Psychology 350: Special Topics
An introduction to R for psychological research
Scoring items to make scales

William Revelle
Northwestern University
Evanston, Illinois USA

https://personality-project.org/courses/350

NORTHWESTERN UNIVERSITY

May, 2024

Outline

Preliminaries

- Using keys

Validity

Items

Validity

A bit of math

Maximize prediction

- Best scales on the bfi

Forming scales to increase reliability

1. We know from our prior discussions of reliability and test theory that forming composites of items will enhance prediction and reliability.
2. Known since Spearman (1904) the the power of aggregation is that error averages out and that signal averages in.
3. How to form aggregates?
   - Item sums
   - item averages
   - These are obviously just transforms of each, unless we have missing data
4. Multiple ways of doing this:
   - Scripts (one off)
   - Write your own function
   - Explore the *psych* functions

The basic script applied to the `ability` data set

R code

```
data <- read.file()  #get the data you want to score
data <- ability
Cscores <- rowSums(data)   #total scores
Mscores <- rowMeans(data)  #mean scores
cor2(Cscores,Mscores)
length(Cscores)
#do it again, but with na.rm=TRUE
Cscores <- rowSums(data, na.rm=TRUE)   #total scores
Mscores <- rowMeans(data,na.rm=TRUE)  #mean scores
cor2(Cscores,Mscores)
plot(Mscores ~ Cscores,main="means versus sum scores")
```

```
Cscores <- rowSums(data)   #total scores
> Mscores <- rowMeans(data)  #mean scores
> cor2(Cscores,Mscores)
[1] 1
> length(Cscores)
[1] 1525
> #do it again, but with na.rm=TRUE
> Cscores <- rowSums(data, na.rm=TRUE)   #total scores
> Mscores <- rowMeans(data,na.rm=TRUE)  #mean scores
> cor2(Cscores,Mscores)
[1] 0.94
```

Using the alpha or scoreFast functions

R code

```
al <- alpha(data)
cor2(al$scores,Mscores)
sf <- scoreFast(list(colnames(data)),data)
cor2(sf,Mscores)
keys <- list(colnames(data))
sf <- scoreFast(keys,data)
```

```
al <- alpha(data)
> cor2(al$scores,Mscores)
[1] 1
sf <- scoreFast(list(colnames(data)),data)
> cor2(sf,Mscores)
   [,1]
-A    1

keys <- list(colnames(data))
sf <- scoreFast(keys,data)
cor2(sf,Mscores)
   [,1]
-A    1
>
```

Why the list called 'keys'?

Using a keys list

1. To score multiple scales, or to specify the direction of keying an item, we can use a 'keys.list"
2. This is just list of key names and the items that go into the scale
3. Consider subscales of the ability data set

R code

```
ability.keys
```

```
 ability.keys
$ICAR16
 [1] "reason.4"  "reason.16" "reason.17" "reason.19" "letter.7"  "letter.33" "letter.34" "lett
 [9] "matrix.45" "matrix.46" "matrix.47" "matrix.55" "rotate.3"  "rotate.4"  "rotate.6"  "rota

$reasoning
[1] "reason.4"  "reason.16" "reason.17" "reason.19"

$letters
[1] "letter.7"  "letter.33" "letter.34" "letter.58"

$matrix
[1] "matrix.45" "matrix.46" "matrix.47" "matrix.55"

$rotate
[1] "rotate.3" "rotate.4" "rotate.6" "rotate.8"
```

Using the keys.list for the ability data set

R code

```
sf5 <- scoreFast(ability.keys,ability)
cor2(sf5,Mscores)
```

```
 sf5 <- scoreFast(ability.keys,ability)
cor2(sf5,Mscores)
            [,1]
ICAR16-A    1.00
reasoning-A 0.77
letters-A   0.78
matrix-A    0.74
rotate-A    0.69
```

This is better than our simple `rowMeans` call because it allows for multiple scale at once.

We could do multiple calls to `rowMeans`

R code

```
multi.score <- data.frame( all = rowMeans(ability, na.rm=TRUE),
            reasoning = rowMeans(ability[,1:4], na.rm=TRUE),
            letters = rowMeans(ability[,5:8], na.rm=TRUE),
             matrix = rowMeans(ability[,9:12], na.rm=TRUE),
             rotate = rowMeans(ability[,13:16], na.rm=TRUE)
                         )
  cor2(multi.score,sf5)
```

```
 cor2(multi.score,sf5)
          ICAR16-A reasoning-A letters-A matrix-A rotate-A
all           1.00        0.77      0.78     0.74     0.69
reasoning     0.77        1.00      0.52     0.45     0.37
letters       0.78        0.52      1.00     0.44     0.36
matrix        0.74        0.45      0.44     1.00     0.34
rotate        0.69        0.37      0.36     0.34     1.00
```

Keys are most useful when we have reverse keyed items

1. Consider the 5 scales of the `bfi`

R code

```
bfi.keys
```

```
bfi.keys
$agree
[1] "-A1" "A2"  "A3"  "A4"  "A5"

$conscientious
[1] "C1"  "C2"  "C3"  "-C4" "-C5"

$extraversion
[1] "-E1" "-E2" "E3"  "E4"  "E5"

$neuroticism
[1] "N1" "N2" "N3" "N4" "N5"

$openness
[1] "O1"  "-O2" "O3"  "O4"  "-O5"
```

Show the items from the dictionary

R code

```
lookupFromKeys(bfi.keys,  bfi.dictionary[,2:3])
```

```
$agree
                                          Item   Giant3
A1- Am indifferent to the feelings of others. Cohesion
A2  Inquire about others' well-being.         Cohesion
A3  Know how to comfort others.               Cohesion
A4  Love children.                            Cohesion
A5  Make people feel at ease.                 Cohesion

$conscientious
                                      Item    Giant3
C1  Am exacting in my work.                Stability
C2  Continue until everything is perfect. Stability
C3  Do things according to a plan.        Stability
C4- Do things in a half-way manner.       Stability
C5- Waste my time.                        Stability

$extraversion
                                      Item     Giant3
E1- Don't talk a lot.                     Plasticity
E2- Find it difficult to approach others. Plasticity
E3  Know how to captivate people.         Plasticity
E4  Make friends easily.                  Plasticity
E5  Take charge.                          Plasticity
```

Some items need to be reversed keyed

1. Could do this mechanically by recoding those items
2. Better is to it automatically by subtracting the item from the max - min + 1
3. i.e. for a 6 point item, subtract from 7

Can find simple sums or means or do some statistics at the same time

1. Conventional statistics include α but also average r , etc.

R code

```
fs <- scoreFast(bfi.keys,bfi)
scales <-  scoreItems(bfi.keys,bfi)
names(scales)
cor2(fs, scales$scores)
```

```
fs <- scoreFast(bfi.keys,bfi)
> scales <-  scoreItems(bfi.keys,bfi)
> names(scales)
 [1] "scores"          "missing"         "alpha"           "av.r"            "sn"
 [6] "n.items"         "item.cor"        "cor"             "corrected"       "G6"
[11] "item.corrected"  "response.freq"   "raw"             "ase"             "med.r"
[16] "keys"            "MIMS"            "MIMT"            "Call"
> cor2(fs, scales$scores)
               agree conscientious extraversion neuroticism openness
agree-A         1.00          0.26         0.46       -0.19     0.15
conscientious-A 0.26          1.00         0.26       -0.23     0.20
extraversion-A  0.46          0.26         1.00       -0.22     0.22
neuroticism-A  -0.18         -0.23        -0.22        1.00    -0.09
openness-A      0.15          0.19         0.21       -0.09     1.00
```

scoreItems gives a great deal of output

R code

```
scales
```

```
scales
Call: scoreItems(keys = bfi.keys, items = bfi)

(Unstandardized) Alpha:
      agree conscientious extraversion neuroticism openness
alpha   0.7          0.72         0.76        0.81      0.6

Standard errors of unstandardized Alpha:
      agree conscientious extraversion neuroticism openness
ASE   0.014         0.014        0.013       0.011    0.017

Average item correlation:
          agree conscientious extraversion neuroticism openness
average.r  0.32          0.34         0.39        0.46     0.23

Median item correlation:
        agree conscientious  extraversion   neuroticism      openness
         0.34          0.34          0.38          0.41          0.22

 Guttman 6* reliability:
         agree conscientious extraversion neuroticism openness
Lambda.6   0.7          0.72         0.76        0.81      0.6

Signal/Noise based upon av.r :
             agree conscientious extraversion neuroticism openness
Signal/Noise   2.3           2.6          3.2         4.3      1.5

Scale intercorrelations corrected for attenuation
 raw correlations below the diagonal, alpha on the diagonal
```

scales (continued)

```
Scale intercorrelations corrected for attenuation
 raw correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
              agree conscientious extraversion neuroticism openness
agree          0.70          0.36         0.63      -0.245     0.23
conscientious  0.26          0.72         0.35      -0.305     0.30
extraversion   0.46          0.26         0.76      -0.284     0.32
neuroticism   -0.18         -0.23        -0.22       0.812    -0.12
openness       0.15          0.19         0.22      -0.086     0.60

Average adjusted correlations within and between scales (MIMS)
              agree cnscn extrv nrtcs opnns
agree          0.32
conscientious  0.22  0.34
extraversion   0.44  0.26  0.39
neuroticism   -0.20 -0.26 -0.28  0.46
openness       0.11  0.15  0.18 -0.08  0.23

 Average adjusted item x scale correlations within and between scales (MIMT)
              agree cnscn extrv nrtcs opnns
agree          0.68
conscientious  0.18  0.69
extraversion   0.33  0.19  0.71
neuroticism   -0.14 -0.18 -0.17  0.76
openness       0.10  0.12  0.14 -0.05  0.62
```

Scale validity

1. How do we assess the validity of a scale
2. Face/Faith The items look right
3. Concurrent The scale correlates with relevant criteria
4. Predictive The scale predicts criteria in the future
5. Construct
    - Convergent The scales correlate with what alternative measures of the same thing
    - Divergent The scales do not correlate with measures of alternative constructs
    - Incremental The scales add to our predictive power

SAPA measures self report

1. How to validate the self reports of the SAPA project
2. SAPA participants were asked to nominate anonymous friends
3. These friends then gave peer ratings
4. Zola et al. (2021) reported the validity of self report personality items from the SAPA personality inventory (SPI) (Condon, 2018) in terms of 30 peer reports on 8 dimensions. Here are the polychoric correlations of these items. spi items were collected using SAPA procedures for 158,631 participants (mean n/item = 18,180), 908 of whom received peer ratings..
5. The Multitrait-multimethod correlations were found from the correlations.

Zola validities

R code

```
scores <- psych::scoreOverlap(zola.keys[c(1:5,33:37)],zola) #MTMM of I
```

```
lowerMat(scores$cor)
                  Agrbl Cnscn Nrtcs Extrv Opnnn Agrbl Cnscn Stblt Extrv IntlO
Agreeableness      1.00
Conscientiousness  0.28  1.00
Neuroticism       -0.12 -0.18  1.00
Extraversion       0.25  0.12 -0.25  1.00
Opennness          0.08  0.05 -0.09  0.13  1.00
Agreeableness      0.47  0.10 -0.01  0.00 -0.09  1.00
Conscientiousness  0.15  0.55 -0.12 -0.01 -0.04  0.18  1.00
Stability          0.13  0.16 -0.58  0.05  0.07  0.25  0.25  1.00
Extraversion       0.23  0.28 -0.27  0.49  0.11  0.07  0.23  0.22  1.00
IntellectOpenness  0.14  0.08 -0.15  0.09  0.30  0.19  0.24  0.27  0.15  1.00

lowerMat(scores$MIMS)  #average item correlations within and between domains
                  Agrbl Cnscn Nrtcs Extrv Opnnn Agrbl Cnscn Stblt Extrv IntlO
Agreeableness      0.33
Conscientiousness  0.10  0.32
Neuroticism       -0.05 -0.07  0.38
Extraversion       0.10  0.05 -0.11  0.39
Opennness          0.03  0.02 -0.03  0.05  0.30
Agreeableness      0.18  0.04  0.00  0.00 -0.03  0.17
Conscientiousness  0.06  0.23 -0.05  0.00 -0.02  0.07  0.26
Stability          0.05  0.07 -0.25  0.02  0.03  0.10  0.11  0.28
Extraversion       0.09  0.11 -0.11  0.21  0.04  0.03  0.10  0.09  0.21
IntellectOpenness  0.05  0.03 -0.06  0.03  0.11  0.07  0.10  0.11  0.06  0.16
```

Reliability and Validity

1. Validity (r_{xy}) is bounded by the square root of reliability (r_{xx}) (Spearman, 1904)
$$r_{xy} \leq \sqrt{r_{xx}}.$$
2. To increase reliability, we form scales by aggregating related items.
3. This is based upon the notion that all measurement is "befuddled with error" (McNemar, 1946).
4. Items in particular are thought to be mainly error with just a little bit of reliable variance.

Items are better than we think

1. Typical belief is that because items are noisy (unreliable) we need to aggregate items to improve the measurement quality of our scale.
2. Classical model of an item considers True Scores and Errors (Spearman, 1904; Lord and Novick, 1968; McDonald, 1999), $X_i = \tau_i + \epsilon_i$
3. A more refined model considers general variance, group variance, specific variance and error (McDonald, 1999).

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{1}$$

4. And we find ω_t and ω_h (McDonald, 1999; Zinbarg et al., 2005)

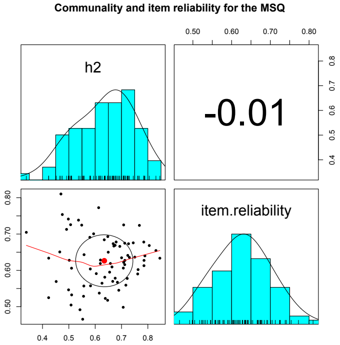
$$\omega_t = \frac{\sigma_X^2 - \Sigma\sigma_i^2 + \Sigma h_i^2}{\sigma_X^2}. \tag{2}$$

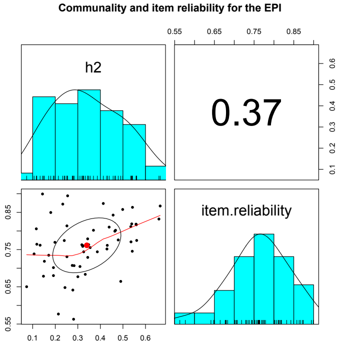
$$\omega_h = \frac{(\Sigma\lambda_i)^2}{\sigma_X^2} = \frac{\mathbf{1cc'1'}}{\sigma_X^2}. \tag{3}$$

But the variance of an item is much more than what is common

1. In the case of one administration, specific and error are confounded.
2. But, if we have repeated measures (t_1, t_2) , we can show that the reliable variance $(r_{t_1 t_2})$ is much greater than the common variance (h^2).
3. Consider the reliability of 75 mood items taken twice $(\overline{r_{12}} = .63)$ and compare with the communality of these items. $\overline{h^2} = .63)$. (Data from the `msqR` data set in *psychTools*).
4. More striking is comparing reliabilities of 57 items from the EPI (Eysenck and Eysenck, 1964) taken several weeks apart $(\overline{r_{12}} = .76)$ with their communalities $(\overline{h^2} = .34)$. (Data from the `epiR` data set in *psychTools*).
5. David Condon reports within test item reliabilities of .6 -.8.

Communalities and item reliabilities for the MSQ and EPI





Using polychoric (msq) or tetrachoric (epi) correlations.

MSQ statistics

	vars	n	mean	sd	median	min	max	range	se
Communality (h2)	1	75	0.63	0.11	0.65	0.34	0.85	0.51	0.01
item reliability	2	75	0.63	0.07	0.63	0.47	0.81	0.34	0.01

EPI statisics

	vars	n	mean	sd	median	min	max	range	se
Communality (h2)	1	57	0.34	0.15	0.32	0.07	0.67	0.59	0.02
item reliability	2	57	0.76	0.07	0.76	0.56	0.90	0.34	0.01

Validity: a very broad concept

1. Until about 1955, validity was how well a test actually predicted something.
2. But in the 1950's, perhaps in a reaction to behaviorism and in reaction to the plethora of empirical scale developed for the MMPI (Hathaway and McKinley, 1943) or the Strong Vocational Interest test (Strong Jr., 1927), validity came to include *construct validity* (Cronbach and Meehl, 1955; Loevinger, 1957).
3. By emphasizing constructs, and the convergent and discriminant patterns of correlations (Campbell and Fiske, 1959), there began a great emphasis upon factorially pure measures.
4. Questions of unidimensionality of scales became more important, and criticisms of standard measures of internal consistency such as α or λ_3 became common (Sijtsma, 2008) as psychometricians recommended more model based estimates.
5. Simple predictive validity was ignored at best and denigrated at worst (Borsboom et al., 2003, 2004).

Let's consider an example

1. Consider four different tests where the items range in their correlations with each (internal consistency) and with a criterion (predictive validity).
2. The four tests have average intercorrelations of .1 to .4 and thus α ranging from .31 to .73 and have item validies of .2 and thus scale validities ranging from .27 to .35
3. The question is which is the better test?

Which set of items (X1..X4) have the highest validity when predicting Y?

A) $\alpha = .73$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.4	1.0			
X3	0.4	0.4	1.0		
X4	0.4	0.4	0.4	1.0	
Y	0.2	0.2	0.2	0.2	1.0

B) $\alpha = .63$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.3	1.0			
X3	0.3	0.3	1.0		
X4	0.3	0.3	0.3	1.0	
Y	0.2	0.2	0.2	0.2	1.0

C) $\alpha = .5$ $R_y = .?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.2	1.0			
X3	0.2	0.2	1.0		
X4	0.2	0.2	0.2	1.0	
Y	0.2	0.2	0.2	0.2	1.0

D) $\alpha = .31$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.1	1.0			
X3	0.1	0.1	1.0		
X4	0.1	0.1	0.1	1.0	
Y	0.2	0.2	0.2	0.2	1.0

Please rank order these four cells in terms of validity.

Which set of items (X1..X4) have the highest validity when predicting Y?

A) $\alpha = .73$ $R_y = .27$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.4	1.0			
X3	0.4	0.4	1.0		
X4	0.4	0.4	0.4	1.0	
Y	0.2	0.2	0.2	0.2	1.0

B) $\alpha = .63$ $R_y = .29$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.3	1.0			
X3	0.3	0.3	1.0		
X4	0.3	0.3	0.3	1.0	
Y	0.2	0.2	0.2	0.2	1.0

C) $\alpha = .5$ $R_y = .32$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.2	1.0			
X3	0.2	0.2	1.0		
X4	0.2	0.2	0.2	1.0	
Y	0.2	0.2	0.2	0.2	1.0

D) $\alpha = .31$ $R_y = .35$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.1	1.0			
X3	0.1	0.1	1.0		
X4	0.1	0.1	0.1	1.0	
Y	0.2	0.2	0.2	0.2	1.0

Validity is higher the lower the internal consistency.

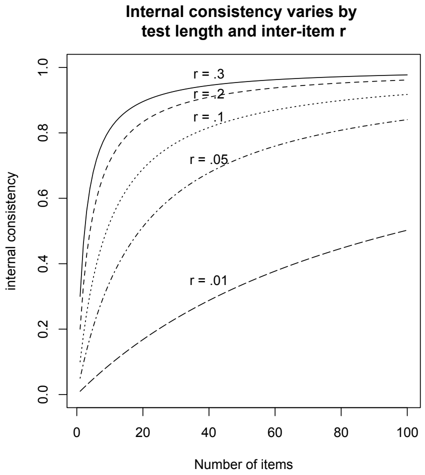
Validity and reliability: a short digression

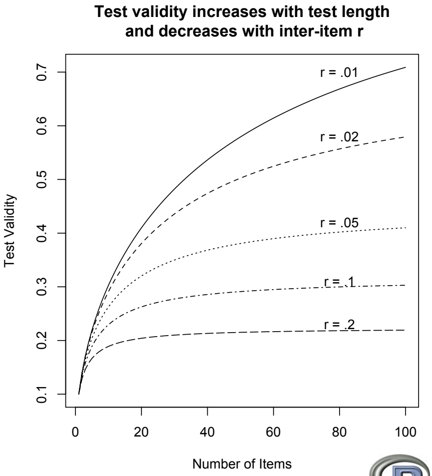
1. Although we know from Spearman that we can correct for reliability to find the "True" relationship between two variables, this does not help us in the real world.
2. Reliability is incorrectly associated with internal consistency which leads to such derivations as coefficients KR20 (Kuder and Richardson, 1937), λ_3 (Guttman, 1945) or α (Cronbach, 1951).
3. Expressed terms of inter-item correlations, this is just $\frac{k\bar{r}}{1+(k-1)\bar{r}}$ and increases with test length (k) and the average interitem correlation ($\bar{r}$).
4. However, *validity* of a k item test (r_{y_k}) or the correlation with an external criterion, Y, also increases with test length, and the average item validity ($\bar{r_y}$) but decreases as the inter-item correlation increases $r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k+k*(k-1)\bar{r}}}$.

Reliability and Validity

1. Lets unpack these two equations.
   Internal consistency varies by number of items and average correlation.
$$\lambda_3 = \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{4}$$
2. But validity varies by number of items, average within test correlation and average item validity
$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k*(k-1)\bar{r}}}. \tag{5}$$

The trade off between test consistency and test validity

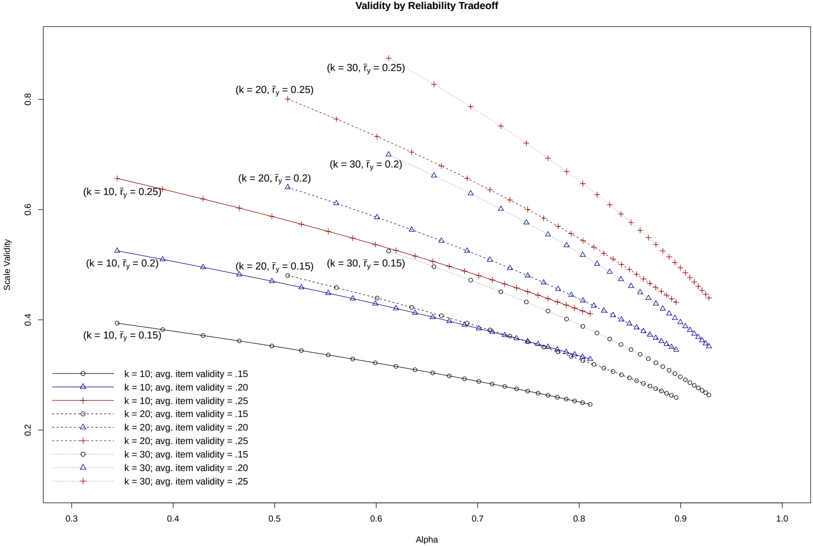




Showing the reliability by validity tradeoff

1. Consider 9 scales formed from
2. 10, 20 or 30 items
3. Average validities of .15, .20, .25
4. Plot scale validity by scale α for $.3 < \alpha < .9$

The trade off between test consistency and test validity



Increasing validity implies increasing the diversity of the item content

1. The goal of construct validity is have pure measures with high internal consistency.
   (Spears that measure one thing well).
2. And highly correlated measures of the same constructs.
3. But if the goal is predictive validity, we should minimize internal consistency and have independent predictors.
4. By emphasizing practical validity, we are ignoring most of what we have been taught (and teach) about reliability (Revelle and Condon, 2018, 2019) and scale construction (Revelle and Garner, 2023).
5. Predictive validity can be enhanced by casting a broader net.
6. Variations on this theme have been discussed before (Condon et al., 2021; Möttus et al., 2020).

Emphasizing predictive validity over reliability

1. The `bestScales` function selects items that most predict a criterion
2. But we need to *Cross Validate* these predictions
3. Without cross validation, we are fooling ourselves.
4. Either use bootstrap cross validation or K-folds
   - Bootstrap takes multiple samples and then aggregates the (bagging)
   - K-folds splits samples into K parts, derives on derives the model on K-1 parts and then validate it on the remaining part

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 et al., 2020) .

Using bestScales on the bfi

R code

```
bs <- bestScales(x=bfi[1:25], criteria = bfi[26:28], dictionary=bfi.di
```

```
Call = bestScales(x = bfi[1:25], criteria = bfi[26:28], dictionary = bfi.dictionary[2:3])
The items most correlated with the criteria yield r's of
          correlation n.items
gender           0.32       9
education        0.14       1
age              0.24      10

The best items, their correlations and content  are
$gender
   gender                                     Item     Giant3
N5   0.21 Panic easily.                           Stability
A2   0.18 Inquire about others' well-being.       Cohesion
A1  -0.16 Am indifferent to the feelings of others. Cohesion
A3   0.14 Know how to comfort others.             Cohesion
A4   0.13 Love children.                          Cohesion
E1  -0.13 Don't talk a lot.                       Plasticity
N3   0.12 Have frequent mood swings.              Stability
O1  -0.10 Am full of ideas.                       Plasticity
A5   0.10 Make people feel at ease.               Cohesion

$education
   education                                     Item   Giant3
A1      -0.14 Am indifferent to the feelings of others. Cohesion

$age
     age                                     Item     Giant3
A1 -0.16 Am indifferent to the feelings of others. Cohesion
C4 -0.15 Do things in a half-way manner.          Stability
A4  0.14 Love children.                           Cohesion
```

Bootstrap 100 times

R code

```
bs1 <-   bestScales(x=bfi[1:25], criteria = bfi[26:28],
             dictionary=bfi.dictionary[2:3], n.iter=100)
```

```
Call = bestScales(x = bfi[1:25], criteria = bfi[26:28], n.iter = 100,
    dictionary = bfi.dictionary[2:3])
          derivation.mean derivation.sd validation.m validation.sd final.valid final.wtd N.wtd
gender               0.32         0.021         0.30         0.032        0.29      0.33    10
education            0.16         0.029         0.13         0.026        0.14      0.17    10
age                  0.25         0.018         0.22         0.024        0.24      0.25    10

 Best items on each scale with counts of replications

 Criterion =  gender
   Freq mean.r sd.r                                      Item      Giant3
N5  100   0.21 0.02 Panic easily.                                Stability
A2  100   0.18 0.02 Inquire about others' well-being.             Cohesion
A1  100  -0.16 0.02 Am indifferent to the feelings of others.     Cohesion
A3   98   0.14 0.02 Know how to comfort others.                   Cohesion
E1   94  -0.13 0.02 Don't talk a lot.                           Plasticity
A4   91   0.13 0.02 Love children.                                Cohesion

 Criterion =  education
   Freq mean.r sd.r                                      Item    Giant3
A1  100  -0.15 0.02 Am indifferent to the feelings of others. Cohesion

 Criterion =  age
   Freq mean.r sd.r                                      Item     Giant3
A1  100  -0.16 0.02 Am indifferent to the feelings of others.  Cohesion
C4   99  -0.15 0.02 Do things in a half-way manner.           Stability
A4  100   0.15 0.02 Love children.                             Cohesion
```

Now try the 10 criteria from the spi

R code

```
bs.spi <- bestScales(spi[11:145],spi[1:10], folds=10, dictionary=spi.c
```

```
Call = bestScales(x = spi[11:145], criteria = spi[1:10], folds = 10,
    dictionary = spi.dictionary[, c(2, 5)])
          derivation.mean derivation.sd validation.m validation.sd final.valid final.wtd N.wtd
age                  0.36        0.0076        0.354         0.055        0.35      0.36    10
sex                  0.35        0.0078        0.350         0.040        0.35      0.35    10
health               0.44        0.0056        0.431         0.042        0.43      0.44    10
p1edu                0.13        0.0181        0.117         0.048        0.12      0.19    10
p2edu                0.11        0.0123        0.082         0.045          NA      0.19    10
education            0.30        0.0112        0.283         0.041        0.26      0.31    10
wellness             0.24        0.0065        0.216         0.050        0.23      0.24    10
exer                 0.31        0.0122        0.296         0.026        0.30      0.32    10
smoke                0.27        0.0052        0.260         0.047        0.27      0.28    10
ER                   0.15        0.0142        0.130         0.029        0.13      0.16    10
```

Try it with max item =20

R code

```
bs.spi <- bestScales(spi[11:145],spi[1:10], folds=10, dictionary=spi.d
```

```
Call = bestScales(x = spi[11:145], criteria = spi[1:10], n.item = 20,
    folds = 10, dictionary = spi.dictionary[, c(2, 5)])
          derivation.mean derivation.sd validation.m validation.sd final.valid final.wtd N.wtd
age                  0.38        0.0073        0.373         0.053        0.37      0.36    10
sex                  0.41        0.0070        0.396         0.035        0.39      0.35    10
health               0.44        0.0060        0.438         0.054        0.44      0.44    10
p1edu                0.13        0.0208        0.119         0.052        0.12      0.19    10
p2edu                0.12        0.0213        0.072         0.042          NA      0.19    10
education            0.33        0.0063        0.306         0.060        0.32      0.31    10
wellness             0.23        0.0040        0.222         0.032        0.24      0.24    10
exer                 0.32        0.0045        0.305         0.050        0.31      0.32    10
smoke                0.27        0.0069        0.250         0.043        0.26      0.28    10
ER                   0.15        0.0103        0.135         0.051        0.13      0.16    10

fm
```

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2):203–219.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4):1061–1071.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(8):81–105.

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*.

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costani, G., Greiff, S., Johnson, W., Lukaszesksi, A., Murray, A., Revelle, W., Wright, A. G., Ziegler, M., and Zimmerman, J. (2021). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Elleman, L. G., McDougald, S., Revelle, W., and Condon, D. (2020). That takes the BISCUIT: a comparative study of predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. *European Journal of Psychological Assessment*, 36(6):948–958.

Eysenck, H. J. and Eysenck, S. B. G. (1964). *Eysenck Personality Inventory*. Educational and Industrial Testing Service, San Diego, California.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Hathaway, S. and McKinley, J. (1943). Manual for administering and scoring the MMPI.

Kuder, G. and Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports Monograph Supplement 9*, 3:635–694.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. The Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co, Reading, Mass.

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates, Mahwah, N.J.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, 43(4):289–374.

Möttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costani, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszeski, A., Murray, A., Revelle, W., Wright, A. G., Yarkoni, T., Ziegler, M., and Zimmerman, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, 34(6).

Revelle, W. and Condon, D. M. (2018). Reliability. In Irwing, P., Booth, T., and Hughes, D. J., editors, *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. John Wily & Sons, London.

Revelle, W. and Condon, D. M. (2019). Reliability: from alpha to omega. *Psychological Assessment*, 31(12):1395–1411.

Revelle, W. and Garner, K. M. (2023). Measurement: Reliability, construct validation, and scale construction. In Harry T. Reis, T. W. and Judd, C. M., editors, *Handbook of Research Methods in Social and Personality Psychology (in press)*.

Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2):201–292.

Strong Jr., E. K. (1927). Vocational interest test. *Educational Record*, 8(2):107–121.

Zinbarg, R. E., Revelle, W., Yovel, I., and Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1):123–133.

Zola, A., Condon, D. M., and Revelle, W. (2021). The Convergence of Self and Informant Reports in a Large Online Sample. *Collabra: Psychology*, 7(1).