Scoring & Decision Making

DeShon - 2007

Scoring Overview
- Once you have administered the test and cleaned the data...
- What number is used to represent the person on the latent variable of interest?
  - What's the right answer
    - Empirical vs. Rational Keying
  - Summing responses
    - Number correct, number endorsed, number checked
    - Weighted summing, non-unique summing
  - Corrections for artifacts (lie scales)
  - Forming Composites of subscales

Correct/Incorrect Measures
- How do you determine correct?
  - Rational Keying
    - Experts agree on the right answer
    - Find right answers in authoritative texts on the topic
  - Empirical Keying
    - Compare correlation of item response alternatives to a criterion of interest
    - Compare existing groups and find items that discriminate between the groups - Discriminant-groups validity model
- Scoring algorithms for scoring items and constructing scales from item responses are often not disclosed
  - Why?
  - Item scores matter
  - Scale construct routines are largely irrelevant unless you must base your interpretation on existing norms

Interpretation and Decision Making
- Once you have scores, how do you interpret test scores and use them for decision making?
  - Ranking/Top-Down decision making
  - Banding
  - Cut scores
  - Norms
    - Z-scores, T-scores, percentiles
- Top-Down/Ranking is very common
  - Decisions based on relative standing in the distribution of test scores
  - Higher scores mean more of the trait
- Hard to demonstrate that higher scores mean higher standing on the latent trait if there is much error in the scores
**Banding**
- Set up ranges of the test scores that are distinguishable based on the standard error of the difference.
- Then select candidates at random or using some other criterion (seniority) within the band.
  - Fixed bands
  - Sliding bands

**Criterion-Referenced Measures**
- Develop a cutoff and the meaning of scores is based on standing relative to the cut score.
- Pass/fail: Usually used for knowledge and achievement tests.
- Many methods available for computing cut scores:
  - Ebel
  - Nedelsky
  - Angoff

**Angoff Method**
- **Subject Matter Experts** (SMEs) evaluate all items and estimate the probability that a minimally qualified person would get the item right.
- The average of the item scores is the **cut score** for the exam.
- A bit more complex than this....

**Norms**
- Raw scores of psychological tests usually have little inherent meaning.
- For normative tests, meaning is derived by comparing scores to other individuals (e.g., other members of a sample or a normative sample).
  - Percentiles
  - Z scores
  - T scores
- Representativeness of the norming sample is crucial!

**Norms - Percentiles**
- Percentile: relative position in the sample or reference group.
- Percentile rank: percentage of people that earned a raw score lower than the given score.
- Percentage of persons, not items.
- Example: GRE scores.

**Norms – Z scores**
- Expresses distance of score from the mean in SD units.
- Advantages of standard scores:
  - Includes information about the person’s standing in the distribution (i.e., percentile rank).
  - Allows comparisons across tests that have different raw metrics.
Norms: T scores
- T scores are linear transformations of Z scores
  - T score = (Z score * 10) + 50
  - Mean = 50, SD = 10
- If normal T-scores will be between 20 and 80
- Why? Easier for lay audiences to interpret
  - For Z scores, half the scores are negative.

Comparison of Norms

<table>
<thead>
<tr>
<th>Z score</th>
<th>T score</th>
<th>Percentile rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>80</td>
<td>99.9</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>97.5</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>84</td>
</tr>
<tr>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>-1</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>-2</td>
<td>30</td>
<td>2.5</td>
</tr>
<tr>
<td>-3</td>
<td>20</td>
<td>.1</td>
</tr>
</tbody>
</table>

Example: MMPI-2
- Designed for routine diagnostic assessments
  - Most frequently used personality test in the US for adults and adolescents
- Empirical keying approach
- 567 true/false items
- 10 clinical scales plus validity scales
- Original Norms
  - 724 Minnesota "normals" and 221 psychiatric patients
- Revised Norms
  - 2600 U.S. residents aged 18-90 (census derived)

Example: MMPI-2
- Empirical/Criterion keying
  - Identify a criterion group (e.g., people diagnosed with schizophrenia)
  - Identify a comparison group (e.g., persons with no mental illness)
  - Administer many, many test items to both groups
  - Identify a group of items that discriminates the two groups, i.e., items endorsed more frequently by the criterion group
  - This group of items becomes the schizophrenia scale

Example: MMPI-2
- Resulting scales are a "mixed bag" of items with generally undesirable measurement properties
  - Scales have heterogeneous item content
  - Often multi-dimensional
  - Item overlap across scales
  - Adds to complexity of interpretation
  - But still appears to have practical use

Example: MMPI-2
- Administered individually or in groups
- Administration time is approximately 1 to 1.5 hours
- Scored by hand or computer
- Separate scoring keys by gender
Validity Scales
- K Scale (Cannot say)
  - number of items left unanswered, if 30 or more items are MT-unchecked the protocol is invalid
- F Scale (Infrequency)
  - 66 Items
  - atypical or deviant response style
  - endorsed by less than 10% of the population
- ? Scale (Cannot say)
  - Extreme elevations indicate invalid profile (100 or higher)
  - No exact cutoff for suspecting an invalid profile

Example Items for the K scale
- I get mad easily then get over it soon. (F)
- At times my thoughts have raced ahead faster than I could (F)
- Most people will use somewhat unfair means to get what they want. (F)
- I do not always tell the truth. (T)
- At times I feel like swearing. (T)
- Once in a while I think about things too bad to talk about. (T)
- I get anxious and upset when I have to make a short trip away from home. (T)
- I have gotten many beatings. (T)

Example Items for L scale
- At times I feel like killing people. (F)
- I am troubled by attacks of nausea and vomiting. (T)
- Evil spirits posses me at times. (T)
- I am liked by most people who know me. (F)
- There is something wrong with my mind. (T)
- I think school is a waste of time. (T)
- I get anxious and upset when I have to make a short trip away from home. (T)
- I have gotten many beatings. (T)
- My table manners are not quite as good at home as when I am out in company.
- I would say a lie in order to gain sympathy and help of others. (F)
- I am troubled by attacks of nausea and vomiting. (T)
- I think school is a waste of time. (T)
- I feel depressed. (F)
- I have very few quarrels with members of my family. (T)
- My parents do not really love me. (T)
- My teachers have it in for me. (F)
- Evil spirits posses me at times. (T)
- I am liked by most people who know me. (F)
- I think school is a waste of time. (T)
- My table manners are not quite as good at home as when I am out in company.
- I would say a lie in order to gain sympathy and help of others. (F)
- I am troubled by attacks of nausea and vomiting. (T)
- I think school is a waste of time. (T)
- I feel depressed. (F)
- I have very few quarrels with members of my family. (T)
- My parents do not really love me. (T)
- My teachers have it in for me. (F)
- Evil spirits posses me at times. (T)
- I am liked by most people who know me. (F)
- I think school is a waste of time. (T)
- My table manners are not quite as good at home as when I am out in company.
- I would say a lie in order to gain sympathy and help of others. (F)
- I am troubled by attacks of nausea and vomiting. (T)
- I think school is a waste of time. (T)
- I feel depressed. (F)
- I have very few quarrels with members of my family. (T)
- My parents do not really love me. (T)
- My teachers have it in for me. (F)
- Evil spirits posses me at times. (T)
- I am liked by most people who know me. (F)
- I think school is a waste of time. (T)
- My table manners are not quite as good at home as when I am out in company.
Interpretation

- Yields individual's clinical profile compared with the normative samples.
- Interpretation is configural in nature and not dependent on any one scale.
  - T-score of 65 or higher is considered a clinically significant elevation for all clinical scales.
  - Clinical scales do not measure the low end; don't interpret low scores.
- Interpreted by qualified professionals.
- Welsh Coding:
  - Record the 10 numbers of the clinical scales in order of T scores, from the highest on the left to the lowest on the right.
  - When adjacent scores are within one T score point, they are underlined; when they have the same T score, they are placed in the ordinal sequence found on the profile sheet and underlined.

Example: MMPI-2

<table>
<thead>
<tr>
<th>Scale</th>
<th>T-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRIN</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Hs</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Hy</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Pd</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Mf</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Pa</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>Pt</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Sc</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>Ma</td>
<td>23</td>
</tr>
<tr>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Si</td>
<td>25</td>
</tr>
</tbody>
</table>