Dart Boards vs. Fishing Nets: Alternative metaphors for validity

Part of a discussion with Mijke Rhemtulla

SMEP 2023

William Revelle and Kayla Garner
Northwestern University
Evanston, Illinois USA

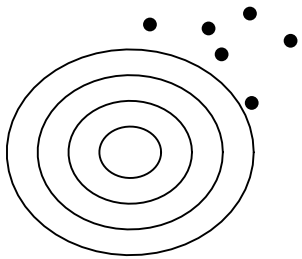
NORTHWESTERN UNIVERSITY

Outline

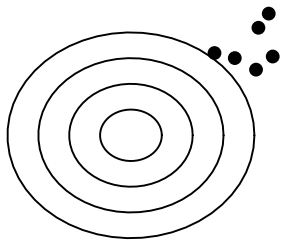
Introduction to the question

1. In a brilliant manuscript which I had the good fortune to review, Mijke Rhemtulla developed the "Dart Board" validity/reliability metaphor.
    - This was based on a strong assumption that validity can be defined as what a factor measures.
    - That is, validity is factorial validity.
    - Reliability is just how well we measure the construct.
    - Validity is the ratio of internal consistency to test-retest reliability.
2. Dartboard validity wants scales to be internally consistent measures of single constructs.
3. Dartboard validity equates validity with how well the test measures a construct.

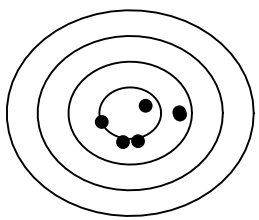
Reliability and Validity as dart throwing



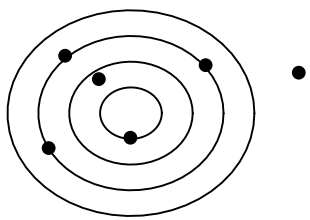
Unreliable and Invalid



Reliable and Invalid



Reliable and Valid



Unreliable but Valid

1. Unfortunately for Mijke, I had just given a keynote address at ISSID entitled "The seductive beauty of latent variables" (Revelle, 2023)
   - That paper was an attack on our beloved application of latent variable models and argued that we should worry more about prediction than factorial homegeneity.
   - I even suggested that to believe in latent variables was akin to believing in the Easter Bunny or the Tooth Fairy.
2. In addition, I had recently published an article with Alice Eagly "Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees" (Eagly & Revelle, 2022) which examined the effect of aggregation on reliability and validity.
   - That paper showed that while aggregation could increase reliability, aggregating unrelated concepts could increase validity.
   - It rediscovered Gulliksen (1950).

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

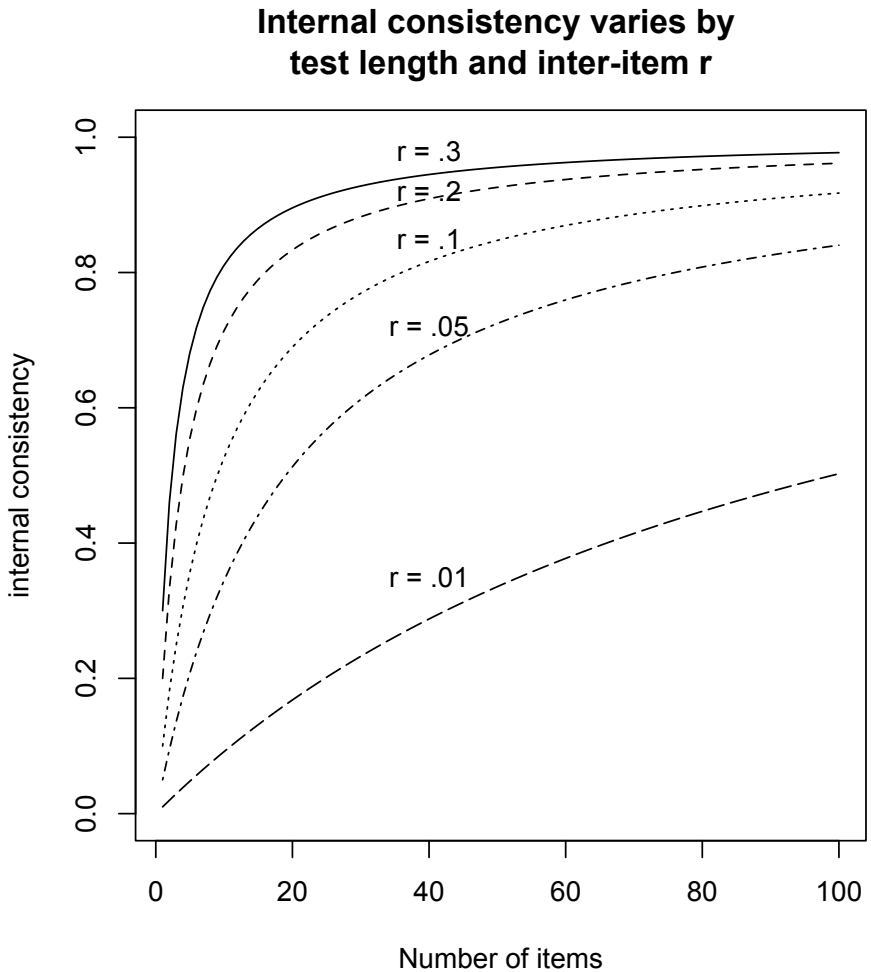
1. Lets unpack these two equations.
   Internal consistency

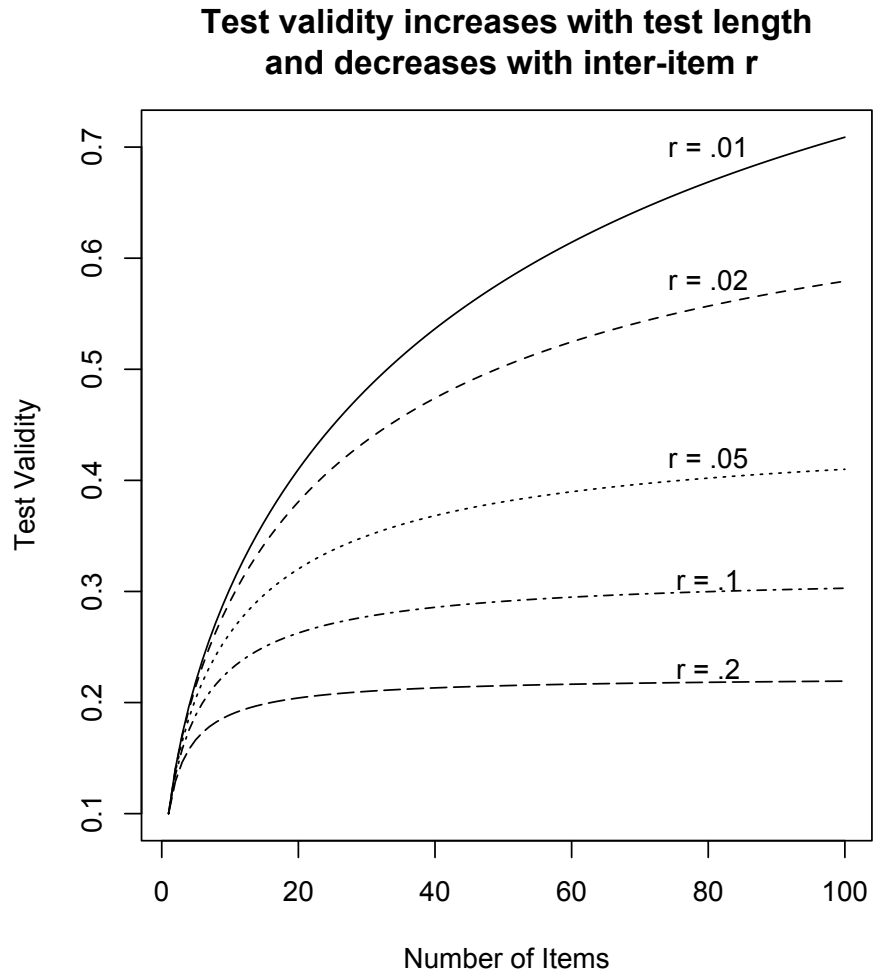
$$\lambda_3 = \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{1}$$

2. but validity

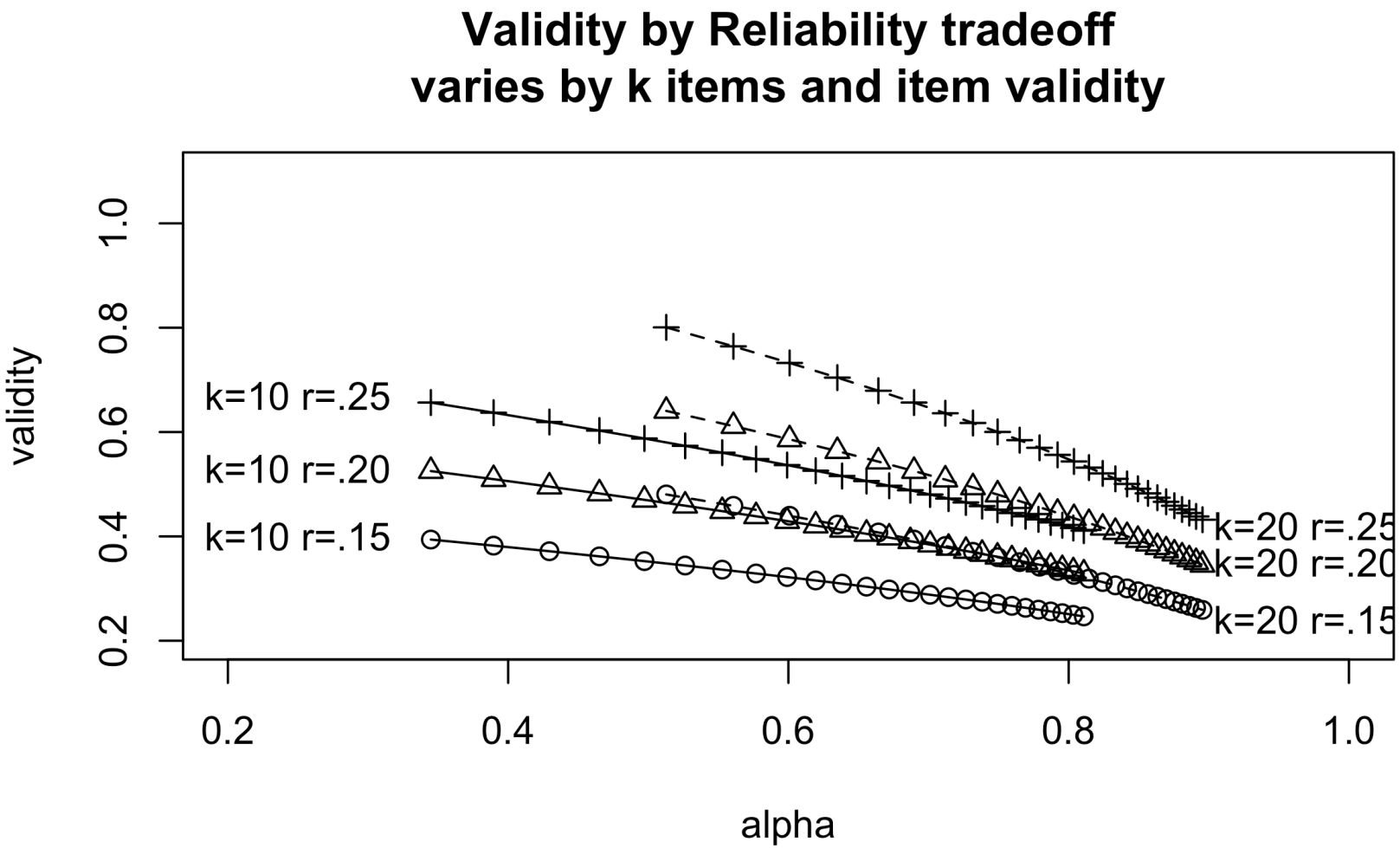
$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k * (k-1)\bar{r}}}. \tag{2}$$

The trade off between test consistency and test validity





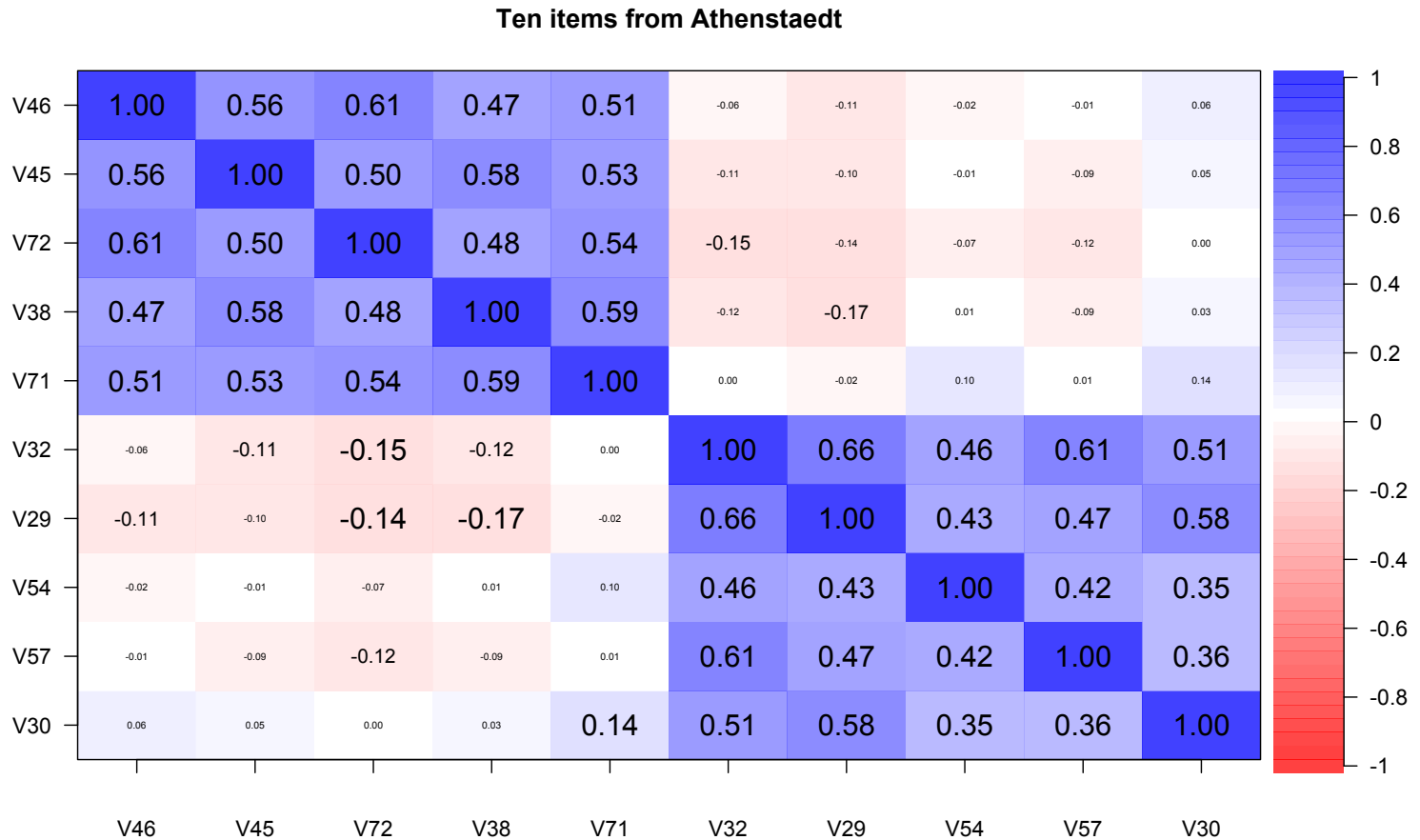
The trade off between test consistency and test validity



Increasing validity implies increasing the diversity of the item content

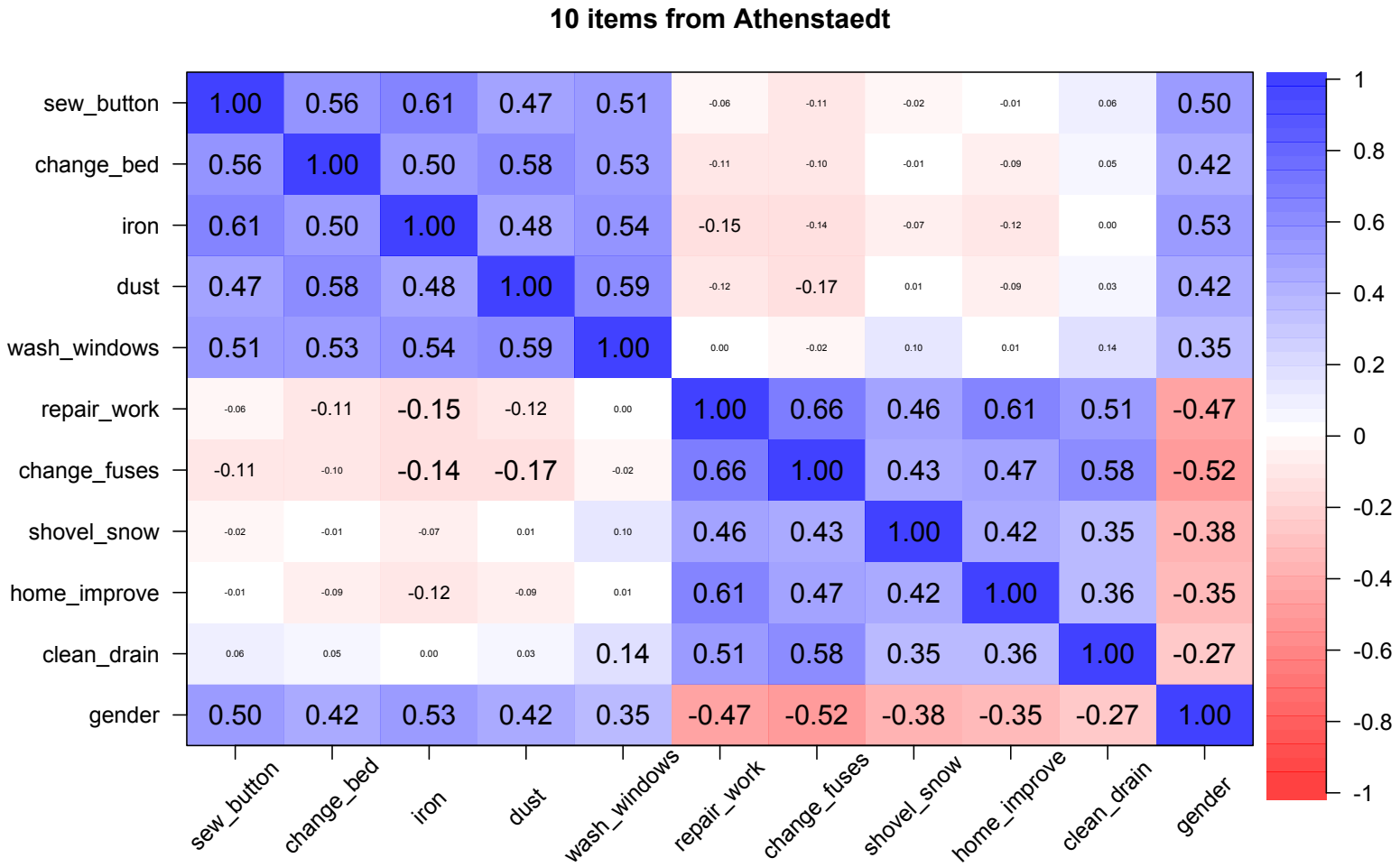
1. The goal of construct validity is have pure measures with high internal consistency. (Measure one thing well).
2. And highly correlated measures of the same constructs.
3. But if the goal is predictive validity, we should minimize internal consistency and have independent predictors.
4. By emphasizing 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. Variations on this theme have been discussed before by (Condon, Wood, Möttus, Booth, Costani, Greiff, Johnson, Lukaszeksi, Murray, Revelle, Wright, Ziegler & Zimmerman, 2021; Möttus, Wood, Condon, Back, Baumert, Costani, Epskamp, Greiff, Johnson, Lukaszeksi, Murray, Revelle, Wright, Yarkoni, Ziegler & Zimmerman, 2020).

10 items from Athenstaedt (2003)



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

10 items from Athenstaedt (2003) predict gender



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

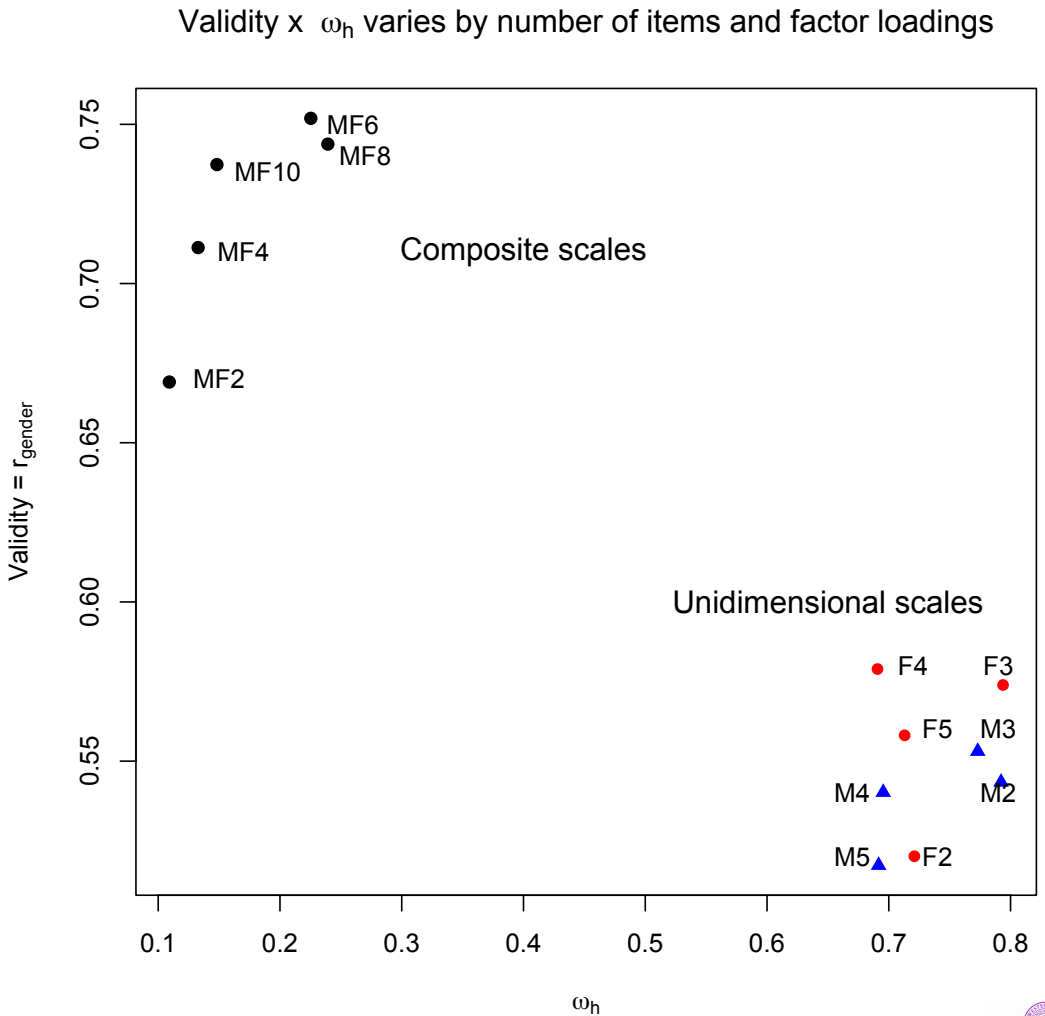
Form various short scales

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



Darts or Fishing Spears versus Fishing Nets

1. The M and F scales are sharper spears (more internally consistent) and have a clear one factor solution.
2. And the mixed composite scales are looser (less internally consistent), less clear construct (multifactorial) and more net like.
3. But Fishing Nets catch more fish (have higher validities) than do Spears.
4. Perhaps it is time to not focus on construct validity or factorial purity but rather on predictive validity.

And now for an alternative opinion

Mijke Rhemtulla

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.

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costani, G., Greiff, S., Johnson, W., Lukaszeksi, 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.

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.

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.

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

Möttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costani, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszeksi, 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. (2023). The seductive beauty of latent variables: ISSID award for distinguished contribution to the study of individual differences. Belfast. International Society for the Study of Individual Differences.

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. & 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)*.

R code

```
library(psych) ; library(psychTools)
select <- cs(V46,V45,V72,V38, V71, V32,V29,V54,V57,V30,gender)
R <- corPlot(Athenstaedt[select])  #from psychTools
#make up a set of scoring keys
keys <- list(
F2 =cs(V46,V45),
F3 = cs(V46,V45,V72),
F4= cs(V46,V45,V72,V38),
F5 = cs(V46,V45,V72,V38, V71),

M2 =  cs(-V32,-V29),
M3 =  cs(-V32,-V29,-V54),
M4 =  cs(-V32,-V29,-V54,-V57),
M5 = cs(-V32,-V29,-V54,-V57,-V30),

MF2 = cs(V46, -V32),
MF4 = cs(V46,V45, -V32, -V29),
MF6= cs(V46,V45,V72, -V32, -V29, -V54),
MF8 =  cs(V46,V45,V72,V38, -V32, -V29, -V54,-V57),
MF10 =  cs(V46,V45,V72,V38, V71, -V32,-V29,-V54,-V57, -V30),
gender=cs(gender)
)
```

```
mf.scores <- scoreOverlap(keys, R)  #find scale validities
mf.om <- reliability(keys,R)  #and reliabilities
mf.df <- data.frame(omega=mf.om$result.df[,1],
     alpha=mf.scores$alpha[1:13],
     valid= mf.scores$cor[14,1:13])

df2latex(mf.df)   #create the table

plot(mf.df[c(1,3)],col=c(rep("red",4),rep("blue",4),
   rep("black",5)),pch=c(rep(16,4),rep(17,4),rep(19,5)),
  main=expression(paste("Validity x  ",
       omega[h]," varies by number of items and factor loadings")),
    xlab =expression(omega[h]), ylab=expression(paste("Validity = "
text(.72,.58,"F4")
text(.79,.58,"F3")
text(.74,.52,"F2")
text(.74,.56,"F5")
text(.79,.54,"M2")
text(.79,.56,"M3")
text(.67,.52,"M5")
text(.67,.54,"M4")

text(.15,.67, "MF2")
```

```
text(.17,.71, "MF4")
text(.26,.75, "MF6")
text(.27,.74, "MF8")
text(.19,.735, "MF10")

text(.4,.71,"Composite scales", cex=1.2)
text(.65,.6,"Unidimensional scales",cex=1.2)
```