Item Formats

- Dichotomous Format
  - Two alternatives
  - True/False
  - MMPI/2, MMPI/A
- Polytomous or Polychotomous Format
  - More than two alternatives
  - Multiple choice
  - Psy427 Midterm, SAT, GRE,

Item Formats

- Distractors
  - Item Formats
    - Incorrect choices on a polychotomous test
    - Best to have three or four
  - BUT -
    - one study (Sidick, Barret, & Doverspike, 1994) found equivalent validity and reliability for a test with two distractors (three items) as one with four distractors (five items).
    - SO, best might be to have two to four (further study is needed)
Should you guess on polytomous tests?

- Depends... Correction for guessing:

\[ \text{Corrected Score} = R - \frac{W}{n-1} \]

- \( R \) is the number correct
- \( W \) is the number incorrect
- \( n \) is the number of polytomous choices
- If no correction for guessing, guess away.
- If there is a correction for guessing, better to leave some blank (unless you can beat the odds)

Other Test Items

- Likert scales
  - On a rating scale of 1-5, or 1-6, 1-7, etc. where
  - 1 = strongly disagree
  - 2 = moderately disagree
  - 3 = mildly disagree
  - 4 = mildly agree
  - 5 = moderately agree
  - 6 = strongly agree
  - rate the following statements....

Other Test Items

- Likert scales
  - Even vs. odd number of choices
    - Even numbers prevents “fence-sitting”
    - Odd numbers allows people to be neutral
  - Likert items are VERY popular measurement items in psychology.
  - Technically ordinal but are often assumed continuous if 5 or more choices
  - With that assumption we can calculate means, factor analyze, etc.
**Other Test Items**

- Category format
  - Like Likert, but with MANY more categories
    - e.g., 10-point scale
  - Best if used with anchors
  - Research supports use of 7-point scales to 21-point scales

**Other Test Items**

- Visual Analogue Scale

  No Headache  |  Worst Headache

- Also used in research
  - dials, knobs
  - time sampling

**Checklists & Q-Sorts**

- Both used in qualitative research as well as quantitative research
- Checklists
  - Present list of words (adjectives)
  - Have person choose to endorse each item
  - Can determine perceptions of concepts using checklists.
Checklists & Q-Sorts

- Adjective Checklists (from http://www.encyclopedia.com/doc/1O87-AdjectiveCheckList.html)
  - In psychometrics, any list of adjectives that can be marked as applicable or not applicable
    - to oneself
    - to one's ideal self
    - to another person, OR
    - to some other entity or concept.

Checklists & Q-Sorts

- Checklists
  - When written with initial uppercase letters (ACL), the term denotes more specifically a measure consisting of a list of 300 adjectives, from absent-minded to zany
  - Selected by the US psychologist Harrison G. Gough (born 1921) and introduced as a commercial test in 1952.
  - The test yields 24 scores, including measures of personal adjustment, self-confidence, self-control, lability, counselling readiness, some response styles, and 15 personality needs, such as achievement, dominance, and endurance.

Checklists & Q-Sorts

- Q-Sorts
  - Introduced by William Stephenson in 1935
    - PhD in physics 1926, PhD in psychology in 1929
    - Student of Charles Spearman
  - Goal: to get a quantitative description of a person's perceptions of a concept
  - Process: give subject a pile of numbered "cards" & have them sort them into piles
  - Piles represent graded degrees of description (most descriptive to least descriptive).
Checklists & Q-Sorts

- Q-Sorts
  - Means of self-evaluation of client's current status
  - The Q-Sort consists of a number of cards, often as many as 40 or 50, even 100 items each consisting of a single trait, belief, or behavior.
  - The goal is to sort these cards into one of five columns ranging from statements such as 'very much like me' to 'not at all like me.'
  - There are typically a specific number of cards allowed for each column, forcing the client to balance the cards evenly.
- Example:
  - California Q-sort, Attachment Q-sort

Example Q-sort

California Q-Sort
**Attachment Q-sort**

- Methods used to evaluate test items.
- What are good items?
- Techniques
  - Item Difficulty (or easiness)
  - Discriminability
  - Extreme Group
  - Item/Total Correlation
  - Item Characteristic Curves
  - Item Response Theory
  - Criterion-Referenced Testing

**Item Difficulty**

- The proportion of people who get a particular item correct or that endorse an item (if there is no "correct" response, e.g. MMPI)
- Often thought of as the item’s easiness because it is based on the number correct/endorsed
**Item Difficulty**

- The difficulty can be given in proportion for or it can be standardized in to a Z-value

\[
Z = \frac{\ln(1 - p) - \ln(p)}{1.7}
\]

**Item Difficulty**

- For example a test with the difficulty of .84

\[
Z = \frac{(\ln(.16) - \ln(.84))}{1.7}
\]

\[
= \frac{(-1.83 + .17)}{1.7}
\]

\[
= -1.66/1.7
\]

\[
= -1.00
\]

(-2 → 2 is typical range)

**Difficult Item (35%)**

If you are taking a criterion referenced test in a social psychology course and you need to score a 92 in order to get an A, the criterion is

a) Social Psychology *
b) Scoring a 92
c) Getting an A
d) Not enough info.
**Difficult Item (35%)**

\[ Z = \frac{[\ln(1 - p)] - \ln(p)}{1.7} \]

\[ Z = \frac{(\ln(.65) - \ln(.35))}{1.7} \]

\[ = \frac{(-.431 + 1.050)}{1.7} \]

\[ = \frac{.619}{1.7} \]

\[ = .364 \]

**Moderate Item (51%)**

The correlation between X and Y is .54. X has a SD of 1.2 and Y has a SD of 5.4. What is the regression coefficient (b) when Y is predicted by X?

a) .12  
b) 2.43*  
c) .375  
d) .45

**Difficult Item (51%)**

\[ Z = \frac{[\ln(1 - p)] - \ln(p)}{1.7} \]

\[ Z = \frac{(\ln(.49) - \ln(.51))}{1.7} \]

\[ = \frac{(-.713 + .673)}{1.7} \]

\[ = \frac{-0.004}{1.7} \]

\[ = -.00235 \]
Easy Item (100%)

- For the following set of data [5 9 5 5 2 4], the mean is
  a) 4
  b) 5
  c) 4.5
  d) 6

Difficult Item (100%)

\[ Z = \frac{\ln(1-p) - \ln(p)}{1.7} \]
\[ Z = \frac{(\ln(0) - \ln(1))}{1.7} \]
\[ = \text{error} \]

Optimum Difficulty

- Mathematically: half-way between chance and 100%.
- Steps (assuming a 5-choice test)
  1. Find half-way between 100% and chance
     - \( \frac{1 - 2}{2} = 0.5 \)
  2. Add this value to chance alone
     - \( 0.5 + 2 = 2.5 \)
- Alternately: Chance + \( \frac{1}{2} \) = optimum difficulty
- A good test will have difficulty values between .30 and .70
**Discriminability**

- Can be defined in 2 ways:
  1. How well does each item distinguish (discriminate) between individuals who are scoring high and low on the test as a whole (e.g. the trait of interest).
  2. Or simply how well is each item related to the trait (e.g. loadings in factor analysis)
    - 1 and 2 are really the same the more an item is related to the trait the better it can distinguish high and low scoring individuals

---

**Discriminability**

- Extreme Group Method
  - First
    - Identify two “extreme” groups
    - Top third vs. bottom third
  - Second
    - Compute “Difficulty” for the top group
    - Compute “Difficulty” for the bottom group
    - Compute the difference between Top Difficulty and Bottom Difficulty
    - Result = Discriminability Index

---

<table>
<thead>
<tr>
<th>Item</th>
<th>High Scores</th>
<th>Low Scores</th>
<th>Discriminability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>12</td>
<td>11</td>
<td>0.05</td>
</tr>
<tr>
<td>2.</td>
<td>13</td>
<td>12</td>
<td>0.06</td>
</tr>
<tr>
<td>3.</td>
<td>14</td>
<td>13</td>
<td>0.07</td>
</tr>
<tr>
<td>4.</td>
<td>15</td>
<td>14</td>
<td>0.08</td>
</tr>
<tr>
<td>5.</td>
<td>16</td>
<td>15</td>
<td>0.09</td>
</tr>
<tr>
<td>6.</td>
<td>17</td>
<td>16</td>
<td>0.10</td>
</tr>
<tr>
<td>7.</td>
<td>18</td>
<td>17</td>
<td>0.11</td>
</tr>
<tr>
<td>8.</td>
<td>19</td>
<td>18</td>
<td>0.12</td>
</tr>
<tr>
<td>9.</td>
<td>20</td>
<td>19</td>
<td>0.13</td>
</tr>
<tr>
<td>10.</td>
<td>21</td>
<td>20</td>
<td>0.14</td>
</tr>
<tr>
<td>11.</td>
<td>22</td>
<td>21</td>
<td>0.15</td>
</tr>
</tbody>
</table>

---

**Discriminability**

- Extreme Group Method
  - First
    - Identify two “extreme” groups
    - Top third vs. bottom third
  - Second
    - Compute “Difficulty” for the top group
    - Compute “Difficulty” for the bottom group
    - Compute the difference between Top Difficulty and Bottom Difficulty
    - Result = Discriminability Index
Discriminability

- Item/Total Correlation
  - Let the total test score "stand in" for the trait of interest; a roughly estimated "factor" of sorts
  - Correlate each item with the total test score; items with higher item/total correlations are more discriminating
  - These correlations are like rough factor loadings

Discriminability

- Point Biserial Method
  - If you have dichotomous scored items (e.g. MMPI) or items with a correct answer
  - Correlate the proportion of people getting each item correct with total test score.
  - One dichotomous variable (correct/incorrect) correlated with one continuous variable (total score) is a Point-Biserial correlation
  - Measures discriminability

Discriminability

- Point Biserial Method

<table>
<thead>
<tr>
<th>Item/Total Correlation</th>
<th>Item</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example:

- Item 1: I enjoy the excitement of a crowd.
- Item 2: I spend most of my spare time by myself.
- Item 3: I am never happier than when alone.
- Item 4: My worries seem to disappear when I get into a crowd.
- Item 5: Whenever possible I avoid being in a crowd.
- Item 6: I like to go to parties and other affairs.
Discriminability

- The discrimination can be standardized into a Z-value as well

\[ Z = \frac{1}{2} \left[ \ln(1 + r) - \ln(1 - r) \right] \]
Selecting items

- Using Difficulty and Discrimination together

![Item Difficulty and Discrimination Graph](image)

Item Characteristic Curves

- A graph of the proportion of people getting each item correct, compared to total scores on the test.
- Ideally, lower test scores should go along with lower proportions of people getting a particular item correct.
- Ideally, higher test scores should go along with higher proportions of people getting a particular item correct.

![Item Characteristic Curves Graph](image)
Item Characteristic Curves

43. I prefer to pass by people I know but have not seen for a long time, unless they speak to me first.
**Item Characteristic Curves**

- "46. I am a very sociable person."
- "82. I like to go to parties and other affairs where there is lots of loud fun."
- "151. It makes me uncomfortable to put on a party even when others are doing the same sort of things."
Item Characteristic Curves

“68: I find it hard to make talk when I meet new people.”

Item Characteristic Curves

“178. I wish I were not so shy.”

Item Characteristic Curves

“240: In a group of people I would not be embarrassed to be called upon to start a discussion or give an opinion about something I know well.”
**Item Characteristic Curves**

"246. I am likely not to speak to people until they speak to me."

**Item Characteristic Curves**

"257. I find it very hard to talk in front of the class."

**Item Characteristic Curves**

"262. I seem to make friends about as quickly as others do."
**Item Characteristic Curves**

“319. I love to go dancing.”

**Cal State Northridge - Psy 427**

---

**Item Characteristic Curves**

“320. I am never happier than when alone.”

**Cal State Northridge - Psy 427**

---

**Item Characteristic Curves**

“331. I enjoy social gatherings just to be with people.”

**Cal State Northridge - Psy 427**

---
Item Characteristic Curves

"335. I enjoy the excitement of a crowd."

```
<table>
<thead>
<tr>
<th>Total Score Groups</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>100.00</td>
</tr>
<tr>
<td>5-9</td>
<td>90.00</td>
</tr>
<tr>
<td>10-14</td>
<td>80.00</td>
</tr>
<tr>
<td>15-19</td>
<td>70.00</td>
</tr>
<tr>
<td>20-24</td>
<td>60.00</td>
</tr>
</tbody>
</table>
```

Item Characteristic Curves

"336. I do not mind meeting strangers."

```
<table>
<thead>
<tr>
<th>Total Score Groups</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>100.00</td>
</tr>
<tr>
<td>5-9</td>
<td>90.00</td>
</tr>
<tr>
<td>10-14</td>
<td>80.00</td>
</tr>
<tr>
<td>15-19</td>
<td>70.00</td>
</tr>
<tr>
<td>20-24</td>
<td>60.00</td>
</tr>
</tbody>
</table>
```

Item Characteristic Curves

"339. My worries seem to disappear when I get into a crowd of lively friends."

```
<table>
<thead>
<tr>
<th>Total Score Groups</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>100.00</td>
</tr>
<tr>
<td>5-9</td>
<td>90.00</td>
</tr>
<tr>
<td>10-14</td>
<td>80.00</td>
</tr>
<tr>
<td>15-19</td>
<td>70.00</td>
</tr>
<tr>
<td>20-24</td>
<td>60.00</td>
</tr>
</tbody>
</table>
```
### Item Characteristic Curves

"408. Some people think it's hard to get to know me."

\[ \text{Total Score Groups} \]

- 0-4
- 5-8
- 9-12
- 13-16
- 17-20
- 21-24

### Other Evaluation Techniques

- Item Response Theory
  - viewing item response curves at different levels of difficulty
  - Looks at standard error at different ranges of the trait you are trying to measure
  - More on this in the next topic
Other Evaluation Techniques

- **Criterion-Referenced Tests**
  - Instead of comparing a score on a test or scale to other respondents’ scores we can compare each individual to what they “should have scored”.
  - Requires that there is a set objective in order to assess whether the objective has been met.
  - E.g. In intro stats students should learn how to run an independent samples t-test a criterion referenced test could be used to test this. This needs to be demonstrated before moving on to another objective.

Other Evaluation Techniques

- **Criterion-Referenced Tests**
  - To evaluate CRT items
    - Give the test to 2 groups one exposed to the material and one that has not seen the material.
    - Distribute the scores for the test in a frequency polygon.
    - The antimode (least frequent value) represents the cut score between those who were exposed to the material and those who weren’t.
    - Scores above the cut score are assumed to have mastered the material, and vice versa.

Criterion Referenced Test

[Image of a Frequency Polygon of a Criterion Referenced Test]
Other Evaluation Techniques

- Criterion-Referenced Tests
  - Often used with Mastery style learning
    - Once a student indicates they’ve “mastered” the material he/she moves on to the next “module” of material
    - If they do not pass the cut score for mastery they receive more instruction until they can master the material