Measures, Reliability, & Validity

Psych 251

10/11/17

Outline

- **Reliability & Validity**
  - Concepts
  - Examples
- **Measures**
  - Types of measures
  - Statistics and measure choice
  - Variables in R

Ground rules of experimental psych

- Interested in psychological behavior, phenomenon, or construct
  - Sources of the construct
  - Effects of other factors
  - Individual variability
- Create an **instrument** that **operationalizes** the construct
  - Usually easy to measure quantitatively
  - Takes behavior out of the real world, brings into the lab where it can be manipulated
- Make argument about how to connect this instrument to the construct of interest

Concepts

- **Reliability**: how well did we measure (w/ the instrument)?
  - What is the error in the measurement?
  - What part of it is due to observers, items individuals, noise?
- **Validity**: how well does this measurement represent the construct?
  - **Face** validity: does it look like the construct?
  - **Internal** validity: are you *actually* measuring something related to the construct?
  - **External** validity: does your measure relate to other measures of the construct?

Reliability

- **Test-retest reliability**
  - Do the same thing again later
  - Perhaps with a marginally different set of "items" operationalizing the same concept?
  - Pearson correlation
- **Inter-observer reliability**
  - Does a different person looking at the same behavior code the same thing?
  - Cohen's kappa
- **Intra-item reliability**
  - Do two different questions group together across participants?
  - Cronbach's alpha

Internal validity

- **Internal validity**
  - How well a study was run
  - how confidently you can conclude effects were produced solely by IV
  - "Was it really the treatment that caused the difference between the subjects in the control and experimental groups?”
  - Experimenter expectancy effects

Experimenter expectancy effects

TABLE 7.8
Expectancy effects in eight areas

Research area	Proportion of results that reached $p < .05$ in the predicted direction	Mean effect size in Cohen's d	Mean effect size in Pearson r
Lab interviews	.38	0.14	.07
Reaction time	.22	0.17	.08
Learning and ability	.29	0.54	.26
Person perception	.27	0.55	.27
Inkblot tests	.44	0.84	.39
Everyday situations	.40	0.88	.40
Psychophysical judgments	.43	1.05	.46
Animal learning	.73	1.73	.65
Median	.39	0.70	.33

How many of the predicted results were significant (meta analysis from 1970s)? How much of this is a correct theory vs. experiment expectancy?

Rosenthal & Rosnow (1978)

Issues in internal validity

- **Stimulus specificity**
  - Is effect general across items?
- **Order effects**
  - Subjects become tired and bored, more or less motivated
- **Testing effects**
  - A pretest can affect subjects' performance on a post-test
- **Selection**
  - Subjects in comparison (e.g., the control and experimental) groups should be functionally equivalent at the beginning of a study.
- **Experimental Mortality/Attrition**
  - If one group has higher dropout than others, may bias selection

Campbell & Stanley (1966)

External validity: some issues

- Is there a relationship between your measure and real-life behaviors that should be controlled/affected by your construct?
  - **Concurrent** validity: measure correlates with some other important measure
  - **Predictive** validity: measure predicts to an outcome

Campbell & Stanley (1966)

Construct Validity

1. A construct is defined implicitly by a network of associations or propositions in which it occurs.
2. Construct validation is possible only when some of the statements in the network lead to predicted relations among observables.
3. The network defining the construct, and the derivation leading to the predicted observation, must be reasonably explicit.
4. Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention.
5. When a predicted relation fails to occur, the fault may lie in the proposed interpretation of the test or in the network.
6. Construct validity cannot generally be expressed in the form of a single simple coefficient.
7. Constructs can vary in nature from those very close to "pure description" to highly theoretical constructs involving hypothesized entities and processes, or making identifications with constructs of other sciences.
8. The investigation of a test's construct validity is not essentially different from the general scientific procedures for developing and confirming theories.

Cronbach & Meehl (1955)

Case study 1: Parent report of child vocab

MacArthur-Bates CDI
Words and Sentences

PART I EARLY WORDS

A. FIRST SIGNS OF UNDERSTANDING

Before children begin to speak, they show signs of understanding language by responding to familiar words and phrases. Below are some common examples. Does your child do any of these?

	Yes	No
1. Respond when name is called (e.g., by turning and looking at source).	●	○
2. Respond to "no no" (by stopping what he/she is doing, at least for a moment).	●	○
3. React to "there's mommy/daddy" by looking around for them.	●	○

B. PHRASES (28)

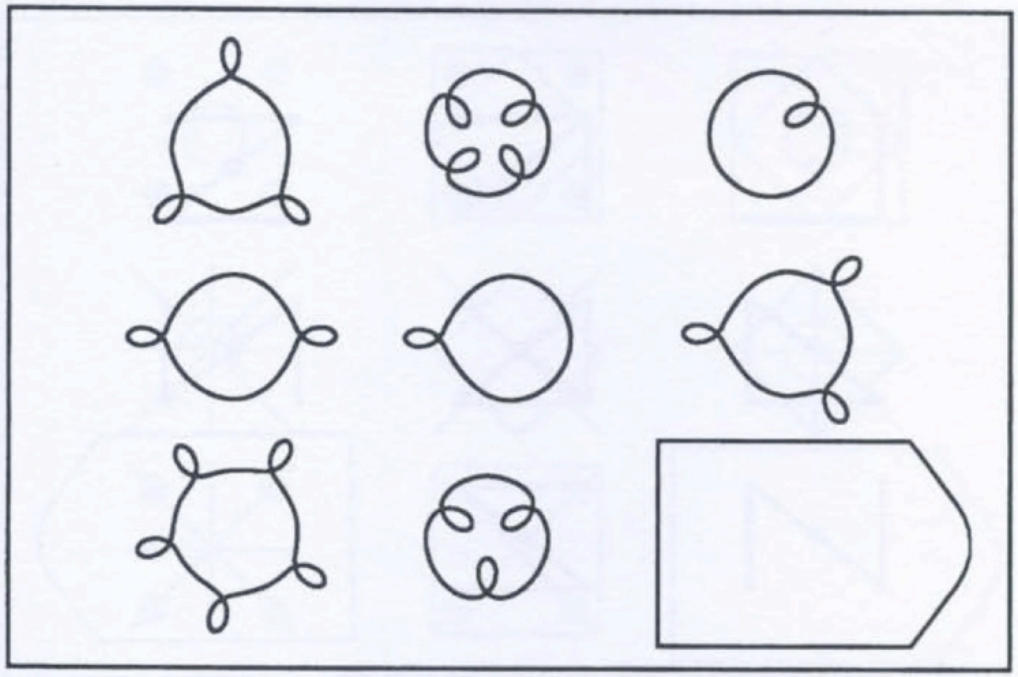
In the list below, please mark the phrases that your child seems to understand.

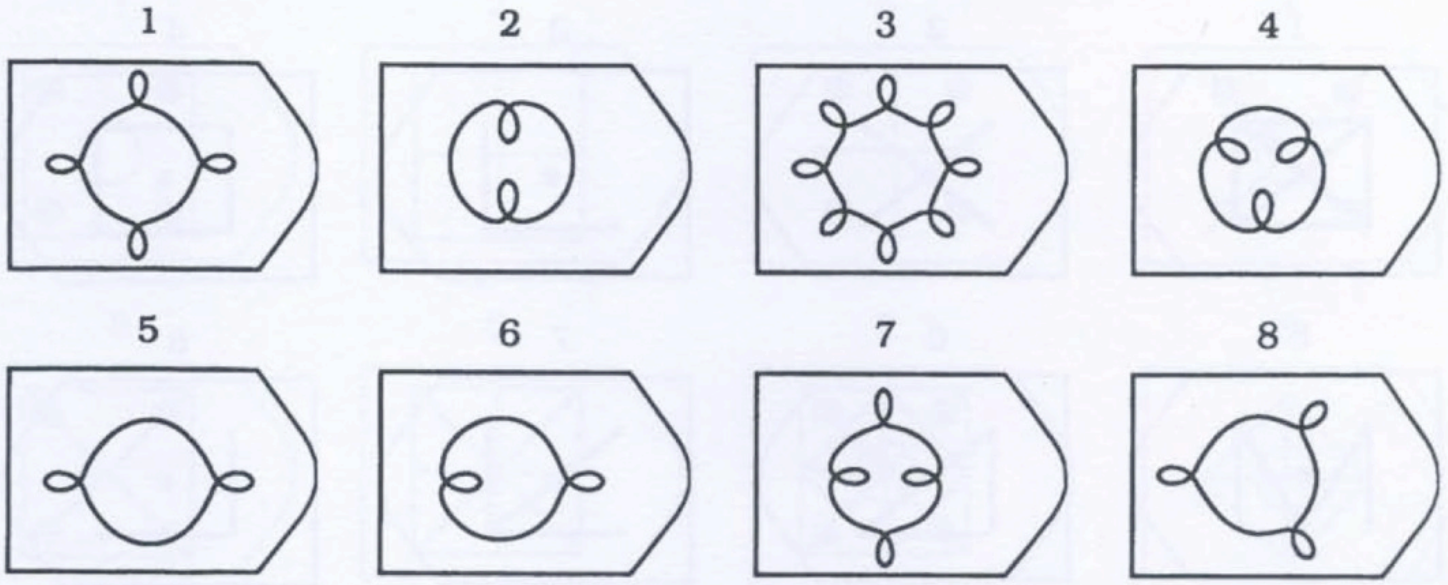
	understands		understands		understands
Are you hungry?	○	Don't touch.	○	Open your mouth.	○
Are you tired/sleepy?	○	Get up.	○	Sit down.	○
Be careful.	○	Give it to mommy.	○	Spit it out.	○
Be quiet.	○	Give me a hug.	○	Stop it.	○
Clap your hands.	○	Give me a kiss.	○	Time to go night night.	○
Change diaper.	○	Go get ____.	○	Throw the ball.	○
Come here/come on.	○	Good girl/boy.	○	This little piggy.	○
Daddy's/mommy's home.	●	Hold still.	○	Want to go for a ride?	○
Do you want more?	●	Let's go bye bye.	○		
Don't do that.	●	Look/look here.	○		

Case study 2: IQ

"There is nothing as important about an individual as his IQ."

Louis Terman




1
2
3
4
5
6
7
8

Is Raven's reliable?

- 576 veterans took Raven's
- Split half correlation between scores = .93
  - So double length would be .96
  - Via Spearman-Brown formula

TABLE 1
RAVEN PROGRESSIVE MATRICES (9): RELIABILITIES AND MEASURES OF CENTRAL TENDENCY FOR AGE RANGES

Ages	*N*	*M*	*Mdn*[a]	r_1I	$r_{21}I$
16-25	103	46	44	.83	.89
26-35	107	41	42	.93	.96
36-45	217	37	38	.94	.96
46-55	69	37	33	.94	.96
56-65	8	37	27	.95	.97
Total	567	40		.93	.96

[a] Median scores for ages 20, 30, 40, 50, 60, given by Raven (9, p. 15) in his table of norms.

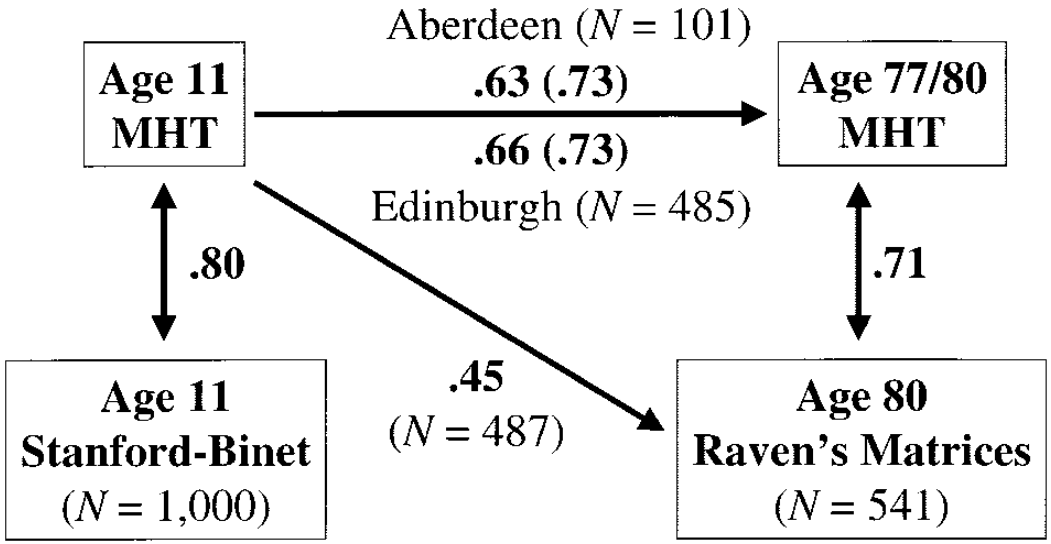
Burke (1972)

Is Raven's internally valid?

- **Confounds** in administration or in design?
  - Unwanted factors affecting results
  - (E.g. item and order effects in other types of experiment)?
- **Demand characteristics**?
  - Unusual for difficult performance-based measures
  - But potentially common for measures of opinion, reaction, etc.

Are IQ measures externally valid?

- Do they measure something outside of themselves?
  - Correlate with other IQ tests
  - Predict life outcomes



Deary et al. (2004)

Sites for IQ critique

- External validity
  - valid across cultures?
- Internal validity
  - compromised by bias against particular populations
  - stereotype threat?
  - may tap multiple constructs at once?
- Etc.

Outline

- **Reliability & Validity**
  - Concepts
  - Examples
- **Measures**
  - Types of measures
  - Statistics and measure choice
  - Variables in R

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

- Nominal: distinct types
- Ordinal: ordered types
- Interval: quantitative
- Ratio: quantitative with a zero point

Why be careful about scales?

- Only some analyses are applicable to some scales
- You wouldn't want to take the mean of an ordinal variable
- Ex: 1: Elementary school education
  2: High school graduate
  3: Some college
  4: College graduate
  5: Graduate degree
- Doesn't make sense to take the mean
  - Midpoints are not interpretable
- Instead you can take the median

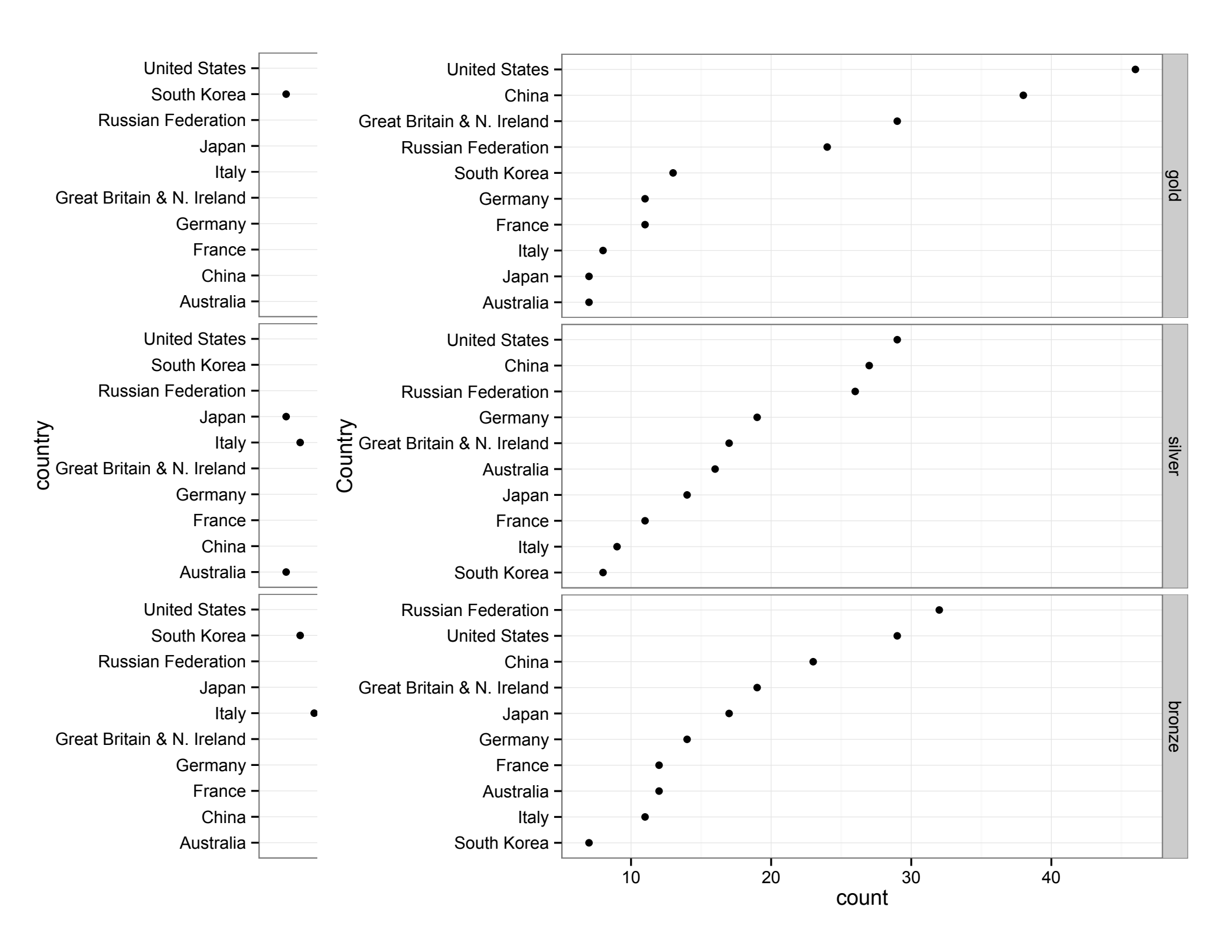
Nominal variables

Sets of things

- Football player numbers
- Subject numbers
- Flavors of ice cream
- Therapy, Drugs, Drugs+Therapy, Control
- Friends, Romans, Countrymen

Nominal variables: Statistics

- How do you find the central tendency?
  - Mode
- How do you test for differences in distribution?
  - Chi-square tests are convenient


United States
South Korea
Russian Federation
Japan
Italy
Great Britain & N. Ireland
Germany
France
China
Australia
country
Country
gold
silver
bronze
10
20
30
40
count

Ordinal variables

Arbitrary orderings

- Rankings (SES, competition results)
- Likert scales (e.g. from 1-7)
  - We'll come back to this

Ordinal (continued)

Likert scales

About ME... A PRE-TYPE III ASSESSMENT SURVEY

Grade: 4 5 6 (circle one)

Sex: Male Female (circle one)

This is a checklist to find out more about you. Some of the sentences describe you better than others. Read each sentence and indicate how much it is like you by putting an **X** in the box that best describes you. In the example below, the person indicated the sentence was **Seldom** like him or her. There are no right or wrong answers. Your answers will be kept secret. Remember to mark one box for each sentence.

	Never like me	**Seldom** like me	**About half of the time** like me	**Usually** like me	**Always** like me
Example: I have taught myself a lot of different things.		X			
1. **I know when someone is happy.**					
2. **When I believe I am right, I am not afraid to let people know what I think.**					

Common Likert Scales

Strongly Agree
Agree
Undecided
Disagree
Strongly Disagree

Very Frequently
Frequently
Occasionally
Rarely
Very Rarely
Never

Extremely Poor
Below Average
Average
Above Average
Excellent

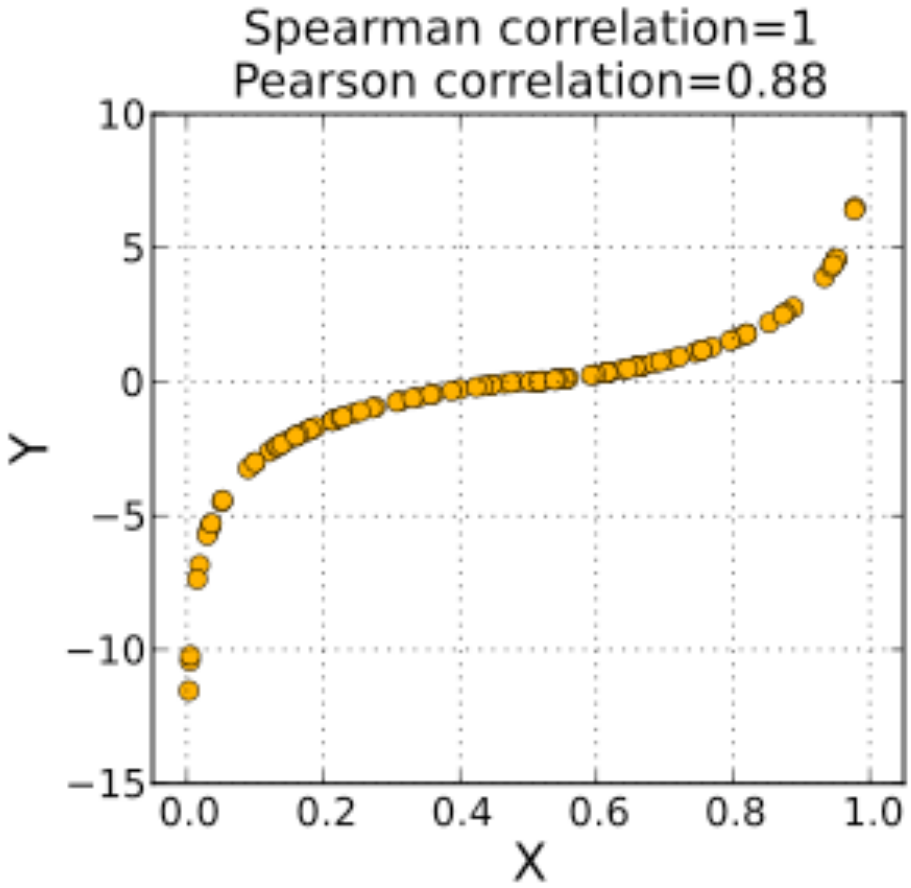
Almost Always True
Usually True
Often True
Occasionally True
Sometimes But Infrequently True
Usually Not True
Almost Never True

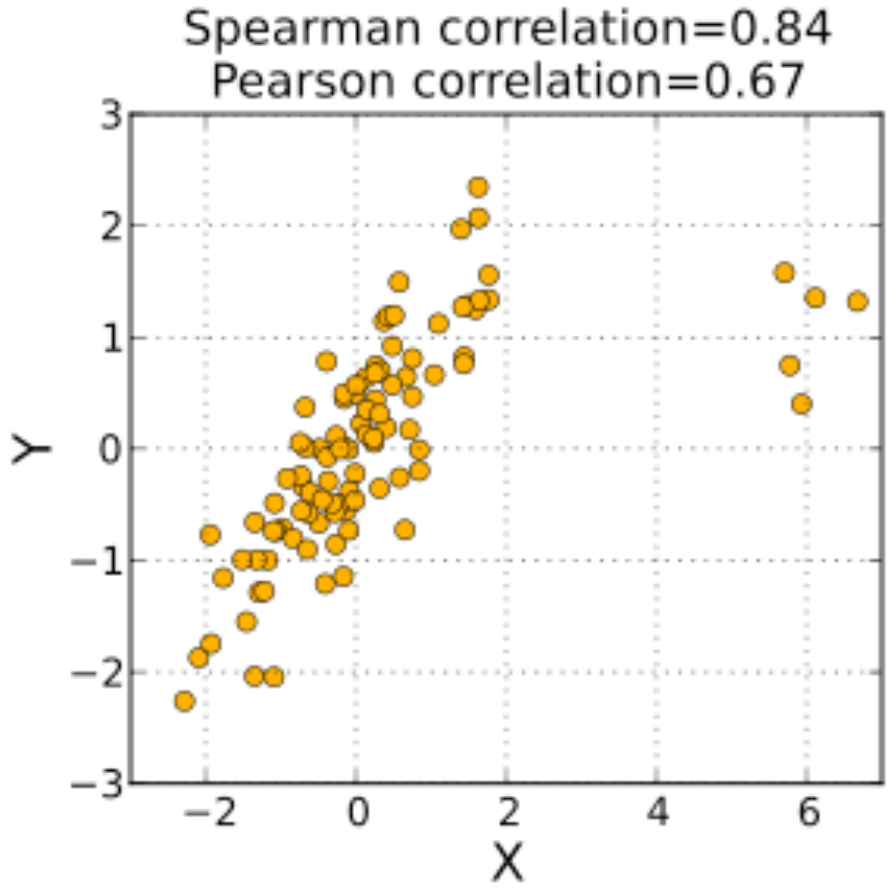
Ordinal variables

- How do you find the central tendency?
  - Median
- How do you measure relationships between ordinal variables?
  - Spearman correlation
- How do you test for differences in distribution?
  - Well… It depends.
  - Definitely non-parametrics, but also (maybe) standard stats for Likert scales?

Correlation on ordinal scales

Spearman correlation measures correlation in ranks





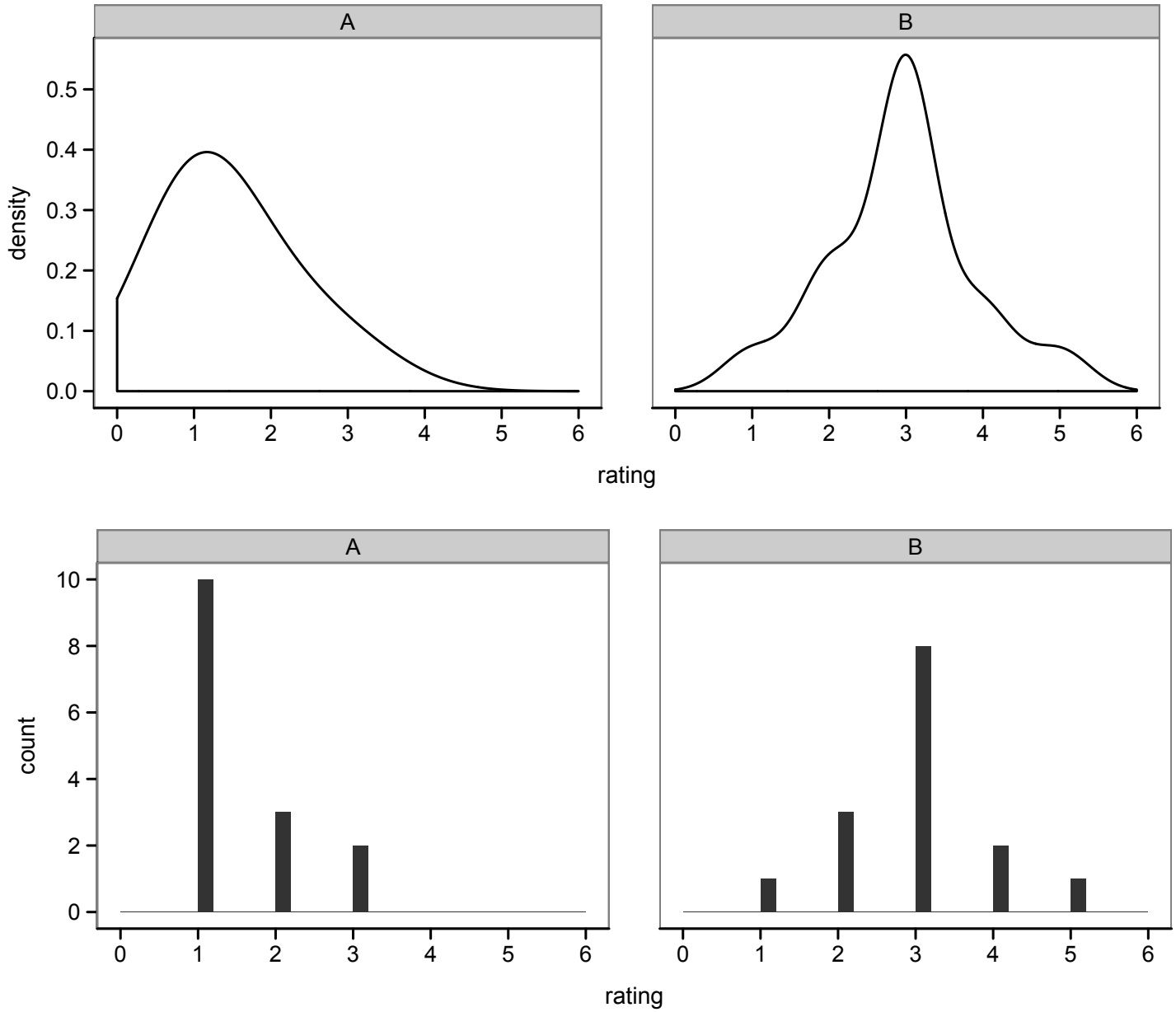
Non-parametric statistics

- General family of statistical tests appropriate for ordinal data
- Assume ordering but not interval
- Also useful for cases in which *data* violate linearity
  - Make fewer distributional assumptions
  - Less power
- Examples (all named after dudes)
  - Mann-Whitney/Wilcoxson (t-test)
  - Kruskal-Wallis (ANOVA)

The case of Likert scales

- Common to assign numeric levels:
  - E.g. strongly agree = 1, strongly disagree = 5
- Can you average these levels?
  - This is a question about whether 1 is as far from 2 as 2 is from 3
  - Not clear, though common practice is to do so
- **One rule of thumb is that stuff is more interval around the middle but gets weird around the edges**

Problems with averaging Likerts



Interval variables

Equal increments but no true zero

- Celsius temperature
- Measuring people
  - Intelligence
  - Fitness
  - BMI
- In the social sciences, we often don't care about zero

Ratio variables

Intervals with a true zero

- Temperature in Kelvin
- Reaction time (sometimes)
- Numbers of things
- ...

Interval and ratio variables: Stats

- How do you find the central tendency?
  - Mean
- How do you measure relationships between ordinal variables?
  - Pearson correlation
- How do you test for differences in distribution?
  - Regression, ANOVA, etc.
- Bonus: What else do you get for ratio variables?
  - Answer: logarithms!

Stevens (1946) classification

Scale Type	Permissible Statistics	Admissible Scale Transformation	Mathematical structure
nominal / categorical	mode, Chi-squared	substitution	unordered
ordinal	median, percentile	monotonic increasing	totally ordered set
interval	mean, standard deviation, correlation, regression, analysis of variance	positive linear	affine line
ratio	above plus geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	one-dimensional vector space

Measure choice considerations

- Spectrum: number of alternatives
  - 2AFC – Many AFC – Free response
- Information content
  - More alternatives = more information
- Task demands
  - Fewer alternatives = faster, easier for kids/impaired populations, automatic responses
  - Fewer alternatives better for RT
- Ease of analysis
  - E.g. free response hard to analyze!
- Diagnosticity
  - More alternatives is more difficult
  - Too easy or difficult and you can't see differences between participants (floor/ceiling effects)

Information content

TABLE 5.3
Minimum number of items required to establish individual judge's accuracy at various levels of statistical significance

Number of alternatives	Chance level	Significance levels (one-tailed)				
		.10	.05	.01	.005	.001
2	.50	4	5	7	9	10
3	.33	3	3	5	5	7
4	.25	2	3	4	4	5
5	.20	2	2	3	4	5
6	.17	2	2	3	3	4
7	.14	2	2	3	3	4
8	.12	2	2	3	3	4
9	.11	2	2	3	3	4
10	.10	1	2	2	3	3
11	.09	1	2	2	3	3
12	.08	1	2	2	3	3

Rosenthal & Rosnow

Variables in R

- Numbers
  - Generalize to arrays
  - `x <- 1` or `x <- c(1,2,3,4)`
  - `x[1]` or `x[1:10]`
- Characters
  - `x <- “hello”` or `x <- c(“hi”,”hey”,”hello”)`
- Factors:
  - nominal or ordinal variable type
  - `abc <- factor(c("x","y","x","z"))`
  - `[1] x y x z`
  - `Levels: x y z`
  - `ordered(ses, levels = c("low", "middle", "high"))`
  - `Levels: low < middle < high`

In R (continued)

- Nominal and ordinal variables are both factors
  - The levels have an ordering, but this ordering can be arbitrary or fixed
  - Up to you to make sure it's reasonable
  - `levels(x) <- c("agree","disagree","neutral")`
  - `levels(x) <- c("disagree","neutral","agree")`
  - Make sure you're not renaming variables though!
- Interval and ratio variables are both numbers
  - It's up to you to decide whether zero is meaningful
- But bad things can happen...
  - E.g. mean of a factor?