Statistical and psychometric methods for measurement: Scale development and validation

Andrew Ho, Harvard Graduate School of Education
The World Bank, Psychometrics Mini Course
Washington, DC. June 11, 2018


Learning Objectives

• How do we “develop and validate” a scale?

  – What is validation?
  – What is reliability?
  – What is factor analysis?
  – What is Item Response Theory?
  – How do we do all this in Stata, and interpret the output accurately?
Some motivating examples

Development and Validation of the Short Grit Scale (Grit–S)

ANGELA LEE DUCKWORTH AND PATRICK D. QUINN

Department of Psychology, University of Pennsylvania

In this article, we introduce brief self-report and informant-report versions of the Grit Scale, which measures trait-level perseverance and passion for long-term goals. The Short Grit Scale (Grit–S) retains the 2-factor structure of the original Grit Scale (Duckworth, Peterson, Matthews, & Kelly, 2007) with 4 fewer items and improved psychometric properties. We present evidence for the Grit–S’s internal consistency, test–retest stability, consensual validity with informant-report versions, and predictive validity. Among adults, the Grit–S was associated with educational attainment and fewer career changes. Among adolescents, the Grit–S longitudinally predicted GPA and, inversely, hours watching television. Among cadets at the United States Military Academy, West Point, the Grit–S predicted retention. Among Scripps National Spelling Bee competitors, the Grit–S predicted final round attained, a relationship mediated by lifetime spelling practice.

Improving Measurement Efficiency of the Inner EAR Scale with Item Response Theory

Annika Jessen, EdM¹, Andrew D. Ho, PhD¹, C. Eduardo Corrales, MD, PhD², Bevan Yueh, MD, MPH³, and Jennifer J. Shin, MD, SM²

Otolaryngology—Head and Neck Surgery
© American Academy of Otolaryngology–Head and Neck Surgery Foundation 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0194599818760528
http://otojournal.org
We say, “\(X\) is important. No one thinks of \(X\). Existing measures of \(X\) are off the mark. \(X\) matters more than everything else. If only we paid attention to \(X\).” If that is your argument, I suggest this research agenda:

1. Establish the **theoretical construct**
   - This measure should exist.

2. Establish the **latent structure**
   - This components of the measure relate as expected.

3. Establish **reliability**
   - The score you estimate should be precise.

4. Establish **predictions** and **intercorrelations**
   - These scores predict outcomes it should. They also predict outcomes things better than, over and above, other scores.

5. Establish **usefulness**
   - Using these scores achieves the intended purposes.
What is validation? My “5 Cs”

1. **Content**
   - Evidence based on tested content, the measured construct.
     - e.g., Alignment studies, theoretical development

2. **Cognition**
   - Evidence based on response processes
     - e.g., Think-aloud protocols

3. **Coherence**
   - Evidence based on internal structure
     - e.g., Reliability analyses

4. **Correlation**
   - Evidence based on relations to other variables
     - e.g., Convergent evidence

5. **Consequence**
   - Evidence based on consequences of testing
     - e.g., Long-term evaluations

---

**Common uses of “validity”, graded:**
- “The measure is valid.” Grade: C-
- “...is a valid and reliable measure” C-
- “...is a validated measure” C
- “X has validated this measure.” C+
- “Has a high validity coefficient” D
- “Validate the score” B
- “Validate the interpretation of the score as...” B+
- “I provide validity evidence for the interpretation of the score as...” A-
- “I provide validity evidence for the use of the score as...” A
Validation (Kane, 2006; 2013)

Validating the Interpretations and Uses of Test Scores

Michael T. Kane
Educational Testing Service

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. In outlining the argument-based approach to validation, this paper makes eight general points. First, it is the proposed score interpretations and uses that are validated and not the test or the test scores. Second, the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made. Third, more-ambitious claims require more support than less-ambitious claims. Fourth, more-ambitious claims (e.g., construct interpretations) tend to be more useful than less-ambitious claims, but they are also harder to validate. Fifth, interpretations and uses can change over time in response to new needs and new understandings leading to changes in the evidence needed for validation. Sixth, the evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable. Seventh, the rejection of a score use does not necessarily invalidate a prior, underlying score interpretation. Eighth, the validation of the score interpretation on which a score use is based does not validate the score use.
An 8-Step Plan

Step 1: Content and Cognition
Step 2: Scoring and Scaling
Step 3: Correlation and Reliability
Step 4: Classical Item Diagnostics
Step 5: Latent Structure Analysis
Step 6: Item Response Theory (IRT)
Step 7: IRT for Efficient Measurement
Step 8: Correlation and Prediction
### Step 1: Content and Cognition (RTFQ)

<table>
<thead>
<tr>
<th>1) New ideas and projects sometimes distract me from previous ones.</th>
<th>5) I often set a goal but later choose to pursue a different one.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2) Setbacks don’t discourage me.</td>
<td>6) I have difficulty maintaining my focus on projects that take more than a few months to complete.</td>
</tr>
<tr>
<td>3) I have been obsessed with a certain idea or project for a short time but later lost interest.</td>
<td>7) I finish whatever I begin.</td>
</tr>
<tr>
<td>4) I am a hard worker.</td>
<td>8) I am diligent.</td>
</tr>
</tbody>
</table>
Step 2: Establishing Scoring and Scaling Rules

* Encode data

```stata
forvalues varnum = 1/8 {
    gen score`varnum' = 0
    replace score`varnum' = 1 if x`varnum' == "Not like me at all"
    replace score`varnum' = 2 if x`varnum' == "Not much like me"
    replace score`varnum' = 3 if x`varnum' == "Somewhat like me"
    replace score`varnum' = 4 if x`varnum' == "Mostly like me"
    replace score`varnum' = 5 if x`varnum' == "Very much like me"
}

* Reverse polarity for items 1, 3, 5, and 6
foreach var of varlist score1 score3 score5 score6 {
    replace `var' = 6-`var'
}

label variable score1 "New ideas... sometimes distract..."
label variable score2 "Setbacks don't discourage..."
label variable score3 "I have been obsessed... but..."
label variable score4 "I am a hard worker."
label variable score5 "I often set a goal but later..."
label variable score6 "I have difficulty maintaining..."
label variable score7 "I finish whatever I begin."
label variable score8 "I am diligent."
```
What are the distributions of item scores for grit?

*New ideas... sometimes distract...

Setbacks don't discourage...

*I have been obsessed... but...

I am a hard worker.

*I often set a goal but later...

*I have difficulty maintaining...

I finish whatever I begin.

I am diligent.
### Inner EAR

**Instructions:** The purpose of this questionnaire is to identify the problems your hearing loss may be causing you. Please answer all the questions based on your experiences over the past 2 weeks. If you use a hearing aid, please answer the way you hear based on normal use of your hearing aid. You may leave any question unanswered.

1. **Over the past 2 weeks,** how would you rate your ability to hear?

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hate it</td>
<td>It's OK</td>
<td>Love it</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. **Over the past 2 weeks,** how would you rate your ability to understand:

   - what family and close friends are saying?
   - speech in a quiet room?
   - what is said by those at your table in a crowded restaurant?
   - what you want to hear, and filter out unwanted noises?
   - telephone conversations?

<table>
<thead>
<tr>
<th></th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Very Good</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

3. **Over the past 2 weeks,** how would you rate your ability to hear:

   - in different listening situations?
   - soft household sounds (car turn signal, clock ticking)?

<table>
<thead>
<tr>
<th></th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Very Good</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

4. **Over the past 2 weeks,** how would you rate your mood based on your ability to hear?

<table>
<thead>
<tr>
<th></th>
<th>Very Bothered</th>
<th>Bothered</th>
<th>A little Bothered</th>
<th>Not Bothered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

5. **Over the past 2 weeks,** how bothered are you:

   - by having to ask people to repeat things?
   - by having limited communication because of your hearing loss?

<table>
<thead>
<tr>
<th></th>
<th>Very Bothered</th>
<th>Bothered</th>
<th>A little Bothered</th>
<th>Not Bothered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Step 2: Establishing Scoring and Scaling Rules

* Loading dataset, Rescale variables to the same scale in new variables

```plaintext
import excel "Raw Data\Export for AH AJ 11232016.xls", sheet("Sheet1") firstrow clear

rename innerear11 ear11

recode innereararray234innerear2a (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2a)
recode innereararray234innerear2b (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2b)
recode innereararray234innerear2c (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2c)
recode innereararray234innerear2d (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2d)
recode innereararray234innerear2e (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear2e)
recode innereararray234innerear3a (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear3a)
recode innereararray234innerear3b (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear3b)
recode innereararray234innerear4 (1=0) (2=25) (3=50) (4=75) (5=100), gen(ear4)
recode innereararray5innerear5a (1=0) (2=33) (3=67) (4=100), gen(ear5a)
recode innereararray5innerear5b (1=0) (2=33) (3=67) (4=100), gen(ear5b)
```

* Rescale variables to the same scale (just ear11)

```plaintext
gen ear11a = .
replace ear11a = 0 if ear11 == 0
replace ear11a = 10 if ear11 == 1
replace ear11a = 20 if ear11 == 2
replace ear11a = 30 if ear11 == 3
replace ear11a = 40 if ear11 == 4
replace ear11a = 50 if ear11 == 5
replace ear11a = 60 if ear11 == 6
replace ear11a = 70 if ear11 == 7
replace ear11a = 80 if ear11 == 8
replace ear11a = 90 if ear11 == 9
replace ear11a = 100 if ear11 == 10
```
What are the distributions of item scores for Inner Ear?
Step 3: Correlations and Cronbach’s alpha

```
pwcorr score1-score8, star(.05)
```

<table>
<thead>
<tr>
<th></th>
<th>score1</th>
<th>score2</th>
<th>score3</th>
<th>score4</th>
<th>score5</th>
<th>score6</th>
<th>score7</th>
<th>score8</th>
</tr>
</thead>
<tbody>
<tr>
<td>score1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score2</td>
<td>-0.0804</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score3</td>
<td>0.3359*</td>
<td>0.1244</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score4</td>
<td>0.0037</td>
<td>0.2302*</td>
<td>0.1233</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score5</td>
<td>0.4149*</td>
<td>-0.0827</td>
<td>0.3990*</td>
<td>0.1977</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score6</td>
<td>0.4684*</td>
<td>-0.0404</td>
<td>0.4775*</td>
<td>0.2997*</td>
<td>0.4930*</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>score7</td>
<td>0.3028*</td>
<td>0.2392*</td>
<td>0.3270*</td>
<td>0.5040*</td>
<td>0.4303*</td>
<td>0.2845*</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>score8</td>
<td>0.1604</td>
<td>0.2625*</td>
<td>0.1677</td>
<td>0.7783*</td>
<td>0.3472*</td>
<td>0.3283*</td>
<td>0.5716*</td>
<td></td>
</tr>
</tbody>
</table>

- **score1**: "New ideas... sometimes distract..."
- **score2**: "Setbacks don't discourage..."
- **score3**: "I have been obsessed... but..."
- **score4**: "I am a hard worker."
- **score5**: "I often set a goal but later..."
- **score6**: "I have difficulty maintaining..."
- **score7**: "I finish whatever I begin."
- **score8**: "I am diligent."

---

© Andrew Ho

Harvard Graduate School of Education
### Table 1. Item Means (SD) and Interitem Correlations (n = 334-353).

<table>
<thead>
<tr>
<th>Item</th>
<th>ear11</th>
<th>ear2a</th>
<th>ear2b</th>
<th>ear2c</th>
<th>ear2d</th>
<th>ear2e</th>
<th>ear3a</th>
<th>ear3b</th>
<th>ear4</th>
<th>ear5a</th>
<th>ear5b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (0-100)</strong></td>
<td>56.23</td>
<td>41.22</td>
<td>50.71</td>
<td>26.91</td>
<td>31.44</td>
<td>47.03</td>
<td>36.15</td>
<td>39.06</td>
<td>47.59</td>
<td>48.36</td>
<td>68.21</td>
</tr>
<tr>
<td>ear11: Ability to hear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family and close friends (ear2a)</td>
<td>0.62</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech in a quiet room (ear2b)</td>
<td>0.49</td>
<td>0.74</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conversation in a crowded restaurant (ear2c)</td>
<td>0.49</td>
<td>0.69</td>
<td>0.59</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter out unwanted noises (ear2d)</td>
<td>0.48</td>
<td>0.72</td>
<td>0.59</td>
<td>0.80</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone conversations (ear2e)</td>
<td>0.41</td>
<td>0.63</td>
<td>0.61</td>
<td>0.51</td>
<td>0.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ear2: Ability to understand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In different listening situations (ear3a)</td>
<td>0.54</td>
<td>0.76</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.68</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft household sounds (ear3b)</td>
<td>0.47</td>
<td>0.65</td>
<td>0.68</td>
<td>0.56</td>
<td>0.61</td>
<td>0.72</td>
<td>0.70</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ear4: Mood based on your ability to hear</td>
<td>0.45</td>
<td>0.59</td>
<td>0.53</td>
<td>0.52</td>
<td>0.59</td>
<td>0.62</td>
<td>0.67</td>
<td>0.64</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ear5: Bother from .</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Need for repetition (ear5a)</td>
<td>0.33</td>
<td>0.52</td>
<td>0.39</td>
<td>0.52</td>
<td>0.52</td>
<td>0.42</td>
<td>0.54</td>
<td>0.48</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Restricted activities because of hearing (ear5b)</td>
<td>0.42</td>
<td>0.47</td>
<td>0.42</td>
<td>0.44</td>
<td>0.44</td>
<td>0.39</td>
<td>0.50</td>
<td>0.46</td>
<td>0.59</td>
<td>0.65</td>
<td>1.00</td>
</tr>
</tbody>
</table>
A response $i$ to item $j$ by person $k$

$$y_{ijk} = \mu + \zeta_j + \zeta_k + \epsilon_{ijk};$$

$$\zeta_j \sim N(0, \psi_1);$$

$$\zeta_k \sim N(0, \psi_2);$$

$$\epsilon_{ijk} \sim N(0, \theta).$$

Note: Only 1 score per person/item combination

- $\mu$ — Overall average score
- $\zeta_j$ — Item location (easiness), $\psi_1$ — variance of item effects
- $\zeta_k$ — Person location (proficiency), $\psi_2$ — variance of person effects
- $\epsilon_{ijk}$ — Person-item interactions and other effects, $\theta$ — error variance
Reliability: What are two relevant intraclass correlations?

- A response $i$ to item $j$ by person $k$: $y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}$;
  - $\zeta_j \sim N(0, \psi_1)$;
  - $\zeta_k \sim N(0, \psi_2)$;
  - $\varepsilon_{ijk} \sim N(0, \theta)$.

- $\mu$ — Overall average score
- $\zeta_j$ — Item location (easiness). Variance: $\psi_1$
- $\zeta_k$ — Person location (proficiency). Variance: $\psi_2$
- $\varepsilon_{ijk}$ — Person-item interactions and other effects. Variance: $\theta$

- Intraclass correlation: $\rho = \frac{\psi_2}{\psi_2 + \theta}$. The correlation between two item responses within persons. The proportion of relative response variation due to persons.

- Intraclass correlation: $\rho_\alpha = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$. Cronbach’s alpha: The correlation between two average (or sum) scores within persons. The proportion of relative score variance due to persons.
Estimation in Stata

- A response $i$ to item $j$ by person $k$: $y_{ijk} = \mu + \zeta_j + \zeta_k + \epsilon_{ijk}$;
  - $\zeta_j \sim N(0, \psi_1)$;
  - $\zeta_k \sim N(0, \psi_2)$;
  - $\epsilon_{ijk} \sim N(0, \theta)$.

\[
\hat{\rho}_\alpha = \frac{\hat{\psi}_{\text{person}}}{\hat{\psi}_{\text{person}} + \frac{\hat{\theta}}{11}} = \frac{416.45}{416.45 + \frac{334.83}{11}} = .93
\]

Relevant intraclass correlation:

\[
\hat{\rho}_\alpha = \frac{416.45}{416.45 + \frac{334.83}{11}} = .93
\]
Cronbach’s alpha directly, in Stata

- A response $i$ to item $j$ by person $k$: $y_{ijk} = \mu + \zeta_j + \zeta_k + \epsilon_{ijk}$;
  \[ \zeta_j \sim N(0, \psi_1); \]
  \[ \zeta_k \sim N(0, \psi_2); \]
  \[ \epsilon_{ijk} \sim N(0, \theta). \]

- Classical computational formula for Cronbach’s alpha:
  \[ \hat{\rho}_\alpha = \frac{n_j}{n_j - 1} \left( 1 - \frac{\sum_j \sigma_{X_j}^2}{\sigma_X^2} \right), \]

- where $\sigma_{X_j}^2$ is the variance of each item score $X_j$, and $\sigma_X^2$ is the variance of a total (summed) score, $X$.

- In Stata:
  \[ . \text{alpha ear11a-ear5b, asis} \]

Test scale = mean(unstandardized items)

Average interitem covariance: 410.7885
Number of items in the scale: 11
Scale reliability coefficient: 0.9304

© Andrew Ho
Harvard Graduate School of Education
How should I think about reliability?

Three Necessary Intuitions

1. Any observed score is **one of many possible replications**.

2. Any observed score is the **sum of a “true score”** (average of all theoretical replications) **and an error term**.

3. Averaging over replications gives us **better estimates** of “true” scores by **averaging over error terms**.
What is the reliability of grit scores?

```
. alpha score1-score8, asis
```

Test scale = mean(unstandardized items)

Average interitem covariance: .2955753
Number of items in the scale: 8
Scale reliability coefficient: 0.7565

### Three Interpretations of Reliability

1. Reliability is the correlation between two sets of observed scores from a replication of a measurement procedure.

2. Reliability is the proportion of “observed score variance” that is accounted for by “true score variance.”

3. Reliability that starts with an average of pairwise part covariances, then increases this average as a function of the number of replications? Why? Because averaging over replications decreases error variance.

Cronbach’s α is a particular type of reliability, one of the most limited, but easy to estimate. Cronbach’s α only considers correlations of scores (or variance) across replications of items.
Spearman-Brown “Prophecy”: How many items do I need for precision?

From some baseline reliability, $\rho$, Spearman-Brown “prophesizes” that increasing the replications (items?) by a multiplicative factor of $K$ will result in reliability ($K$ may be a fraction):

$$\rho_{SB} = \frac{K\rho}{1 + (K - 1)\rho}$$

Note: Given $\rho_\alpha$ for a $J$-item test, and prophecy for a $J'$-item test, you can
1) calculate $K = \frac{J'}{J}$, or
2) use $K_1 = \frac{1}{J}$ for reliability of a 1-item test, then prophesize using $K_2 = J'$.

Using multilevel crossed effects models with person variance $\psi_1$ and error variance $\theta$, we have an equivalent formula:

$$\rho_{SB} = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$$
### Step 4: Classical Item Diagnostics

```
. alpha ear11a-ear5b, asis item
```

Test scale = mean(unstandardized items)

<table>
<thead>
<tr>
<th>Item</th>
<th>Obs</th>
<th>Sign</th>
<th>item-test correlation</th>
<th>item-rest correlation</th>
<th>average interitem covariance</th>
<th>alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>ear11a</td>
<td>334</td>
<td>+</td>
<td>0.6614</td>
<td>0.5945</td>
<td>435.7858</td>
<td>0.9295</td>
</tr>
<tr>
<td>ear2a</td>
<td>353</td>
<td>+</td>
<td>0.8513</td>
<td>0.8168</td>
<td>406.6422</td>
<td>0.9198</td>
</tr>
<tr>
<td>ear2b</td>
<td>353</td>
<td>+</td>
<td>0.7775</td>
<td>0.7277</td>
<td>415.1858</td>
<td>0.9234</td>
</tr>
<tr>
<td>ear2c</td>
<td>353</td>
<td>+</td>
<td>0.7872</td>
<td>0.7430</td>
<td>419.3079</td>
<td>0.9232</td>
</tr>
<tr>
<td>ear2d</td>
<td>353</td>
<td>+</td>
<td>0.8164</td>
<td>0.7748</td>
<td>411.07</td>
<td>0.9216</td>
</tr>
<tr>
<td>ear2e</td>
<td>353</td>
<td>+</td>
<td>0.7683</td>
<td>0.7130</td>
<td>411.6634</td>
<td>0.9240</td>
</tr>
<tr>
<td>ear3a</td>
<td>352</td>
<td>+</td>
<td>0.8720</td>
<td>0.8430</td>
<td>407.1448</td>
<td>0.9191</td>
</tr>
<tr>
<td>ear3b</td>
<td>352</td>
<td>+</td>
<td>0.8139</td>
<td>0.7646</td>
<td>400.3318</td>
<td>0.9215</td>
</tr>
<tr>
<td>ear4</td>
<td>353</td>
<td>+</td>
<td>0.8009</td>
<td>0.7485</td>
<td>401.6557</td>
<td>0.9222</td>
</tr>
<tr>
<td>ear5a</td>
<td>351</td>
<td>+</td>
<td>0.7171</td>
<td>0.6388</td>
<td>407.004</td>
<td>0.9283</td>
</tr>
<tr>
<td>ear5b</td>
<td>349</td>
<td>+</td>
<td>0.7004</td>
<td>0.6099</td>
<td>402.6394</td>
<td>0.9310</td>
</tr>
<tr>
<td>Test scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>410.7885</td>
<td>0.9304</td>
</tr>
</tbody>
</table>
### Step 4: Classical Item Diagnostics

<table>
<thead>
<tr>
<th>Item</th>
<th>Obs</th>
<th>Sign</th>
<th>Item-test correlation</th>
<th>Item-rest correlation</th>
<th>Average interitem covariance</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>ear11a</td>
<td>334</td>
<td>+</td>
<td>0.6614</td>
<td>0.5945</td>
<td>435.7858</td>
<td>0.9295</td>
</tr>
<tr>
<td>ear2a</td>
<td>353</td>
<td>+</td>
<td>0.8513</td>
<td>0.8168</td>
<td>406.6422</td>
<td>0.9198</td>
</tr>
<tr>
<td>ear2b</td>
<td>353</td>
<td>+</td>
<td>0.7775</td>
<td>0.7277</td>
<td>415.1858</td>
<td>0.9234</td>
</tr>
<tr>
<td>ear2c</td>
<td>353</td>
<td>+</td>
<td>0.7872</td>
<td>0.7430</td>
<td>419.3079</td>
<td>0.9232</td>
</tr>
<tr>
<td>ear2d</td>
<td>353</td>
<td>+</td>
<td>0.8164</td>
<td>0.7748</td>
<td>411.07</td>
<td>0.9216</td>
</tr>
<tr>
<td>ear2e</td>
<td>353</td>
<td>+</td>
<td>0.7683</td>
<td>0.7130</td>
<td>411.6634</td>
<td>0.9240</td>
</tr>
<tr>
<td>ear3a</td>
<td>352</td>
<td>+</td>
<td>0.8720</td>
<td>0.8430</td>
<td>407.1448</td>
<td>0.9191</td>
</tr>
</tbody>
</table>

These are **diagnostics that explain item functioning** and sometimes, with additional analysis, warrant item adaptation or exclusion. However, **no item should be altered or excluded on the basis of these statistics alone.**

- **Item-Test Correlation** is a simple correlation between each item response and total test scores (the higher the better). This correlation is sometimes called “classical item discrimination.” Think of it as “item information.”

- **Item-Rest Correlation** is the similar, but the total test score excludes the target item (the higher the better). This avoids a part-whole confounding in correlation.

- **Interitem Correlation** shows the would-be interitem covariance if the item were excluded (the lower the better).

- **Alpha (excluded-item alpha)** shows the would-be $\rho_\alpha$ estimate if the item were excluded (the lower the better).
Step 5: Latent Structure Analysis – Principal Factor Analysis

1) Replace diagonals with an estimate of reliability
2) Conduct a principal components analysis.

<table>
<thead>
<tr>
<th>ear11: Ability to hear</th>
<th>ear2: Ability to understand...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>ear11a</td>
<td>1.00</td>
</tr>
<tr>
<td>ear2a</td>
<td></td>
</tr>
<tr>
<td>ear2b</td>
<td></td>
</tr>
<tr>
<td>ear2c</td>
<td></td>
</tr>
<tr>
<td>ear2d</td>
<td></td>
</tr>
<tr>
<td>ear2e</td>
<td></td>
</tr>
<tr>
<td>ear3a</td>
<td></td>
</tr>
<tr>
<td>ear3b</td>
<td></td>
</tr>
<tr>
<td>ear4</td>
<td></td>
</tr>
<tr>
<td>ear5a</td>
<td></td>
</tr>
<tr>
<td>ear5b</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>correlation coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family and close friends (ear2a)</td>
</tr>
<tr>
<td>Speech in a quiet room (ear2b)</td>
</tr>
<tr>
<td>Conversation in a crowded restaurant (ear2c)</td>
</tr>
<tr>
<td>Filter out unwanted noises (ear2d)</td>
</tr>
<tr>
<td>Telephone conversations (ear2e)</td>
</tr>
<tr>
<td>In different listening situations (ear3a)</td>
</tr>
<tr>
<td>Soft household sounds (ear3b)</td>
</tr>
<tr>
<td>ear4: Mood based on your ability to hear</td>
</tr>
<tr>
<td>ear5: Bother from...</td>
</tr>
<tr>
<td>Need for repetition (ear5a)</td>
</tr>
</tbody>
</table>
Factor analysis/correlation
Method: principal factors
Rotation: (unrotated)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Eigenvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor1</td>
<td>6.15209</td>
</tr>
<tr>
<td>Factor2</td>
<td>0.54844</td>
</tr>
<tr>
<td>Factor3</td>
<td>0.35753</td>
</tr>
<tr>
<td>Factor4</td>
<td>0.18585</td>
</tr>
<tr>
<td>Factor5</td>
<td>0.01411</td>
</tr>
<tr>
<td>Factor6</td>
<td>-0.02968</td>
</tr>
<tr>
<td>Factor7</td>
<td>-0.04374</td>
</tr>
<tr>
<td>Factor8</td>
<td>-0.08464</td>
</tr>
<tr>
<td>Factor9</td>
<td>-0.12194</td>
</tr>
<tr>
<td>Factor10</td>
<td>-0.12772</td>
</tr>
<tr>
<td>Factor11</td>
<td>-0.18856</td>
</tr>
</tbody>
</table>

LR test: independent vs. saturated: \( \text{chi2}(55) = 2464.79 \) Prob>\( \text{chi2} = 0.0000 \)
Step 5: Latent Structure Analysis – Principal Factor Analysis

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor1</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>ear11a</td>
<td>0.6229</td>
<td>0.6120</td>
</tr>
<tr>
<td>ear2a</td>
<td>0.8410</td>
<td>0.2927</td>
</tr>
<tr>
<td>ear2b</td>
<td>0.7546</td>
<td>0.4306</td>
</tr>
<tr>
<td>ear2c</td>
<td>0.7760</td>
<td>0.3978</td>
</tr>
<tr>
<td>ear2d</td>
<td>0.8081</td>
<td>0.3470</td>
</tr>
<tr>
<td>ear2e</td>
<td>0.7298</td>
<td>0.4674</td>
</tr>
<tr>
<td>ear3a</td>
<td>0.8641</td>
<td>0.2533</td>
</tr>
<tr>
<td>ear3b</td>
<td>0.7817</td>
<td>0.3889</td>
</tr>
<tr>
<td>ear4</td>
<td>0.7460</td>
<td>0.4435</td>
</tr>
<tr>
<td>ear5a</td>
<td>0.6314</td>
<td>0.6013</td>
</tr>
<tr>
<td>ear5b</td>
<td>0.6218</td>
<td>0.6134</td>
</tr>
</tbody>
</table>
```
. sem (ear11a-ear5b <- ETA), var(ETA@1) standardized

<table>
<thead>
<tr>
<th>Standardized</th>
<th>OIM</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>ear11a</td>
<td>.6265437</td>
<td>.0353141</td>
<td>17.74</td>
<td>.000</td>
<td>.5266186    .7006721</td>
<td></td>
</tr>
<tr>
<td>ETA_cons</td>
<td>2.362817</td>
<td>.1073507</td>
<td>22.01</td>
<td>.000</td>
<td>1.804841    2.178095</td>
<td></td>
</tr>
<tr>
<td>ear2a</td>
<td>.8454712</td>
<td>.0179031</td>
<td>47.22</td>
<td>.000</td>
<td>.3394463    .0339581</td>
<td></td>
</tr>
<tr>
<td>ETA_cons</td>
<td>1.642187</td>
<td>.0844864</td>
<td>19.44</td>
<td>.000</td>
<td>.3965921    .0371671</td>
<td></td>
</tr>
<tr>
<td>ear2b</td>
<td>.7639354</td>
<td>.0249605</td>
<td>30.61</td>
<td>.000</td>
<td>.6569394    .0446471</td>
<td></td>
</tr>
<tr>
<td>ETA_cons</td>
<td>1.991468</td>
<td>.0952197</td>
<td>20.91</td>
<td>.000</td>
<td>1 (constrained)</td>
<td></td>
</tr>
</tbody>
</table>

LR test of model vs. saturated: chi2(44) = 335.69, Prob > chi2 = 0.0000
The comparative fit index (Bentler 1990) is computed as
\[
CFI = 1 - \left[ \frac{(\chi^2_{ms} - df_{ms})}{\max\{(\chi^2_{bs} - df_{bs}), (\chi^2_{ms} - df_{ms})\}} \right]
\]

The Tucker–Lewis index (Bentler 1990) is computed as
\[
TLI = \frac{(\chi^2_{bs}/df_{bs}) - (\chi^2_{ms}/df_{ms})}{(\chi^2_{bs}/df_{bs}) - 1}
\]

What percent of worst possible (baseline vs. saturated) fit does my model account for? Here, .881 is 88.1% of the bad fit. Around .9 is generally “okay.”
SEM Goodness of Fit (briefly): Population Error

| Population error |  
|------------------|---
| RMSEA            | 0.142 Root mean squared error of approximation  
| 90% CI, lower bound | 0.128  
| upper bound      | 0.156  
| pclose           | 0.000 Probability RMSEA <= 0.05

\[
RMSEA = \sqrt{\frac{\chi^2 - df}{N \times df}}
\]

- Favors simpler models and larger sample sizes. Lower the better.
- Can we be somewhat sure that the standardized distance (badness of fit) is low? (lower bound of 90% CI less than .05)
- And can we be somewhat sure that the standardized distance (badness of fit) is not high? (upper bound of 90% CI less than .10).
An 8-Step Plan

Step 1: Content and Cognition
Step 2: Scoring and Scaling
Step 3: Correlation and Reliability
Step 4: Classical Item Diagnostics
Step 5: Latent Structure Analysis
Step 6: Item Response Theory (IRT)
Step 7: IRT for Efficient Measurement
Step 8: Correlation and Prediction
Step 6: Item Response Theory. Why IRT?

- Item response theory (IRT) supports the vast majority of large-scale educational assessments.
  - State testing programs
  - National and international assessments (NAEP, TIMSS, PIRLS, PISA).
  - Selection testing (SAT, ACT)

- Many presentations of IRT use unfamiliar jargon and specialized software.
  - We will try to connect IRT to other more flexible statistical modeling frameworks.
  - We will use Stata.
Classical Test Theory vs. Item Response Theory

- CTT: A response $i$ to item $j$ by person $k$:
  \[ y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}; \]
  \[ \zeta_j \sim N(0, \psi_1); \]
  \[ \zeta_k \sim N(0, \psi_2); \]
  \[ \varepsilon_{ijk} \sim N(0, \theta). \]

- IRT: A response $i$ to item $j$ by person $k$:
  \[
  \log \left( \frac{P(y_{ijk} = 1)}{1 - P(y_{ijk} = 1)} \right) = \alpha_j + \zeta_k; \\
  \zeta_k \sim N(0,1). 
  \]
  - A logistic model vs. a linear model.
  - Fixed item effects ($\alpha_j$) vs. random item effects ($\zeta_j$).
  - Both models have random effects for persons.
  - IRT extends to a fixed slope coefficient for items, $\beta_j$, on the random slope:
    \[
    \log \left( \frac{P(y_{ijk} = 1)}{1 - P(y_{ijk} = 1)} \right) = \alpha_j + \beta_j \zeta_k; \\
    \zeta_k \sim N(0,1) 
    \]
**Slope-Intercept Parameterization**

We’re familiar with logistic regression models of the form:

\[
\log \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X
\]

\(\beta_0\) is the \(y\)-intercept, and \(\beta_1\) is the slope.

IRT models can have a similar parameterization:

\[
\log \left( \frac{P(X = 1)}{1 - P(X = 1)} \right) = \alpha \theta - \beta
\]

Note \(\beta\) is the negative \(y\)-intercept, corresponding to difficulty, and \(\alpha\) is the slope. See the figure at right. Notice that \(\beta\) is on the logit scale, the \(-y\)-intercept.

**Discrimination-Difficulty Parameterization**

In contrast, in IRT, we prefer to think of difficulty on the same scale as \(\theta\), as the \(x\)-intercept.

So, we use the parameterization:

\[
\log \left( \frac{P(X = 1)}{1 - P(X = 1)} \right) = a(\theta - b)
\]

The slope here, \(a\), is equal to \(\alpha\) in the slope-intercept parameterization, but \(b = \beta / \alpha\) and \(\beta = ab\).

- In slope-intercept parameterization, \(\beta\) is the log-odds of getting an item wrong when \(\theta = 0\).
- In discrimination-difficulty parameterization, \(b\) is the \(\theta\) you need for even odds (50%) of a correct answer.
1 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)
2 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)

\[ \log \left( \frac{P_i(\theta_p)}{1 - P_i(\theta_p)} \right) = a_i(\theta_p - b_i); \theta_p \sim N(0,1) \]
Item Characteristic Curve (ICC) Slider Questions

- What happens when we increase $a$ for the blue item? Which item is more discriminating?
- What happens when we increase $b$ for the blue item? Which item is more difficult?
- What happens when we increase $c$ for the blue item? Which item is more discriminating?
- Try setting blue to .84, 0, .05 and red to .95,.3, .26. Why might the $c$ parameter be the most difficult to estimate in practice?
- Given this overlap, comparisons of items in terms of item parameters instead of full curves will be shortsighted.
  - Difficulty for which $\theta$? Discrimination for which $\theta$?
- For reference, the probability of a correct response when $\theta_p = b_i$ is $\frac{1+c_i}{2}$. The slope at this inflection point is $\frac{a_i(1-c_i)}{4}$.
1 Parameter Logistic (1PL) Model: The Rasch Model

* Fit the item response model
irt 1pl item1-item42
estimates store mcas1pl

* List items by difficulty
estat report, sort(b) byparm

* Plotting Item Characteristic Curves
irtgraph icc, legend(off)
irtgraph ic. estat report, sort(b) byparm

* Estimation: One-parameter logistic model
predict theLog likelihood = -108475.31

|       | Coef.    | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|-------|----------|-----------|-------|------|----------------------|
|       |          |           |       |      |                      |
| Discr | 1.084066 | .0131721  | 82.30 | 0.000| 1.058249             | 1.109882             |

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>item1</td>
<td>-2.58876</td>
<td>.0586296</td>
<td>-44.15</td>
<td>0.000</td>
<td>-2.703671</td>
</tr>
<tr>
<td>item17</td>
<td>-2.491368</td>
<td>.0566288</td>
<td>-43.99</td>
<td>0.000</td>
<td>-2.602359</td>
</tr>
<tr>
<td>item20</td>
<td>-2.177048</td>
<td>.0508108</td>
<td>-42.85</td>
<td>0.000</td>
<td>-2.276635</td>
</tr>
<tr>
<td>item31</td>
<td>-2.075762</td>
<td>.0491291</td>
<td>-42.25</td>
<td>0.000</td>
<td>-2.172053</td>
</tr>
<tr>
<td>item22</td>
<td>-2.073947</td>
<td>.0490997</td>
<td>-42.24</td>
<td>0.000</td>
<td>-2.170181</td>
</tr>
</tbody>
</table>
1 Parameter Logistic (1PL) Item Characteristic Curves (ICCs)
1 Parameter Logistic (1PL) ICCs in Logit Space (Linear)

Items 1-8
The Rasch (1PL) Scale Transformation

- Compression of central scores, stretching of extremes.
- Relative error initially greater in central scores, afterwards greater at extremes.
- Information initially concentrated at extreme score points, afterwards concentrated centrally.
The 2-Parameter Logistic (2PL) IRT Model

\[
\log \left( \frac{P_i(\theta_p)}{1 - P_i(\theta_p)} \right) = a_i(\theta_p - b_i)
\]

Discrimination is the difference in the log-odds of a correct answer for every SD distance of \(\theta_p\) from \(b_i\).

Likelihood ratio test: reject null hypothesis that discrimination parameters are jointly equal. 2PL fits better than 1PL.
2-Parameter Logistic (2PL) ICCs
The 3-Parameter Logistic (3PL) IRT Model

\[ P_i(\theta_p) = c + \frac{(1 - c) \left( \exp \left( a_i (\theta_p - b_i) \right) \right)}{1 + \exp \left( a_i (\theta_p - b_i) \right)} \]

<table>
<thead>
<tr>
<th>item14</th>
<th>-.2654169</th>
<th>.045632</th>
<th>-5.82</th>
<th>0.000</th>
<th>-.3548539</th>
<th>-.1759799</th>
</tr>
</thead>
<tbody>
<tr>
<td>item18</td>
<td>-.1773411</td>
<td>.0421862</td>
<td>-4.20</td>
<td>0.000</td>
<td>-.2600245</td>
<td>-.0946577</td>
</tr>
<tr>
<td>item11</td>
<td>-.1407813</td>
<td>.0463368</td>
<td>-3.04</td>
<td>0.002</td>
<td>-.2315998</td>
<td>-.0499628</td>
</tr>
<tr>
<td>item27</td>
<td>-.1040767</td>
<td>.0468187</td>
<td>-2.22</td>
<td>0.026</td>
<td>-.1958397</td>
<td>-.0123137</td>
</tr>
<tr>
<td>item30</td>
<td>.2031855</td>
<td>.0406553</td>
<td>5.00</td>
<td>0.000</td>
<td>.1235025</td>
<td>.2828685</td>
</tr>
<tr>
<td>item7</td>
<td>.2036091</td>
<td>.0339582</td>
<td>6.00</td>
<td>0.000</td>
<td>.1370522</td>
<td>.270166</td>
</tr>
<tr>
<td>item8</td>
<td>.3319137</td>
<td>.0322655</td>
<td>10.29</td>
<td>0.000</td>
<td>.2686745</td>
<td>.3951529</td>
</tr>
<tr>
<td>Guess</td>
<td>.056637</td>
<td>.0092217</td>
<td>6.14</td>
<td>0.000</td>
<td>.0385629</td>
<td>.0747112</td>
</tr>
</tbody>
</table>

The common \( c \) parameter estimate is an estimated lower asymptote, the pseudo-guessing parameter. Estimated in common across items \( c \) rather than \( c_i \) due to considerable estimation challenges in practice.

```
. lrtest mcas2pl mcas3pl
```

Likelihood-ratio test

( Assumption: mcas2pl nested in mcas3pl )

<table>
<thead>
<tr>
<th>LR chi2(1)</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>31.08</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Likelihood ratio test: reject null hypothesis that the common pseudo-guessing parameter is zero.
3-Parameter Logistic (3PL) ICCs

Item Characteristic Curves

© Andrew Ho
Harvard Graduate School of Education
Graphical Goodness of Fit for Items 1 and 8

Predicted mean (item1)

Predicted mean (item8)
eicc1
eicc8
• Rasch (1PL): 20 items and 200 examinees.
• Hulin, Lissak, and Drasgow:
  – 2PL: 30 items, 500 examinees.
  – 3PL: 60 items, 1000 examinees.
  – Tradeoffs, maybe 30 items and 2000 examinees.
• Swaminathan and Gifford:
  – 3PL: 20 items, 1000 examinees.
• Low scoring examinees needed for 3PL.
• Large samples (above 3500) needed for polychotomous items (scored 0/1/2/...), particularly high or low difficulty items that will have even higher or lower score points.
Estimating $\theta_p$ via Empirical Bayes (EAP)

Empirical Bayes means for Theta

```
predict theta1pl, latent se(th1plse)
```
Dichotomous vs. Polytomous IRT

• Define $k = 0,1, \ldots, K$ categories, where $K = 1$ is the dichotomous case (responses scored 0 or 1) and $K \geq 2$ is the polytomous case.

  – Note that $K$ refers to the number of category boundaries or “cut scores.”

Dichotomous IRT:

$$P(X_{pi} = 1|a_i, b_i, \theta_p) = \frac{1}{1 + \exp(-a_i(\theta_p - b_i))}$$

$$\log\left(\frac{P_i(\theta_p)}{1 - P_i(\theta_p)}\right) = a_i(\theta_p - b_i)$$

Polytomous (Graded Response Model):  

$$P(X_{pi} \geq k|a_i, b_{ik}, \theta_p) = \frac{1}{1 + \exp(-a_i(\theta_p - b_{ik}))}$$

$$\log\left(\frac{P_{ik}(\theta_p)}{1 - P_{ik}(\theta_p)}\right) = a_i(\theta_p - b_{ik})$$

Graded Response Model (GRM), Slope-Intercept Parameterization:

$$P(X_{pi} \geq k|\alpha_i, \beta_{ik}, \theta_p) = \frac{1}{1 + \exp\left(-(\alpha_i \theta_p - \beta_{ik})\right)}$$

$$\log\left(\frac{P_{ik}(\theta_p)}{1 - P_{ik}(\theta_p)}\right) = \alpha_i \theta_p - \beta_{ik}; \quad a_i = \alpha_i; \quad b_{ik} = \frac{\beta_{ik}}{\alpha_i}$$
### Table 2. Item Information and Location Parameter Estimates Based on a Graded Response Model (n = 353).a

<table>
<thead>
<tr>
<th>Item Code</th>
<th>Information Parameter Estimates: a</th>
<th>Location Parameter Estimates</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
<th>b5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ear11 (b1-b5)</td>
<td>1.63</td>
<td></td>
<td>-4.24</td>
<td>-2.39</td>
<td>-1.92</td>
<td>-1.30</td>
<td>-0.70</td>
</tr>
<tr>
<td>ear11 (b6-b10)</td>
<td>0.28</td>
<td></td>
<td>0.45</td>
<td>0.79</td>
<td>1.55</td>
<td>2.57</td>
<td></td>
</tr>
<tr>
<td>ear2a</td>
<td>3.79</td>
<td></td>
<td>-1.37</td>
<td>-0.10</td>
<td>0.94</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>ear2b</td>
<td>2.72</td>
<td></td>
<td>-1.78</td>
<td>-0.65</td>
<td>0.62</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td>ear2c</td>
<td>2.70</td>
<td></td>
<td>-0.65</td>
<td>0.62</td>
<td>1.72</td>
<td>2.90</td>
<td></td>
</tr>
<tr>
<td>ear2d</td>
<td>3.05</td>
<td></td>
<td>-0.84</td>
<td>0.37</td>
<td>1.42</td>
<td>2.32</td>
<td></td>
</tr>
<tr>
<td>ear2e</td>
<td>2.33</td>
<td></td>
<td>-1.55</td>
<td>-0.42</td>
<td>0.79</td>
<td>1.72</td>
<td></td>
</tr>
<tr>
<td>ear3a</td>
<td>4.52</td>
<td></td>
<td>-1.17</td>
<td>0.20</td>
<td>1.16</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>ear3b</td>
<td>2.63</td>
<td></td>
<td>-0.97</td>
<td>-0.09</td>
<td>0.98</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td>ear4</td>
<td>2.45</td>
<td></td>
<td>-1.43</td>
<td>-0.42</td>
<td>0.79</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>ear5a</td>
<td>1.69</td>
<td></td>
<td>-1.23</td>
<td>-0.05</td>
<td>1.59</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>ear5b</td>
<td>1.86</td>
<td></td>
<td>-1.60</td>
<td>-0.73</td>
<td>0.10</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>

aThe ear11 item has 11 score points and thus 10 location parameters that delineate adjacent score points. The ear2, ear3, and ear4 items have 5 score points and thus 4 location parameters. The ear5 items have 4 score points and thus 3 location parameters. See Table 1 for item prompts, score means, and standard deviations.
• We can define item information as the ratio of the squared slope of the logistic curve to the conditional variance (think of a Bernoulli trial) as follows:

\[ I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \]

• For 1PL and 2PL:

\[ I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta) \]

• Maximized when \( P_i = .5 \).

• The steeper the slope of the ICC, the greater the information.
Visualizing Information from an ICC

\[ I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta) \]
Intuition for Item Information

<table>
<thead>
<tr>
<th>Par</th>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

\[
I_i(\theta) = \frac{a_i^2 (1 - c_i)}{(c_i + \exp(a_i(\theta - b_i))) (1 + \exp(-a_i(\theta - b_i)))^2}
\]

\[
\theta_{max} = b_i + \frac{1}{a_i} \log \left( \frac{1 + \sqrt{1 + 8c_i}}{2} \right)
\]
\[ I(\theta) = \sum_i I_i(\theta) \] Test information is the simple sum of item information at a particular \( \theta \)
Conditional Standard Error of Measurement (CSEM)

\[ SE(\hat{\theta}|\theta) = \frac{1}{\sqrt{I(\theta)}} \]

This “U-shaped” IRT CSEM above contrasts with the CSEM for simple sum scoring or %-correct scoring. If conventional scores are a proportion (a simplification), the error is binomial \[ \sqrt{\frac{\phi(1-\phi)}{n_i}} \]. Is conventional error greatest for central or extreme scores?
Step 7: Efficient Measurement - Item Information

![Graph showing item information functions for different theta values.](image-url)
Step 7: Efficient Measurement – Test Information
**Table 3.** Correlation Coefficients for Transformed Variables in Analysis 1.²

<table>
<thead>
<tr>
<th></th>
<th>WRS’</th>
<th></th>
<th>PTA’’</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
<td>Right</td>
<td>Left</td>
<td>Right</td>
</tr>
<tr>
<td>(n = 54)</td>
<td>(n = 52)</td>
<td></td>
<td>(n = 53)</td>
<td>(n = 55)</td>
</tr>
<tr>
<td>3-item scale</td>
<td>0.29</td>
<td>0.45</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.03-0.52</td>
<td>0.11-0.58</td>
<td>0.15-0.60</td>
<td>0.13-0.59</td>
</tr>
<tr>
<td>6-item scale</td>
<td>0.37</td>
<td>0.50</td>
<td>0.40</td>
<td>0.42</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.20-0.64</td>
<td>0.26-0.68</td>
<td>0.24-0.66</td>
<td>0.25-0.67</td>
</tr>
<tr>
<td>8-item scale</td>
<td>0.40</td>
<td>0.48</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.05-0.54</td>
<td>0.15-0.61</td>
<td>0.16-0.62</td>
<td>0.14-0.60</td>
</tr>
<tr>
<td>11-item scale</td>
<td>0.39</td>
<td>0.49</td>
<td>0.39</td>
<td>0.43</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.10-0.57</td>
<td>0.18-0.62</td>
<td>0.19-0.63</td>
<td>0.18-0.62</td>
</tr>
</tbody>
</table>

²All correlations were significant at $P < .05$. WRS’: transformed word recognition scores, $x' = -\log(102 - x)$. Higher scores indicate higher levels of auditory effectiveness. PTA’’’: pure tone audiometry, $x = \log(x)$. Higher scores indicate lower levels of auditory effectiveness.
An 8-Step Plan

Step 1: Content and Cognition
Step 2: Scoring and Scaling
Step 3: Correlation and Reliability
Step 4: Classical Item Diagnostics
Step 5: Latent Structure Analysis
Step 6: Item Response Theory (IRT)
Step 7: IRT for Efficient Measurement
Step 8: Correlation and Prediction
We say, “\(X\) is important. No one thinks of \(X\). Existing measures of \(X\) are off the mark. \(X\) matters more than everything else. If only we paid attention to \(X\).” If that is your argument, I suggest this research agenda:

1. Establish the **theoretical construct**
   - This measure should exist.

2. Establish the **latent structure**
   - This components of the measure relate as expected.

3. Establish **reliability**
   - The score you estimate should be precise.

4. Establish **predictions** and **intercorrelations**
   - These scores predict outcomes it should. They also predict outcomes things better than, over and above, other scores.

5. Establish **usefulness**
   - Using these scores achieves the intended purposes.