is almost always the way to assess measurements of behavior or subjective ratings of patient characteristics
# Classical vs. Generalizability Theory

<table>
<thead>
<tr>
<th>Classical Test Theory</th>
<th>Generalizability Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>One source of error at a time</td>
<td>Multiple sources of error and interactions at same time</td>
</tr>
<tr>
<td>Little information about how to improve reliability</td>
<td>Can make judgments and do what-if scenarios to improve reliability</td>
</tr>
<tr>
<td>Reliability coefficient</td>
<td>Generalizability coefficient</td>
</tr>
</tbody>
</table>
Additional considerations

• Use published valid, reliable tools if you can

• IRB issues
  – If you think you may ever publish your tool, apply for IRB approval early
  – Exemption (de-identified data), expedited, full
  – Informed consent

• Analysis
  – Collaborate with a statistician (psychometrics if possible) in design phase
Summary

• Worksheet provides systematic method for establishing validity and reliability
• Remember there are different forms of validity
  – Usually need to build a case
• Reliability focuses on determining error in measurement
  – Undesirable variability
• There are three approaches to measurement theory (only 2 for most of our needs)