Alternative metaphors for validity: Spear fishing versus fishing nets

William Revelle and Kayla M. Garner

Department of Psychology
Northwestern University
Evanston, Illinois USA

World Conference on Personality, Curacao

April 2024
Slides available at https://personality-project.org/sapa

Abstract

In contrast to the current belief that high internal consistency is an important aspect of a test, we show that predictive validity may be negatively associated with internal consistency. We show that broader tests with lower internal consistency (fishing nets) out perform tests with high internal consistency (sharp spears).
The evaluation of personality scales has been seduced by a belief in latent variables and the importance of construct validity at the cost of actually being useful for prediction (Revelle, 2024). We will examine the tradeoff between internal consistency and validity with examples from gender, pro-envioronmental attitudes, and beliefs about gun control (Garner, 2024) In all of these domains, validity was non-monotonically related to internal consistency: Less internally consistent scales were more valid than were ones with a cleaner factor structure and higher internal consistency.
Although these ideas are not new (Gulliksen, 1950), they seem to have been forgotten. It is time for personality measurement to focus on predicting real things rather than emphasizing theoretically pure but vacuous measures.

We will review the historic use of "dust bowl empiricism" in scale construction and consider the use of simple machine learning algorithms (e.g., `bestScales` in *psych*) to develop predictive instruments.

Outline

Reliability and validity

Items

Validity

A bit of math

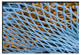
A few examples

Conclusion

Additional slides

Predictive validity versus construct validity

1. Many would agree that theory development is more fun than the hard work of making personality research actually useful.
2. Emphasis upon theory development has led to an emphasis on construct validity (Cronbach & Meehl, 1955; Loevinger, 1957) at the cost of predictive validity (Hogan, 2009; Hogan & Sherman, 2020; Gough, 1965; Gulliksen, 1950).
3. An alternative framework considers the importance of items in prediction real world criteria without focussing upon construct validity (Revelle, 2024).
4. Here we elaborate on the power of items and suggest that when constructing scales we should focus on the breadth of our measures (*fishing nets*) rather than high levels of internal consistency measurement (*spear fishing*).
5. For we catch more fish with fishing nets than spears.



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 & 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).

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

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

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

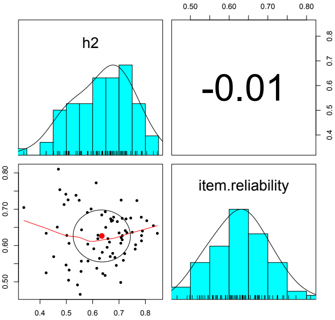
$$\omega_h = \frac{(\Sigma\lambda_i)^2}{\sigma_X^2} = \frac{\boldsymbol{1cc'1'}}{\sigma_X^2}. \tag{3}$$

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

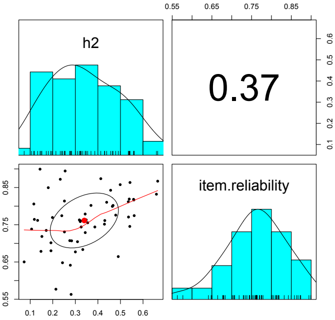
1. In the case of one administration, specific and error are confounded.
2. But, if we have repeated measures (t_1, t_2) , we can show that the reliable variance $(r_{t_1 t_2})$ is much greater than the common variance (h^2).
3. Consider the reliability of 75 mood items taken twice $(\overline{r_{12}} = .63)$ 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 & 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

Communality and item reliability for the MSQ



Communality and item reliability for the 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 & McKinley, 1943) or the Strong Vocational Interest test (Strong Jr., 1927), validity came to include *construct validity* (Cronbach & Meehl, 1955; Loevinger, 1957).
3. By emphasizing constructs, and the convergent and discriminant patterns of correlations (Campbell & 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.

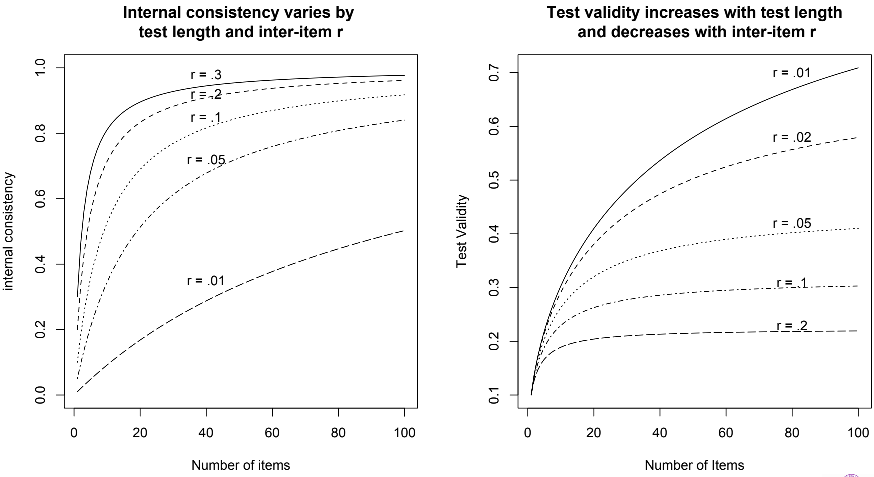
Validity and reliability: a short digression

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 & 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
   $$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k+k*(k-1)\bar{r}}}. \tag{5}$$

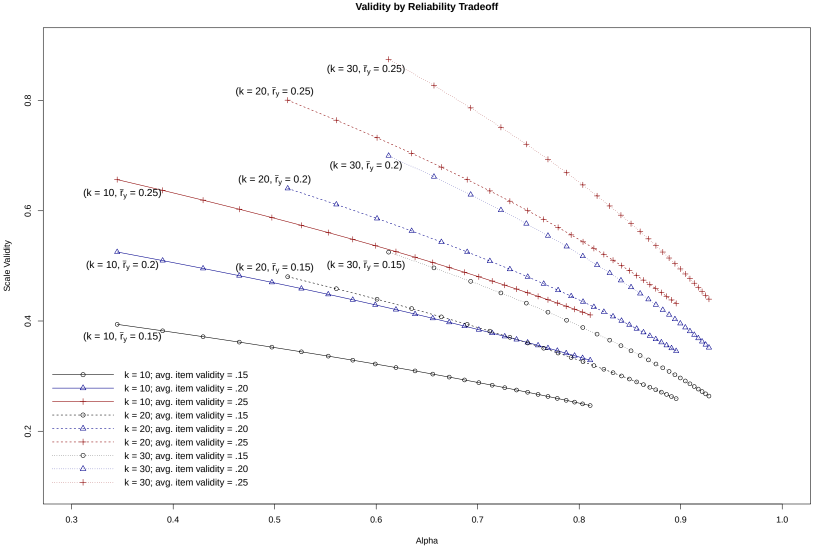
The trade off between test consistency and test validity



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 $.3 < \alpha < .9$

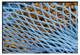
The trade off between test consistency and test validity



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 & Condon, 2018, 2019) and scale construction (Revelle & 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).

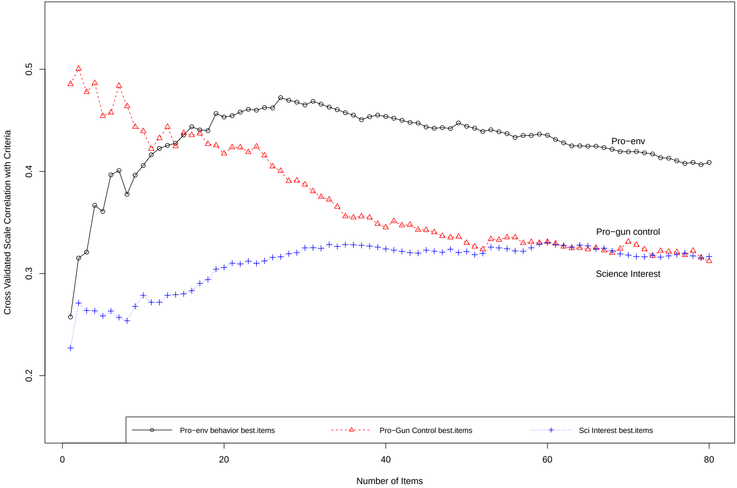
Three examples

1. Aggregating items to predict pro-environmental behaviors (as discussed by Garner (2024).
2. Aggregating items to predict attitudes toward gun control Garner (2024).
3. Aggregating items to predict interest in science Garner (2024).
4. For these three examples, we find scales using items from the SAPA data set (Condon, 2018) which used Massively Missing Completely at Random (MMCAR) data collection with volunteer participants ($N > 200,000$).
5. We compare the predictive validity of cross validated multiple regressions for five broad personality dimensions (SPI-5), 27 narrower facets (SPI-27), and empirically chosen scales from the SPI-135.

Selecting the most valid items

1. Simple dust-bowl empiricism (aka machine learning) allows us to select (and cross validate) those items that best predict a criterion.
2. Using the `bestScales` function from *psych* we found the items that best predicted pro-gun control attitudes, interest in science and pro-environmental behaviors.
3. The cross validated validities achieved their maximum with 2 (gun control), and $\approx$ 30 items for science and the environment.

Predictive validity is a non-monotonic function of number of items



Comparing 5 high level, 27 lower level and best scales solutions

Table: Multiple Rs predicting 3 criteria

Predictors	Gun control	Interest in Science	Green behavior
spi "Big 5" scales	0.21	0.28	0.33
spi 27 facet scales	0.50	0.36	0.48
best5 items	0.61	0.26	0.38
best10 items	0.57	0.27	0.48
best15 items	0.51	0.29	0.48
best20 items	0.50	0.31	0.48

Table: Items that best predict each criteria

SAPA item	Correlation	Content of item
		Attitudes towards Gun Control
q_1825	0.57	Tend to vote for liberal political candidates.
q_1824	-0.46	Tend to vote for conservative political candidates.
q_379	0.26	Believe that people are basically moral.
q_1328	-0.21	Like to stand during the national anthem.
q_4289	0.19	Trust people to mainly tell the truth.
		Interest in Science
q_1392	0.22	Love to think up new ways of doing things.
q_422	0.20	Can handle a lot of information.
q_2745	0.18	Am able to come up with new and different ideas.
q_240	0.18	Am quick to understand things.
q_128	0.16	Am full of ideas.
		Environmental behaviors
q_1825	0.27	Tend to vote for liberal political candidates.
q_348	0.25	Believe in the importance of art.
q_607	-0.22	Do not enjoy going to art museums.
q_1303	0.22	Like to begin new things.
q_1132	0.22	Have read the great literary classics.

Summary and conclusions

1. Predicting behavior is hard.
2. Items have meaningful variance over and beyond what they have in common with other items in factorially pure scales.
3. Forming scales based upon validity coefficients (and then cross validating these scales) leads to higher validities than simple regressions based upon pure factors.
4. We encourage you to fish with broad nets not sharp spears.

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.

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

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

Campbell, D. T. & 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 /sc4p9/.

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.

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.

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

Garner, K. (2024). The forgotten trade-off between internal consistency and validity. *Multivariate Behavioral Research*.

Gough, H. G. (1965). Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology*, *70*(4), 294–302.

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons, Inc.

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

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

Hogan, R. (2009). John Holland.

Hogan, R. & Sherman, R. A. (2020). Personality theory and the nature of human nature. *Personality and Individual Differences*, *152*, 109561.

Kuder, G. & 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. & 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.

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., & 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. (2024). The seductive beauty of latent variable models: Or why i don't believe in the easter bunny. *Personality and Individual Differences*, *221*, 112552.

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. & 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)*.

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.

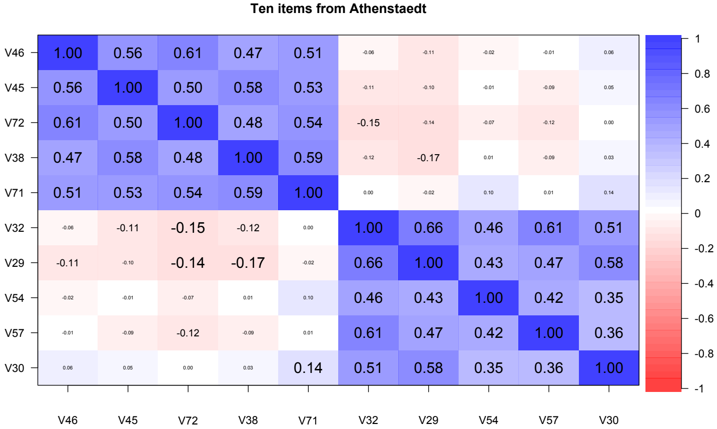
Strong Jr., E. K. (1927). Vocational interest test. *Educational Record*, *8*(2), 107–121.

Zinbarg, R. E., Revelle, W., Yovel, I., & 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.

Other examples

1. Several other examples of the power of aggregating items without focussing on internal consistency have already been published:
2. Revelle et al. (2021) compare Big 5, Little 27 and empiricallly based items for 10 criteria.
3. Revelle (2024) and Eagly & Revelle (2022) demonstrated advantages of aggregation to predict gender

10 items from Athenstaedt (2003)

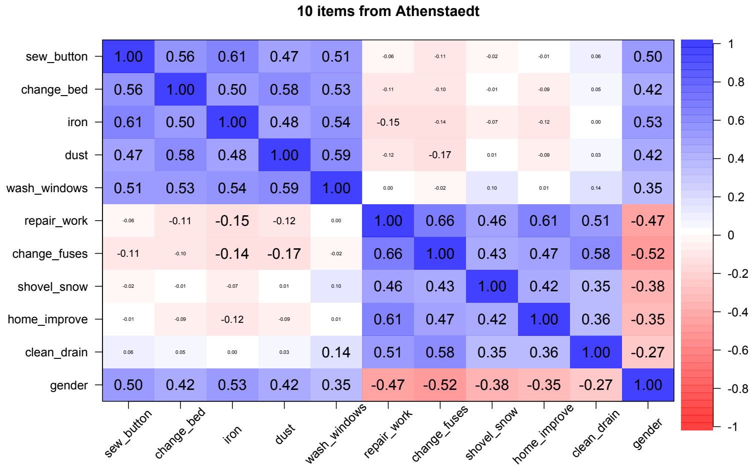


Ten items from Athenstaedt

	V46	V45	V72	V38	V71	V32	V29	V54	V57	V30
V46	1.00	0.56	0.61	0.47	0.51	-0.06	-0.11	-0.02	-0.01	0.06
V45	0.56	1.00	0.50	0.58	0.53	-0.11	-0.10	-0.01	-0.09	0.05
V72	0.61	0.50	1.00	0.48	0.54	-0.15	-0.14	-0.07	-0.12	0.00
V38	0.47	0.58	0.48	1.00	0.59	-0.12	-0.17	0.01	-0.09	0.03
V71	0.51	0.53	0.54	0.59	1.00	0.00	-0.02	0.10	0.01	0.14
V32	-0.06	-0.11	-0.15	-0.12	0.00	1.00	0.66	0.46	0.61	0.51
V29	-0.11	-0.10	-0.14	-0.17	-0.02	0.66	1.00	0.43	0.47	0.58
V54	-0.02	-0.01	-0.07	0.01	0.10	0.46	0.43	1.00	0.42	0.35
V57	-0.01	-0.09	-0.12	-0.09	0.01	0.61	0.47	0.42	1.00	0.36
V30	0.06	0.05	0.00	0.03	0.14	0.51	0.58	0.35	0.36	1.00

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

10 items from Athenstaedt (2003) predict gender



	sew_button	change_bed	iron	dust	wash_windows	repair_work	change_fuses	shovel_snow	home_improve	clean_drain	gender
sew_button	1.00	0.56	0.61	0.47	0.51	-0.06	-0.11	-0.02	-0.01	0.06	0.50
change_bed	0.56	1.00	0.50	0.58	0.53	-0.11	-0.10	-0.01	-0.09	0.05	0.42
iron	0.61	0.50	1.00	0.48	0.54	-0.15	-0.14	-0.07	-0.12	0.00	0.53
dust	0.47	0.58	0.48	1.00	0.59	-0.12	-0.17	0.01	-0.09	0.03	0.42
wash_windows	0.51	0.53	0.54	0.59	1.00	0.00	-0.02	0.10	0.01	0.14	0.35
repair_work	-0.06	-0.11	-0.15	-0.12	0.00	1.00	0.66	0.46	0.61	0.51	-0.47
change_fuses	-0.11	-0.10	-0.14	-0.17	-0.02	0.66	1.00	0.43	0.47	0.58	-0.52
shovel_snow	-0.02	-0.01	-0.07	0.01	0.10	0.46	0.43	1.00	0.42	0.35	-0.38
home_improve	-0.01	-0.09	-0.12	-0.09	0.01	0.61	0.47	0.42	1.00	0.36	-0.35
clean_drain	0.06	0.05	0.00	0.03	0.14	0.51	0.58	0.35	0.36	1.00	-0.27
gender	0.50	0.42	0.53	0.42	0.35	-0.47	-0.52	-0.38	-0.35	-0.27	1.00

Clearly a two factor solution but with some interesting correlations with gender.

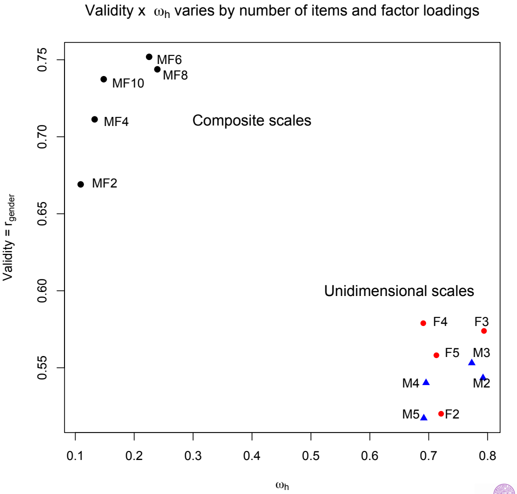
Form various short scales

1. It is easy to form 2 ... 5 item short and factorially pure scales from these items. (F2 ... F5, or M2 ... M5)
2. Equally easy to form 2 .. 10 item composite scales mixing M and F content (MF2 ... MF10)
3. Just M or just F scales are very internally consistent ($\omega_h = .72$ ... $.85$) and reasonably valid ($r_{gender} = .52$ ... $.58$)
4. But the composite (MF) scales are much less internally consistent ($\omega_h = .11$ ... $.23$, $\alpha = .11$ ... $.77$) and more valid ($r_{gender} = .67$ ... $.75$)

Reliability and Validity for Short M, F, and MF scales

Relability and Validity

Scale	ω_h	α	r_{gender}
F2	0.72	0.72	0.52
F3	0.79	0.79	0.57
F4	0.69	0.82	0.58
F5	0.71	0.85	0.56
M2	0.79	0.79	0.54
M3	0.77	0.76	0.55
M4	0.70	0.81	0.54
M5	0.69	0.82	0.52
MF2	0.11	0.11	0.67
MF4	0.13	0.59	0.71
MF6	0.23	0.69	0.75
MF8	0.24	0.75	0.74
MF10	0.15	0.77	0.74


Validity x ωh varies by number of items and factor loadings
Validity = rgender
ωh
MF6
MF8
MF10
MF4
MF2
Composite scales
Unidimensional scales
F4
F3
F5
M3
M4
M2
M5
F2