Psychology 405: Psychometric Theory

Validity

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

May, 2020
Outline

Preliminaries

Predictions and Decisions
  Classics in prediction and selection

The VA study
  Interviews

Predicting important outcomes

Interviews

Measuring individual differences: the power of items
  PWAS
  spi data
Observed Variables

\[ X \]

- \( X_1 \)
- \( X_2 \)
- \( X_3 \)
- \( X_4 \)
- \( X_5 \)
- \( X_6 \)

\[ Y \]

- \( Y_1 \)
- \( Y_2 \)
- \( Y_3 \)
- \( Y_4 \)
- \( Y_5 \)
- \( Y_6 \)
Latent Variables

\[ \xi \quad \eta \]

\[ \xi_1 \quad \eta_1 \]

\[ \xi_2 \quad \eta_2 \]
Theory: A regression model of latent variables

\[ \xi \quad \eta \]
A measurement model for $X$ – Correlated factors

\[
\delta \quad X \quad \xi
\]
A measurement model for Y - uncorrelated factors

\[ \eta \rightarrow Y_1 \rightarrow \epsilon_1 \]
\[ \eta_1 \rightarrow Y_2 \rightarrow \epsilon_2 \]
\[ \eta_1 \rightarrow Y_3 \rightarrow \epsilon_3 \]
\[ \eta_2 \rightarrow Y_4 \rightarrow \epsilon_4 \]
\[ \eta_2 \rightarrow Y_5 \rightarrow \epsilon_5 \]
\[ \eta_2 \rightarrow Y_6 \rightarrow \epsilon_6 \]
A complete structural model

\[ \delta \quad X \quad \xi \quad \eta \quad Y \quad \epsilon \]

\[ \delta_1 \rightarrow X_1 \quad \xi_1 \quad \eta_1 \rightarrow Y_1 \quad \epsilon_1 \]
\[ \delta_2 \rightarrow X_2 \quad \xi_1 \quad \eta_1 \rightarrow Y_2 \quad \epsilon_2 \]
\[ \delta_3 \rightarrow X_3 \quad \xi_1 \quad \eta_1 \rightarrow Y_3 \quad \epsilon_3 \]
\[ \delta_4 \rightarrow X_4 \quad \xi_2 \quad \eta_2 \rightarrow Y_4 \quad \epsilon_4 \]
\[ \delta_5 \rightarrow X_5 \quad \xi_2 \quad \eta_2 \rightarrow Y_5 \quad \epsilon_5 \]
\[ \delta_6 \rightarrow X_6 \quad \xi_2 \quad \eta_2 \rightarrow Y_6 \quad \epsilon_6 \]
Types of Validity

Face/Faith

- $X_1$ to $Y_1$
- $X_2$ to $Y_2$

Concurrent

- $X_2$ to $Y_2$
- $X_3$ to $Y_3$

Predictive

- $X_3$ to $Y_3$
- $X_4$ to $Y_4$

Convergent

- $X_4$ to $Y_4$
- $X_5$ to $Y_5$

Discriminant

- $X_5$ to $Y_5$
- $X_6$ to $Y_6$

Convergent

- $X_6$ to $Y_6$
- $X_7$ to $Y_7$

Construct

- $X_7$ to $Y_7$
- $X_8$ to $Y_8$
Face Validity

Face/Faith

Representative Content

Seeming relevance
Concurrent Validity

Does a measure correlate with the criterion?

Need to define the criterion.

Assumes that what correlates now will have predictive value.
Predictive Validity

\[ X_3 \xrightarrow{\text{Predictive}} Y_3 \]

Does a measure correlate with the criterion?

Need to define the criterion.

Allow time to pass
Prediction

1. Continuous predictor, continuous criterion
   - Regression, multiple regression, correlation
   - Slope of regression implies how much change for unit change in predictor

2. Continuous predictor, dichotomous criterion
   - Point bi-serial correlation

3. Dichotomous predictor, dichotomous outcome
   - Phi
   - The Taylor-Russell tables (Taylor & Russell, 1939) and the problem of Selection Ratios and Base Rates

\[
\phi = \frac{VP - BR \times SR}{\sqrt{(BR)(1 - BR)(SR)(1 - SR)}} \tag{1}
\]

- Therefore, the number of valid positives is

\[
VP = BR \times SR + \phi \sqrt{(BR)(1 - BR)(SR)(1 - SR)} \tag{2}
\]
Tetrachoric and phi as function of cut points

\[ \rho = 0.4 \]
\[ \phi = 0.26 \]
A decision theoretic approach

Valid Positives as function of False Positives

Decision Theory
Tetrachoric and phi as function of cut points .5,0
A decision theoretic approach with low beta

Valid Positives as function of False Positives

Decision Theory
Tetrachoric and phi as function of cut points 1,0

\[
\begin{align*}
X < \tau & \quad Y > \Upsilon \\
X > \tau & \quad Y > \Upsilon \\
X > \tau & \quad Y < \Upsilon \\
X < \tau & \quad Y < \Upsilon \\
\end{align*}
\]

\[
\begin{align*}
\rho = 0.4 & \quad \phi = 0.21
\end{align*}
\]
A decision theoretic approach with high beta

Valid Positives as function of False Positives

Decision Theory
Tetrachoric and phi as function of cut points 2,0
A decision theoretic approach with high beta

Valid Positives as function of False Positives

Decision Theory

Probability of observation
A decision theoretic analysis with 4 different cut points
Applying decision theory to a prediction problem: the case of predicting future psychiatric diagnoses from military inductees. (Data from Danielson & Clark (1954) as discussed by Wiggins (1973).

<table>
<thead>
<tr>
<th>Predicted Positive</th>
<th>Predicted Negative</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>49</td>
<td>40</td>
</tr>
<tr>
<td>True Negative</td>
<td>79</td>
<td>336</td>
</tr>
<tr>
<td>Column Totals</td>
<td>118</td>
<td>376</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted Positive</th>
<th>Predicted Negative</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>.097</td>
<td>.079</td>
</tr>
<tr>
<td>True Negative</td>
<td>.157</td>
<td>.667</td>
</tr>
<tr>
<td>Column Totals</td>
<td>.234</td>
<td>.746</td>
</tr>
</tbody>
</table>

Accuracy = \( .097 + .667 = .76 \)
Sensitivity = \( .097/(.097 + .079) = .55 \)
Specificity = \( .667 / (.667+.157) = .81 \)
\( \Phi = \frac{.097-.196*.234}{\sqrt{.196*.804*.234*.747}} = .32 \)
The Danielson and Clark data set as a decision problem

Valid Positives as function of False Positives

Decision Theory
Classics in Prediction and selection

1. Gideon’s selection of soldiers
2. OSS and Army Air Corps selection studies
3. Kelly & Fiske (1950) (1950) selection of psychology students (Kelly & Fiske, 1951)
4. Astronaut selection
5. Peace Corps selection
Gideon’s assessment
The assessment of pilots – how to show a .45 correlation makes a difference
Predicting clinical psychologists – Kelly and Fiske

1. Multiple predictors of graduate school performance: Kelly and Fiske (1950), Multiple predictors
2. Ability, Interests, temperament (each with $r \approx 0.2 - 0.25$) have multiple $R$ of 0.4–0.5
3. Are they able, interested and stable?
VA study: overview

- Researchers
  - nearly 40 cooperating clinical training programs
  - ≈ 75 psychologists on research staff

- Participants
  - 3/4 of those entering graduate training in 1946, 1947, 1948
  - N = 160, 128, 545 (selected down to 98)

- Measures
  - Objective tests
  - Clinical assessments
Objective instruments

- Ability
  - Millers Analogy Test
  - Thurstone Tests of Primary Mental Abilities
- Temperament and Character
  - Minnesota Multiphasic Personality Inventory
  - Guildord Martin Battery of Personality Inventories
- Interests, Values
  - Allport-Vernon Scale of Values
  - Strong Vocational Interest Blank
  - Kuder Preference Record
Assessment ratings

- Seven days of tests, interviews and “other” procedures
  - Three raters spent a week studying 4 trainees
  - Staff time devoted to each candidate was at least 7 man-days
- Ratings based on interviews, projective tests, role playing
  - Ratings on:
    - 22 descriptive variables (e.g., cooperativeness, talkativeness)
    - 10 evaluative variables (e.g., social adjustment, emotional expression)
    - 11 predictive variables (e.g. academic, diagnostician, overall suitability)
Criterion variables after 2 years

- Training status (Failure, still in Training, Ph.D. obtained)
- 2nd year evaluations
  - Skill in clinical diagnosis
  - Skill in individual psychotherapy
  - Skill in Research
  - Preference for hiring
- Generally high correlations among all the criteria
High correlations among the criteria

*Intercorrelations among selected criterion evaluations*

*N = 130 P-3 trainees evaluated in the spring of 1949, for whom all evaluation measures were available.*

<table>
<thead>
<tr>
<th></th>
<th>Clinical Diagnosis</th>
<th>Individual Therapy</th>
<th>Research</th>
<th>Preference for Hiring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univ.¹</td>
<td>Univ.</td>
<td>Univ.</td>
<td>Univ.</td>
</tr>
<tr>
<td>Clinical Diagnosis:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univ.</td>
<td>72</td>
<td>81</td>
<td>55</td>
<td>88</td>
</tr>
<tr>
<td>Instal.</td>
<td>47</td>
<td>60</td>
<td>54</td>
<td>65</td>
</tr>
<tr>
<td>Individual Therapy:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univ.</td>
<td>62</td>
<td>73</td>
<td>48</td>
<td>78</td>
</tr>
<tr>
<td>Instal.</td>
<td>38</td>
<td>63</td>
<td>37</td>
<td>74</td>
</tr>
<tr>
<td>Research:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univ.</td>
<td>11</td>
<td>28</td>
<td>65</td>
<td>76</td>
</tr>
<tr>
<td>Instal.</td>
<td>30</td>
<td>31</td>
<td>54</td>
<td>56</td>
</tr>
<tr>
<td>Preference for Hiring:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univ.</td>
<td>44</td>
<td>34</td>
<td>40</td>
<td>71</td>
</tr>
<tr>
<td>Instal.</td>
<td>38</td>
<td>70</td>
<td>31</td>
<td>56</td>
</tr>
</tbody>
</table>

¹ University staff evaluations.
² Installation staff evaluations.
The more they know about you, the more they will judge you

<table>
<thead>
<tr>
<th>Assessor</th>
<th>Information on Which Predictions Were Based</th>
<th>Criterion Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projectivist</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Proj. Integration</td>
<td>X X X X</td>
<td>X</td>
</tr>
<tr>
<td>Initial Interviewer</td>
<td>X X X X X</td>
<td>X</td>
</tr>
<tr>
<td>Intensive Interviewer</td>
<td>X X X X X</td>
<td>X</td>
</tr>
<tr>
<td>Pre-Conference</td>
<td>X X X X X X X X</td>
<td>X</td>
</tr>
<tr>
<td>Situationists (Pooled Rating)</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Prelim. Pooled</td>
<td>X X X X X X X X</td>
<td></td>
</tr>
<tr>
<td>Final Pooled</td>
<td>X X X X X X X X</td>
<td>X</td>
</tr>
</tbody>
</table>
Objectives are just as good

<table>
<thead>
<tr>
<th>B. Objective Test Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller Analogies</td>
</tr>
<tr>
<td>24* 15 06 05 24* 23* 23* 18</td>
</tr>
<tr>
<td>Strong Test</td>
</tr>
<tr>
<td>Psychologist—1938</td>
</tr>
<tr>
<td>29* 14 20 21 35** 31** 27* 19</td>
</tr>
<tr>
<td>Psychologist—1948 (Kriedt)</td>
</tr>
<tr>
<td>36** 20 28* 21 41** 32** 31** 20</td>
</tr>
<tr>
<td>Psychologist, Clinical, 1948 (Kriedt)</td>
</tr>
<tr>
<td>22 30** 18 16 07 07 17 09</td>
</tr>
<tr>
<td>Psychologist, VA Clinical (This Project)</td>
</tr>
<tr>
<td>36** 33** 35** 32** 33** 27* 36** 25*</td>
</tr>
<tr>
<td>Allport-Vernon Theoretical</td>
</tr>
<tr>
<td>23* -03 16 04 25* 15 23* -02</td>
</tr>
<tr>
<td>Guilford-Martin</td>
</tr>
<tr>
<td>C—Lack of Cycloid Disposition</td>
</tr>
<tr>
<td>29* 24* 29* 17 19 11 28* 16</td>
</tr>
<tr>
<td>N—Lack of Nervous Tenseness and Irritability</td>
</tr>
<tr>
<td>28* 23* 21 24* 21 06 22 20</td>
</tr>
</tbody>
</table>
the finding that the interview did not add to, but actually tended to decrease, the validity of clinical judgments made in the 1947 assessment program was confirmed by submitting the paper and pencil materials on these same candidates to a later assessment staff which made predictions without any face-to-face contact with the assessee. Under these conditions, the new staff made predictions with slightly higher validities than those made by the staff in 1947, who had the additional data from the interview, situation tests, etc.
Interests matter

The VA Clinical Psychologist key, developed by this project on the basis of the responses of full time VA psychologists, regularly yields relatively high correlations with all criterion evaluations, and compares favorably with the best predictions based on assessment ratings. Other psychologist keys, including the original (1938) general psychologist key and two developed by Kriedt (2), do fairly well. Not shown in the table is a correlation of .61 (N = 44) between scores based on the psychologist key (1938) and the scores made on the objective test of Knowledge of Clinical Psychology three years later. Thus, scores from a single objective test obtainable by mail, at little cost, predicted each of several criteria as well as any of the clinical judgments made in the entire assessment program.
Motivation

Our findings suggest that, in selection for professional training, more attention might well be given to the role of motivation. Perhaps at the level of graduate training, we need establish only a minimal cutting score on tests of intellectual aptitudes; beyond that point, the strength of motivation and the absence of conflicting drives may be the determining factors in success in professional training, and even in the conduct of professional duties.
Faith validity of interviews

Many who have seen our results have been disturbed by the findings regarding the validity for this selection problem of specific techniques which are felt by many professional psychologists to have a high degree of face-validity (or is it faith validity?). Thus, it was the firm conviction of the staff of the OSS assessment program that the global evaluation of a person permits much more accurate predictions of his future performance than can possibly be achieved by a more segmental approach. Unfortunately, the OSS data did not provide a conclusive answer to this question. Our own findings to date serve to raise doubts concerning the validity of this general proposition.
We must evaluate our judgments

Evidence such as that accumulating in this project serves to remind us of the fallibility of the human being both as a measuring device and as an integrator of data. In laboratories, in factories, and in accounting offices, it has been found necessary to supplement his sensory and perceptual capacities with an elaborate array of measuring instruments and computing devices. Pending the gradual development of better measures of psychological variables and comparable aids for combining them, we must continue to rely heavily on human judgment. In so doing, however, we must be continually aware of the magnitude of the errors of such judgments. These errors can be minimized by placing greatest reliance on measures of demonstrated reliability and validity.
Putting it together

We are, in fact, rather encouraged at the probability of being able to predict such criteria with a multiple R of around .50 on the basis of an inexpensive test battery which may be administered without requiring the applicant to present himself at the university of his choice.
More recent prediction studies

1. Terman & Oden (1947, 1959); Oden (1968)
2. Kuncel, Campbell & Ones (1998); Kuncel, Hezlett & Ones (2001); Kuncel & Hezlett (2007) and graduate school prediction
Kuncel et al. meta analysis predicting graduate school performance

Table 2
Meta-Analysis of GRE and UGPA Validities: Total Sample

<table>
<thead>
<tr>
<th>Predictor</th>
<th>N</th>
<th>k</th>
<th>r&lt;sub&gt;obs&lt;/sub&gt;</th>
<th>SD&lt;sub&gt;obs&lt;/sub&gt;</th>
<th>SD&lt;sub&gt;res&lt;/sub&gt;</th>
<th>p</th>
<th>SD&lt;sub&gt;p&lt;/sub&gt;</th>
<th>90% credibility interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GGPA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>14,156</td>
<td>103</td>
<td>.23</td>
<td>.14</td>
<td>.10</td>
<td>.34</td>
<td>.15</td>
<td>.09 to .59</td>
</tr>
<tr>
<td>Quantitative</td>
<td>14,425</td>
<td>103</td>
<td>.21</td>
<td>.11</td>
<td>.06</td>
<td>.32</td>
<td>.08</td>
<td>.19 to .45</td>
</tr>
<tr>
<td>Analytical</td>
<td>1,928</td>
<td>20</td>
<td>.24</td>
<td>.12</td>
<td>.04</td>
<td>.36</td>
<td>.06</td>
<td>.26 to .46</td>
</tr>
<tr>
<td>Subject</td>
<td>2,413</td>
<td>22</td>
<td>.31</td>
<td>.12</td>
<td>.05</td>
<td>.41</td>
<td>.07</td>
<td>.30 to .52</td>
</tr>
<tr>
<td>UGPA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9,748</td>
<td>58</td>
<td>.28</td>
<td>.13</td>
<td>.10</td>
<td>.30</td>
<td>.11</td>
<td>.12 to .48</td>
</tr>
<tr>
<td><strong>1st-year GGPA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>45,615</td>
<td>1,231</td>
<td>.24</td>
<td>.19</td>
<td>.09</td>
<td>.34</td>
<td>.12</td>
<td>.14 to .54</td>
</tr>
<tr>
<td>Quantitative</td>
<td>45,618</td>
<td>1,231</td>
<td>.24</td>
<td>.19</td>
<td>.08</td>
<td>.38</td>
<td>.12</td>
<td>.18 to .58</td>
</tr>
<tr>
<td>Analytical</td>
<td>36,325</td>
<td>1,080</td>
<td>.24</td>
<td>.19</td>
<td>.06</td>
<td>.36</td>
<td>.09</td>
<td>.21 to .51</td>
</tr>
<tr>
<td>Subject</td>
<td>10,225</td>
<td>98</td>
<td>.34</td>
<td>.11</td>
<td>.03</td>
<td>.45</td>
<td>.04</td>
<td>.38 to .52</td>
</tr>
<tr>
<td>UGPA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>42,193</td>
<td>1,178</td>
<td>.30</td>
<td>.18</td>
<td>.10</td>
<td>.33</td>
<td>.10</td>
<td>.17 to .49</td>
</tr>
<tr>
<td><strong>Comprehensive exam scores&lt;sup&gt;b&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>1,198</td>
<td>11</td>
<td>.34</td>
<td>.16</td>
<td>.12</td>
<td>.44</td>
<td>.15</td>
<td>.19 to .69</td>
</tr>
<tr>
<td>Quantitative&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1,194</td>
<td>11</td>
<td>.19</td>
<td>.11</td>
<td>.04</td>
<td>.26</td>
<td>.06</td>
<td>.16 to .36</td>
</tr>
<tr>
<td>Subject&lt;sup&gt;d&lt;/sup&gt;</td>
<td>534</td>
<td>4</td>
<td>.43</td>
<td>.07</td>
<td>.00</td>
<td>.51</td>
<td>.00</td>
<td>.51 to .51</td>
</tr>
<tr>
<td>UGPA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>592</td>
<td>6</td>
<td>.12</td>
<td>.05</td>
<td>.00</td>
<td>.12</td>
<td>.00</td>
<td>.12 to .12</td>
</tr>
<tr>
<td><strong>Faculty ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>4,766</td>
<td>35</td>
<td>.23</td>
<td>.12</td>
<td>.08</td>
<td>.42</td>
<td>.14</td>
<td>.19 to .65</td>
</tr>
<tr>
<td>Quantitative</td>
<td>5,112</td>
<td>34</td>
<td>.25</td>
<td>.10</td>
<td>.02</td>
<td>.47</td>
<td>.04</td>
<td>.40 to .54</td>
</tr>
<tr>
<td>Analytical</td>
<td>1,982</td>
<td>9</td>
<td>.23</td>
<td>.05</td>
<td>.00</td>
<td>.35</td>
<td>.00</td>
<td>.35 to .35</td>
</tr>
<tr>
<td>Subject</td>
<td>879</td>
<td>12</td>
<td>.30</td>
<td>.16</td>
<td>.11</td>
<td>.50</td>
<td>.18</td>
<td>.20 to .80</td>
</tr>
<tr>
<td>UGPA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3,695</td>
<td>22</td>
<td>.25</td>
<td>.12</td>
<td>.10</td>
<td>.35</td>
<td>.14</td>
<td>.12 to .58</td>
</tr>
<tr>
<td><strong>Degree attainment&lt;sup&gt;a&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>6,304</td>
<td>32</td>
<td>.14</td>
<td>.14</td>
<td>.12</td>
<td>.18</td>
<td>.16</td>
<td>-.08 to .44</td>
</tr>
<tr>
<td>Quantitative</td>
<td>6,304</td>
<td>32</td>
<td>.14</td>
<td>.17</td>
<td>.15</td>
<td>.20</td>
<td>.20</td>
<td>-.13 to .53</td>
</tr>
<tr>
<td>Analytical</td>
<td>1,233</td>
<td>16</td>
<td>.08</td>
<td>.25</td>
<td>.22</td>
<td>.11</td>
<td>.30</td>
<td>-.38 to .60</td>
</tr>
<tr>
<td>Subject</td>
<td>2,575</td>
<td>11</td>
<td>.32</td>
<td>.16</td>
<td>.14</td>
<td>.39</td>
<td>.17</td>
<td>.11 to .67</td>
</tr>
<tr>
<td>UGPA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6,315</td>
<td>33</td>
<td>.12</td>
<td>.17</td>
<td>.16</td>
<td>.12</td>
<td>.16</td>
<td>-.14 to .38</td>
</tr>
</tbody>
</table>
Table 9
GRE and UGPA Unit-Weighted Composite Predicting
GGPA and Faculty Ratings

<table>
<thead>
<tr>
<th>Predictor set</th>
<th>Predictive validity of unit-weighted composite</th>
<th>Predictive validity of composite plus UGPA (unit weighted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal</td>
<td>.41</td>
<td>.48</td>
</tr>
<tr>
<td>Quantitative</td>
<td>.42</td>
<td>.50</td>
</tr>
<tr>
<td>Analytical</td>
<td>.38</td>
<td>.46</td>
</tr>
<tr>
<td>Subject</td>
<td>.49</td>
<td>.54</td>
</tr>
<tr>
<td>Verbal + Quantitative</td>
<td>.46</td>
<td>.53</td>
</tr>
<tr>
<td>Verbal + Quantitative + Analytical</td>
<td>.45</td>
<td>.50</td>
</tr>
<tr>
<td>Verbal + Quantitative + Subject</td>
<td>.52</td>
<td>.56</td>
</tr>
<tr>
<td>Verbal + Quantitative + Analytical+ Subject</td>
<td>.50</td>
<td>.54</td>
</tr>
</tbody>
</table>

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average.
Benbow and Lubinski: Beyond the threshold

Beyond the Threshold Hypothesis

**Outcome**
- Any Doctorate (PhD, MD, JD): OR = 2.7*
- Any Peer-reviewed Publication: OR = 4.5*
- STEM Publications (≥1): OR = 5.9*
- STEM Doctorates: OR = 18.2*
- Patents (≥1): OR = 6.1*
- Income in 95th Percentile: OR = 3.3*
- STEM Tenure (Top 50): OR = 7.7*

---

**Proportion of Quartile With Outcome**

<table>
<thead>
<tr>
<th>Quartile</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Any Doctorate (PhD, MD, JD)</td>
</tr>
<tr>
<td>Q2</td>
<td>Any Peer-reviewed Publication</td>
</tr>
<tr>
<td>Q3</td>
<td>STEM Publications (≥1)</td>
</tr>
<tr>
<td>Q4</td>
<td>STEM Doctorates</td>
</tr>
</tbody>
</table>

---

**Age 13 SAT Math Score**

- 400
- 500
- 600
- 700
- 800

---

References

[Benbow and Lubinski: Beyond the threshold](#)
Deary: the Scottish sample and test retest reliability

Figure 3. Scattergram of age-corrected Moray House Test (MHT) scores at age 11 and age 80 for participants in the Lothian Birth Cohort 1921 of the Scottish Mental Survey 1932.
Deary: the Scottish sample and mortality
Deary: the Scottish sample and mortality: a model

![Diagram](image)

*Figure 6.* Some possible influences and pathways linking mental ability in childhood and survival. From *Brain and Longevity: Perspectives in Longevity* (p. 162, Figure 3), by C. Finch, J.-M. Robine, & Y. Christen (Eds.), 2003, Berlin: Springer. Copyright 2003 by Springer. Adapted with permission.
The persistent myth of the validity of the interview

1. It has been known since 1950 that interviews are appealing but do not work.

2. Everyone relies on their feeling that they work, remembering successes, forgetting failures.

3. Clinical versus actuarial prediction Dawes, Faust & Meehl (1989)

4. Experience is not a good teacher when the feedback is slow Dawes (1989)

5. Belief in the unstructured interview Dana, Dawes & Peterson (2013)

6. Summarized very well in Dawes (2009)
1. As discussed by Dawes (2009), DeVaul, Jervey, Chappell, Caver, Short & O’Keefe (1987) examined the effect of interview ratings on later success in medical school.

2. Of 2200 applicants to medical school, 800 were invited to interview and were interviewed.

3. Of these, 150 of the top 350 were offered positions.

4. The state then provided funding for an additional 50 students.

5. Only the 700-800 ranked students were still available.

6. After four years: “Even when the top 50 students in committee preference were compared with the 50 applicants, there were no differences. Thus, the least desirable candidates performed as well as the most desirable.”
How useful are items?

1. The common observation/belief is that items have low correlations with other items.

2. From a classical reliability perspective:
   \[ \text{Item variance} = \text{general} + \text{group} + \text{specific} + \text{error}. \]

3. The “gospel” is that items are mainly error variance.

4. This is true from a latent variable perspective, but less true if we actually examine item variance.

5. Perhaps 20% of an item is general factor variance, another 10-20% group variance but about 40% is specific and reliable variance.

6. We can see this by doing a variance decomposition of items that are repeated across time (Revelle & Condon, 2018)

7. So what?

8. Let's look at the correlates of items.
**Items as analogous to SNPs in GWAS studies**

1. In Genome Wide Association Studies one examines phenotypic variation as it correlates with differences in SNP frequencies across the genome.

2. We can do the same by examining phenotypic variation and correlation across the persome (Mõttus, Sinick, A.Terracciano, Hřebíčkova, Kandler & Jang, 2019)

3. A typical approach is to show the correlations and their probability values (corrected for multiple tests)
   - Typically displayed in “Manhattan Plots” across the genome. We do this across the “Persome”.

4. First show plots for an open source data set (spi) available in the *psych* package (Revelle, 2020).
   - This is a set of 135 temperament items (Condon, 2018), with 10 criteria for 4,000 subjects.

5. Then do the same for items from the Big 5, then an extended set (the little 27) then for a bigger data set with even more items.

6. Finally, we show the profile correlations of college majors and occupations for 255K participants across 900 items.
**Sapa Personality Inventory (spi data set)**

**Table:** Descriptive statistics for the 10 criteria variable taken from the spi data set.

<table>
<thead>
<tr>
<th>Variable</th>
<th>vars</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>trmmd</th>
<th>mad</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>skew</th>
<th>krtss</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>4000</td>
<td>26.90</td>
<td>11.49</td>
<td>23</td>
<td>25.02</td>
<td>7.41</td>
<td>11</td>
<td>90</td>
<td>79</td>
<td>1.45</td>
<td>1.8</td>
</tr>
<tr>
<td>sex</td>
<td>2</td>
<td>3946</td>
<td>1.60</td>
<td>0.49</td>
<td>2</td>
<td>1.62</td>
<td>0.00</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-0.39</td>
<td>-1.8</td>
</tr>
<tr>
<td>health</td>
<td>3</td>
<td>3536</td>
<td>3.51</td>
<td>0.98</td>
<td>4</td>
<td>3.54</td>
<td>1.48</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>-0.25</td>
<td>-0.4</td>
</tr>
<tr>
<td>p1edu</td>
<td>4</td>
<td>3051</td>
<td>4.72</td>
<td>2.39</td>
<td>5</td>
<td>4.77</td>
<td>4.45</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>-0.11</td>
<td>-1.3</td>
</tr>
<tr>
<td>p2edu</td>
<td>5</td>
<td>2896</td>
<td>4.33</td>
<td>2.32</td>
<td>5</td>
<td>4.28</td>
<td>4.45</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>0.09</td>
<td>-1.3</td>
</tr>
<tr>
<td>education</td>
<td>6</td>
<td>3330</td>
<td>4.10</td>
<td>2.21</td>
<td>3</td>
<td>4.00</td>
<td>1.48</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>0.41</td>
<td>-1.0</td>
</tr>
<tr>
<td>wellness</td>
<td>7</td>
<td>3311</td>
<td>1.54</td>
<td>0.50</td>
<td>2</td>
<td>1.55</td>
<td>0.00</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-0.17</td>
<td>-1.9</td>
</tr>
<tr>
<td>exer</td>
<td>8</td>
<td>3310</td>
<td>3.57</td>
<td>1.60</td>
<td>4</td>
<td>3.60</td>
<td>1.48</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>-0.35</td>
<td>-1.0</td>
</tr>
<tr>
<td>smoke</td>
<td>9</td>
<td>3348</td>
<td>2.19</td>
<td>2.04</td>
<td>1</td>
<td>1.70</td>
<td>0.00</td>
<td>1</td>
<td>9</td>
<td>8</td>
<td>1.83</td>
<td>2.1</td>
</tr>
<tr>
<td>ER</td>
<td>10</td>
<td>3347</td>
<td>1.16</td>
<td>0.48</td>
<td>1</td>
<td>1.03</td>
<td>0.00</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3.42</td>
<td>12.7</td>
</tr>
</tbody>
</table>

spi items are taken from Condon (2018)
Scale development and cross validation

1. Weights based upon data are best fits for those data
2. Need to “Cross Validate” on a different set
3. Original cross validation technique was to split the sample into 2, derive on first half, report the validities on the second half
4. KFold cross validation splits the data into K parts, derives the model on K-1 parts and then validate it on the remaining part. Repeat this K times (folds) and then average across folds.
5. Boot Strap Aggregation (“bagging”) takes many (100 - 1000) bootstrap samples and then aggregates across the hold out sample. Bootstrap automatically produces a hold out since 62.3% of subjects are in the derivation sample and 37.7% are in the holdout for each iteration.
6. The bestScales function does either K-fold or bagging and produces the Best Items Scale that is Cross-validated Unit-weighted, Informative and Transparent (Elleman, McDougald, Revelle & Condon, 2020).
### Internal consistency and correlations of the Big 5

**Table:** Reliabilities and correlations of the Big 5 scales from the spi data set. $\alpha$ reliability is on the diagonal of the correlations.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\omega_{total}$</th>
<th>$\omega_h$</th>
<th>Agree</th>
<th>Consc</th>
<th>Neuro</th>
<th>Extra</th>
<th>Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td>0.91</td>
<td>0.61</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.89</td>
<td>0.61</td>
<td>0.24</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.93</td>
<td>0.71</td>
<td>-0.12</td>
<td>-0.19</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extraversion</td>
<td>0.92</td>
<td>0.70</td>
<td>0.23</td>
<td>0.07</td>
<td>-0.20</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>Openness</td>
<td>0.88</td>
<td>0.72</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.12</td>
<td>0.13</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Item Validities: Manhattan plots

Manhattan Plot of health

Manhattan Plot of exer

Manhattan Plot of smoke
Cross validation of predictions

1. 10 criteria from the spi data set.
2. Linear regression using the spi-Big 5 (14 item scales for each of 5 dimensions)
3. bestScales solution choosing items from the spi to predict criteria
4. The “little 27” lower level factors of the spi
5. Multiple regression using all 135 items
6. All models developed on random subset of 2,000 subjects, cross validated on the remaining 2,000.
Profile Correlations

1. Normally, we examine the correlations of scales with criteria and criteria with criteria.

2. Analogous to the “genetic correlation” which is the correlation of phenotypes across the genome, we can find the “persome correlation” which is the correlation of the criteria across the persome.

3. These reflect the amount that the predictable variance in one outcome is the same as the predictable variance in another outcome.

4. The next slide compares phenotypic correlations with persomic correlations.

See Revelle, Dworak & Condon (2020) for more examples and the data and code for doing these examples.
Cross validated correlations predictions 10 different criteria.

Cross validation of multiple regression on SPI data

- 135
- little27
- bestScales
- Big5
Comparing phenotypic to profile correlations


