

# Psychometrics

## Lecture 7

### Multivariate statistics

Psychology 613 – Spring 2022

*With content by **Rose Maier,**  
**Allison Tackman, and Kathryn Iurino***

# What is (are?) psychometrics?

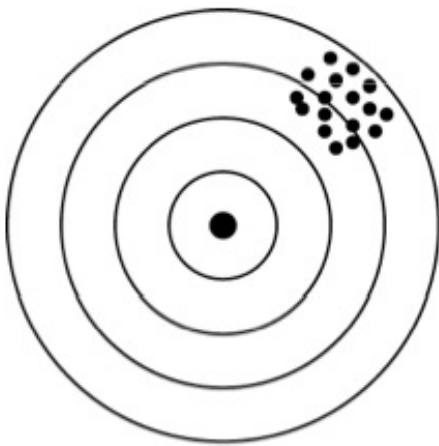
- **How to construct and evaluate measurement methods**
- Suppose you have this idea that there is this thing called “extraversion”, which refers to a person being outgoing, gregarious, assertive, positive, and warm
- How would you develop a measure for it?  
What standards would you use to evaluate a measure of it?

# What makes a measure “good”?

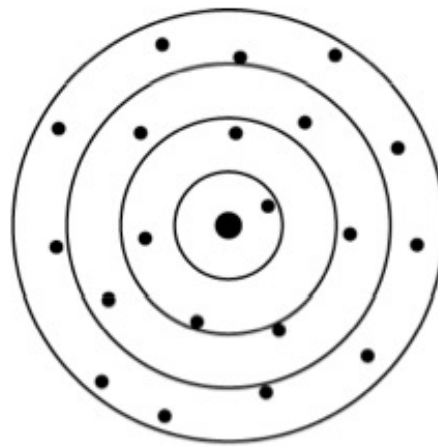
## - Many different criteria...

- **Validity:** measuring what we say we are measuring
- **Reliability:** relative absence of measurement error
- **Unidimensionality:** measure is really measuring just 1 thing
- **Generalizability:** works well across populations and contexts
- **Generativity:** use of measure leads to good rather than bad consequences
- Other standards you can think of? These are just some

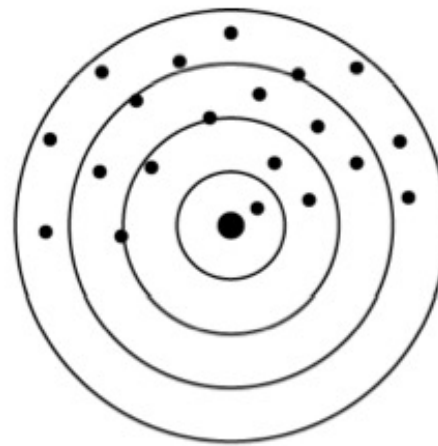
# First, the distinction between reliability and validity



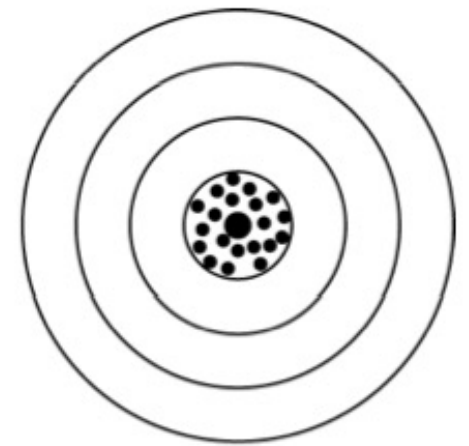
Reliable  
Not valid



Valid  
Not reliable



Not reliable  
Not valid



Reliable  
Valid

# Validity

- How do we determine whether a measure is evaluating what its supposed to and not some other construct?
  - In personality research, no 'gold standard'
  - Evidence for validity comes from diverse sources

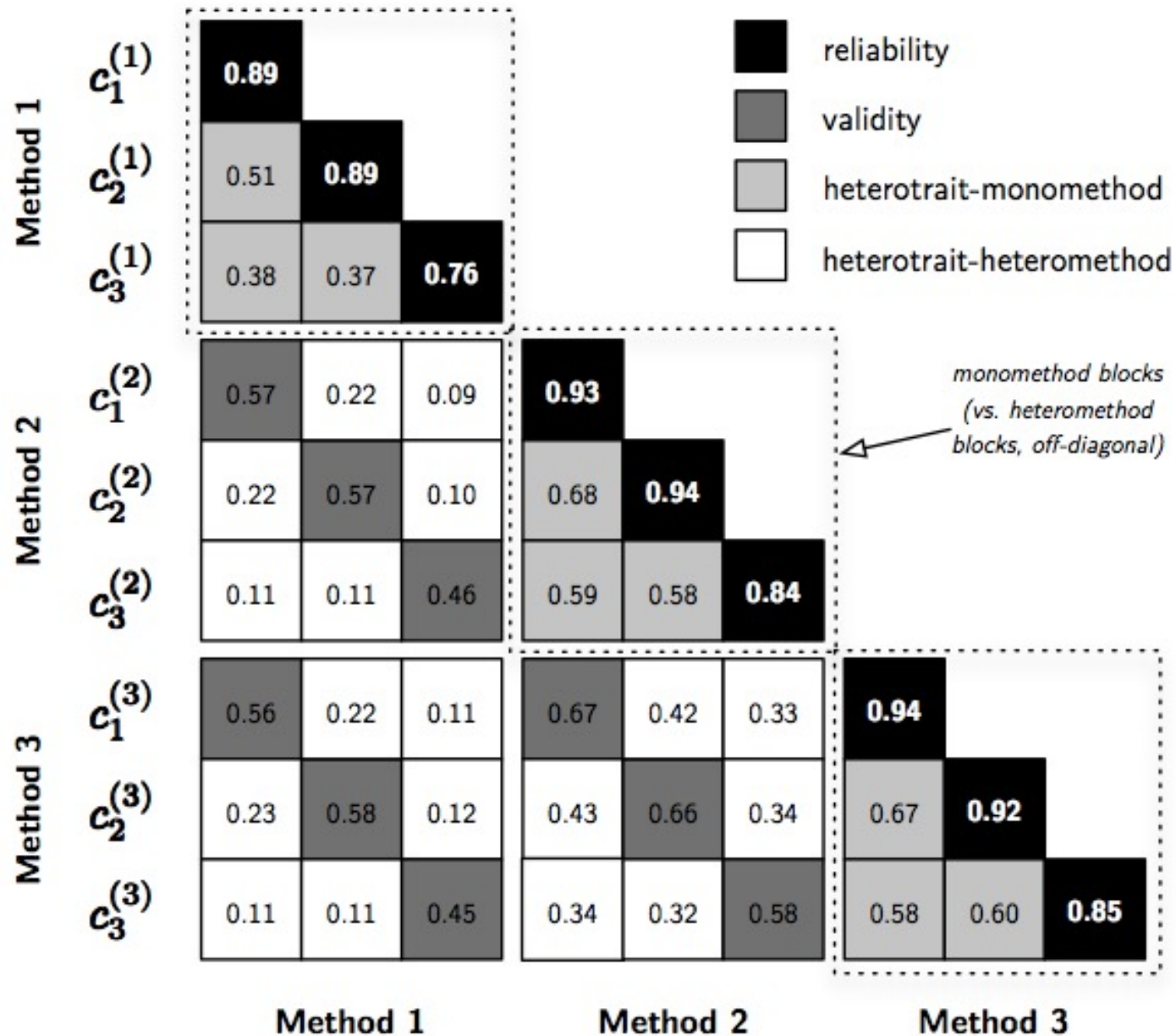
# Some types of validity

- **Content validity:**
  - Evaluate by converging expert opinion, focus groups
  - Are the items making up the measure appropriate, clear and comprehensive? (assumption is that they are randomly selected from a universe of relevant items)
- **Criterion-related Validity:**
  - Evaluate by looking at correlations with other measures (“criteria”)
  - Predictive validity: does it predict outcomes in the future?
  - Convergent/Discriminant Validity: Does the measure relate to constructs theory predicts it should relate to, and NOT relate to constructs theory would predict it shouldn’t relate to?
  - Construct validity: Have alternative explanations for responses on the measure been ruled out?

# Example: convergent/discriminant validity ~ Multi-method multi-trait matrix

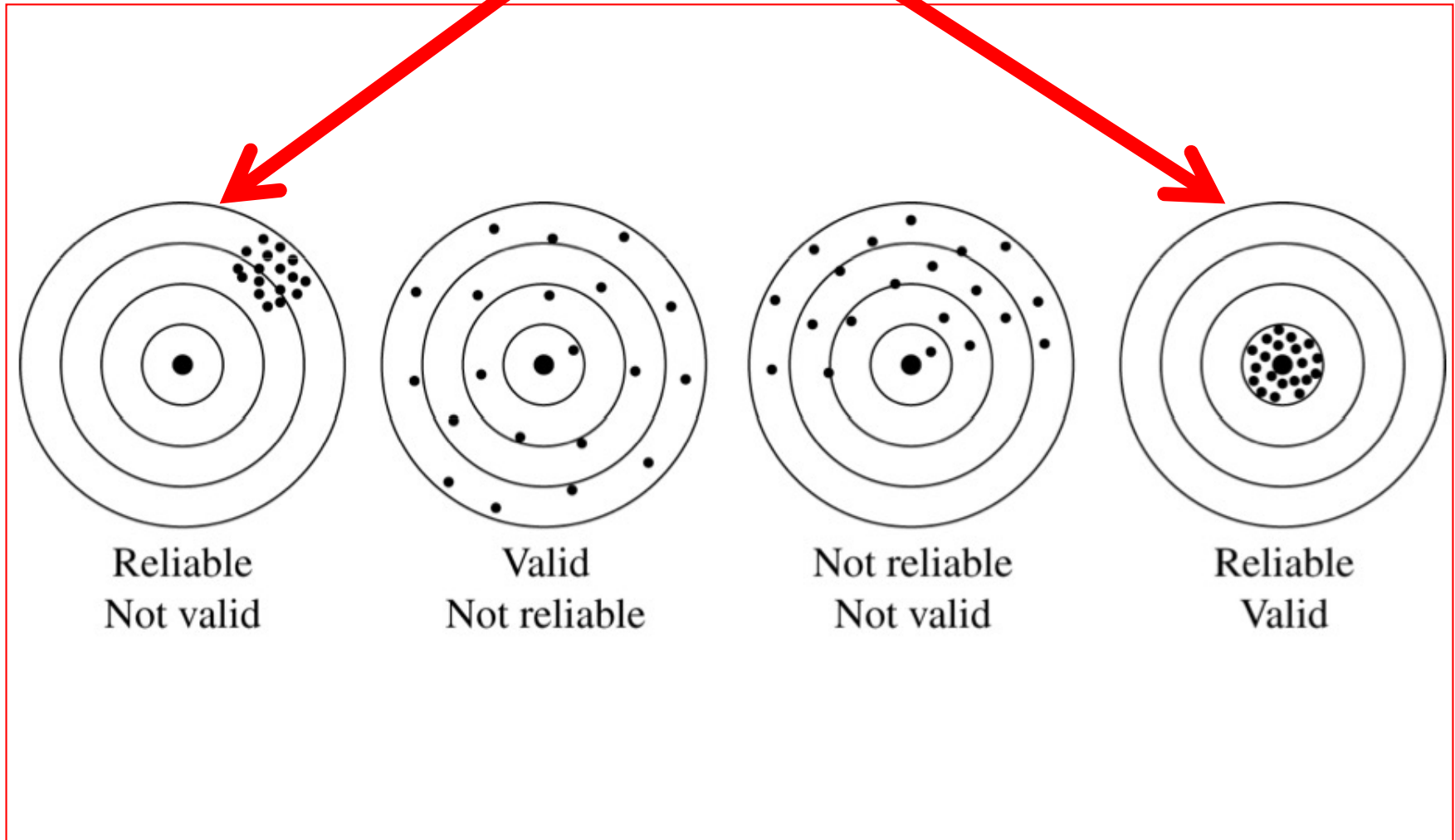
- Let's say we wanted to measure 3 relatively independent personality traits (A, O, E)
- Using 3 different methods (self-report, informant-report, behavioral data)
- Correlation between any pair of observations should be affected by shared variance due to *method*, and *trait*
- Hopefully, shared variance due to trait > shared variance due to method
- We can get better estimates of method variance and trait variance if we have all possible combinations of traits/methods

# Multi-method multi-trait matrix





# Now onto reliability



# Reliability (CTT)

- Relative absence of measurement error
- Observed score=true score variance + error variance [X=T +E]
- Reliability= True score variance/Observed score variance
- Hence, Reliability=

$$\frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

# Types of reliability [ $X=T + E$ ]

- Internal consistency
  - Alpha tells you the error associated with choosing a particular set of items to measure a construct (rather than some other randomly selected set from the universe of eligible items)
  - If you just want to know about how closely items are related can look at mean inter-item correlation
- Test-retest (stability)
  - Tells you the error associated with differences in testing context and time
- Inter-rater/inter-judge
  - Tells you the error associated with differences across judges or observers

# Reliability: Cronbach's alpha

$$\frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Cronbach's alpha:

# of items

Mean of inter-item covariances

$$\alpha = \frac{K\bar{c}}{K\bar{c} + (\bar{v} - \bar{c})}$$

Mean of item variances

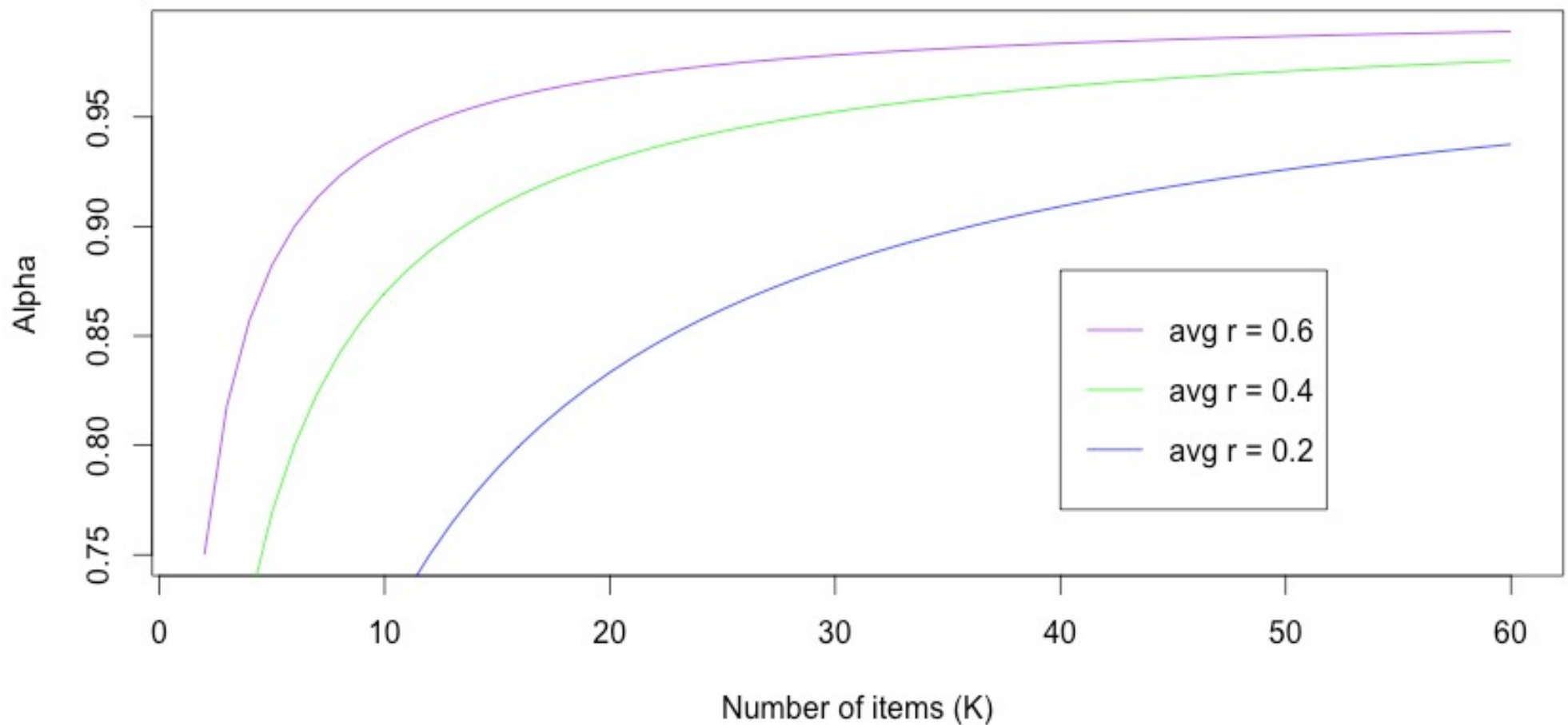
$$\alpha_{\text{standardized}} = \frac{K\bar{r}}{K\bar{r} + (1 - \bar{r})}$$

Mean of inter-item correlations

Rearranging formula for alpha...

$$\alpha_{\text{standardized}} = \frac{K\bar{r}}{(1 + (K - 1)\bar{r})}$$

Alpha as a function of K



# Why does reliability matter?

- Estimates of correlations between your measure and other measures are constrained by reliability
  - Upper limit of correlation with another measure =  $\sqrt{\alpha}$ , i.e., the correlation of that measure with itself
  - If two measures X and Y have reliability of .70 (common benchmark in personality research) true correlation of .70 between X and Y would be reduced to .49; true correlation of .30 would be reduced to .27 (John & Soto, 2007)
  - If X and Y have  $\alpha = .5$ , true correlation of .70 would become .35; true correlation of .3 would become .15 (John & Soto, 2007)
- So creating reliable measures is important for getting accurate estimates of correlations with other variables

# Limitations of relying on Cronbach's alpha exclusively

Doesn't tell you if scale is unidimensional

– For ex, these two scales have the same alpha (from John & Soto, 2007)

Scale B: 6 items, mean interitem correlation = .52,  $\alpha = .87$

	1	2	3	4	5	6
1	—					
2	.52	—				
3	.52	.52	—			
4	.52	.52	.52	—		
5	.52	.52	.52	.52	—	
6	.52	.52	.52	.52	.52	—

Scale C: 6 items, mean interitem correlation = .52,  $\alpha = .87$

	1	2	3	4	5	6
1	—					
2	.70	—				
3	.70	.70	—			
4	.40	.40	.40	—		
5	.40	.40	.40	.70	—	
6	.40	.40	.40	.70	.70	—

→ ALWAYS examine structure of inter-item correlations as well; don't just report alpha!

→ Variance of inter-item correlations is a good proxy for unidimensionality; can also use factor analysis

# Limitations of relying on Cronbach's alpha exclusively

- Can increase Cronbach's alpha by adding redundant items
- This results in a scale that has high inter-item correlations - so high alpha - but limited conceptual breadth and therefore limited predictive validity



# Final notes: Cronbach's alpha

What is considered a good alpha value?

A rule of thumb is .70 or higher, but this varies by field. Ask yourself: how much measurement error am I willing to tolerate?

In summary, keep in mind...

1. Short measures (scales with fewer items) will have lower alphas by definition of the metric.
2. Measures that attempt to cover the broad range of a construct (e.g., extraversion) will have lower alphas – a very high alpha can indicate that your items are redundant!

# Psychometrics in R

## Step 1: Quality (*sanity*) check.

*Are all of your items within the response range?*

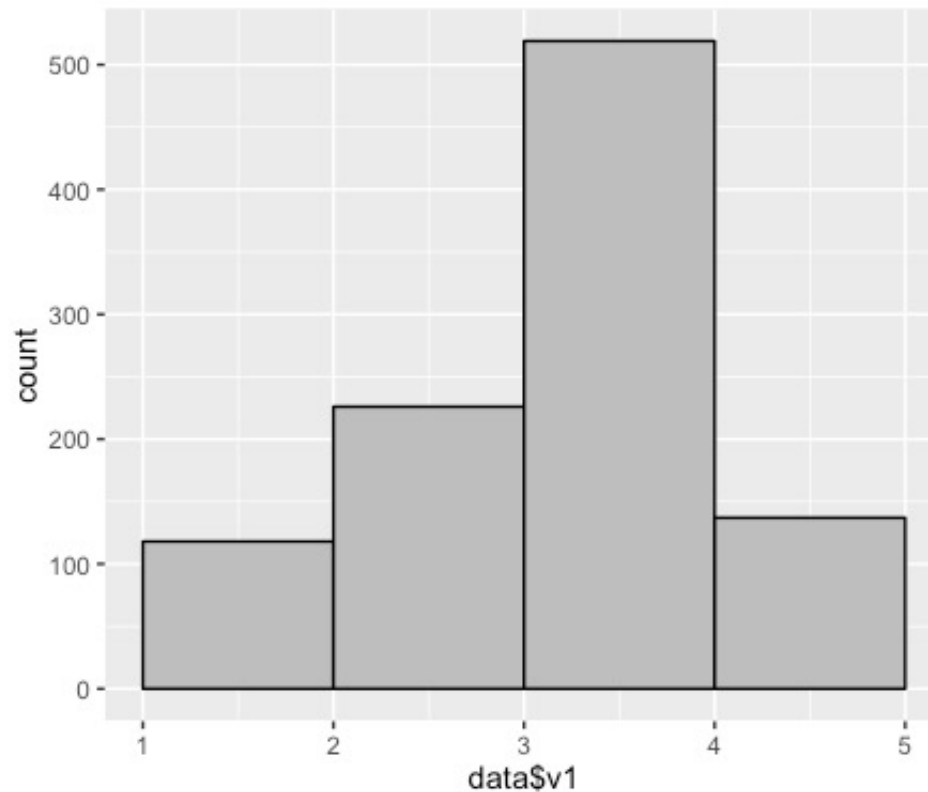
```
> describe(data)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
v1	1	1000	3.66	0.89	4	3.72	0.00	1	5	4	-0.65	0.17	0.03
v2	2	1000	1.78	0.84	2	1.64	1.48	1	5	4	1.22	1.63	0.03
v3	3	1000	4.05	0.79	4	4.13	0.00	1	5	4	-0.92	1.38	0.03
v4	4	1000	1.98	0.95	2	1.85	1.48	1	5	4	0.98	0.68	0.03
v5	5	1000	4.12	0.82	4	4.22	1.48	1	5	4	-1.02	1.31	0.03
v6	6	1000	2.37	1.07	2	2.29	1.48	1	5	4	0.56	-0.45	0.03
v7	7	1000	4.04	0.88	4	4.16	1.48	1	5	4	-1.02	1.05	0.03
v8	8	1000	3.74	0.77	4	3.78	0.00	1	5	4	-0.85	1.01	0.02

# Psychometrics in R

## Step 1: Sanity check.

*Check response category usage in the histogram.*



# Psychometrics in R

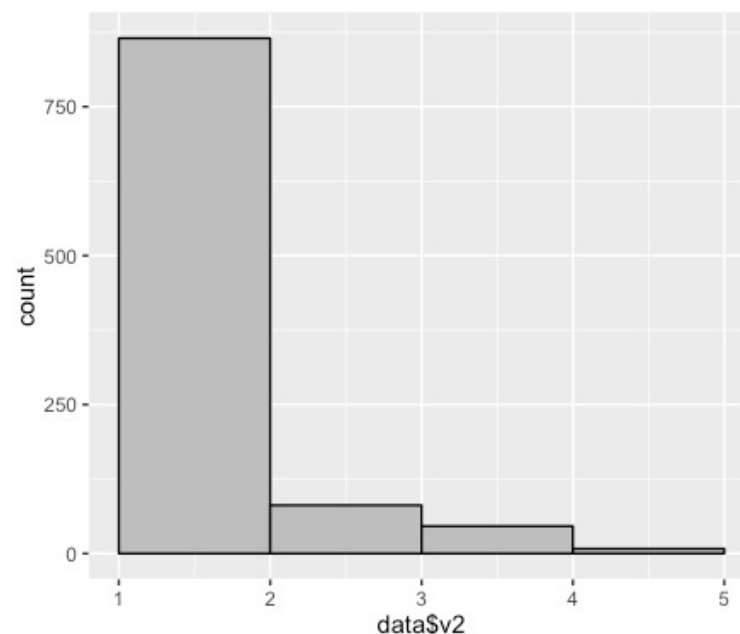
## Step 1: Sanity check.

*Check response category usage by running frequencies.*

*Item 2: “Is reserved”  
...is positively skewed*

```
> table(data$v2)
```

1	2	3	4	5
419	446	81	46	8



# Psychometrics in R

## Step 2: Recode any reverse-keyed/scored variables.

```
keys = c(1,-1,1,-1,1, -1, 1, 1) #reverse the 2nd, 4th, and 6th items  
dataRev <- reverse.code(keys,data,mini=rep(1,8),maxi=rep(5,8))
```

```
> head(data)
```

	v1	v2	v3	v4	v5	v6	v7	v8
1	4	2	5	3	5	2	4	4
2	4	1	4	1	5	2	4	4
3	4	1	4	1	4	2	4	4
4	4	2	2	3	4	3	3	4
5	4	4	4	1	4	2	4	4
6	4	1	3	3	5	2	4	4

```
> head(dataRev)
```

	v1	v2-	v3	v4-	v5	v6-	v7	v8
[1,]	4	4	5	3	5	4	4	4
[2,]	4	5	4	5	5	4	4	4
[3,]	4	5	4	5	4	4	4	4
[4,]	4	4	2	3	4	3	3	4
[5,]	4	2	4	5	4	4	4	4
[6,]	4	5	3	3	5	4	4	4

Handy, huh? The reversed items are indicated in the output with a -

# Psychometrics in R

Step 2: Recode any reverse-keyed/scored variables. *Note: You can often check for reverse scoring by examining the correlation matrix.*

```
> cor(data)
```

	v1	v2	v3	v4	v5	v6	v7	v8
v1	1.0000000	-0.2787045	0.3426767	-0.3122959	0.3117518	-0.2504850	0.1665468	0.1551161
v2	-0.2787045	1.0000000	-0.2466038	0.2833671	-0.5267426	0.2536206	-0.2410860	-0.2031503
v3	0.3426767	-0.2466038	1.0000000	-0.4964671	0.2610490	-0.3377161	0.1877486	0.2048907
v4	-0.3122959	0.2833671	-0.4964671	1.0000000	-0.2972296	0.2842200	-0.2376081	-0.2016817
v5	0.3117518	-0.5267426	0.2610490	-0.2972296	1.0000000	-0.3824939	0.3027439	0.2840431
v6	-0.2504850	0.2536206	-0.3377161	0.2842200	-0.3824939	1.0000000	-0.1120111	-0.2043906
v7	0.1665468	-0.2410860	0.1877486	-0.2376081	0.3027439	-0.1120111	1.0000000	0.3051465
v8	0.1551161	-0.2031503	0.2048907	-0.2016817	0.2840431	-0.2043906	0.3051465	1.0000000



```
> cor(data)
```

	v1	v2	v3	v4	v5	v6	v7	v8
v1	1.0000000	-0.2787045	0.3426767	-0.3122959	0.3117518	-0.2504850	0.1665468	0.1551161
v2	-0.2787045	1.0000000	-0.2466038	0.2833671	-0.5267426	0.2536206	-0.2410860	-0.2031503
v3	0.3426767	-0.2466038	1.0000000	-0.4964671	0.2610490	-0.3377161	0.1877486	0.2048907
v4	-0.3122959	0.2833671	-0.4964671	1.0000000	-0.2972296	0.2842200	-0.2376081	-0.2016817
v5	0.3117518	-0.5267426	0.2610490	-0.2972296	1.0000000	-0.3824939	0.3027439	0.2840431
v6	-0.2504850	0.2536206	-0.3377161	0.2842200	-0.3824939	1.0000000	-0.1120111	-0.2043906
v7	0.1665468	-0.2410860	0.1877486	-0.2376081	0.3027439	-0.1120111	1.0000000	0.3051465
v8	0.1551161	-0.2031503	0.2048907	-0.2016817	0.2840431	-0.2043906	0.3051465	1.0000000

```
> cor(dataRev)
```

	v1	v2-	v3	v4-	v5	v6-	v7	v8
v1	1.0000000	0.2787045	0.3426767	0.3122959	0.3117518	0.2504850	0.1665468	0.1551161
v2-	0.2787045	1.0000000	0.2466038	0.2833671	0.5267426	0.2536206	0.2410860	0.2031503
v3	0.3426767	0.2466038	1.0000000	0.4964671	0.2610490	0.3377161	0.1877486	0.2048907
v4-	0.3122959	0.2833671	0.4964671	1.0000000	0.2972296	0.2842200	0.2376081	0.2016817
v5	0.3117518	0.5267426	0.2610490	0.2972296	1.0000000	0.3824939	0.3027439	0.2840431
v6-	0.2504850	0.2536206	0.3377161	0.2842200	0.3824939	1.0000000	0.1120111	0.2043906
v7	0.1665468	0.2410860	0.1877486	0.2376081	0.3027439	0.1120111	1.0000000	0.3051465
v8	0.1551161	0.2031503	0.2048907	0.2016817	0.2840431	0.2043906	0.3051465	1.0000000

# Psychometrics in R

## Step 3: Create the scale.

### *Option 1: The mean*

```
> BFile <- rowMeans(dataRev) # from the “matrix” package
```



# Psychometrics in R

## Step 3: Create the scale.

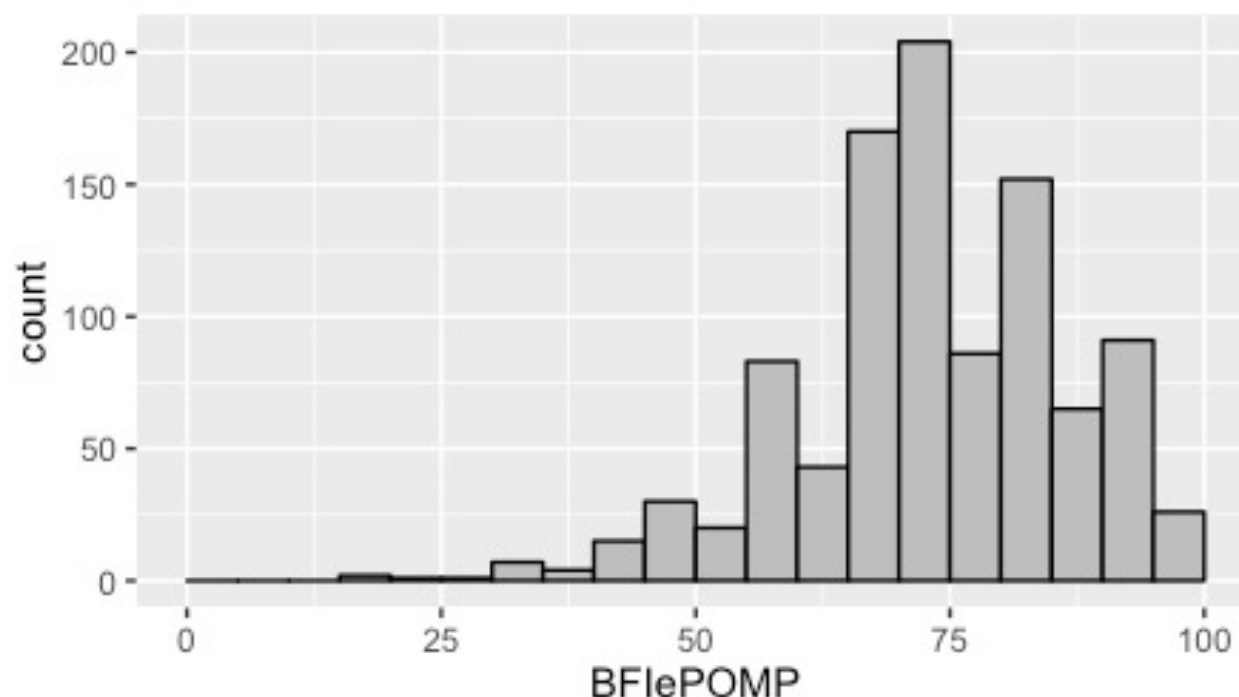
*Option 2: POMP score (percent of maximum possible)*

$\text{BFlePOMP} = 100 * (\text{rowMeans}(\text{dataRev}) - 1) / 4$

```
> BFlePOMP = 100*(rowMeans(dataRev)-1)/4
```

```
> head(BFlePOMP)
```

```
[1] 78.125 84.375 81.250 59.375 71.875 75.000
```



# Psychometrics in R

## Step 4: Actually run the reliability analyses!

*Our main question: How well do these items measure the psychological construct (extraversion)?*

```
library(psych)  
alpha(data)
```

```
> alpha(data)
```

Some items ( v2 v4 v6 ) were negatively correlated with the total scale and probably should be reversed.  
To do this, run the function again with the 'check.keys=TRUE' option

Reliability analysis

Call: alpha(x = data)

*It checks keying for you!!*

# Psychometrics in R

```
library(psych)  
alpha(dataRev)
```

## Reliability analysis

Call: `alpha(x = dataRev)`

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd
0.75	0.75	0.75	0.27	3	0.018	3.9	0.53

lower	alpha	upper	95% confidence boundaries
0.71	0.75	0.78	

This is the primary value to report in a research paper as the reliability of the raw scale score.


The mean raw score, or grand mean

$$\alpha = \frac{K\bar{c}}{(\bar{v} + (K - 1)\bar{c})}$$

$$\alpha_{\text{standardized}} = \frac{K\bar{r}}{(1 + (K - 1)\bar{r})}$$

# R Output


```
Reliability if an item is dropped:
      raw_alpha std.alpha G6(smc) average_r S/N alpha se
v1         0.73      0.73    0.72      0.28 2.7   0.020
v2-        0.72      0.72    0.71      0.27 2.6   0.020
v3         0.71      0.72    0.70      0.27 2.5   0.020
v4-        0.71      0.72    0.70      0.26 2.5   0.020
v5         0.70      0.70    0.69      0.25 2.4   0.021
v6-        0.73      0.73    0.72      0.28 2.7   0.019
v7         0.74      0.74    0.73      0.29 2.9   0.019
v8         0.74      0.74    0.73      0.29 2.9   0.019
```



What the reliability of the questionnaire would be if the item were excluded from the questionnaire

# R Output

The correlation between the score on the individual item and the mean of the scores on the remaining items. For example, the correlation between v7 and the mean of the scores on the remaining items is .34



Item statistics							
	n	raw.r	std.r	r.cor	r.drop	mean	sd
v1	1000	0.59	0.58	0.48	0.42	3.7	0.89
v2-	1000	0.62	0.63	0.56	0.47	4.2	0.84
v3	1000	0.63	0.64	0.57	0.50	4.1	0.79
v4-	1000	0.65	0.64	0.58	0.49	4.0	0.95
v5	1000	0.69	0.70	0.66	0.57	4.1	0.82
v6-	1000	0.62	0.58	0.49	0.42	3.6	1.07
v7	1000	0.52	0.53	0.41	0.34	4.0	0.88
v8	1000	0.51	0.53	0.41	0.35	3.7	0.77

# Another type of Reliability: Inter-rater Reliability

-The question is similar: How similar are these different measurements? But now the error in measurement is attributable to differences across raters, rather than to differences across sets of items

*Note: You can use Cronbach's alpha for interrater reliability, but remember that it only captures covariance, regardless of scale. With interrater reliability, we often care about similarity of **magnitude** of scores (i.e. scale matters), and Cronbach's alpha can't help there.*

# Interrater Reliability in R

Raters treated as “items” in  
ICC analysis: 1 per column

Subjects  
being rated  
on the rows

	SID	R1	R2	R3	R4	R5	R6	R7	R8
1	1	4	4	4	4	4	4	4	3
2	2	2	2	2	1	2	2	1	1
3	3	3	3	3	2	2	3	2	2
4	4	0	0	0	0	0	0	0	0
5	5	0	0	0	0	0	0	0	0
6	6	1	1	1	1	1	2	1	1
7	7	1	2	2	1	1	0	2	1
8	8	1	1	1	1	0	0	1	0
9	9	0	1	0	0	0	0	0	0
10	10	3	2	2	2	99	2	3	3
11	11	3	3	3	2	3	3	3	3
12	12	0	1	1	1	1	1	1	0
13	13	2	2	2	2	2	3	2	2
14	14	1	1	1	1	1	1	1	1



# Interrater Reliability in R

```
data <- read.spss("Lecture 7_psychometrics_extraversionICC.sav", to.data.frame=TRUE)
data <- data[,2:9]
library(psych)
ICC(data)
```

```
> ICC(data)
```

```
Call: ICC(x = data)
```

Intraclass correlation coefficients

	type	ICC	F	df1	df2	p	lower bound	upper bound
Single_raters_absolute	ICC1	0.42	6.7	39	280	0	0.29	0.56
Single_random_raters	ICC2	0.45	27.6	39	273	0	0.21	0.65
Single_fixed_raters	ICC3	0.77	27.6	39	273	0	0.68	0.85
Average_raters_absolute	ICC1k	0.85	6.7	39	280	0	0.77	0.91
Average_random_raters	ICC2k	0.87	27.6	39	273	0	0.68	0.94
Average_fixed_raters	ICC3k	0.96	27.6	39	273	0	0.94	0.98

Number of subjects = 40

Number of Judges = 8

The one you want is single measures, and random raters.



# Interrater Reliability in R

## Why the “single measures” line?

The single measures ICC is the amount of variance that is due to between-subject variance, relative to total variance (variance between raters plus variance between subjects).

Higher numbers indicate that more observed variance is attributed to the subjects than to error caused by differences in the judges' ratings.

If all of your raters agreed perfectly, the ICC would be 1.

# Interrater Reliability in R

## Comparing ICC and Cronbach's alpha:

Intraclass correlation coefficients

	type	ICC	F	df1	df2	p	lower bound	upper bound
Single_raters_absolute	ICC1	0.42	6.7	39	280	0	0.29	0.56
Single_random_raters	ICC2	0.45	27.6	39	273	0	0.21	0.65

Reliability analysis

Call: alpha(x = data)

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd
0.96	0.98	0.98	0.86	49	0.039	2.4	1.4

lower alpha	upper	95% confidence boundaries
0.89	0.96	1.04

R1	R2	R3	R4	R5	R6	R7	R8
3	0	0	0	0	0	1	2
3	0	1	0	0	0	0	2
4	1	1	2	0	0	1	2
3	1	1	2	1	1	1	2
4	0	1	2	1	0	0	4
4	1	1	2	1	1	0	2
4	1	0	2	0	1	1	4
4	1	1	2	1	1	1	2

Raters seem to agree about the ranking of subjects (subj 1 and 2 are lower than subj 3), but they're clearly using the scale differently. This is reflected in the ICC, but not in Cronbach's alpha.

# Response Biases: Stuff you should know

Threats to the reliability and  
validity of questionnaires  
*(From Allison Tackman)*

# Some Common Response Biases

1. *Acquiescence (“yea or nay-saying”)*
2. *Social desirability (“faking good”)*
3. *Malingering (“faking bad”)*
4. *Carelessness or random responding*

# Acquiescence

I agree a lot	I agree a little	I neither agree nor disagree	I disagree a little	I disagree a lot
1	2	3	4	5

## *LOT (Life Orientation Test) Questionnaire*

- 1. In uncertain times, I usually expect the best*
- 2. I'm always optimistic about my future*
- 3. Overall, I expect more good things to happen to me than bad*

## *Sense of Power Questionnaire*

- 1. I can get others to do what I want*
- 2. I think I have a great deal of power*
- 3. If I want to, I get to make the decisions*

# Acquiescence

Person	Acquiescence?	LOT Score	SOP Score
Fozzy	Y	4.7	4.6
Kermit	N	3.3	2.7
Rolph	Y	4.7	4.9
Sweetums	N	3.3	4.7
Gonzo	N	1.6	3.7
Piggy	N	3.4	4.9

$$r = .48$$

# Acquiescence

Person	Acquiescence?	LOT Score	SOP Score
Kermit	N	3.3	2.7
Sweetums	N	3.3	4.7
Gonzo	N	1.6	3.7
Piggy	N	3.4	4.9

$$r = .23$$

# Social Desirability

Agree

Disagree

## *BPI (Berkeley Puppet Interview)*

- 1. I have lots of friends at school*
- 2. My parents' fights are about me*
- 3. I am a smart boy/girl*
- 4. I'm lonely a lot*
- 5. Kids say mean things to me*

Even to a 5-year-old, it's pretty clear what the "right" answer is to each question...



# Social Desirability



*BPI (Berkeley Puppet Interview)*

*Iggy: I have lots of friends at school.*

*Ziggy: I don't have lots of friends at school. How about you?*

# Tips for Avoiding/Detecting Response Bias

- Acquiescence

- Include both + keyed and – keyed items

I agree a lot 1	I agree a little 2	I neither agree nor disagree 3	I disagree a little 4	I disagree a lot 5
--------------------	-----------------------	--------------------------------------	--------------------------	-----------------------

## *LOT (Life Orientation Test) Questionnaire*

- 1. In uncertain times, I usually expect the best*
- 2. I'm always optimistic about my future*
- 3. Overall, I expect more good things to happen to me than bad*

I agree a lot 1	I agree a little 2	I neither agree nor disagree 3	I disagree a little 4	I disagree a lot 5
--------------------	-----------------------	--------------------------------------	--------------------------	-----------------------

## *LOT (Life Orientation Test) Questionnaire*

- 1. In uncertain times, I usually expect the best*
- 2. If something can go wrong for me, it will*
- 3. I'm always optimistic about my future*
- 4. I hardly ever expect things to go my way*
- 5. I rarely count on good things happening to me*
- 6. Overall, I expect more good things to happen to me than bad*

# Tips for Avoiding/Detecting Response Biases

- Social Desirability/Malingering
  - Make responses anonymous
  - Can use special scales to detect (see for example the BIDR from Paulhus)
  - Don't use overly evaluative items if possible

I agree a lot 1	I agree a little 2	I neither agree nor disagree 3	I disagree a little 4	I disagree a lot 5
--------------------	-----------------------	--------------------------------------	--------------------------	-----------------------

## LOT (Life Orientation Test) Questionnaire

*Please be as honest and accurate as you can throughout. Try not to let your response to one statement influence your responses to other statements. There are no "correct" or "incorrect" answers. Answer according to your own feelings, rather than how you think "most people" would answer.*

*1. In uncertain times, I usually expect the best*

*2. It's easy for me to relax*

*3. If something can go wrong for me, it will*

*4. I'm always optimistic about my future*

*5. I enjoy my friends a lot*

*6. It's important for me to keep busy*

*7. I hardly ever expect things to go my way*

*8. I don't get upset too easily*

*9. I rarely count on good things happening to me*

*10. Overall, I expect more good things to happen to me than bad*

# Tips for Avoiding/Detecting Response Biases

- Carelessness/Random responding
  - Minimize participant fatigue
  - Write items in clear and straightforward manner
  - Detect and eliminate random responders by dendrogram analysis