The seductive beauty of latent variables

ISSID Award for Distinguished Contribution to the Study of Individual Differences

William Revelle
Northwestern University
Evanston, Illinois USA
NORTHWESTERN UNIVERSITY

International Society for the Study of Individual Differences
Belfast, N.Ireland
July 18, 2023
Slides at https://personality-project.org/sapa/issid.pdf

Outline

Abstract

Seduced by their mathematical beauty, psychologists have been using latent variable models for more than a century. Whether discussing a general factor of cognitive ability, personality, or psychopathology there has been an unfortunate tendency to reify hierarchical structures without examining the utility of alternative models. To some of us, the use of latent variables was an unfortunate mistake. By emphasizing internal consistency rather than validity, parsimony of fit rather than function, the use of latent variables has led psychological measurement and theory down a beautifully seductive garden path rather than focusing on the real problem of actually being useful. I will address some of these alternatives and suggest that it is time to think more critically of the use of latent variable models in our theorizing and applications.

The importance of luck over a lifetime

1. As I have said at another occasion reviewing my career the secrets for success over a lifetime are very easy:
2. Good luck
3. Great mentors
4. Great colleagues
5. Great students
6. Live long enough
7. Good luck (Having just read an article about the importance of luck in a career (Pluchino, Biondo & Rapisarda, 2018)) I emphasize this again.

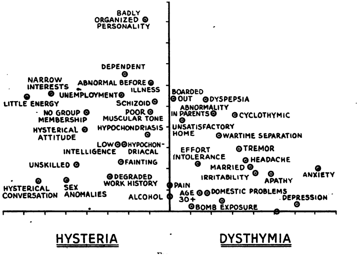
ISSID and latent variables

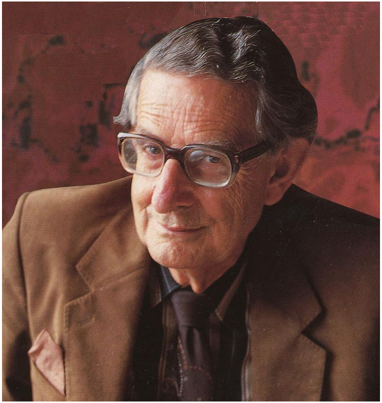
1. To challenge latent variable models at an ISSID meeting is a daunting (foolish?) task.
2. The three prior recipients of this award were leaders in promoting the power of latent variable models.
3. Hans Eysenck, Arthur Jensen were and Ian Deary is truly giants in our field and their contributions to the study of individual differences has been enormous.

The first meeting of ISSID was held in London 40 years ago. A few of us here remember the excitement of that meeting. (Photo from Robert Stelmack).

Hans Eysenck

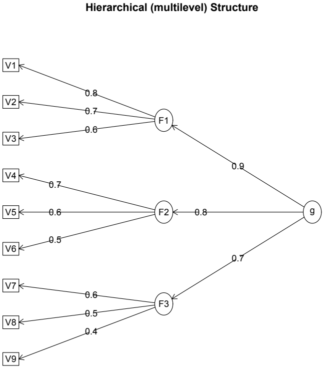
Hans Eysenck, as a student of Cyril Burt, searched for the latent variables of personality. One of his earliest studies was of the factor structure of behavioral measures among hospitalized soldiers (Eysenck, 1944). He was the founder of ISSID.

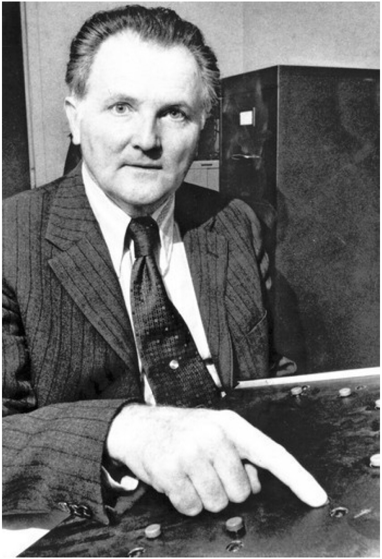




Arthur Jensen

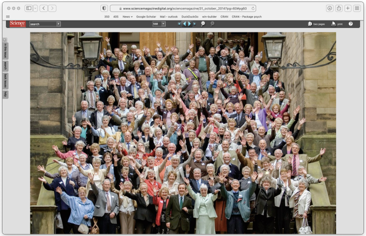
Arthur Jensen (1998) emphasized the g factor of cognitive ability and discussed what makes a good g factor (Jensen & Weng, 1994).



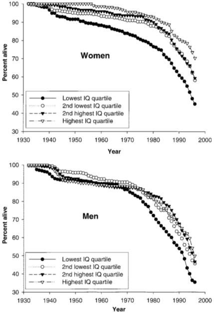


Ian Deary

Ian Deary (2001) remains a leader in intelligence research, with his collaborators on the MidLothian study of cognition over the life span. He is both a critic and a supporter of factorial models of cognition.



The Midlothian participants



Latent variable models have a long and distinguished history

1. The distinction between manifest and latent variables was perhaps first made by Plato (nd) in his Allegory of the Cave.
   - Prisoners confined in a cave observe only shadows on the wall.
   - These shadows represent unseen people moving in front of an unseen fire.
   - As the people move closer to the fire, shadow lengths increase.
2. By "correcting" for the attenuation due to unreliability (Spearman, 1904, p 253) converted observed correlations into estimates of the "true" (latent) correlation between various measures of cognitive ability.

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'_1 p'_2} r_{q'_1 q'_2}}} \tag{1}$$

The factor model

"We therefore bring our general theorem to the following form. Whenever branches of intellectual activity are at all dissimilar, then their correlations with one another appear wholly due to their being all variously saturated with some common fundamental Function (or group of Functions) (Spearman, 1904, 273)

This has become known as the *factor model*, which models Covariance matrices by

$$\boldsymbol{C} \approx \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}^2. \qquad (2)$$

Latent variables and test theory in the 20th century

The basic model of test theory for the next 120 years continued to use Spearman's idea of observed scores, true scores, and error.

- X = T + E

This led to the the various derivations of test reliability (internal consistency) as a function of item intercorrelation with formulae known as KR-20 (Kuder & Richardson, 1937), λ_3 (Guttman, 1945), and α (Cronbach, 1951).

- Because of computational constraints, although derivations of these formulae were based upon covariances, the calculations were done on test and item variances.

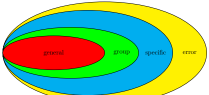
Latent variables and test theory in the 21st century

1. With the introduction of "high speed" computers, these techniques were elaborated with factor models of the correlations and coefficients such as ω_h, and ω_t (McDonald, 1999) could be calculated.
2. By using computers, Item Response Theory went from a theory (Lord, 1952; Lord & Novick, 1968) to an explosion of computer algorithms.

- Although initially seen as an alternative to factor analysis, IRT could be shown to be just "non-linear" factor analysis (McDonald, 1999) applied to tetrachoric or polychoric correlation matrices.

Test theory

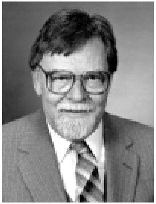
1. Classical and "Modern" psychometrics (IRT) treats items as mainly noise.
2. Our goal is to estimate the True (latent score) from manifest (observed) scores which are "befuddled with error".
3. Item variance = True Score Variance + Error (Spearman, 1904)
4. Item variance = General Factor variance + Group Variance + Error (McDonald, 1999) $(\sigma_e^2 = 1 - h^2)$
5. Item variance = General Factor variance + Group Variance + Specific Variance + Error
    - Specific and error are confounded unless we have test-retest measures.
    - Items tend to have communalities of .2 - .3 but short term stabilities of $\approx .8$, thus specific variance $\approx .5$.



Construct Validity as an extension of True Score Theory

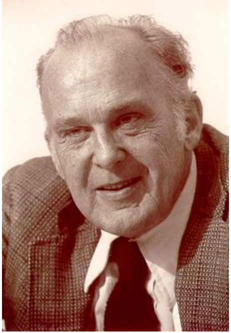
1. Construct validity in terms of the structure of latent variables was introduced by Cronbach & Meehl (1955). This was probably partly as a counter to behaviorism.
2. Elaborated by Loevinger (1957) who dismissed the idea of mere "practical" validity.
3. Construct validity could be conceptualized
    - Convergent: different measures of the same construct should go together
    - Divergent: measures of different constructs should not go together
    - Incremental: a measure should add something .

A test should be defined by what it measures and what it does not measure



Construct validity and the "Nomological Net"

1. Tests did not have validity, they were part of a network of validity.
2. Best exemplified in the Multi-Trait – Multi-Method Matrix of Campbell & Fiske (1959).




Agreement between Self Report and Peer Ratings
An example of a Multi-Trait–Multi-Method Matrix

Table: Self report and peer report from the SAPA-project. Correlations reported by Zola et al. (2021). Reliabilities on the main diagonal. Raw correlations below the diagonal. Correlations corrected for reliability above the diagonal. Upper left quadrant reflects SAPA Personality Inventory scores (Condon, 2018) for 158,631 participants, mean n/item = 18,180. Other quadrants reflect 908 peer rated participants. Data from the `zola` dataset in the *psychTools* package.

	Self Report					Peer Ratings				
Variable	Agrbl	Cnscn	Nrtcs	Extrv	Opnnn	Agrbl	Cnscn	Stblt	Extrv	IntlO
Agreeableness	**0.87**	0.32	-0.14	0.28	0.09	**0.75**	0.21	0.18	0.34	0.22
Conscientiousness	0.28	**0.87**	-0.20	0.13	0.06	0.16	**0.78**	0.22	**0.42**	0.13
Neuroticism	-0.12	-0.18	**0.90**	-0.28	-0.10	-0.01	-0.16	**-0.78**	-0.40	-0.25
Extraversion	0.25	0.12	-0.25	**0.90**	0.14	0.01	-0.01	0.07	**0.71**	0.14
Opennness	0.08	0.05	-0.09	0.13	**0.86**	-0.14	-0.06	0.10	0.17	**0.49**
Agreeableness	**0.47**	0.10	-0.01	0.00	-0.09	**0.45**	0.36	**0.47**	0.15	**0.44**
Conscientiousness	0.15	**0.55**	-0.12	-0.01	-0.04	0.18	**0.58**	**0.42**	**0.41**	**0.47**
Stability	0.13	0.16	**-0.58**	0.05	0.07	0.25	0.25	**0.60**	0.38	**0.52**
Extraversion	0.23	0.28	-0.27	**0.49**	0.11	0.07	0.23	0.22	**0.52**	0.32
IntellectOpenness	0.14	0.08	-0.15	0.09	0.30	0.19	0.24	0.27	0.15	**0.44**

The unfortunate emphasis upon construct validity reduced the emphasis upon the practical use of tests

1. In a response to operationalism, construct validity was in strong contrast to three other approaches.
2. Constructs, as embedded in nomological networks, were seen as theoretical concepts and could only be evaluated in terms of the pattern of correlations.
3. Criterion-oriented validation procedures, on the other hand, harkened back to the operational definitions of behaviorism.
    - Concurrent validity is the correlation with a current criterion.
    - Predictive validity is the correlation with a future criterion.
4. Content validity was established by showing that the test items were a sample of a universe in which the investigator is interested.

Loevinger and the boiling of eggs

1. Favorably quoting Marschak, Loevinger said: (p 641) "A theory provides us with solutions which are potentially useful for a large class of decisions.
2. Hence, the more we know about its properties the better. If we merely want to know how long it takes to boil an egg, the best is to boil one or two without going into the chemistry of protein molecules. The need for chemistry is due to our want to do other and new things "
3. She goes on to say "The argument against classical criterion-oriented psychometrics is thus two-fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry.
4. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful."

In case we did not understand

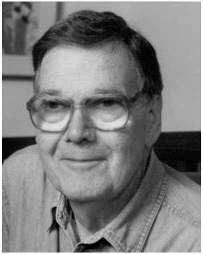
"the most fruitful direction for the development of psychometric devices, and hence of psychometric theory, is toward measurement of traits which have real existence in some sense;
that this orientation is antithetical to one which places first emphasis on prediction, decisions, or "utility;" that most decision-oriented psychometric studies would be more fruitfully formulated as trait-oriented studies;
and that such legitimately pressing decisions as must inevitably be made will also best be served by a predominantly trait-oriented psychometrics." Loevinger (1957)

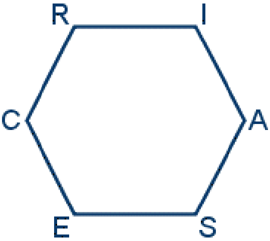
Believing in latent variables is like believing in the Easter Bunny

1. Although the Cronbach & Meehl (1955) and Loevinger (1957) papers were (and are) required reading in most personality programs
2. Not everyone accepted that construct validity was the ultimate goal.
3. The pragmatically oriented wanted tests to be useful.
4. To them, to believe in latent variables and construct validity was to believe in the Easter Bunny.

In defense of predictive validity

1. Perhaps reflecting their training at the University of Minnesota where the Strong Vocational Interest (Strong Jr., 1927) and the MMPI (Hathaway & McKinley, 1943) were developed. Harrison Gough and John Holland developed criterion based tests.
2. Gough developed the CPI and the Adjective Check List.
3. Holland is well known for his theory of vocational interests.



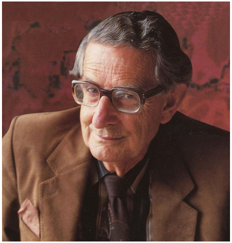


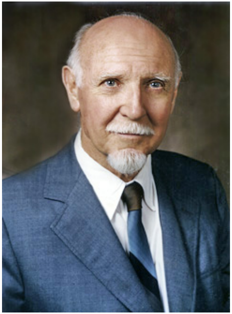
Successful application of predictive validity

1. The California Psychological Inventory (Gough, 1957), the Adjective Check List (Gough, 1960) and the Holland Coding system of the RIASEC have been very successful publications.
2. More well known to members of ISSID is the success of the Hogan Personality Inventory (Hogan & Hogan, 1995)

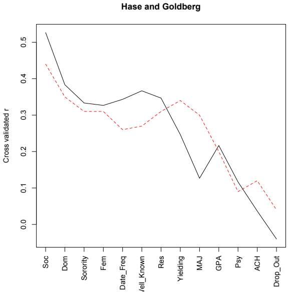
Test construction with and without latent variables

1. An unfortunate amount of effort was devoted in the 1950's, 60's and 70's to the debate about how to construct tests.
2. Proponents of factor based tests (e.g., the EPI, EPQ, the 16PF) included Hans and Sybil Eysenck (Eysenck & Eysenck, 1964) and Raymond Cattell (Cattell & Stice, 1957)
3. Proponents of empirically keyed tests, included e.g., the Strong, (Strong Jr., 1927), the MMPI (Hathaway & McKinley, 1943) and the JPI (Jackson, 1983) .
4. Rational based tests were developed by Gough (e.g., 1957).



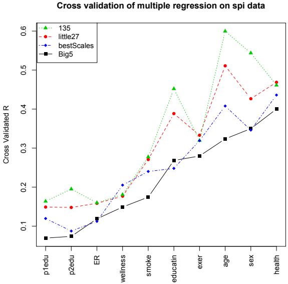


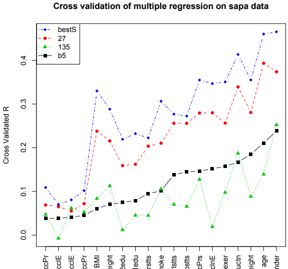
1. In a monumental work comparing these various approaches, Lew Goldberg compared these methods, as well as random keying.
2. The initial report (Hase & Goldberg, 1967) suggested no difference on the average of keying methods when comparing the validity for 13 criteria for factor, empirical, rational and theoretically derived keys.
3. In followup, however, Goldberg (1972) found that factorial procedures were best for easily predicted criteria, but worst for hard to predict criteria.



Comparing the power of items to factorially derived scales

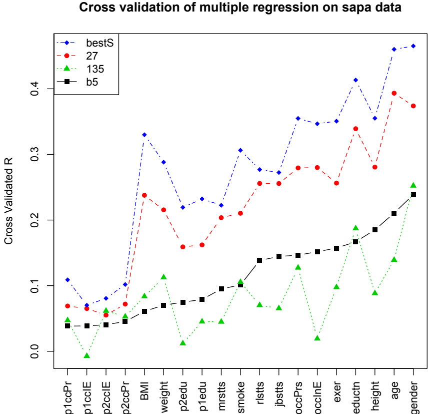
1. We recently compared factor based (5 and 27) scales to empirically derived scales for 135 items (N=4,000) as well as 696 items (N= 126,884) (Revelle, Dworak & Condon, 2021)
2. Using data from the `spi` dataset in the *psych* package in R and from Dataverse (Condon & Revelle, 2015; Condon, Roney & Revelle, 2017a,b), we compared cross validated multiple regressions for factor based, Machine Learning based (i.e., the `bestScales` algorithm), and raw items.





Empirical scales dominate latent factor score estimates

1. Cross validated latent factor based scores for 5 factors were uniformly dominated by those for 27 lower level factor based scales which were in turn dominated by just using the empirically based scales themselves.



But what are the items that predict?

1. The supposed advantage of factor based scores, is that we can understand the prediction.
2. And when we do the multiple regression, we can examine the β weights to "understand" the predictive model.

Table: Standardized coefficients from the Big 5 with `spi` data

Variable	p1edu	p2edu	ER	wllns	smoke	edctn	exer	age	sex	helth
((Intercept)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Agree	0.02	0.01	-0.03	0.06	-0.08	**0.12**	-0.01	**0.16**	**0.16**	0.01
Consc	-0.03	-0.05	0.02	**0.11**	-0.08	0.06	**0.16**	**0.13**	**0.10**	**0.17**
Neuro	-0.04	-0.03	**0.13**	0.03	0.06	**-0.15**	**-0.12**	**-0.14**	**0.29**	**-0.27**
Extra	0.05	0.06	0.05	0.09	0.08	-0.09	0.09	**-0.11**	0.09	**0.14**
Open	0.06	0.06	-0.01	0.00	0.09	**0.14**	0.07	**0.12**	**-0.12**	0.01
R	0.10	**0.11**	**0.13**	**0.17**	**0.18**	**0.26**	**0.27**	**0.31**	**0.36**	**0.41**

We can also examine the items that are most predictive.

R code

```
bs <- bestScales(spi[11:145],spi[1:10],
        dictionary=spi.dictionary[,c(2,6),drop=FALSE])
```

```
        health                                         item         L27
q_820    0.35             Feel comfortable with myself.    WellBeing
q_2765   0.34                      Am happy with my life.    WellBeing
q_811   -0.34 Feel a sense of worthlessness or hopelessness.    WellBeing
q_578   -0.34                            Dislike myself.    WellBeing
q_1371   0.31                                 Love life.    WellBeing
q_56     0.28            Am able to control my cravings. SelfControl
q_1505  -0.27                             Panic easily.      Anxiety
q_4249  -0.26        Would call myself a nervous person.      Anxiety
q_808   -0.26                        Fear for the worst.      Anxiety
q_1452  -0.25                         Neglect my duties.     Industry

          exer                                         item              L27
q_1024 -0.24                 Hang around doing nothing.    EasyGoingness
q_1052 -0.22                Have a slow pace to my life.    EasyGoingness
q_1444 -0.20                  Need a push to get started.        Industry
q_1452 -0.20                          Neglect my duties.        Industry
q_811  -0.19 Feel a sense of worthlessness or hopelessness.       WellBeing
q_1371  0.18                                  Love life.       WellBeing
q_2765  0.18                       Am happy with my life.       WellBeing
q_820   0.18              Feel comfortable with myself.       WellBeing
q_56    0.17             Am able to control my cravings.     SelfControl
q_1662  0.17                             Seek adventure. SensationSeeking
```

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 & 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

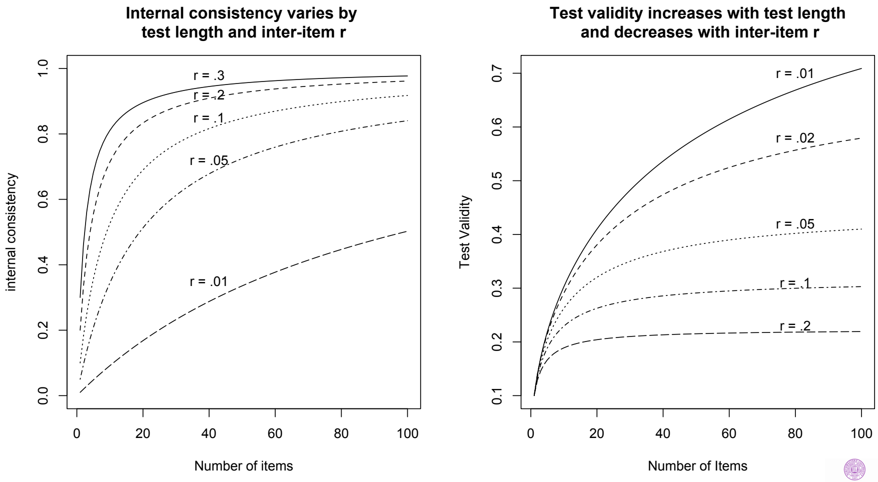
1. Lets unpack these two equations.
   Internal consistency

$$\lambda_3 = \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{3}$$

2. but validity

$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k*(k-1)\bar{r}}}. \tag{4}$$

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
2. And highly correlated 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 & Condon, 2018, 2019) and scale construction (Revelle & Garner, 2023).
5. Variations on this theme have been discussed before by (Condon, Wood, Möttus, Booth, Costani, Greiff, Johnson, Lukaszeski, Murray, Revelle, Wright, Ziegler & Zimmerman, 2021; Möttus, Wood, Condon, Back, Baumert, Costani, Epskamp, Greiff, Johnson, Lukaszeski, Murray, Revelle, Wright, Yarkoni, Ziegler & Zimmerman, 2020).

Two real life example of validity with low internal consistency

1. Alice Eagly and I (Eagly & Revelle, 2022) have recently reviewed how there are multiple perspectives to sex/gender differences.
   - There is a great number of studies showing small gender differences on many different measures.
   - There are also many studies showing large differences.
2. How can this be?
3. We suggest one should look both at the trees as well as the forests of sex differences.
4. That is, by properly aggregating data at the item level (the trees which show small differences) into composite scales (the forest), we showed some very large differences.

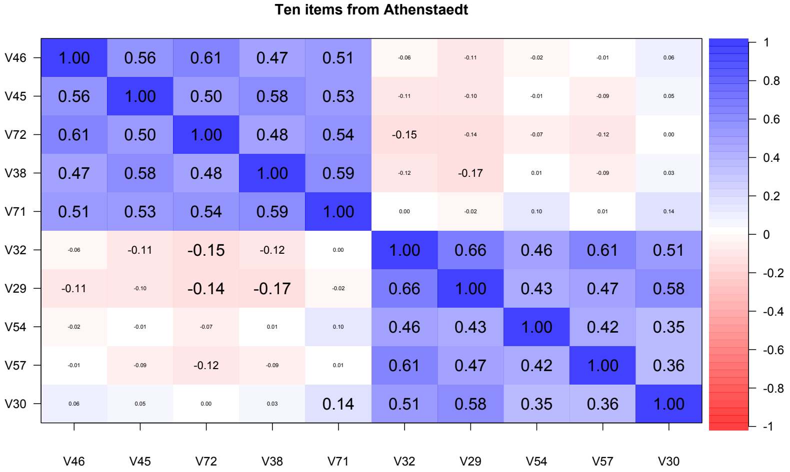
The power of aggregation

1. But this effect was not merely a large effect associated with aggregation.
2. The utility of aggregation has been known since (Spearman, 1910; Brown, 1910), and rediscovered by Fishbein & Ajzen (1974) with respect to attitudes and by Epstein (1983) with respect to personality.
3. (The power of aggregation was not forgotten by members of ISSID, but seems to have been forgotten by many others.)
4. We used data from Athenstaedt (2003) and Gruber, Distlberger, Scherndl, Ortner & Pletzer (2020) to show this effect most clearly.
5. Both sets of investigators asked gender specific questions.

The first 20 items from Athenstaedt (2003)

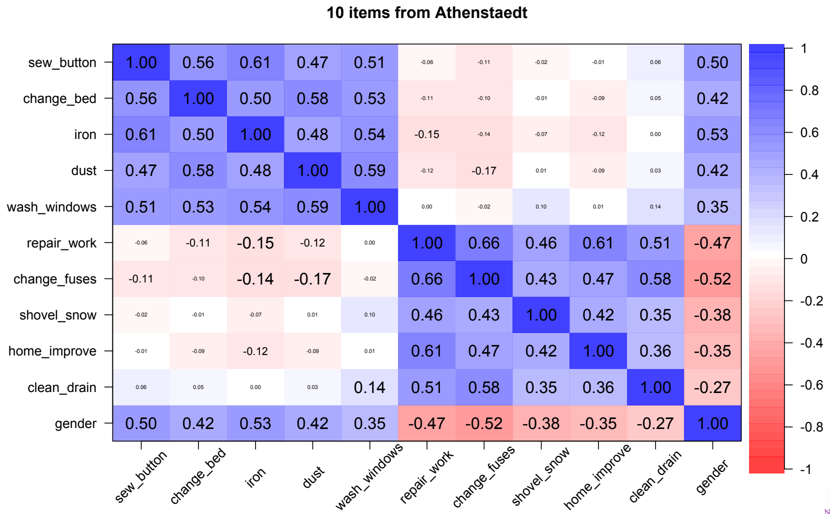
Var #	Item
V1	To pay attention to ones appearance in the office
V2	Offer fire to somebody
V3	Paint an Apartment
V4	Mow the Lawn
V5	Make the Bed
V6	Hold the Door Open for your Partner
V7	Do the Dishes
V8	Do Extreme Sports
V9	Tinker with the Car
V10	Talk about Sports
V11	Assemble Prefabricated Furniture
V12	Drive a Car in a Risky Way
V13	Listen Attentively to Others
V14	Tell your Partner about Problems at Work
V15	Play on a Computer
V16	Set the Table
V17	Watch ones Weight
V18	Care for a Partner if he/she is Ill
V19	Play Chess
V20	Meet with friends at a Regulars Table

10 items from Athenstaedt (2003)



Clearly a two factor solution (using the inter-ocular trauma test).

10 items from Athenstaedt (2003) with gender



```
     Factor Analysis using method =  minres
Call: fa(r = select.items[1:10], nfactors = 2)
Standardized loadings (pattern matrix) based upon correlation matrix
     MR1   MR2   h2  com                       Item
V71  0.76  0.12 0.58 1.05             Wash Windows
V45  0.74 -0.03 0.55 1.00        Change Bed Sheets
V46  0.73  0.01 0.53 1.00          Sew on a Button
V38  0.72 -0.05 0.53 1.01       Dust the Furniture
V72  0.72 -0.09 0.53 1.03           Do the Ironing
V32 -0.03  0.84 0.72 1.00           Do Repair Work
V29 -0.06  0.78 0.63 1.01             Change Fuses
V57 -0.02  0.66 0.43 1.00 Do Home Improvement Jobs
V30  0.14  0.65 0.42 1.10            Clean a Drain
V54  0.06  0.57 0.32 1.02              Shovel Snow

                       MR1  MR2
SS loadings           2.72 2.52
Proportion Var        0.27 0.25
Cumulative Var        0.27 0.52
Proportion Explained  0.52 0.48
Cumulative Proportion 0.52 1.00

 With factor correlations of
     MR1  MR2
MR1  1.0 -0.1
MR2 -0.1  1.0

Mean item complexity =  1
Test of the hypothesis that 2 factors are sufficient.
```

Find the regression of the gender on the factor score estimates – using biserial correlations

R code

```
select <- selectFromKeys(Athenstaedt.keys$MF10)
f2 <- fa(Athenstaedt[select],2)
temp <- cbind(gender=Athenstaedt$gender, f2$scores)
R <- mixedCor(temp)$rho
lmCor(gender ~ MR1 + MR2, data=R)  #biserial
lmCor(gender ~ MR1 + MR2, data=temp) #Pearson
```

```
Call: lmCor(y = gender ~ MR1 + MR2, data = R)

Multiple Regression from matrix input

 DV =  gender
    slope  VIF Vy.x
MR1  0.63 1.01 0.44
MR2 -0.61 1.01 0.41

 Multiple Regression
          R   R2  Ruw R2uw
gender 0.93 0.86 0.93 0.86
```

But biserial correlation is actually a latent model

What would the correlation be if Y were not dichotomous but rather a latent normal distribution. Same problem with logistic regression, which is also a latent model.

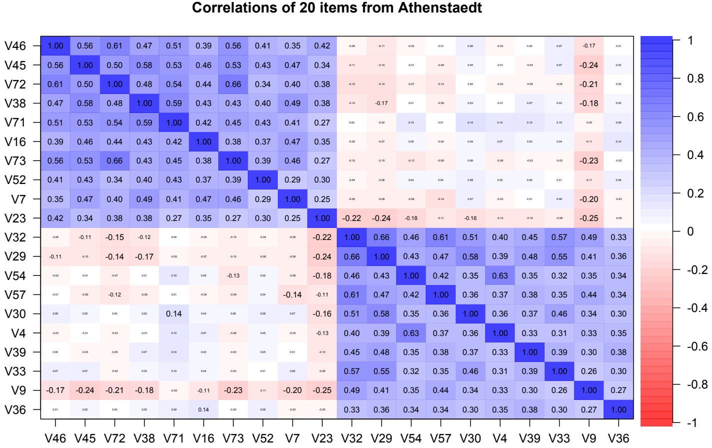
Try just normal regression.

R code

```
lmCor(gender ~ MR1 + MR2, data=temp)
```

```
Multiple Regression from raw data

 DV =  gender
            slope   se      t       p lower.ci upper.ci  VIF Vy.x
(Intercept)  0.00 0.03   0.00 1.0e+00    -0.06     0.06 1.00 0.00
MR1          0.51 0.03  17.85 5.2e-57     0.45     0.56 1.01 0.28
MR2         -0.49 0.03 -17.13 2.0e-53    -0.54    -0.43 1.01 0.26

Residual Standard Error =  0.68  with  573  degrees of freedom

 Multiple Regression
          R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
gender 0.74 0.55 0.74 0.55        0.54     0.03    343.73   2 573 8.14e-99
```

20 items $\omega_h = .14, \alpha = .85, \omega_t = .89, r_v = .77$



Clearly a two factor solution (using the inter-ocular trauma test).

2 factors of the Athenstaedt (2003) data

	MR1	MR2	h2	com	Item
V45	0.75	-0.01	0.57	1.00	Change Bed Sheets
V72	0.73	-0.07	0.55	1.02	Do the Ironing
V46	0.73	0.01	0.53	1.00	Sew on a Button
V71	0.72	0.14	0.53	1.08	Wash Windows
V38	0.72	-0.01	0.52	1.00	Dust the Furniture
V73	0.70	-0.06	0.50	1.02	Do the Laundry
V16	0.62	0.05	0.38	1.01	Set the Table
V7	0.60	-0.03	0.36	1.00	Do the Dishes
V52	0.56	0.06	0.31	1.02	Take Care of Flowers
V23	0.47	-0.20	0.28	1.37	Wrap Presents Beautifully
V32	-0.05	0.80	0.65	1.01	Do Repair Work
V29	-0.08	0.77	0.61	1.02	Change Fuses
V30	0.13	0.65	0.42	1.08	Clean a Drain
V57	-0.04	0.65	0.43	1.01	Do Home Improvement Jobs
V54	0.03	0.63	0.40	1.00	Shovel Snow
V33	0.10	0.62	0.39	1.06	Change Light Bulbs
V4	0.05	0.60	0.36	1.01	Mow the Lawn
V39	0.07	0.60	0.35	1.03	Buy Electric Appliances
V9	-0.21	0.54	0.36	1.29	Tinker with the Car
V36	0.10	0.51	0.26	1.08	Cook Meat on the Grill

	MR1	MR2
SS loadings	4.54	4.22
Proportion Var	0.23	0.21
Cumulative Var	0.23	0.44
Proportion Explained	0.52	0.48
Cumulative Proportion	0.52	1.00

With factor correlations of

	MR1	MR2
MR1	1.00	-0.09
MR2	-0.09	1.00

Wait, aren't we doing latent variable modeling? I thought we didn't believe in the Easter Bunny

1. We were finding two latent factors of 20 items.
2. Showing how these two linear composites could predict the criterion better than either separately.
3. Lets try simple "dustbowl empiricism" (now known as "Supervised Machine Learning").
4. We have known since the 50's (e.g., Cureton, 1950) that we need to cross validate all regression models for they overfit the data.
5. We use the `bestScales` function (aka BISCUIT (Elleman, McDougald, Revelle & Condon, 2020)) to do K-fold cross validation.
6. BISCUIT: Best Item Scales that are Cross-validated, Unit-weighted, Informative and Transparent.

BISCUIT algorithm is a very simple procedure

1. For k = e.g. 10, and N subjects, find the items that most correlate with the criterion for $\frac{N}{k-1}$, and then validate these on the remaining $\frac{N}{k}$.
2. Repeat this k times.
3. Count how many times each item is the top n.items.
4. Form a scale of those top items.
5. Elleman et al. (2020) compared BISCUIT to more elegant Machine Learning Algorithms (e.g. Lasso, Elastic Net, and Random Forest) (James, Witten, Hastie & Tibshirani, 2022).
6. BISCUIT is particularly appropriated for the case of large amounts of missing data.
7. BISCUIT is implemented as `bestScales` in the *psych* package.

R code

```
bestScales(Athenstaedt[,2:75],"gender",n.item=20,folds=10,
           dictionary=Athenstaedt.dictionary)
```

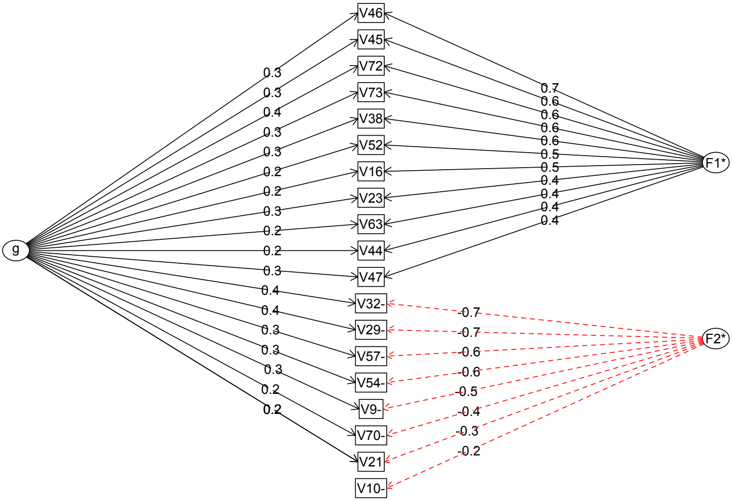
```
Number of iterations set to the number of folds =  10
Call = bestScales(x = Athenstaedt[, 2:75], criteria = "gender", n.item = 20,
    folds = 10, dictionary = Athenstaedt.dictionary)
       derivation.mean derivation.sd validation.m validation.sd final.valid final.wtd N.w
gender            0.83        0.0078         0.82         0.021        0.82      0.81

 Best items on each scale with counts of replications
 Criterion =  gender
    Freq mean.r sd.r ItemLabel                                        Item
V23   10   0.53 0.01       V23                   Wrap Presents Beautifully
V72   10   0.53 0.01       V72                             Do the Ironing
V29   10  -0.52 0.01       V29                              Change Fuses
V46   10   0.50 0.01       V46                            Sew on a Button
V47   10   0.48 0.01       V47                                Do Aerobics
V73   10   0.48 0.01       V73                             Do the Laundry
V32   10  -0.47 0.01       V32                             Do Repair Work
V70   10  -0.45 0.01       V70 Help your Partner Put on His or Her Coat
V9    10  -0.43 0.01        V9                         Tinker with the Car
V38   10   0.42 0.02       V38                         Dust the Furniture
V45   10   0.42 0.02       V45                          Change Bed Sheets
V44   10   0.41 0.01       V44                 Do Handiwork (e.g. Knitting)
V52   10   0.38 0.01       V52                       Take Care of Flowers
V54   10  -0.38 0.01       V54                                 Shovel Snow
V21   10   0.37 0.01       V21                          Watch Soap Operas
V63   10   0.36 0.01       V63             Decorate the Office with Flowers
V16   10   0.36 0.01       V16                               Set the Table
V10   10  -0.36 0.01       V10                          Talk about Sports
```

The best scales solution

1. What are the items that predict the criterion?
2. Add them up (or find the mean of the items for each person) and find the correlation of this composite with the criterion.
3. The resulting scale is not necessarily (and probably not) univocal. It is not a good "latent" variable but is a good manifest predictor of the criterion ($r = .83$).
4. The structure of this scale shows that the general factor saturation $\omega_h = .3$ meaning that it is definitely not measuring one latent variable.

$\omega_h = .3, \alpha = .85, \omega_t = .88, r_v = .83$ for BISCUIT scale

omega of the BISCUIT based best 20 items for the Athenstaedt data



The data from the Athenstaedt (2003) data set

Table: Choosing the best k items to predict sex in the Athenstaedt (2003) data set. r is the correlation of a k-item scale with sex, avrg is the average correlation with the predictor set, alpha is the alpha reliability of the predictor set. Means show the average validity of the items used in the scale.

A table from the psych package in R

k	scale validity	avrg.	alpha	mean validity
5	0.66	0.14	0.49	0.43
10	0.74	0.13	0.62	0.40
20	0.77	0.11	0.72	0.35
30	0.76	0.10	0.77	0.32
40	0.76	0.09	0.80	0.29
50	0.75	0.08	0.81	0.26
60	0.75	0.06	0.81	0.24
70	0.72	0.05	0.79	0.21

The prior analysis was choosing items in terms of their validity. That is to say, we take the cream first. Compare this to just randomly choosing items. In this case, as the number of items being aggregated increases, the validity increases as predicted by Equation 4.

Table: The item and scale statistics when scales are formed from random subsets of domain items. The ratio is just the average validity/sqrt(average item correlation).

Reliability and validity of various length scales when items are chosen randomly.

Variable	N.items	alpha	validity	average.r	item.validity	ratio	modeled
r.five	5	0.20	0.43	0.05	0.21	0.96	0.43
r.ten	10	0.51	0.53	0.09	0.23	0.75	0.53
r15	15	0.68	0.56	0.12	0.24	0.69	0.56
r20	20	0.73	0.61	0.12	0.25	0.72	0.61
r30	30	0.80	0.64	0.12	0.25	0.72	0.64
r40	40	0.85	0.68	0.12	0.26	0.74	0.68
r50	50	0.88	0.68	0.13	0.26	0.73	0.68
all.56	56	0.89	0.70	0.12	0.26	0.75	0.70

Another data set (GERAS) is from Gruber et al. (2020)

1. Gruber et al. (2020) report on the psychometric properties of a multifaceted Gender Related Attributes Survey.
2. They included 3 domains (Personality, Cognition and Activities and Interests) in their study 2.
3. The data are included in *psych* as the GERAS data set. N =471.
4. Try BISCUIT again.

The data from the Gruber et al. (2020) data set

Table: Exploring the benefits and costs of aggregation. Although reliability will increase, because the items were chosen in order of their validity, scale validity is non-monotonic with the number of items (see figure). The ratio is just the average validity/sqrt(average item correlation.

Reliability and validity of various length scales when items are chosen by their validity.

Variable	N.items	alpha	validity	average.r	item.validity	ratio	modeled
five	5	0.67	0.65	0.29	0.43	0.80	0.65
ten	10	0.76	0.71	0.24	0.40	0.82	0.71
fifteen	15	0.80	0.73	0.21	0.37	0.81	0.73
twenty	20	0.82	0.74	0.19	0.35	0.82	0.74
thirty	30	0.85	0.73	0.16	0.32	0.79	0.73
fourty	40	0.87	0.73	0.14	0.29	0.78	0.73
fifty	50	0.88	0.72	0.13	0.27	0.77	0.72
fiftysix	56	0.89	0.70	0.12	0.26	0.75	0.70

R code

```
reliability(GERAS.keys,GERAS.items)
```

```
Measures of reliability
reliability(keys = GERAS.keys, items = GERAS.items)
       omega_h alpha omega.tot  Uni r.fit fa.fit max.split min.split mean.r med.r n.items
M.pers    0.01  0.66      0.65 0.31  0.46   0.66      0.81      0.29   0.16  0.14      10
F.pers    0.25  0.80      0.84 0.68  0.73   0.93      0.86      0.65   0.28  0.26      10
M.cog     0.30  0.73      0.84 0.42  0.62   0.68      0.83      0.33   0.28  0.20       7
F.cog     0.42  0.70      0.76 0.65  0.70   0.93      0.80      0.57   0.25  0.20       7
M.act     0.48  0.75      0.78 0.83  0.89   0.94      0.82      0.65   0.27  0.26       8
F.act     0.48  0.75      0.79 0.79  0.87   0.91      0.81      0.59   0.27  0.26       8
Pers      0.09  0.77      0.81 0.25  0.43   0.59      0.89      0.35   0.14  0.11      20
Cog       0.05  0.67      0.74 0.07  0.17   0.39      0.83      0.03   0.13  0.09      14
Act       0.09  0.75      0.79 0.34  0.57   0.59      0.86      0.29   0.16  0.16      16
M         0.25  0.81      0.83 0.36  0.57   0.62      0.89      0.60   0.15  0.13      25
F         0.23  0.83      0.85 0.47  0.62   0.76      0.90      0.66   0.17  0.14      25
MF.all    0.26  0.85      0.86 0.23  0.43   0.52      0.91      0.69   0.10  0.09      50
```

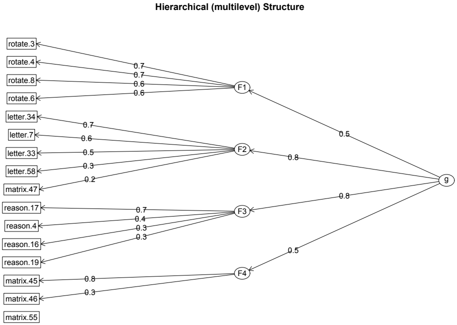
Table: For the three subdomains as well as Masculine and Feminine subscales of the Gruber et al. (2020) data we found three estimates of reliability ($\omega h, \alpha, \omega_t$), average within scale item correlation ($\bar{r}_i$), average item validity ($\bar{r_y}$), observed scale validity, average item Cohen d ($\bar{d}$) and Cohen d for the aggregated scale.

Variable	ω_h	α	ω_t	$\bar{r}$	items	$\bar{r_y}$	Validity	$\bar{d}$	C_d
M.pers	0.01	0.66	0.65	0.16	10	0.14	0.28	0.28	0.57
F.pers	0.25	0.80	0.84	0.28	10	-0.23	-0.39	-0.48	-0.86
M.cog	0.30	0.73	0.84	0.28	7	0.19	0.31	0.39	0.65
F.cog	0.42	0.70	0.76	0.25	7	-0.12	-0.20	-0.24	-0.40
M.act	0.48	0.75	0.78	0.27	8	0.23	0.38	0.48	0.83
F.act	0.48	0.75	0.79	0.27	8	-0.38	-0.63	-0.82	-1.61
Pers	0.09	0.77	0.81	0.14	20	0.18	0.42	0.38	0.93
Cog	0.05	0.67	0.74	0.13	14	0.15	0.36	0.31	0.77
Act	0.09	0.75	0.79	0.16	16	0.30	0.65	0.65	1.73
M	0.25	0.81	0.83	0.15	25	0.18	0.44	0.37	0.97
F	0.23	0.83	0.85	0.17	25	-0.25	-0.58	-0.52	-1.41
MF.all	0.26	0.85	0.86	0.10	50	0.21	0.63	0.45	1.61

But what about g?

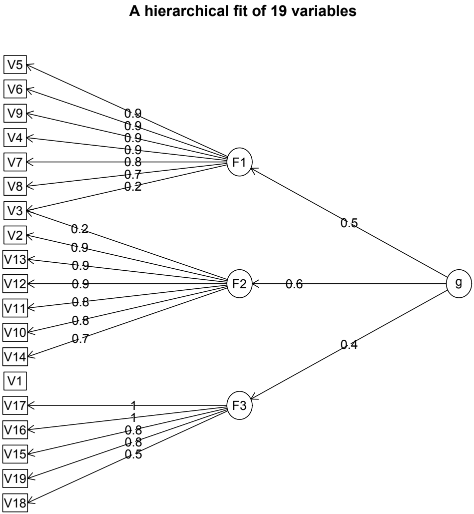
1. One of Spearman's great contributions was recognizing that measures of cognitive ability are all correlated positively.
2. However, this positive manifold does not imply a general causal factor.
3. It is important to remember that factors are (convenient) fictions Revelle (1983); Revelle & Ellman (2016); Schonemann (1990)
4. Alternative explanations of the positive manifold include:
    - g: a general factor (Spearman, 1904)
    - Multiple, independent "bonds" (Bartholomew, Deary & Lawn, 2009; Thomson, 1916)
    - Growth influences multiple processes (Kovacs & Conway, 2016, 2019)
    - "cluster" psychometrics

Higher order factors of the ICAR-16



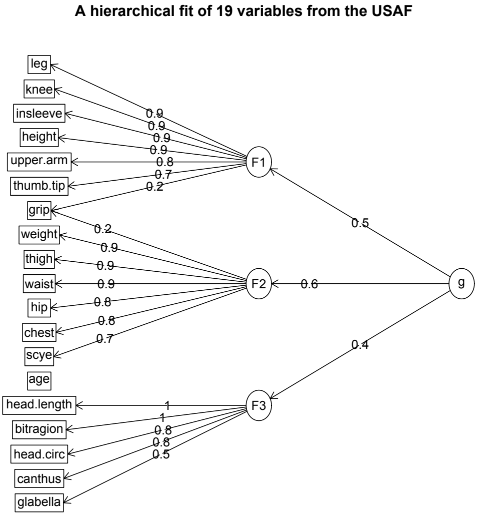
1. Clear higher order factor from the ICAR (Condon & Revelle, 2014)
2. Data are in the `ability` dataset
3. 4000 participants and 16 items from ICAR-16 taken from the SAPA study

19 variables show a clear general factor



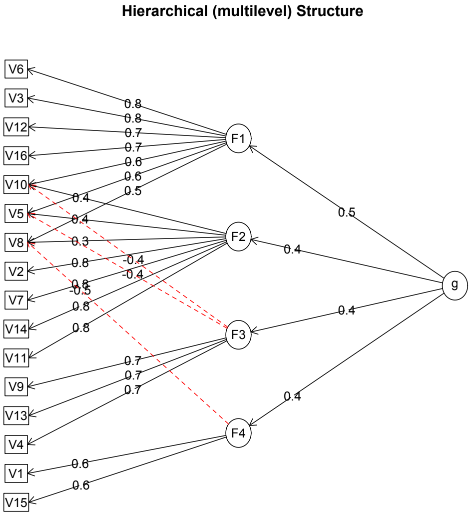
1. Clear higher order factor
2. Just another example of ‘g’
3. g models are typically shown as g causing the lower order factors
4. But is this just one mathematical model?

19 measures from the United States Air Force USAF



1. Clear higher order factor
2. Is being big a ‘g’ for size?
3. Do we really think that being big causes longer arms?

A general factor from the `sim.bonds` model



1. Simulation using the `sim.bonds` model of multiple independent causes.
2. $\omega_h = .4, \alpha = .86, \omega_t = .91$
3. See (Bartholomew et al., 2009; Thomson, 1916) for elaboration.

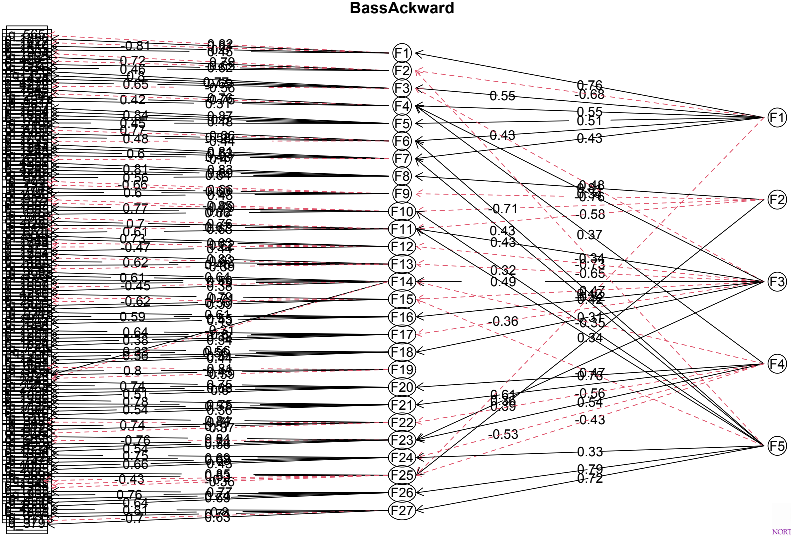
The search for structure is misled by the factor model

1. In cognitive psychology, the search for a single higher order factor led to such models as the Carroll-Horn-Cattell (CHC) model (McGrew, 2009, 2023)
   - This was an attempt to organize the many lower order factors of cognitive ability
   - But process models account for the structure just as well.
2. In personality we have seen hierarchical models including
   - Eysenck's well known hierarchy
   - Attempts to find higher order factors of the Big 5 (Digman, 1997; Loehlin, 2012; Loehlin & Martin, 2011; Revelle & Wilt, 2013)
   - Lower order "aspects" of the big 5 (DeYoung, Quilty & Peterson, 2007)

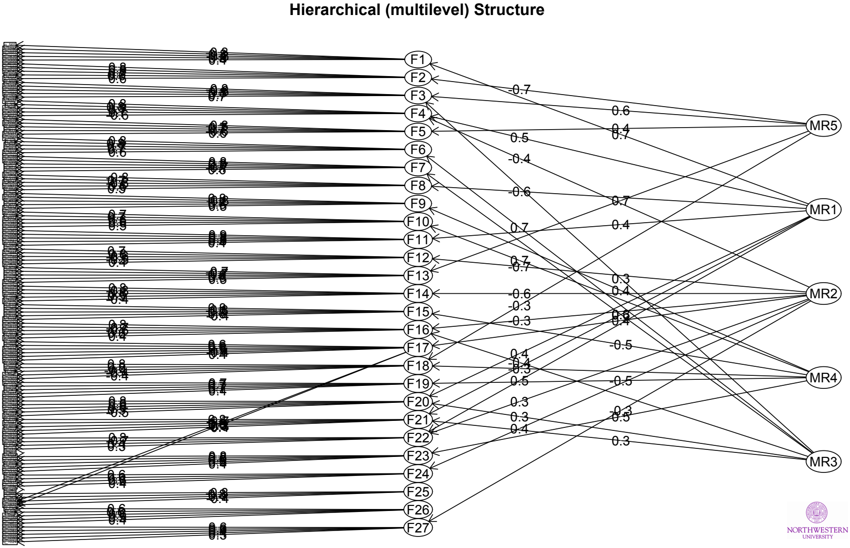
Higher order factors versus fewer factors

1. Confusion about hierarchical versus multiple representations
2. Goldberg (2006); Waller (2007) consider ways to represent multiple factor solutions in terms of factor/component correlations.
3. This produces what Goldberg described as the Bass-Ackwards solution .
4. Unfortunately some confuse Bass-Ackwards with a hierarchical (factors of factors) solution. (But not Waller.)
5. These are not the same.
    - Consider 135 items from the SPI (Condon, 2018) `spi` data set
    - Bass-ackwards for 27 and 5 factors
    - Hierarchical solution for 27 and 5 factors

bassAckwards solution of the spi of the spi items



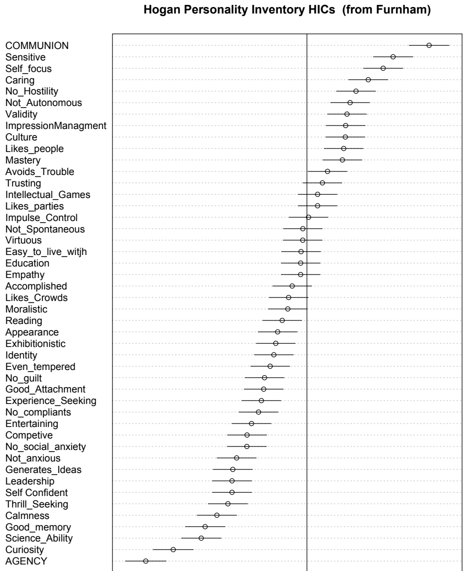
fa.hierarchical solution of the spi of the spi items



Revelle, Furnham & Eagly (2023) have started a joint project

1. Adrian has collected voluminous data sets with various Hogan Instruments
2. Has reported small sex differences on individual scales (Furnham & Treglown, 2021)
3. But when empirically (not factor analytically) combining these scales to find sex differences, two higher order dimensions appear
4. These can loosely be associated the dimensions of interpersonal agency and communion.
5. This solution was not a consequence of a higher order model, but rather going back to the original data.
6. The Agency and Communion dimensions are formed from lower level HICS but not from any one higher order factor.
7. Agency is also known as Getting Ahead; Communion as Getting Along

HICs sorted by Cohen d. From Revelle et al. (2023)



1. Cohen d (standardized mean differences) for sex differences on 44 HICs
2. Data for Hogan Personality Inventory (Hogan & Hogan, 1995, 2007) from Furnham & Treglown (2021).

Table: Cohen ds and 95% confidence intervals select from 44 Hogan HICs. The *Communion* and **Agency composite scores are also included.**

Variable	lower	Cohen d	upper	Name of HIC	Domain
H1a	0.11	0.22	0.32	Validity	
H2a	-0.14	-0.03	0.07	Empathy	Adjustment
H5a	-0.59	-0.48	-0.38	**Calmness**	
H10a	-0.43	-0.32	-0.22	**Competive**	Ambition
H11a	-0.51	-0.40	-0.30	**Self Confident**	
H13a	-0.51	-0.40	-0.30	**Leadership**	
H16a	-0.05	0.06	0.16	Likes parties	Sociability
H21a	-0.14	-0.03	0.07	Easy to live with	Interpersonal Sensivity
H22a	0.36	0.46	0.57	*Sensitive*	
H23a	0.22	0.33	0.43	*Caring*	
H25a	0.16	0.26	0.37	*No Hostility*	
H26a	-0.21	-0.10	0.00	Moralistic	Prudence
H29a	0.13	0.23	0.34	*Not Autonomous*	
H33a	-0.67	-0.57	-0.46	**Science Ability**	Inquisitive
H34a	-0.83	-0.72	-0.61	**Curiosity**	
H35a	-0.53	-0.42	-0.32	**Thrill Seeking**	
H37a	-0.50	-0.40	-0.29	**Generates Ideas**	
H38a	0.10	0.21	0.31	Culture	
H39a	-0.14	-0.03	0.07	Education	Learning Approach
H40a	-0.65	-0.55	-0.44	**Good memory**	
Agency	-0.97	-0.87	-0.76	AGENCY	
Communion	0.55	0.66	0.76	COMMUNION	

Summary of Agency and Communion

Table: Summary of Cohen d for Agency and Communion from each study. Compare these aggregated scores with the absolute Cohen d. The last three columns shows the Mahalobinis distances (MD), the common language effect size (CL), U3 statistic and the correlation of agency and communion for each study.

Inventory	Agency	Communion	$\bar{\lvert d \rvert}$	MD	CL	U3	r_{AC}
NEO1	-0.20	0.63	0.21	1.40	0.84	0.92	
NEO2	-0.28	0.62	0.23	1.45	0.85	0.93	
HPI	-0.87	0.66	0.24	1.53	0.86	0.94	
16PF	-0.31	0.90	0.20	1.2	0.80	0.88	-.25
HPTI	0.54	0.49	0.17	0.96	0.75	0.83	-.29
MVPI	-0.67	0.35	0.21	1.30	0.82	0.90	
MVPI2	-0.65	0.40	0.20	1.21	0.80	0.89	
HDS1	-0.23	0.14	0.08	0.50	0.64	0.69	
HDS2	-0.43	0.40	0.19	0.71	0.69	0.76	
HDS3	-0.40	0.34	0.17	0.64	0.67	0.74	
Mean	-0.47	0.44	0.19	1.09	0.77	0.85	

Summary

1. Although emphasized for 120 years, latent variable models are not particularly helpful.
2. By forcing a true score model they prevent us from discovering how item/scale specific variance relates to criteria.
3. By emphasizing theoretical purity over empirical utility, they cause us to ignore the real power and utility of personality and ability: To predict behavior.
4. It is time for us to abandon our beliefs in the tooth fairy and the easter bunny and develop a real science of prediction.

Conclusions

1. Luck, persistence and thinking about thing differently are important to success.
2. The contributions of many students and colleagues is greatfully acknowledged.
3. Many of those colleagues are from ISSID and other obscure societies such as SMEP.
4. Although latent variables have had a long and honored history, perhaps we should focus on what we we could do well which is predict behavior.

Athenstaedt, U. (2003). On the content and structure of the gender role self-concept: Including gender-stereotypical behaviors in addition to traits. *Psychology of Women Quarterly*, *27*(4), 309–318.

Bartholomew, D., Deary, I., & Lawn, M. (2009). A new lease of life for Thomson's bonds model of intelligence. *Psychological review*, *116*(3), 567–579.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322.

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

Cattell, R. B. & Stice, G. (1957). *Handbook for the Sixteen Personality Factor Questionnaire.* Champaign, Ill.: Institute for Ability and Personality Testing.

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

Condon, D. M. & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.

Condon, D. M. & Revelle, W. (2015). Selected personality data from the SAPA-Project: 08dec2013 to 26jul2014. *Harvard Dataverse*.

Condon, D. M., Roney, E., & Revelle, W. (2017a). Selected personality data from the sapa-project: 22dec2015 to 07feb2017. [48,350 participant data file and codebook]. *Harvard Dataverse*.

Condon, D. M., Roney, E., & Revelle, W. (2017b). Selected personality data from the sapa-project: 26jul2014 to 22dec2015. [54,855 participant data file and codebook]. *Harvard Dataverse*.

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costani, G., Greiff, S., Johnson, W., Lukaszeski, A., Murray, A., Revelle, W., Wright, A. G., Ziegler, M., & 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. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, *10*(1), 94–96.

Deary, I. J. (2001). *Intelligence: A very short introduction*. OUP Oxford.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, *93*(5), 880–896.

Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*, 1246–1256.

Eagly, A. H. & Revelle, W. (2022). Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees. *Perspectives on Psychological Science*, *17*(5), 1339–1358.

Elleman, L. G., McDougald, S., Revelle, W., & 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.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360 – 392.

Eysenck, H. J. (1944). Types of personality: a factorial study of seven hundred neurotics. *The British Journal of Psychiatry*, *90*(381), 851–861.

Eysenck, H. J. & Eysenck, S. B. G. (1964). *Eysenck Personality*

Inventory. San Diego, California: Educational and Industrial Testing Service.

Fishbein, M. & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological review*, *81*(1), 59.

Furnham, A. & Treglown, L. (2021). Sex differences in personality scores on six scales: Many significant, but mostly small, differences. *Current Psychology*, 1–11.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs. No 72-2*, *7*.

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40*(4), 347 – 358.

Gough, H. G. (1957). Manual for the California Psychological Inventory.

Gough, H. G. (1960). The adjective check list as a personality assessment research technique. *Psychological Reports*, *6*(1), 107–122.

Gruber, F. M., Distlberger, E., Scherndl, T., Ortner, T. M., & Pletzer, B. (2020). Psychometric properties of the multifaceted gender-related attributes survey (GERAS). *European Journal of Psychological Assessment*, *36*(4), 612–623.

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

Hase, H. D. & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, *67*(4), 231–248.

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

Hogan, R. & Hogan, J. (1995). *The Hogan personality inventory manual (2nd. ed.)*. Tulsa, OK: Hogan Assessment Systems.

Hogan, R. & Hogan, J. (2007). *The Hogan Personality Inventory (3rd Editionr)*. Tulsa, OK: Hogan Assessment Systems.

Jackson, D. N. (1983). *JPI-R: Jackson Personality Inventory, Revised*. Sigma Assesment Systems, Incorporated.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *An introduction to statistical learning with applications in R:* (2 ed.). New York: Springer Science and Business Media.

Jensen, A. R. (1998). The g factor: The science of mental ability. *Westport, CT: Prager*.

Jensen, A. R. & Weng, L.-J. (1994). What is a good g? *Intelligence*, *18*(3), 231–258.

Kovacs, K. & Conway, A. R. (2019). A unified cognitive/differential approach to human intelligence: Implications for iq testing. *Journal of Applied Research in Memory and Cognition*, *8*(3), 255 – 272.

Kovacs, K. & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151 – 177.

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

Loehlin, J. (2012). How general across inventories is a general factor of personality? *Journal of Research in Personality*, *46*(3), 258–263.

Loehlin, J. C. & Martin, N. G. (2011). The general factor of personality: Questions and elaborations. *Journal of Research in Personality*, *45*(1), 44–49.

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

Lord, F. (1952). A theory of test scores (Psychometric Monograph No. 7). *Iowa City, IA: Psychometric Society*.

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

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

McGrew, K. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10.

McGrew, K. S. (2023). Carroll'sthree-stratum (3s) cognitive ability theory at 30 years: Impact, 3s-chc theory clarification, structural replication, and cognitive achievement. psychometric network analysis extension. *Journal of Intelligence*, *11*(2).

Mõttus, 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., & 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).

Plato (n.d.). *Plato The Republic : the complete and unabridged Benjamin Jowett translation (1892)* (3rd ed.). Oxford: Oxford Univeristy Press.

Pluchino, A., Biondo, A. E., & Rapisarda, A. (2018). Talent versus luck: the role of randomnss in success and failure. *Advances in Complex Systems*, *21*.

Revelle, W. (1983). Factors are fictions, and other comments on individuality theory. *Journal of Personality*, *51*(4), 707–714.

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

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

Revelle, W., Dworak, E. M., & Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, *169*.

Revelle, W. & Ellman, L. G. (2016). Factors are still fictions [peer commentary on "towards more rigorous personality trait–outcome research," by R. Mõttus]. *European Journal of Personality*, *30*, 324–325.

Revelle, W., Furnham, A. F., & Eagly, A. H. (2023). Further thoughts on understanding the magnitude of sex/gender differences in psychology. *working draft (in prep)*.

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

Revelle, W. & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality*, *47*(5), 493–504.

Schonemann, P. H. (1990). Facts, fictions, and common sense about factors and components. *Multivariate Behavioral Research*, *25*(1), 47–51.

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

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.

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

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology*, *8*, 271–281.

Waller, N. (2007). A general method for computing hierarchical component structures by Goldberg's Bass-Ackwards method. *Journal of Research in Personality*, *41*(4), 745 – 752.

Zola, A., Condon, D. M., & Revelle, W. (2021). The Convergence of Self and Informant Reports in a Large Online Sample. *Collabra: Psychology*, *7*(1). 25983.

NORTHWESTERN UNIVERSITY