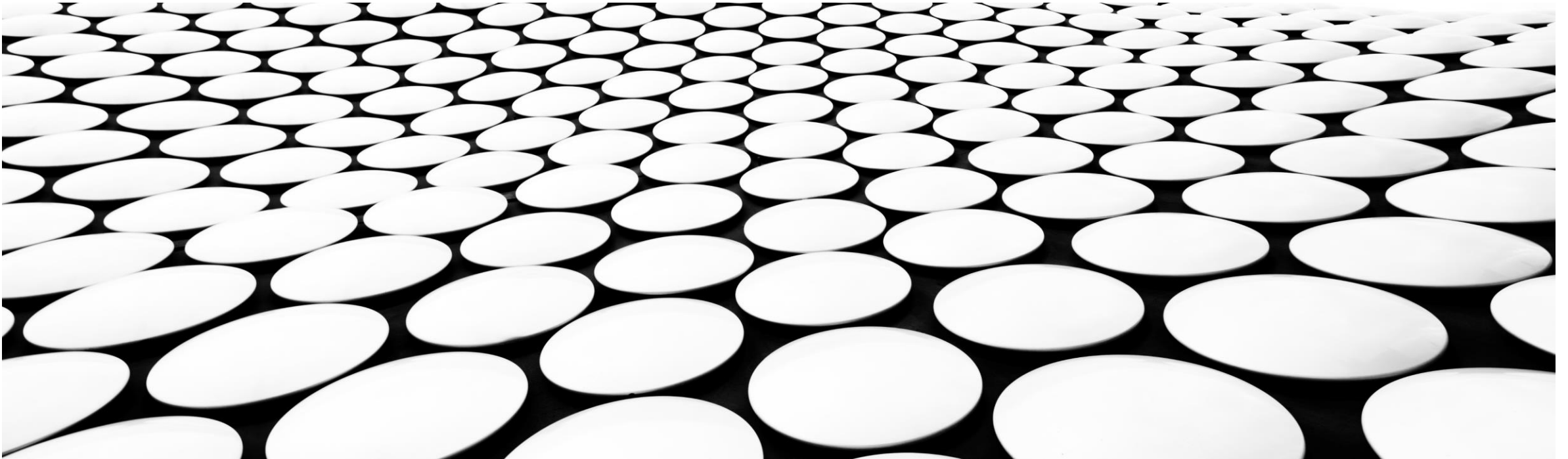


---

# MEASUREMENT NOT SCHMEASUREMENT

WINTER SCHOOL 4MS



# WELCOME! THE OPENING PLAN

- Introductions
- Goals and plan for the day
- Digging in...
  - What is measurement?



<https://128.pl/zR83s>

# INTRODUCTIONS, WHO AM I?

- [Jessica Kay Flake, PhD](#)
  - BS Psychology
  - MA Quantitative Psychology
  - PhD Measurement
  - Post-doc Quantitative and Educational Psychology
  - Associate Prof in Quantitative Psychology
- [Research on measurement development, use, and practice](#)
  - Latent variable modeling
    - More recently focusing on analysis planning
  - Instrument development
  - Open science and large-scale collaboration





# INTRODUCTIONS, WHO ARE YOU?

- What is your name?
- What is a key construct you measure in your research?

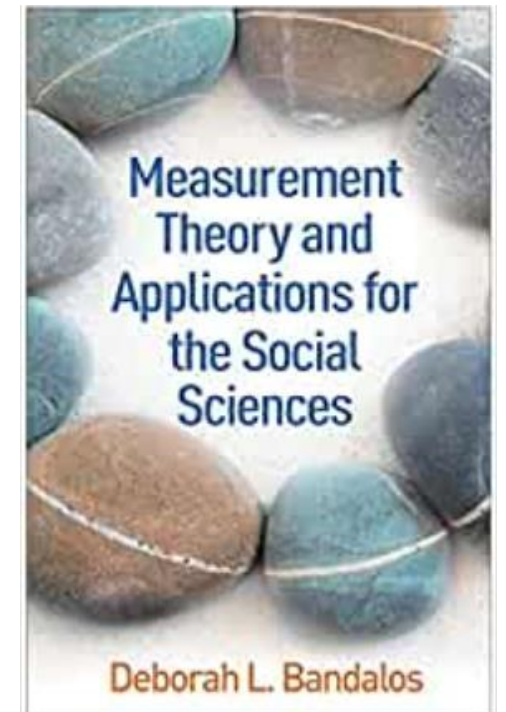


## WHAT ARE THE GOALS OF TODAY?

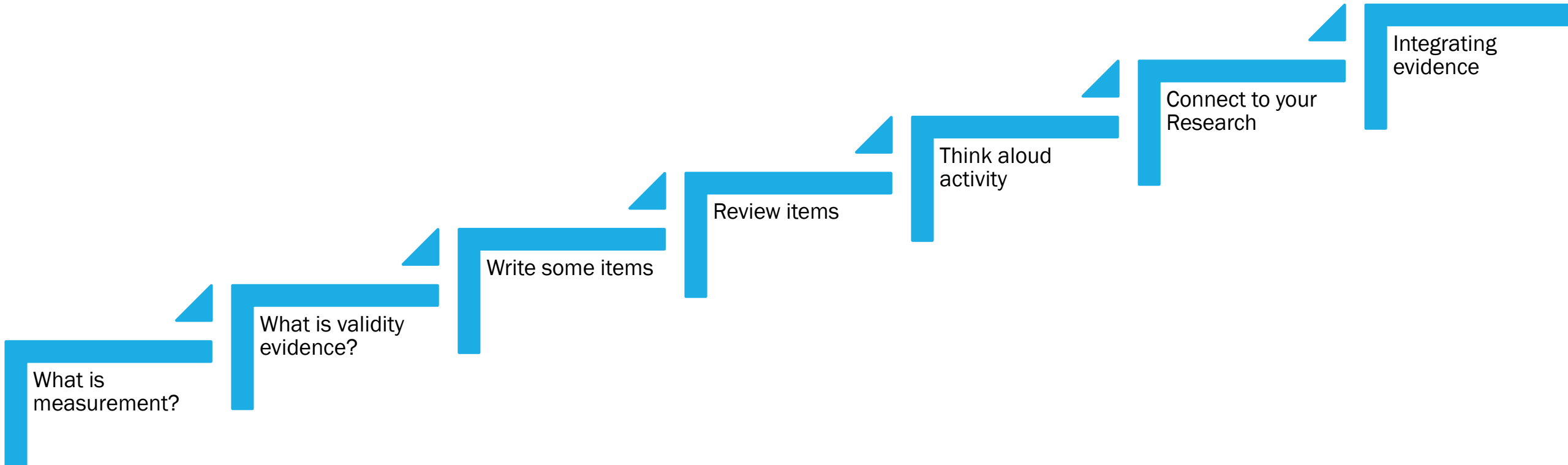
- **Be able to define construct validity and describe different sources of validity evidence**
- **List examples of methodologies for different sources of validity evidence**
- **Evaluate scale items for poor, confusing, or problematic wording**
- **List qualitative approaches to review item content**
- Integrate qualitative and quantitative information to evaluate item properties
- Evaluate multiple sources of validity evidence to select a scale
- Evaluate multiple sources of validity evidence to develop or refine a scale

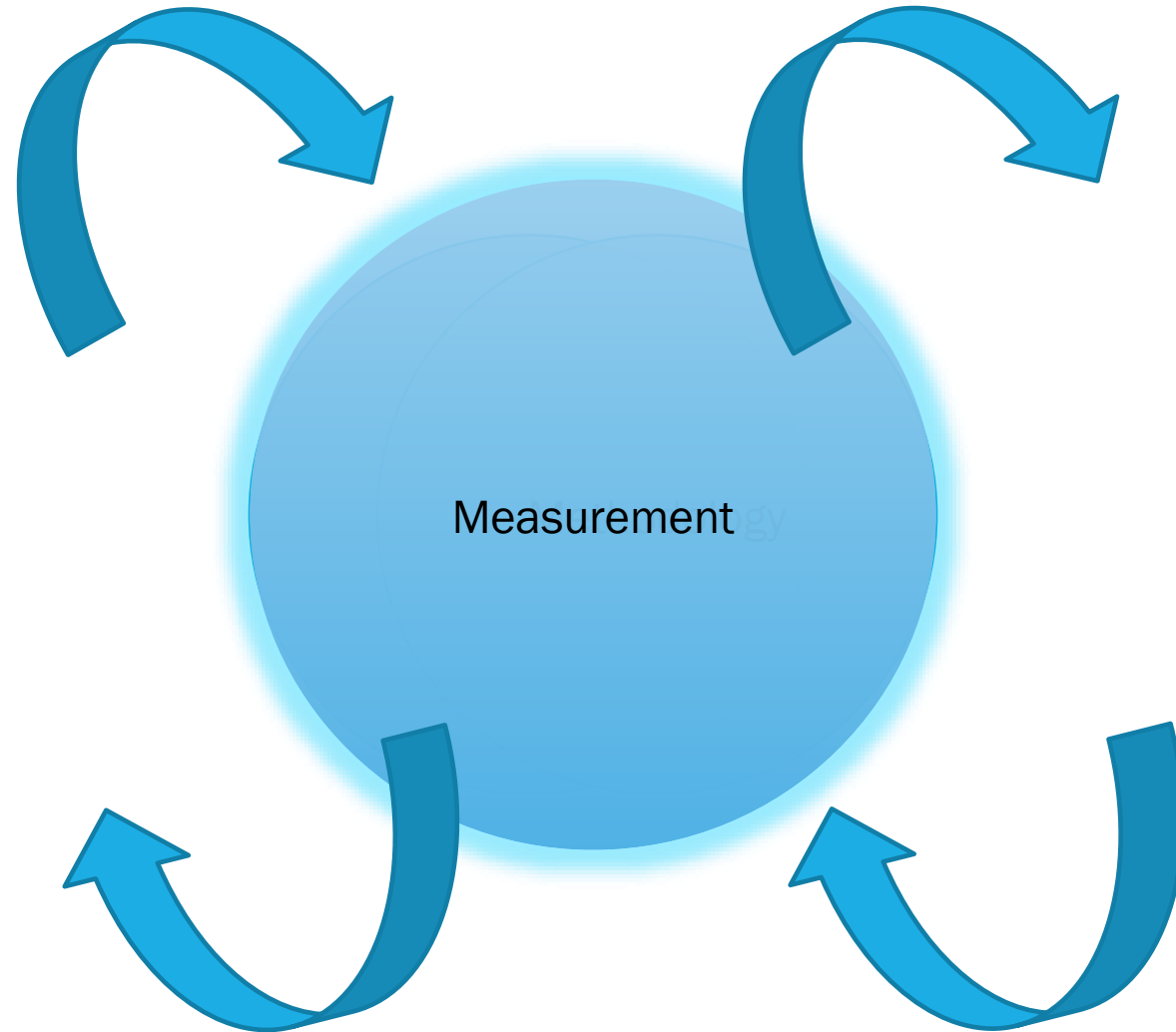
# DISCLAIMERS

- The content is largely drawn from Debbi's book
- We will only scrape the surface, you can get a PhD in measurement!
  - And then a post doc, and then a job, and then a whole career...
- Goal today is to gain conceptual grounding and overview that will facilitate further study
- This isn't a data analysis workshop and we won't focus any specific software, we will focus on concepts and QUALITATIVE methods
- Importantly, we get to spend the day thinking about and discussing measurement things!



# GOALS FOR TODAY







# WHAT ARE LATENT CONSTRUCTS AND WHAT IS MEASUREMENT?

- **Construct** – a theoretical entity that we hypothesize exists to account for certain characteristics or behaviors
  - Also called factors, unobserved variables, latent constructs
  - Attitudes, personality, abilities, motivation, and ideologies are all constructs
- **Instrument** – a procedure used to elicit the behaviors that are assumed to be caused by the construct and to infer a person’s level or status on the construct
  - also called tests, surveys, scales, measures
  - The GRE, student evaluations, intake form for counselling services
  - Can you think of other types of measures psychologists use?
- Implied causal theory
  - There is some underlying construct, and it causes the responses to the items, tasks, or measurement stimuli
    - Verbal ability is what **causes** a student to get an item correct on the vocabulary section of the GRE, not something else
  - Not all measures follow this theory (other theories you know of?), but our example will focus on it



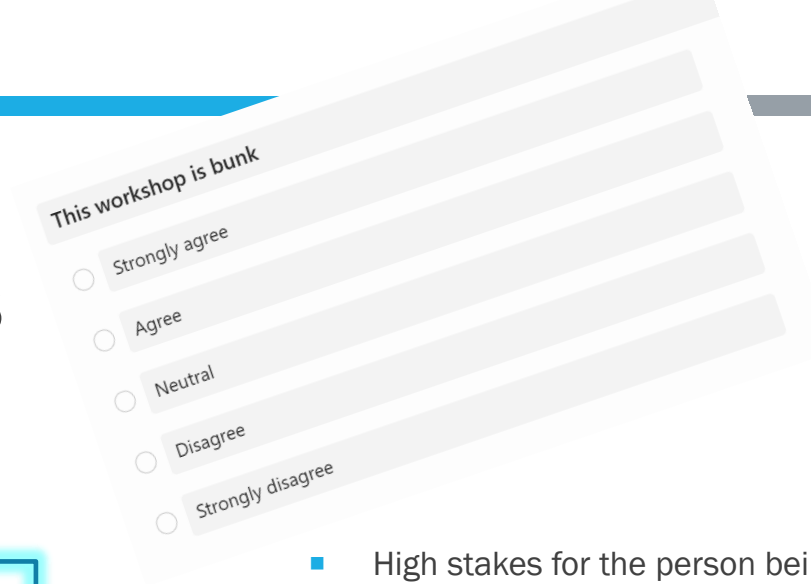


# WHAT IS MEASUREMENT?

- **Measurement** – giving a number to quantify the properties of an object (usually people, but sometimes animals!) according to some rules
- What are some instruments you have taken or administered?
  - What scores do these produce, what does it mean?

# MEASURE TYPES AND USES

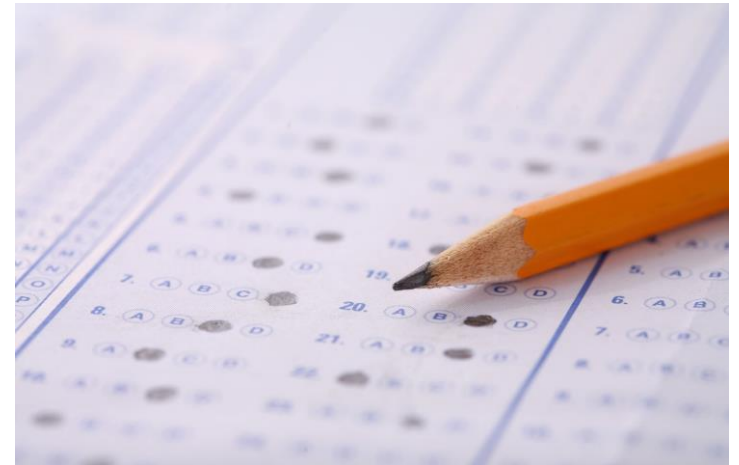
- Tests
  - Multiple choice, essay, and in-between
  - Surveys of attitudes or beliefs
    - Agree-disagree, yes/no
  - Clinical and diagnostic assessments
    - Frequency of behaviors, severity
  - Personality assessments
    - Like me/ not like me, forced choice between options
  - Tasks or performances
    - Judges rate a performance, task completion correct or incorrect
  - Implicit assessments
    - Reaction time
  - Other stuff?



- High stakes for the person being assessed
  - Selection
  - Evaluation
  - Determination of eligibility
- Low to medium stakes for the person being assessed (potential high stakes for society)
  - Inform research
  - Inform policy
  - Make decisions for businesses
  - Introspection
  - Self-improvement


# WHAT COULD GO WRONG AND WHY DO WE NEED A MEASUREMENT THEORY?

- Think back to when you took intro stats exams
  - Construct – statistical knowledge
  - Instrument – an exam, higher scores should mean more stats knowledge
  - What are some of your gripes about exams you've taken?
  - As the prof, I need to create some tasks that elicit the construct, knowledge
    - E.g., Define sampling error in one sentence
    - Ideas for some approaches to eliciting this knowledge?
    - They are limitless, any ones that you pick will only be a **sample\***
    - There are many potential sources of measurement error, sampling the wrong content is just one of them



# WHAT COULD GO WRONG AND WHY DO WE NEED A MEASUREMENT THEORY

- Measurement error
  - Standardized settings are used to reduce errors
    - Scenario 1: students take the stats exam at home, have one week to complete it with little oversight
    - Scenario 2: students take the exam in the gym, get 1 page of a notes sheet, can not work with others and the exam is proctored
  - Even with standardization, there is always some chance that in the given context, the responses elicited are not caused by the construct of interest
- What are some of the problems that might cause errors?

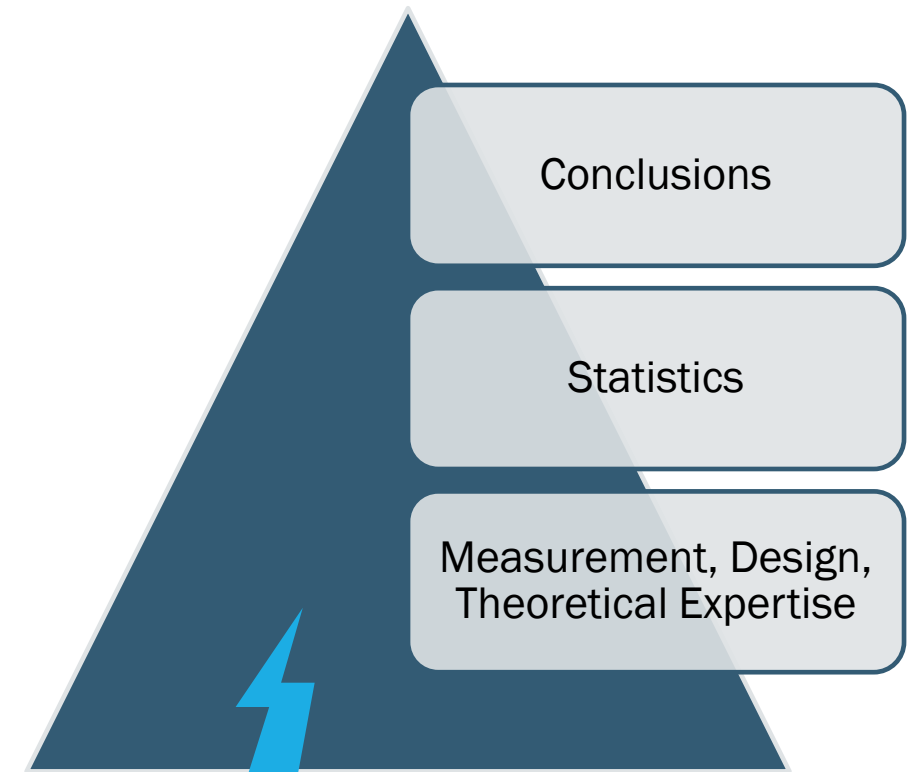


Measurement theory and psychometrics are a field in psychology that is dedicated to developing the theory, methods, and practices used to develop and improve instruments, we study how to reduce measurement error\*

*\*And estimate, describe, and model it*

# THE ROLE OF MEASUREMENT AND VALIDITY THEORY

- Measurement theory is at the foundation of psychological science
  - Psychologists want to make conclusions about the human mind and behavior
  - They do so using data (and often fancy statistics)
  - But first they have to figure out how to collect the data
- When we design experiments and/or measure people, we generate numbers, we trust that those numbers have the meaning we assume
- Instruments might need to be differentially developed or evaluated for different purposes
  - We use the process of construct validation to evaluate the numbers instruments produce and how they will be used



# IF YOU REMEMBER SOMETHING FROM TODAY...

- That a score from a measurement process represents what you intend it to is a scientific claim
- The claim, “higher scores on this extraversion measure indicate more extraversion” is a scientific claim
- You need evidence of that claim in the same way you need evidence for the claim that,
  - “this treatment is effective”
  - “this experience predicts this outcome”
  - “this groups thinks this differently than this other group”
- You need to be comprehensive and creative in gathering that evidence, the more evidence, the stronger you can make your claim
  - Just like any study you work on

# WHAT IS VALIDITY?

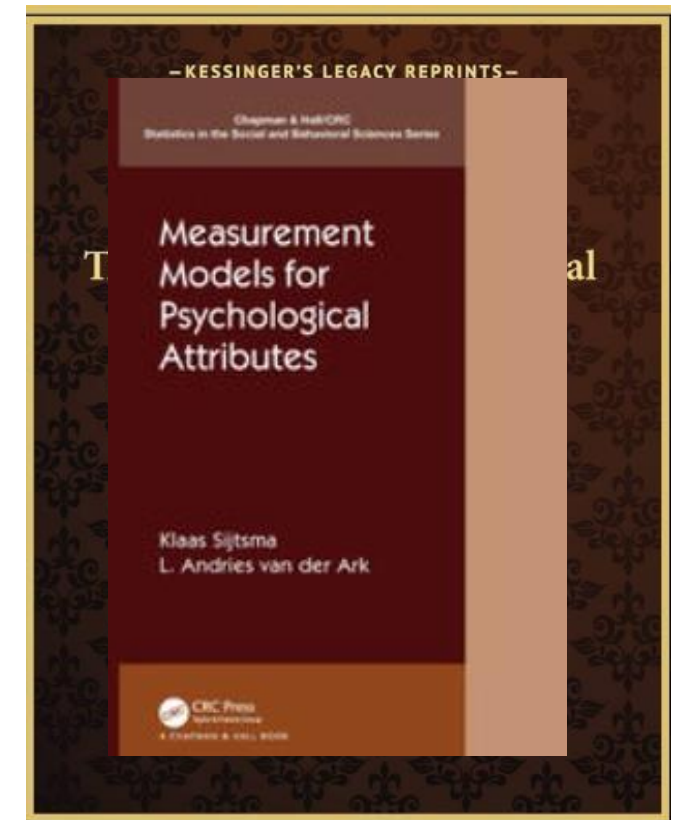
- When you hear the word validity what do you think?
  - Research design terms versus measurement terms





# WHAT IS VALIDITY IN MEASUREMENT?

- In this workshop we will think about a specific validity that pertains to measurement
- Defining measurement validity is a difficult and debated theoretical entity in the field
- **“Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests”**



## CONSTRUCT VALIDITY INTRODUCED IN 1955

- “Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim.” [C&M \(1955\)](#)
- Contemporary theory (since the 80s) places all previous validity types under construct validity
- Different types of validity are now discussed as different sources of validity evidence



# CONTEMPORARY VALIDITY THEORY

- Using validity evidence to develop a validity argument for your intended use and purpose
  - Evidence is specific to the interpretation
  - Validation is a program of research, the more evidence the better
  - Evidence needs to be collected in an on-going way
    - Particularly as instruments are used for different purposes and in different contexts

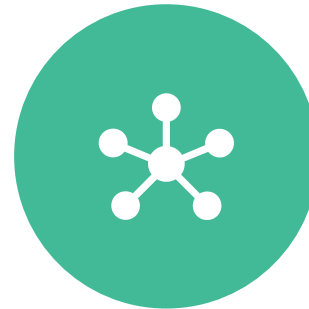
# VALIDATION SHOULD BE A PROGRAM OF RESEARCH THAT GATHERS SOURCES OF EVIDENCE (LOEVINGER, 1957)



**SUBSTANTIVE PHASE – BUILD THEORY ABOUT WHAT THE CONSTRUCT IS**



**STRUCTURAL PHASE – COLLECT EMPIRICAL EVIDENCE THAT SUPPORT THE ITEMS MEASURE THE CONSTRUCT**



**EXTERNAL PHASE – SEE IF CONSTRUCT CONNECTS TO OTHERS AS YOU EXPECT**

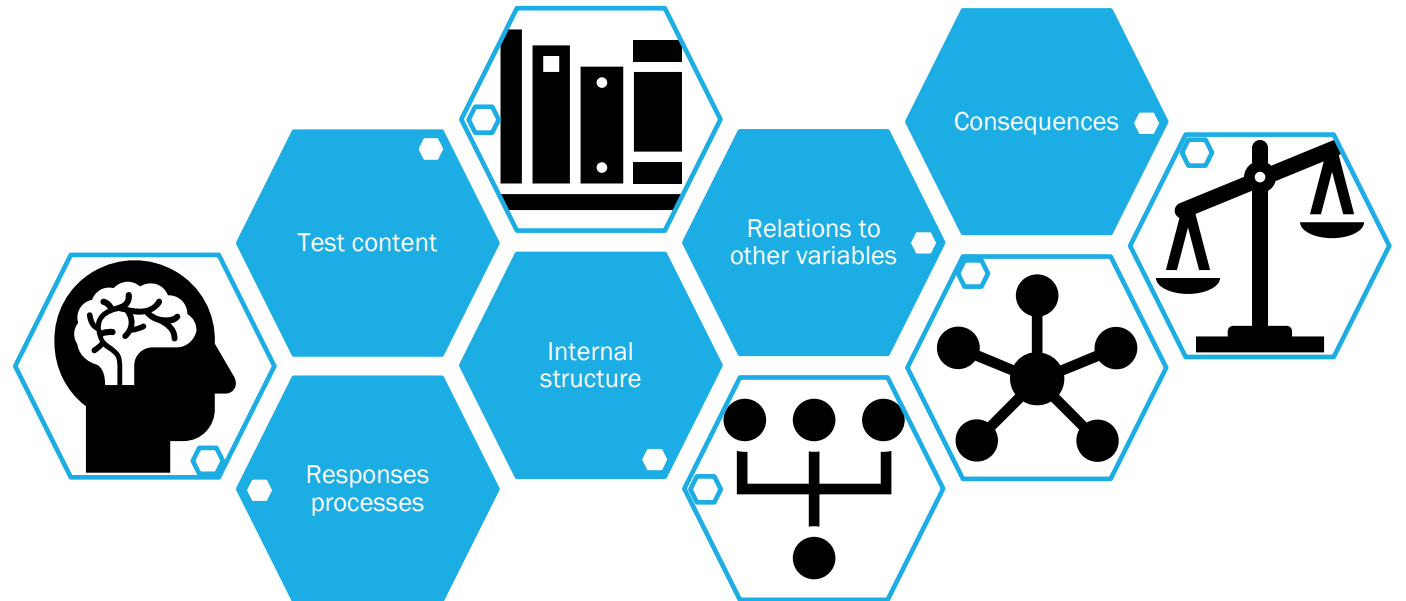


**EVIDENCE HAS BE GATHERED WHEN INSTRUMENTS ARE USED IN NEW CONTEXTS, FOR NEW PURPOSES, AND ACROSS TIME**

My meta science work shows a lot of research focuses here, with little reporting and transparency about the earlier steps

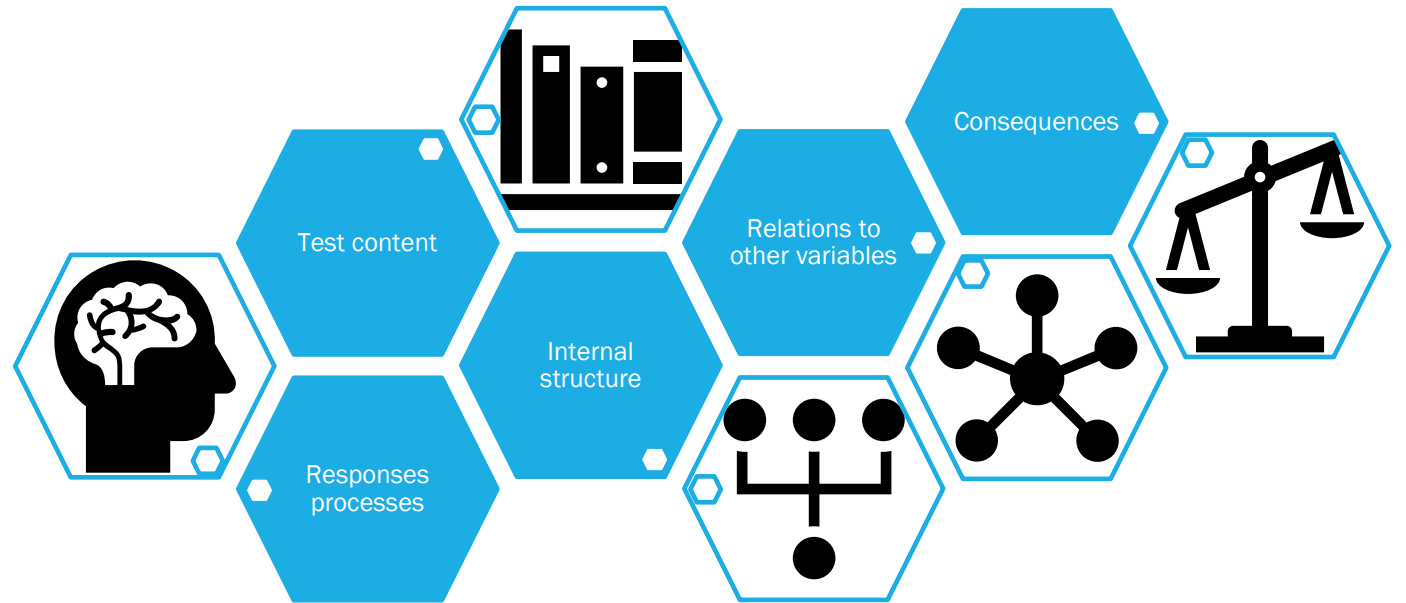
# TYPES OF VALIDITY EVIDENCE

- Validity types are now reworked as sources of validity evidence, evidence of...
  - Content (content validity)
  - Responses processes (construct validity, cognitive validity)
  - Internal structure (construct validity, psychometric validity)
  - Relations to other variables (criterion related or predictive validity)
  - Consequences of testing (this wasn't an original validity type)



# TYPES OF VALIDITY EVIDENCE

The ideal instrument development process would be designed around collecting these various sources of evidence

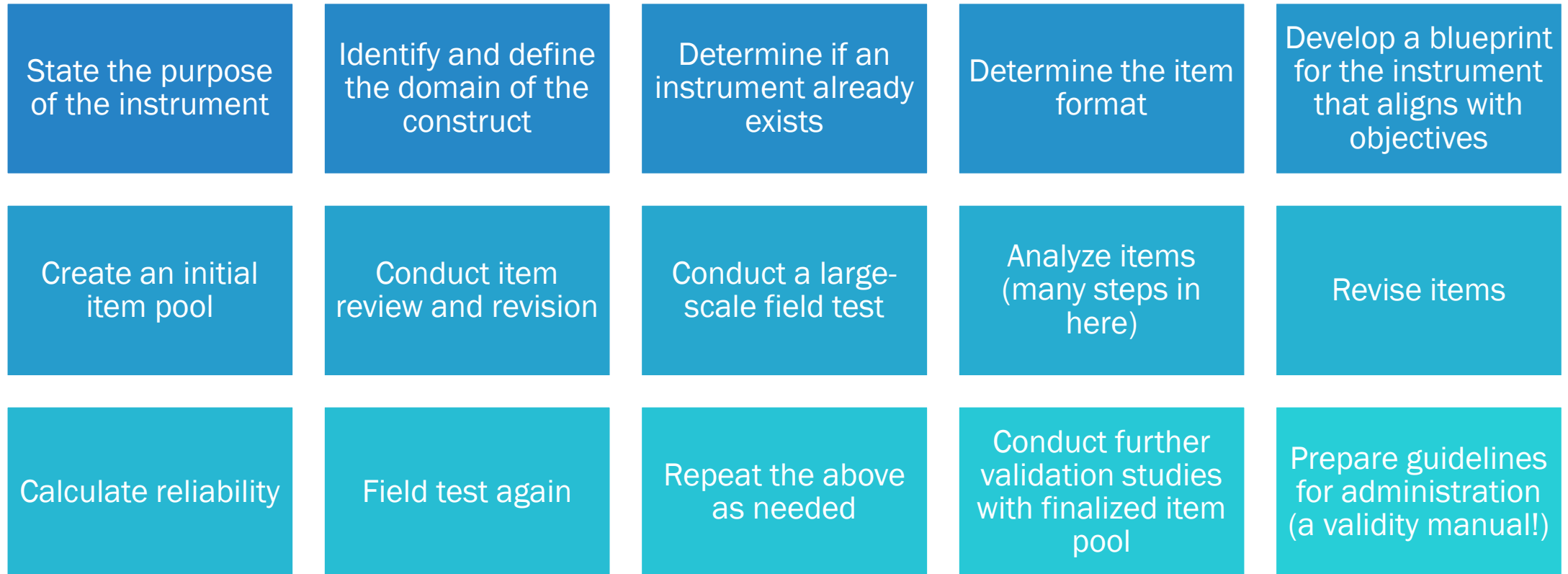


We are going to focus on some more conceptual aspects that I see lacking in practice. This frontend work builds a strong foundation for the rest of the process of validation.

## BIGGIE TAKEAWAY SLIDE

Argument	Validity Evidence	
The instrument contains the necessary items for measuring the breadth the construct	Content	★
The items tap into the intended cognitive/thought processes	Responses processes	★
Relations among items is consistent with theory of construct	Internal structure	
Relations with scores and other constructs are consistent with theory	Relations to other variables	
Intended consequences are realized, unintended consequences are not from invalidity	Consequences	

# INSTRUMENT DEVELOPMENT TIMELINE







# TIME CHECK

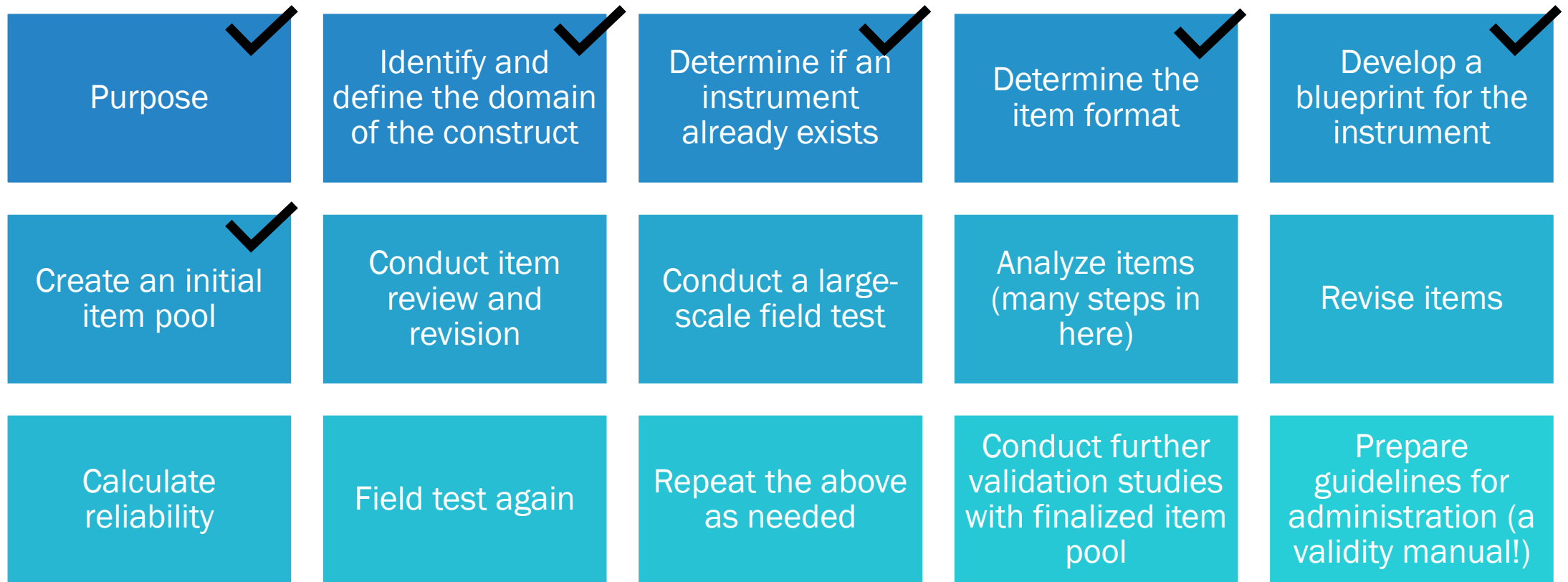
- Next up is working with an example

## OPEN (REAL) DATA SET FROM FAKE SCENARIO: PICKLE FANATICISM

- Market researchers at EvilCorp have been hired by a pickle company to rebrand and revamp their pickle products. They hope to enter the global market by identifying pickle fanatics who can be recruited to become pickle influencers.
- Focus groups identified many issues with the current marketing strategy. For example, key information about the pickle, that is it portly and hearty or spicy and sassy, is hard to read. The cartoon was found to be off-putting, to say the least, in pilot research.
- Initial literature review identified pickle fanaticism as an aspect of one's personality and is defined as a general zeal for pickled products. A few potential facets of pickle fanaticism: a strong desire to eat pickled products, the extreme liking of pickled products, and the general feeling of needing to evangelize to others about the benefits of pickled products.
- EvilCorp has hired us to conduct review of initial items and analyses their team conducted



## WE CAN IMAGINE THAT EVILCORP ALREADY DID THESE STEPS



## LET'S TRY WRITING SOME ITEMS

- Write 2 items for each facet of pickle fanaticism
  - Desire: A strong desire to eat pickle products
    - E.g., I have a strong desire to eat pickles
  - Liking: An extreme liking of pickle products
    - E.g., I like pickles extremely
  - Evangelism: A general feeling of needing to tell others positive things about pickle products
    - E.g., I feel a need to tell others about the benefits of pickles

I know I haven't taught you how to write items yet, we will try it, learn some best practices, then revise!

## ITEM RESPONDING

- Even with the best of efforts and intentions people may not respond to items as you expect
  - Items written to measure the same thing could elicit vastly different responses (partisan pollsters know this)
  - The process these items provoke is different despite that they intend to capture the same view
    - Support for affirmative action
- All in all, do you favor or oppose affirmative action programs in employment for blacks, provided there are no rigid quotas?
    - 74% favor
  - Do you think blacks and other minorities should receive preference in hiring to make up for past inequalities or not?
    - 21% agree
  - Do you favor or oppose affirmative action programs for blacks and other minorities, which do not have rigid quotas?
    - 51% favor

# COGNITIVE PROCESS MODEL OF RESPONDING



Optimizers make a sincere effort to engage in this full process to the best of their ability



Satisficers skip some of the steps or put forth less than their best effort

# ITEM INTERPRETATION

- Figure out what the item means
  - Some terms might be understood differently by different people
  - “several” 2-20
  - “children” babies to 20-year olds depending on who you ask
    - Be specific!
  - Incorporation of context
    - Teacher fills out a survey at work about children, reports they have 20
    - Teacher fills out a survey at home about children, reports they have 2
  - Using other items and response scale as context
    - Cover the full range of responses
    - Avoid negative numbers
      - -5 to 5 and 0-10 produce different response distributions
  - Ask personal information at the end so it doesn’t influence how they respond throughout



Satisficers are cool with a superficial understanding of the item, or use the meaning of the first few items to determine their response to all the items



Optimizers might go back and reread other questions or use the response options to help themselves understand the meaning

# RESPONSE GENERATION

- People with experience or knowledge of the concept/issue/attitude can respond readily
- People without experience may lean on related experience or knowledge
  - e.g., stance on abortion – someone involved in a pro-life campaign has chronically accessible views and can answer easily
  - Someone who hasn't thought about it much, but is conservative, might go with conservative norms, which are pro-life
- People with more experience and knowledge respond more consistently
- People look to other items if they are confused
  - “I use research to make decisions”
    - If previous questions mentioned published papers versus personal experience
- Order randomly, review items for confusion, give plenty of time and encourage thoughtfulness, consider if your population of interest has the experience



Optimizers will try to figure it out, satisficers may just pick “agree,” respond randomly, or choose neutral



# FORMATTING AND REPORTING A RESPONSE

- People will quickly come up with a response in their head, but then they have to report it on a given response scale
- Make sure the response scale is appropriate
  - Don't use a never to always scale or a position statement
  - Avoid using sometimes or often, be specific
  - Label response options (more reliable)
    - Different people will interpret them differently
  - Be aware that the extremes provide an anchor for the response, choose wisely

How would you rate the 2020 Republican National Convention?

Historic  
Great  
Good  
Other

How would you rate President Trump's acceptance speech?

Historic  
Great  
Good  
Other

Did you watch the dark and depressing Democrat Convention?

Yes  
No

How would you rate the Democrat Convention?

Terrible  
Awful  
Bad  
Other

# EDITING THE RESPONSE

- Understand a question one way, read a few others, then edit
- Realize you aren't representing yourself accurately, then edit
- Realize you are making yourself look bad, then edit
- Realize you might not get what you want and then edit
- Socially desirable responding
  - Desired traits vary across cultures and people
  - Try to put people at ease, confirm confidentiality or anonymity
  - Collect a social desirability measure and use it to exclude participants or control for it
  - Consider other response formats beyond Likert (forced choice)



Examples?



## SO HOW DO YOU WRITE ITEMS?

- It isn't easy!
- Next slides have smaller text/long list of guidelines

# TIPS FOR WRITING LIKERT TYPE ITEMS

1. Avoid statements that refer to the past
  1. Bad example, “there is more support for marijuana use today than 30 years ago”
    1. Even those of us who are old enough (I’m not!) might not know or remember
2. Avoid factual statements
  1. Bad example, “marijuana is legal in Quebec”
    1. People will agree or disagree, you won’t know why
3. Avoid statements that can have more than one interpretation
  1. Bad example, “I often use marijuana”
    1. Daily smokers think multiple times a day is often, whereas never users think once a week is often
4. Avoid irrelevant statements
  1. I know you want a good long list of pilot items, but don’t wild and just start writing stuff that isn’t quite relevant

# WRITING LIKERT ITEMS

5. Avoid very extreme statements no one or everyone would agree with
  5. Bad example, “Use of marijuana should be punishable by death”
    5. this item will have no variance
6. BUT, have statements that span the whole range from positive to negative on the construct, avoid neutral statements
  5. Bad example, “I don’t care if marijuana is legal or not”
    5. If people feel neutral, they will agree, which will make their score higher, even though they aren’t higher on the construct
7. Use clear, direct, and simple language
8. Use short statements
9. The statements should include only one thought (i.e., double barreled items)
  5. Bad example, “I support marijuana use and legalization”
10. Avoid universals – all, none, always, never
  5. Bad example, “No one is immune to the effects of marijuana”
    5. People will try to think of an exception!

# WRITING LIKERT ITEMS

11. Avoid leading words only, merely, just
  11. Bad example, “marijuana is just a natural substance”
    11. Also has some facty-ness, watch out!
12. Use simple sentences, not complex or compound
  11. Bad example, “marijuana use is should be supported by the government, but there should be age restrictions, as well as oversight”
13. Avoid advanced vocabulary and jargon (imagine a 6<sup>th</sup> grader reading it)
  11. Bad example, “marijuana is okay to use if one had adequate metacognition to monitor their use”
14. Use an equal number of positive and negative oriented items
  11. It won't prevent carelessness, but it might help you spot it
15. Avoid negative phrasing (but have negative orientation)
  11. Bad example, “I do not support marijuana legalization” – if they do support it, they would have to choose “disagree” confusing!
  12. Better example, “marijuana should be illegal” – if they support it, saying disagree here is easier on the brain
16. Avoid positive and negative orientation in the same sentence
  11. Bad example, “It is okay for people to use marijuana if they buy it from the government”

# BIGGIE OVERVIEW SLIDE

Avoid referring to the future	Short statements, less than 20 words
Avoid facts or seeming facts	One complete thought only
Avoid statements with more than one interpretation	Avoid universals (all, never)
Avoid irrelevant statements	Avoid leading (merely, just, only)
Avoid extreme statements (everyone will agree or disagree)	Avoid complex and compound sentences
Cover the whole range of the construct (positive to negative), don't include neutral statements	Avoid higher than 6 <sup>th</sup> grade vocab
Simple, clear, direct language	Include equal number of positive and negative statements
	Clearly positive OR negative, not a blend of both



## LET'S REVISE ITEMS

- Take a look over your items, using what we've just discussed revise a few of them
- Pick 1 item to share
  - What was the original?
  - What was the revision and why?



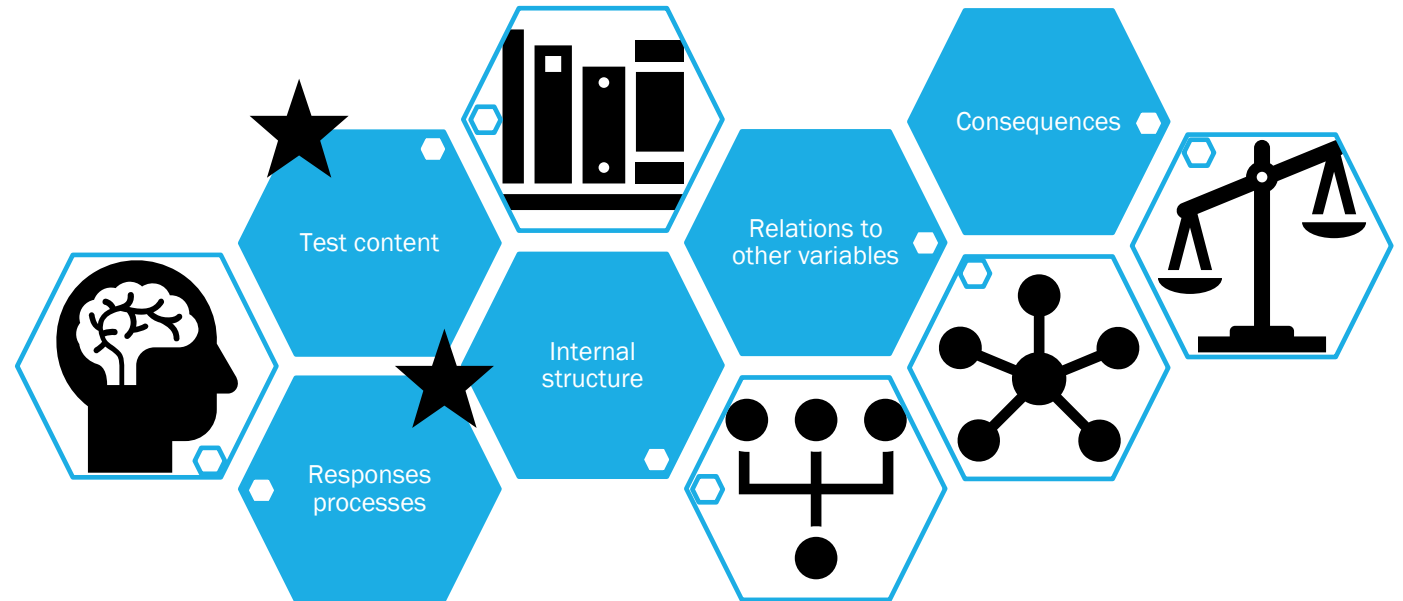


# TIME CHECK

- Next up is learning about some concrete item review approaches
- If close to lunch just go through content alignment slides and do activity after lunch
- If 30+ minutes start activity

# ITEM REVIEWS PROVIDE VALIDITY EVIDENCE

- Initial item review
- Content alignment
- Response processes
- After this, much of the evidence is quantitative
  - “poor” quantitative evidence is difficult to understand and interpret without theoretical and qualitative background work



# CONTENT ALIGNMENT

- Give a group (4-8?) of experts your items
- Ask them to map or categorize the items into the facets or components you wrote them to measure
  - Sometimes called a backwards translation or Q-sort
- Incorporate an open-ended question for other feedback about item wording
- Calculate basic agreement and frequency, review for item wording feedback
  - E.g., if at least on expert mapped the item wrong, we will revise or remove
  - E.g., if a majority of experts map the item incorrectly, we will revise or remove
  - E.g., if at least one experts thinks an item is not clear we will remove



# CONTENT ALIGNMENT EXAMPLE FROM FLAKE ETAL 2015

	A	B	C	D	E	F	G	H	I	J
	<p><b>Directions:</b> Please read each item, and then decide which dimension of motivation the item belongs to (from <b>Effort Related</b>, <b>Effort Not Related</b>, <b>Loss of Valued Alternatives</b>, and <b>Psycho/Emotional</b>). Place a “1” in the column to indicate the dimension you are mapping the item to. If you feel an item does not belong to any of the dimensions, place a “1” in the “none” column. If you select none, please indicate what motivation construct you think the item is measuring in the comments space. If you feel that the item maps to numerous dimensions, place a “1” in all columns that apply. If you map an item to numerous dimensions, please explain why in the comments space. However, if you map an item to <b>ONLY one dimension</b> please rate how sure you were about your mapping in the <b>Certainty</b> column and rate how important you think the item is for measuring the dimension in the <b>Relevance</b> column. <b>If you mapped to numerous dimensions, you do not need to rate the Certainty or Relevance.</b> Finally, I have provided space for you to offer feedback or suggestions for revision. Your additional thoughts about specific items is greatly appreciated.</p>									
		<b>Item</b>	<b>Effort Related Costs-</b> Negative appraisals about the amount of effort, time, or energy required by the class	<b>Effort Unrelated Costs-</b> Negative appraisals about the amount of effort, time, or energy required by other things outside the task, which in turn compete and limit the amount of effort a student has to put into the class	<b>Loss of Valued Alternative Costs-</b> Negative appraisals about having to give up other tasks to engage in the class, like not being able to do other desired things or missing out on activities	<b>Psychological/Emotional Costs-</b> Negative appraisals about how one feels psychologically or emotionally in the class. Feelings or mental states that are a result of the class	<b>None-</b> Item doesn't seem to fit anywhere	<b>Certainty-</b> How certain/sure are you about where you chose to map the item? 1=Very Uncertain 2=Fairly Uncertain 3= Fairly Certain 4= Very Certain	<b>Relevance-</b> How relevant do you think this item is to measuring the dimension (i.e., does it do a good job of capturing the dimension?) 1= Very Irrelevant 2= Fairly Irrelevant 3= Fairly Relevant 4=Very Relevant	<b>Comments/Suggested Revisions</b> Would you revise this item? How so? Thoughts about this item? Please use this space for comments
2	1)	This class makes me feel bad				1		4	4	
3	2)	Because of other things that I do, I don't have time to put into this class		1				4	4	
4	3)	This class is too much work	1					4	4	
5	4)	This class is stressful				1		4	4	
5	5)	Taking this class makes me unhappy				1		4	4	

# ITEM REVIEW AND MAPPING ACTIVITY

- Access worksheet in shared folder
- EvilCorp wrote items to measure three factors
  - Strong Desire to Eat Pickled Products
  - Liking of Pickled Products
  - Pickle Evangelism
- Read each item, and consider which factor it measures and how relevant you think it is to the factor
- Attempt to identify if/how the item violates guidelines
- Leave any other notes

# BIGGIE OVERVIEW SLIDE OF GUIDELINES

Avoid referring to the future	Short statements, less than 20 words
Avoid facts or seeming facts	One complete thought only
Avoid statements with more than one interpretation	Avoid universals (all, never)
Avoid irrelevant statements	Avoid leading (merely, just, only)
Avoid extreme statements (everyone will agree or disagree)	Avoid complex and compound sentences
Cover the whole range of the construct (positive to negative), don't include neutral statements	Avoid higher than 6 <sup>th</sup> grade vocab
Simple, clear, direct language	Include equal number of positive and negative statements
	Clearly positive OR negative, not a blend of both



# ITEM REVIEW DISCUSSION

- Any good items?
- Any clear problems?
- Any confusions?
- Any takeaways?

Response Scale				
Strongly Disagree =1	Disagree	Neither agree nor disagree	Agree	Strongly Agree = 5



## INITIAL ITEM POOL GIVEN TO EVILCORP FOR REVIEW

Strong desire to eat pickled products	Extreme liking of pickled products	Feels the need to evangelize about the benefits of pickles
I think about eating pickles at most meals	Pickles are made with vinegar	I have many friends who like pickles
Over the past 30 years I have eaten thousands of pickles	I only eat pickles	Weekly I make sure to tell a friend about pickles, the different kinds of pickles, where you can buy them, and how much they cost
I often contemplate the role of pickles in a post-modern society	I always eat pickles for breakfast	My family should eat pickles regularly
I dream about pickles	I do not like pickles much	I don't like my friends who don't eat pickles
I prefer to eat other snacks over pickles	I like pickles less than other foods	Everyone should eat pickles
My go to snack is a pickle	Pickles taste extremely good	Pickles are great gifts
I would rather eat dill pickles than sweet pickles.	I don't mind pickles.	I like to tell people about my favorite pickles
I eat pickles often.	Pickles are a unique food.	I want to share my love of pickles with the world
Pickles are delectable sustenance	Pickles taste bad	My friends should not eat pickles
I would like to eat pickles everyday	Pickles are delicious	I'm secretive about my pickle habits
I avoid eating pickles	I like pickles a lot	I recommend pickles to people I know



# THE THINK ALOUD PROTOCOL

- Also called cognitive interviewing or response process evaluation
  - Respondent says their thoughts out loud as they respond to the measure with minimal intervention from the interviewer
- This can be a useful exercise even with a small number of people
- Focus on the intended population
- Recording and transcription of responses can then be used to code responses for themes or rate against validity criteria
  - There are many approaches to qualitative data analysis
    - code for themes you think you might find (a priori), or code for themes and see what you find (grounded theory)
  - Rate items for their validity
    - Ahead of data collection, for each item develop criteria – if the item is valid, what **should** they be thinking? Then categorize responses against the criteria

## EXAMPLE INSTRUCTIONS

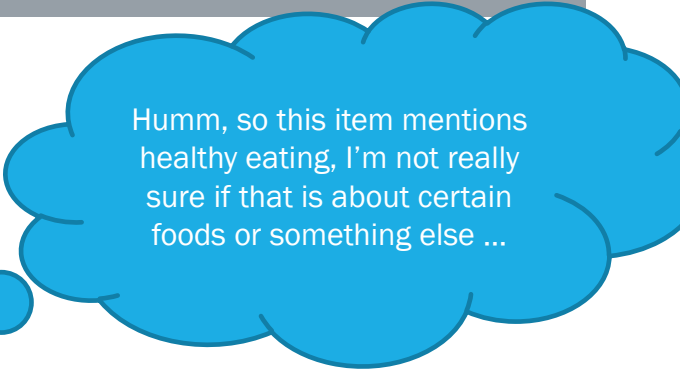
- “I’m going to give you a short questionnaire to complete and ask you to do something called a think aloud while you answer the questions. What that means is that, as you are responding to each statement, I’d like you to think out loud and say all the things that go through your mind as you’re choosing your answer. The reason I’m asking you to do this is so that I can have some insight into the process by which you reach your answers. I’ll ask you to read the item out loud, say everything you’re thinking as you decide on your answer, and say your chosen answer out loud as well as indicating it on the sheet.
- I will demonstrate and then you can give it a try...”
- Then the interviewer does a test item, and has the participants do a test item to practice, before starting the data collection phase
- Test items should be items not relevant to the instrument

# DEMONSTRATION


	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
I ate healthy this week.					
I enjoy eating health foods.					

## DURING THE INTERVIEWS...

- Interfere as little as possible, non-directive probes only (e.g. “please remember to say your thoughts out loud”)
- Don’t answer questions about meaning! Have deflections prepared.
- Don’t correct their response, even if it directly contradicts what they’re saying
- Retrospective session
  - If you are interested in something specific, you can ask!
  - Examples
    - Whether they interpreted health food as positive or negative in the original item
    - What their most and least preferred alternative version was



Humm, so this item mentions healthy eating, I'm not really sure if that is about certain foods or something else ...



Well, I never eat healthy so I'll put "neutral"



## LET'S TRY A MINI THINKALOUD!

- Access worksheet
- Write in two of the items you wrote from before, to go with the existing items
- Each person pick one of the existing items to think aloud on so you don't both do the same item
- Take turns, one person think aloud for all the items and respond, while the other person takes notes, then switch



# SHARE RESULTS

- Any insights gained you didn't have from your own review?
- Unexpected and/or expected results?
- Did you interpret items differently than the person you interviewed?



## **END OF SECTION ON ITEM REVIEW**

- Check the time, is there an hour left?
- If yes, discuss item analysis, if no go to open office hours

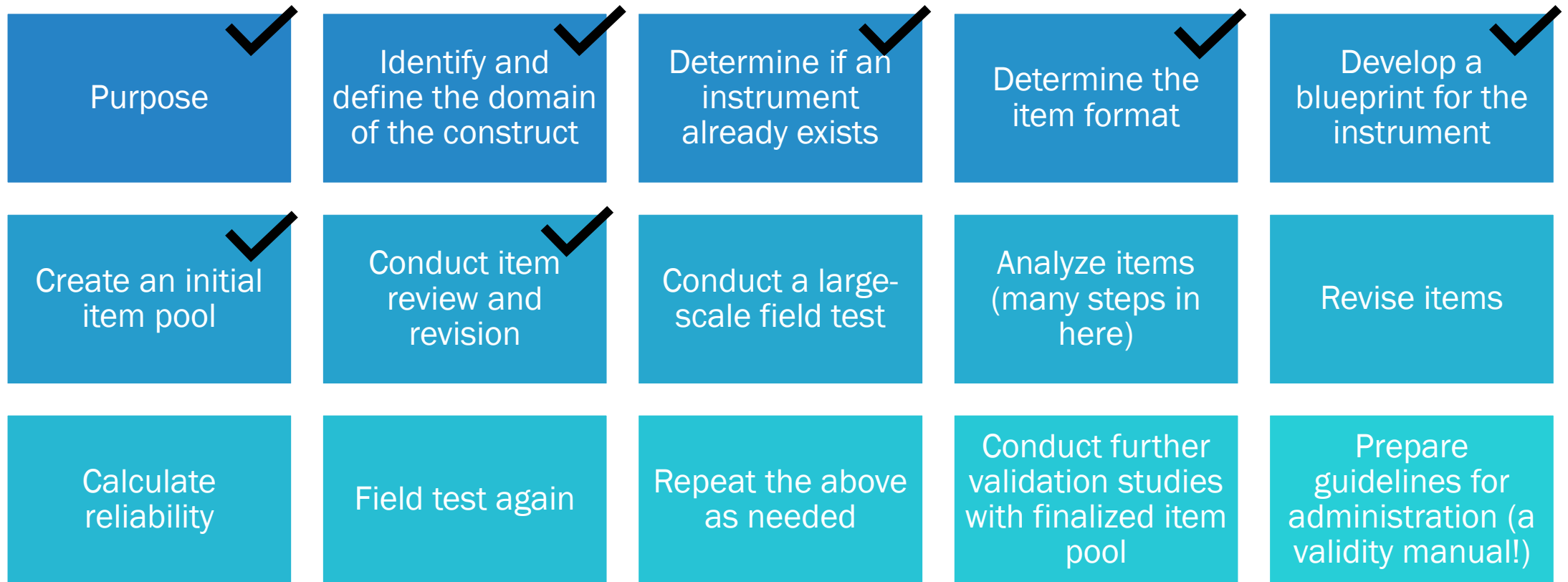


# SOME TIME FOR DISCUSSION

- I'm going to give you all a few minutes to consider questions and discussion points from the day
- How does this connect to your research?
- Is there a challenge or idea you have?



# INSTRUMENT DEVELOPMENT PROCESS STATUS



# GET SOME ITEM RESPONSES!

- You can collect data online, via scantron, or on paper forms.
- Once you have item responses data you can do A LOT of things, but first we will talk about what you can do with descriptive statistics
  - An “item analysis”
- Most of item analysis is descriptive statistics, things you learned in intro stats, put to a different purpose
- This step can be glossed over, but it saves a lot of headache later to get to know the item response data before anything more complex

# ITEM ANALYSIS FOR SURVEY ITEMS

Most standard software packages include reliability analysis as a part of item analysis, but reliability should be considered after factor structure is confirmed

## 1. Review

- Determine which items are positive and negative
- Determine scoring method and interpretation
- Through qualitative methods, identify any possible problematic items

## 2. Descriptives

- Investigate descriptive statistics and graphs for each item
- Look for skew, sparseness, aberrant items

## 3. Correlation Matrix

- Identify weakly correlated items
- Identify potential subsets
- Calculate mean interitem correlation (will do this when we get to reliability)

## 4. Item-totals

- Investigate item-total correlations (will do this when we get to reliability)
- Investigate if-item-deleted (will do this when we get to reliability)

# 1. ITEM REVIEW

- Decide how we will score and interpret scores
  - A total score: sum up all of the item responses such that those with higher scores have higher levels of fanaticism
- For this to work, we need to reverse score negatively worded items

Website Satisfaction					
1. The website has a user friendly interface	Strongly Disagree	Disagree	Neutral	<u>Agree</u>	Strongly Agree
2. The website is usual my first choice for research	Strongly Disagree	Disagree	Neutral	Agree	<u>Strongly Agree</u>
3. I have difficulty using the website	Strongly Disagree	<u>Disagree</u>	Neutral	Agree	Strongly Agree
4. It is easy to upload new images to the website	Strongly Disagree	Disagree	Neutral	Agree	<u>Strongly Agree</u>
5. The website has a pleasant color scheme	Strongly Disagree	Disagree	<u>Neutral</u>	Agree	Strongly Agree
6. The website has a good selection of images	Strongly Disagree	Disagree	Neutral	<u>Agree</u>	Strongly Agree
7. The website is ugly	<u>Strongly Disagree</u>	Disagree	Neutral	Agree	Strongly Agree

## Website Satisfaction

1. The website has a user friendly interface

Strongly Disagree      Disagree      Neutral      Agree      Strongly Agree

2. The website is usual my first choice for research

Strongly Disagree      Disagree      Neutral      Agree      Strongly Agree

3. I have difficulty using 4 website

Strongly Disagree      Disagree      Neutral      Agree      Strongly Agree

4. It is easy to upload new images to the website

Strongly Disagree      Disagree      Neutral      Agree      5  
Strongly Agree

5. The website has a pleasant color scheme

Strongly Disagree      Disagree      3  
Neutral      Agree      Strongly Agree

6. The website has a good selection of images

Strongly Disagree      Disagree      Neutral      Agree      Strongly Agree

7. The 5 site is ugly

Strongly Disagree      Disagree      Neutral      Agree      Strongly Agree

Total score without reverse coding = 11/20  
(not high, even though they generally like the website)

Reverse score: 5=1, 4=2, 3=3, 2=4, 5=1  
Total score with reverse coding = 17/20  
(makes more sense)

## NEGATIVELY WORDED ITEMS

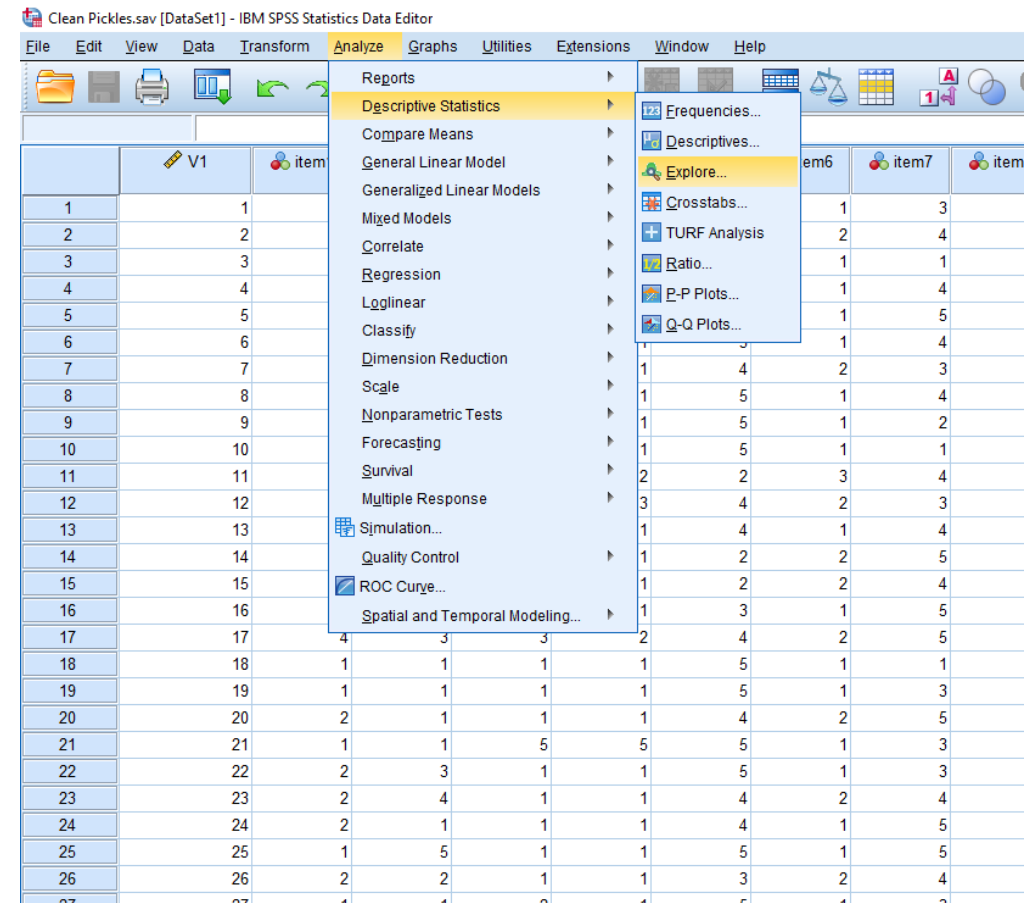
Strong desire to eat pickled products	Extreme liking of pickled products	Feels the need to evangelize about the benefits of pickles
I think about eating pickles at most meals	Pickles are made with vinegar	I have many friends who like pickles
Over the past 30 years I have eaten thousands of pickles	I only eat pickles	Weekly I make sure to tell a friend about pickles, the different kinds of pickles, where you can buy them, and how much they cost
I often contemplate the role of pickles in a post-modern society	I always eat pickles for breakfast	My family should eat pickles regularly
I dream about pickles	I do not like pickles much*N	I don't like my friends who don't eat pickles
I prefer to eat other snacks over pickles*N	I like pickles less than other foods*N	Everyone should eat pickles
My go to snack is a pickle	Pickles taste extremely good	Pickles are great gifts
I would rather eat dill pickles than sweet pickles.	I don't mind pickles.	I like to tell people about my favorite pickles
I eat pickles often.	Pickles are a unique food.	I want to share my love of pickles with the world
Pickles are delectable sustenance	Pickles taste bad*N	My friends should not eat pickles*N
I would like to eat pickles everyday	Pickles are delicious	I'm secretive about my pickle habits*N
I avoid eating pickles*N	I like pickles a lot	I recommend pickles to people I know

## 2. DESCRIPTIVES; SPSS EXAMPLE

EXAMINE VARIABLES=item1 item2 item3 item4 item5 item6  
item7 item8 item9 item10 item11 item12 item13  
item14 item15 item16 item17 item18 item19 item20  
item21 item22 item23 item24 item25 item26 item27  
item28 item29 item30 item31 item32 item33

/PLOT HISTOGRAM  
/STATISTICS DESCRIPTIVES  
/CINTERVAL 95  
/MISSING LISTWISE  
/NOTOTAL.

This gives you  
basic descriptive  
stats and  
histograms



The screenshot shows the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the 'Descriptive Statistics' submenu is selected, with 'Explore...' highlighted. The background shows a data table with columns labeled 'V1', 'item6', 'item7', and 'item'. The data table contains 27 rows of numerical data.

## 2. DESCRIPTIVES; R EXAMPLE

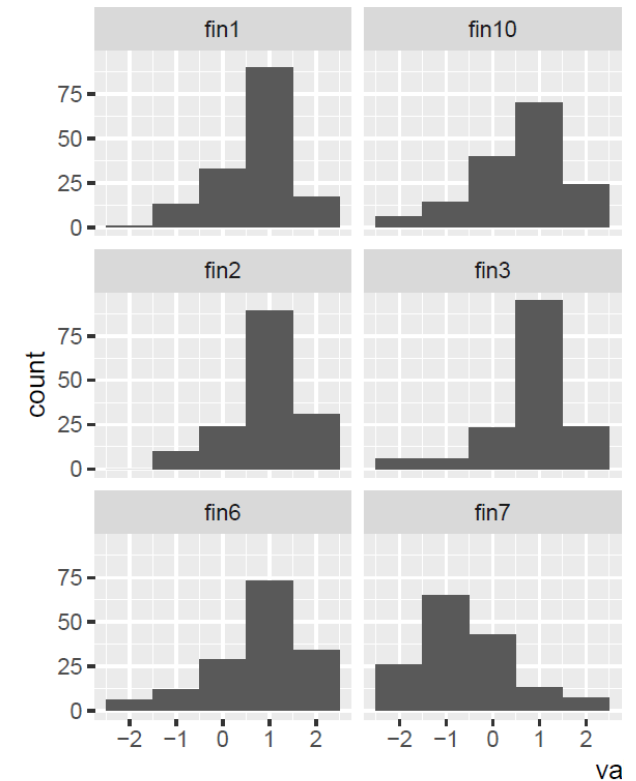
describe gives you basic  
descriptives for all vars in the data  
you feed it

```
describe(finItems)
```

##	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	
##	fin1	1	154	0.71	0.80	1	0.76	0.00	-2	2	4	-0.73	0.50	0.06
##	fin2	2	154	0.92	0.78	1	0.98	0.00	-1	2	3	-0.66	0.35	0.06
##	fin3	3	154	0.81	0.88	1	0.91	0.00	-2	2	4	-1.32	2.29	0.07
##	fin4	4	154	0.75	0.93	1	0.85	0.00	-2	2	4	-0.90	0.69	0.08
##	fin5	5	154	0.65	0.95	1	0.71	1.48	-2	2	4	-0.59	-0.05	0.08
##	fin6	6	154	0.76	1.01	1	0.87	1.48	-2	2	4	-0.87	0.45	0.08
##	fin7	7	154	-0.58	1.01	-1	-0.66	1.48	-2	2	4	0.64	0.10	0.08
##	fin8	8	154	0.71	0.91	1	0.78	1.48	-2	2	4	-0.59	0.01	0.07
##	fin9	9	154	-0.05	1.14	0	-0.07	1.48	-2	2	4	0.14	-0.93	0.09
##	fin10	10	154	0.60	0.99	1	0.67	1.48	-2	2	4	-0.68	0.18	0.08
##	fin11	11	154	0.38	1.17	1	0.45	1.48	-2	2	4	-0.46	-0.73	0.09
##	fin12	12	154	0.73	1.03	1	0.85	1.48	-2	2	4	-0.85	0.40	0.08

Ggplot for some histograms

```
finItems %>%  
gather() %>%  
ggplot(aes(value)) +  
  facet_wrap(~ key, scales = "fixed") +  
  geom_histogram(bins = 5)
```

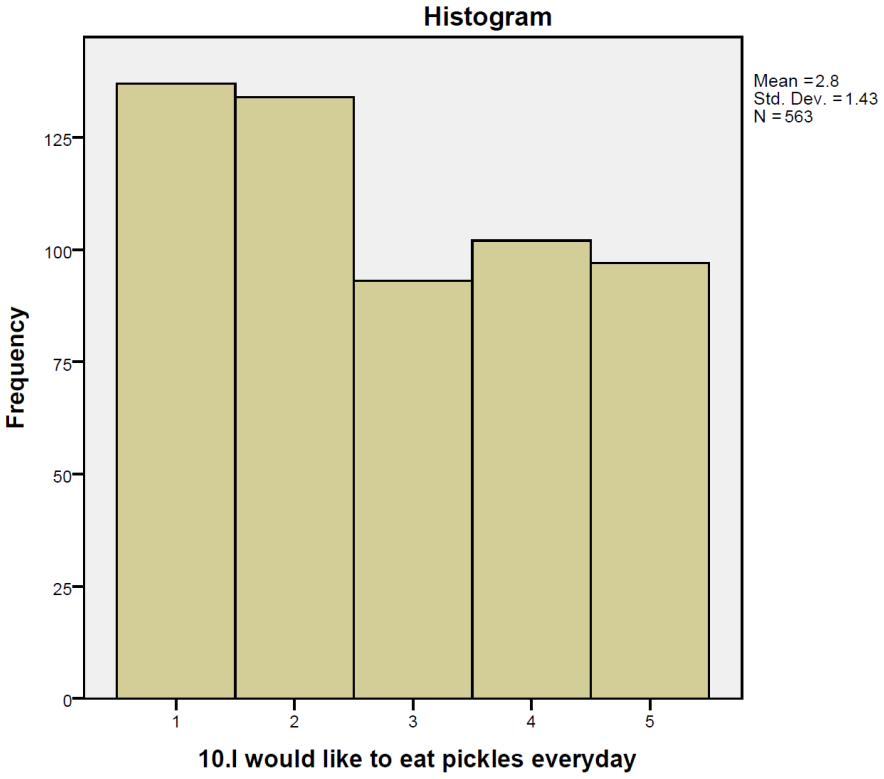
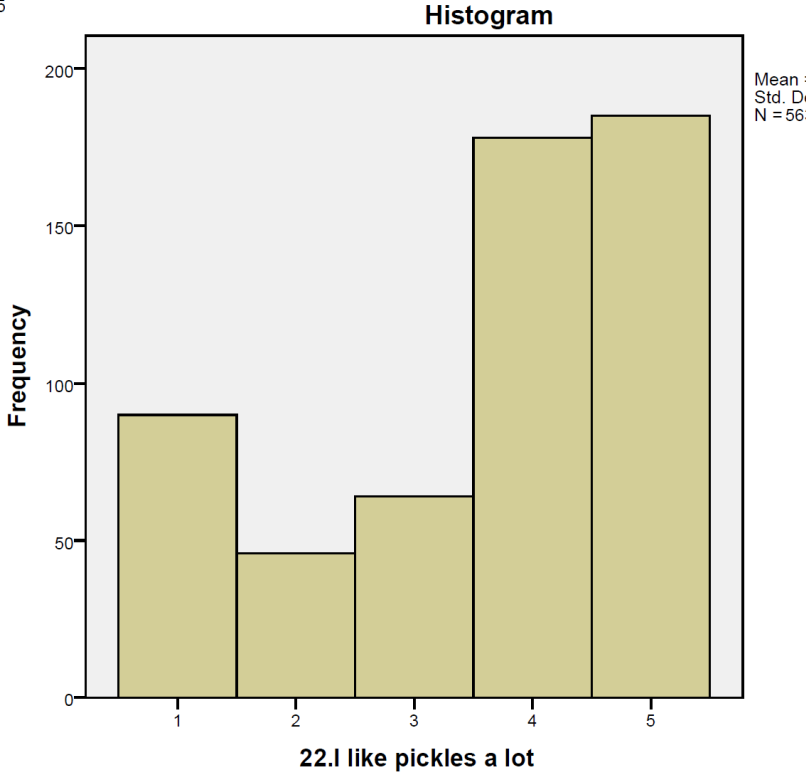
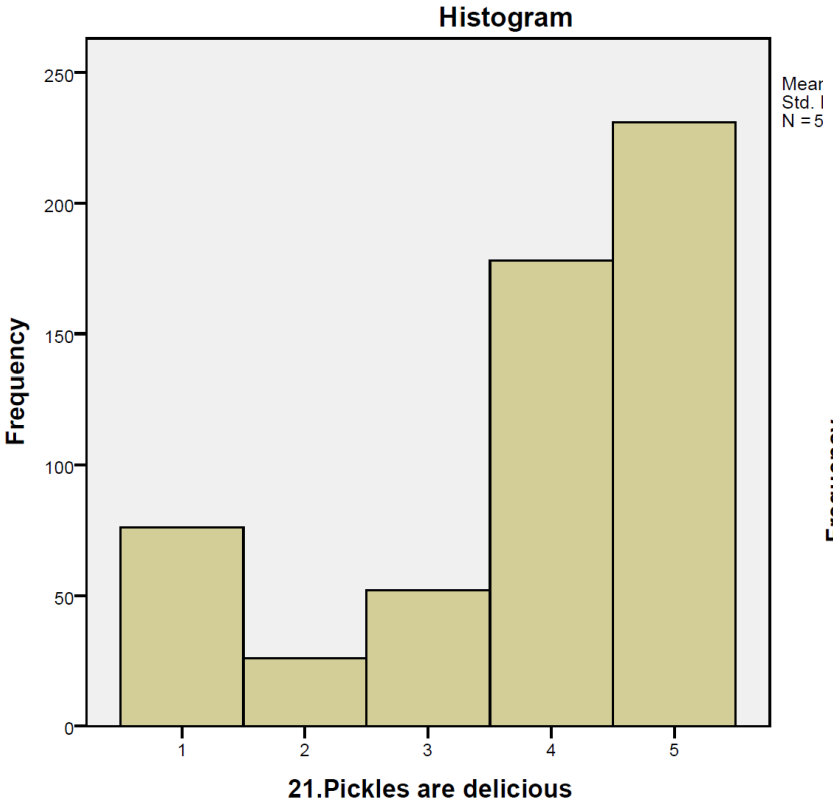




## 2. DESCRIPTIVE INVESTIGATIONS

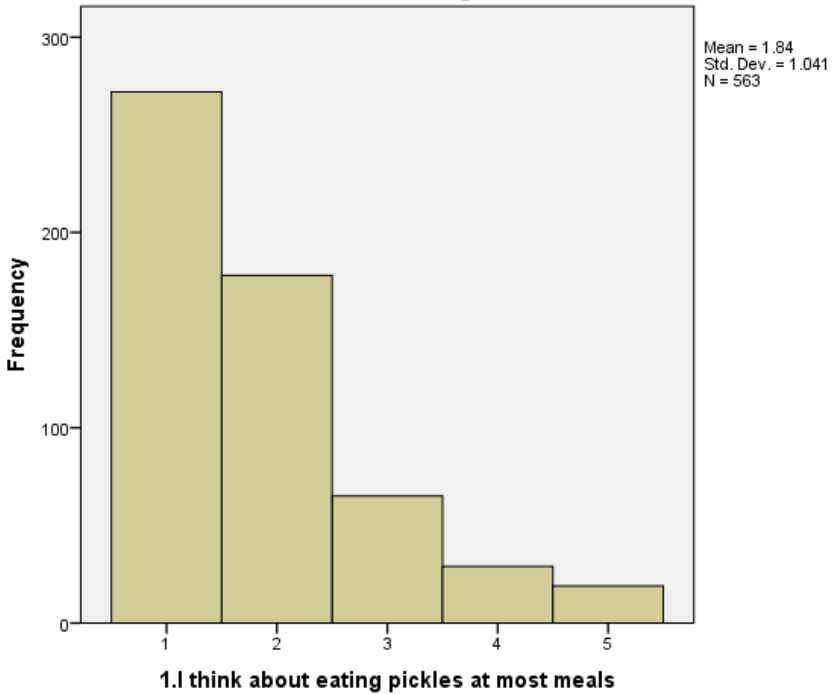
- Investigate the responses patterns
- What are the min, max?
  - Are items skewed?
  - Are some categories not used and/or have sparse responses?
- Response patterns can be hard to interpret on their own, that is why item review is an important first step

# SIMPLY WORDED, FACE VALID ITEMS

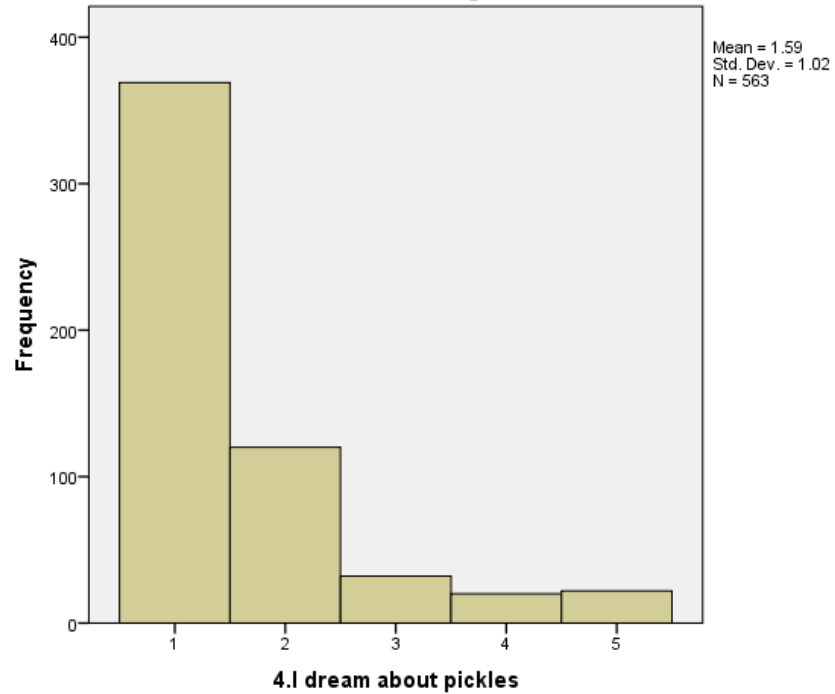


# TOO EXTREME, SPARSE RESPONDING AND SKEW

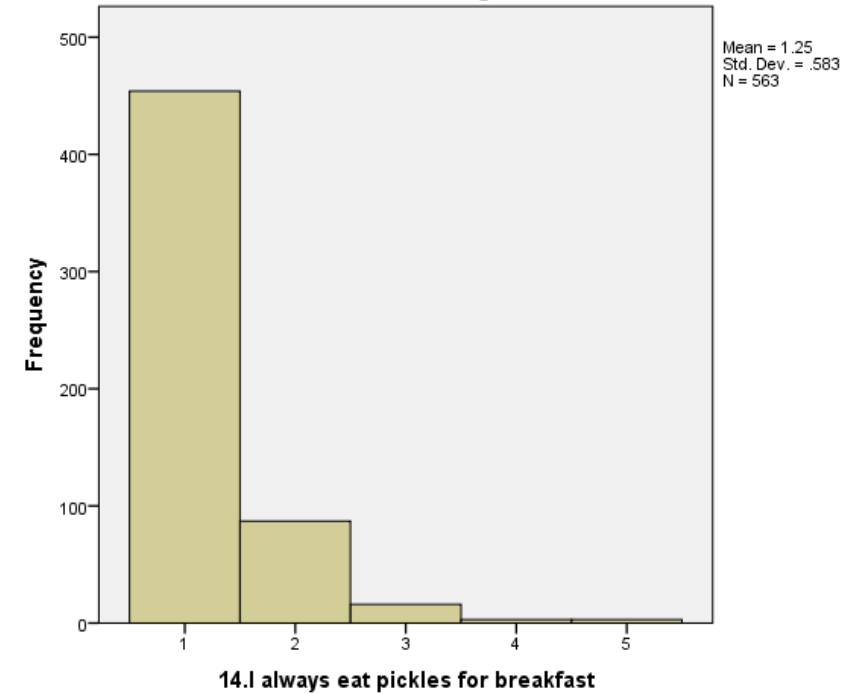
Histogram



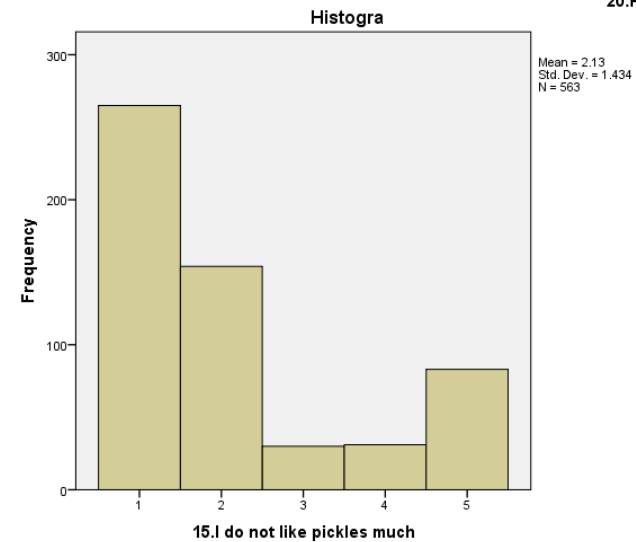
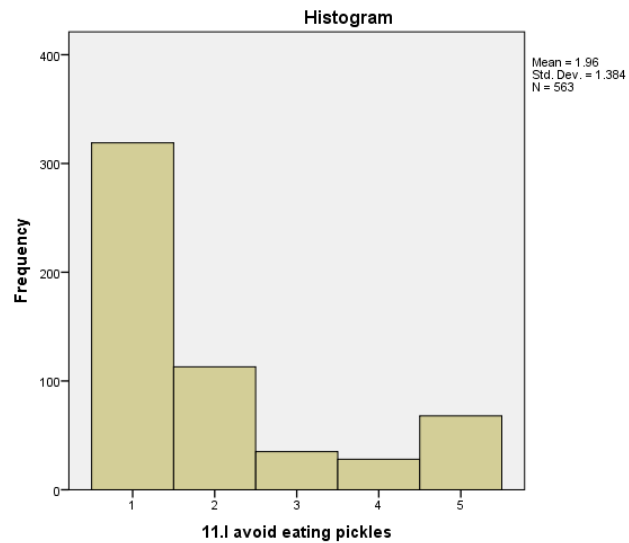
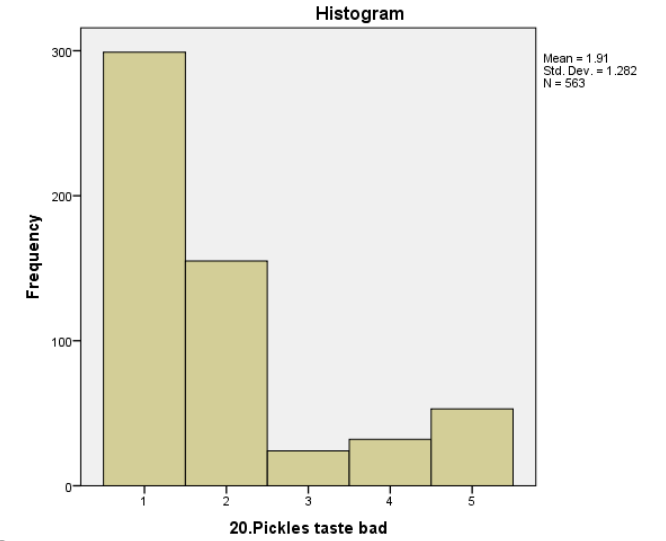
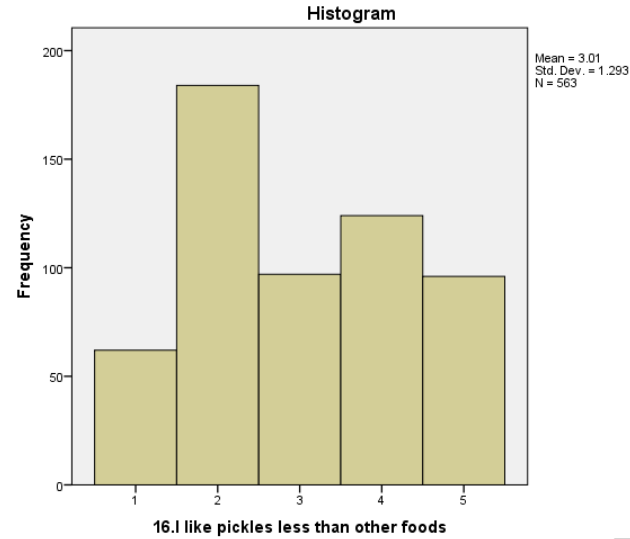
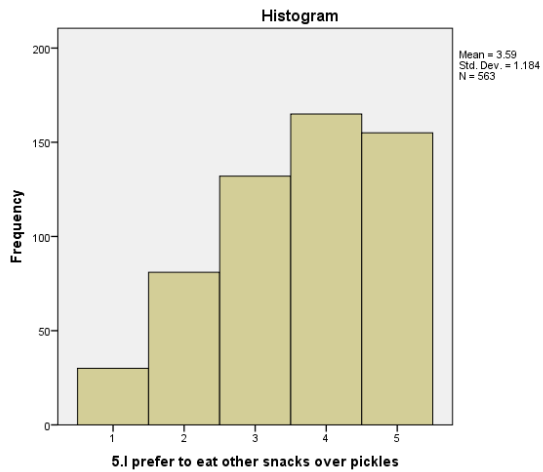
Histogram



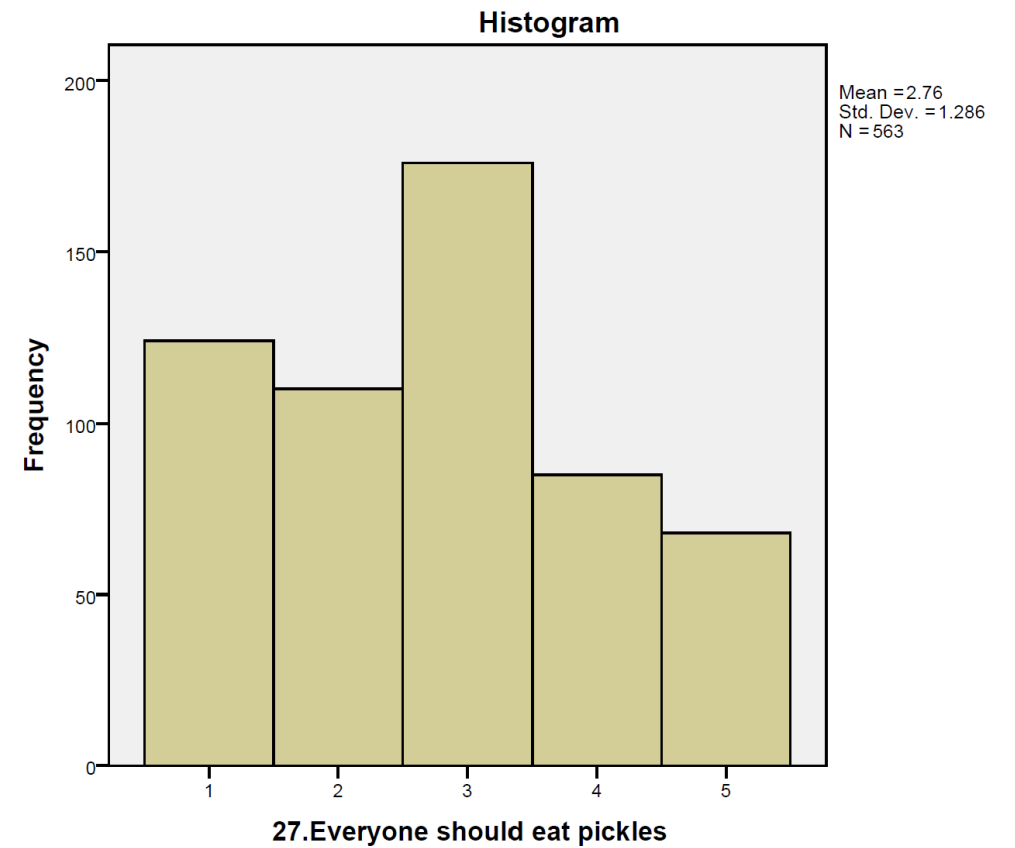
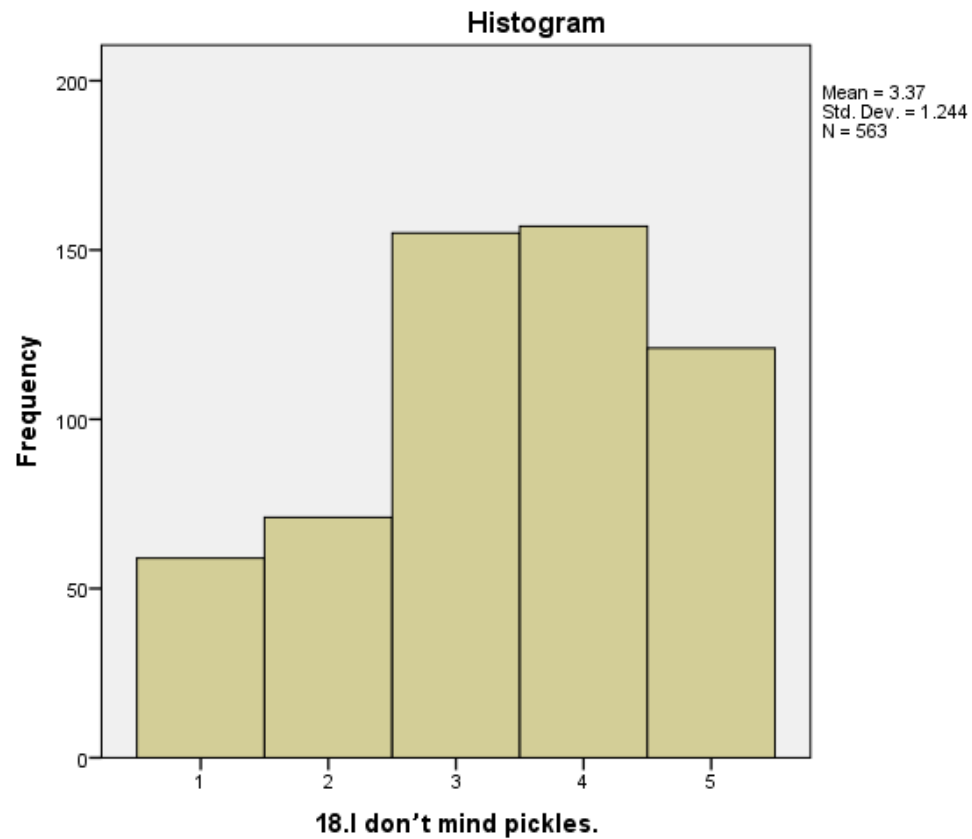
Histogram



# NEGATIVELY WORDED ITEMS



# ITEMS FROM THINK-ALOUD EXERCISE



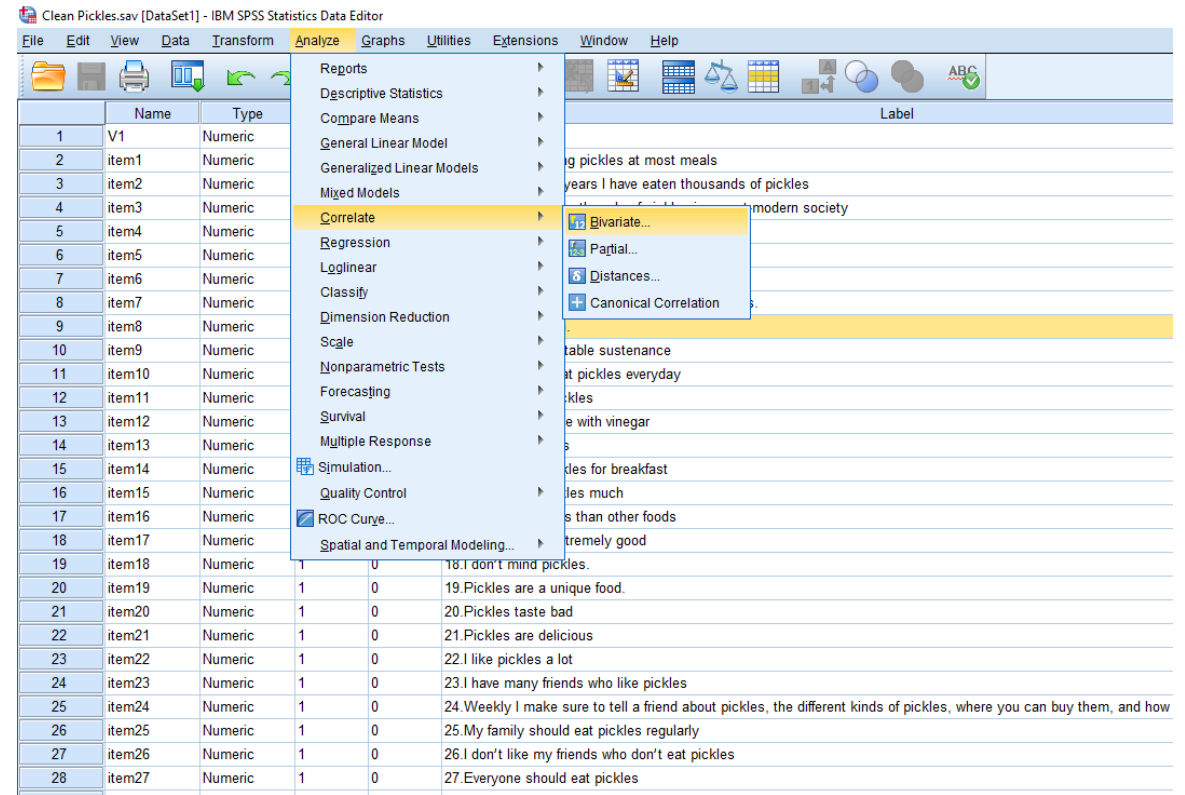
## WHAT ARE THE TAKE-AWAYS HERE?

- Add the evidence together
  - Item review, content mapping, think aloud, aberrant response pattern
- When items don't behave as you expect, this is sign...
  - That they don't measure what you expect
  - Without qualitative review you are left guessing, but mixed methods can be very informative

# 3. CORRELATIONS; SPSS EXAMPLE

## CORRELATIONS

```
/VARIABLES=item1 item2 item3 item4 item5 item6 item7  
item8 item9 item10 item11 item12 item13  
item14 item15 item16 item17 item18 item19 item20  
item21 item22 item23 item24 item25 item26 item27  
item28 item29 item30 item31 item32 item33  
/PRINT=TWOTAIL NOSIG  
/MISSING=PAIRWISE.
```



### 3. CORRELATIONS, R GGLPOT EXAMPLE

You can't just open R and do this,  
you must install packages and read  
the data in properly at first  
BUT, if you do you get a nice matrix  
that is color coded!

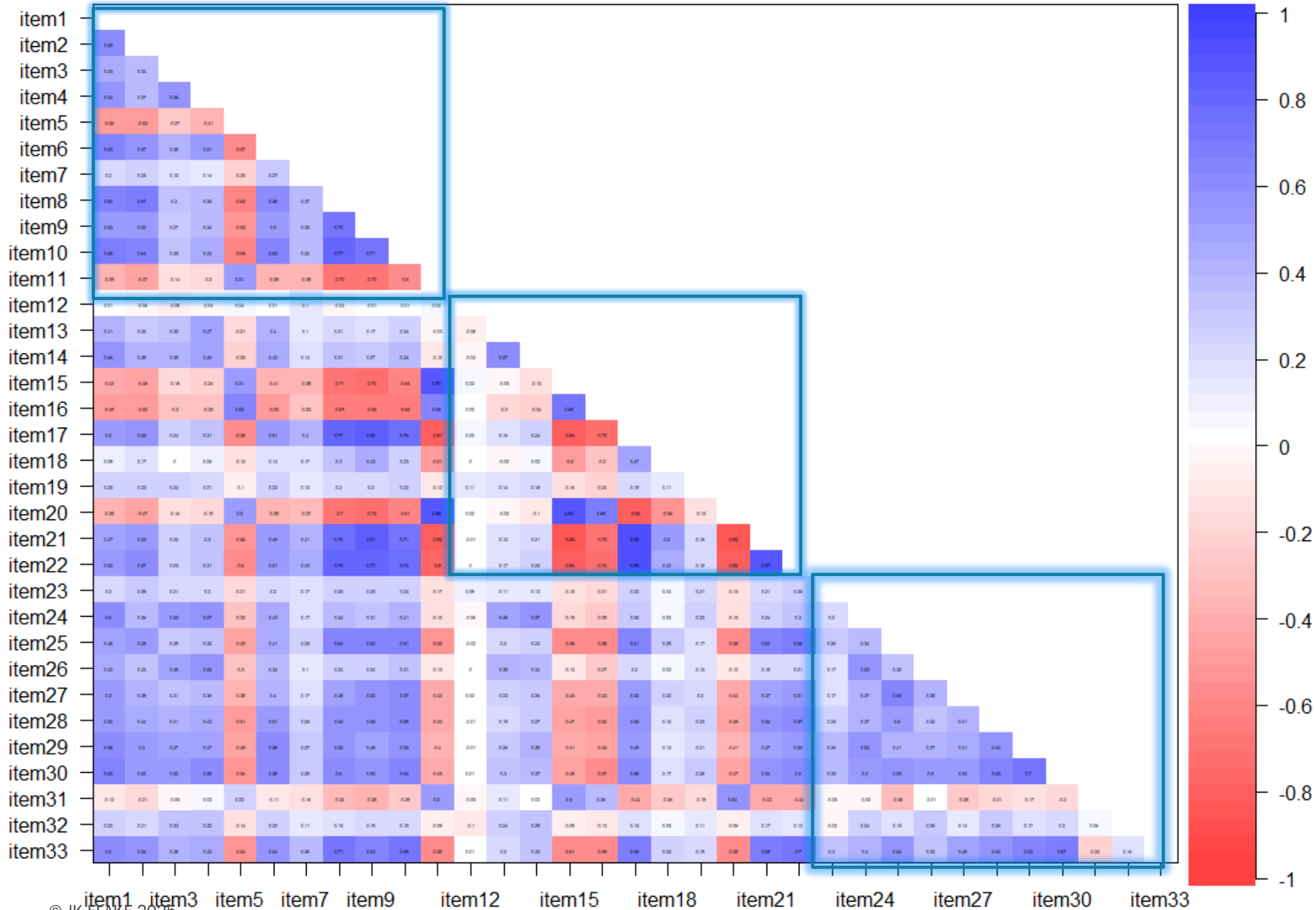
```
items <- select(DATASET, starts_with("item"))
```

```
corPlot(cor(items), numbers = T, colors = T, upper = F, diag = F)
```



### 3. CORRELATIONS

- If the items are measuring the same thing, they should be strongly correlated
  - Negatively worded items should be strongly negatively correlated to positively worded items
- If the instrument has facets or subfactors, items within a facet should be more strongly correlated than across facets
  - PF pilot work proposed three facets, e.g., evangelism items should be more strongly correlated to one another than to pickle liking items
- Generally you want to...
  - Identify items that are weakly correlated
  - Identify clusters of items that are strongly correlated
  - Are negatively worded items correlated as you would expect?



I12: Pickles are made with vinegar  
 I18: I don't mind pickles  
 I23: I have many friends who like pickles  
 I26: I don't like my friends who don't like pickles  
 I31: My friends should not eat pickles  
 I32: I'm secretive about my pickle habits



## NEXT SECTION ON FACTOR ANALYSIS

- After having reviewed the items qualitatively and quantitatively, you're ready for quantitative psychometric methods
- Will not get to this in this workshop, slides for reference



# OPEN OFFICE HOURS AND ACTIVITIES

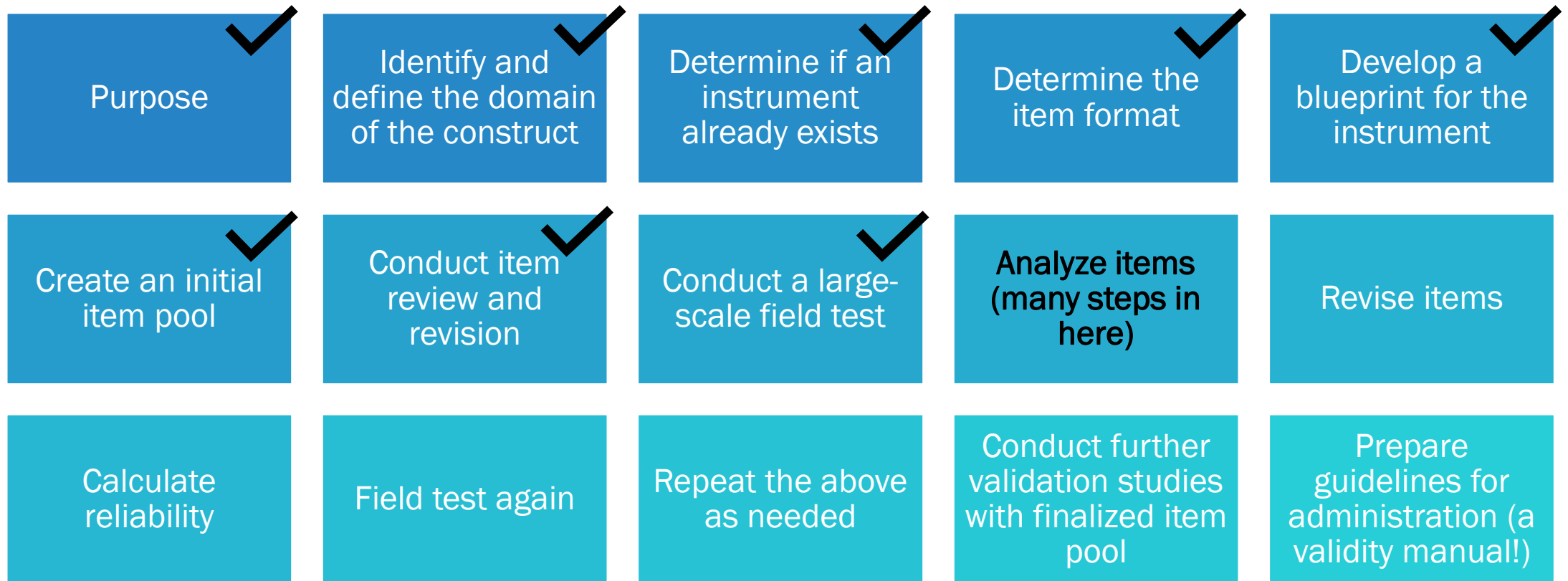
- Afternoon can be used for open office hours and discussion
- Time to review instruments you're using in your research
  - How do they adhere to guidelines?
  - What different sources of evidence are there for them?
  - What could you learn from item review or thinkaloud protocols?
- If you don't use many survey style instruments, can anything you learn today be applied to your kind of measure?



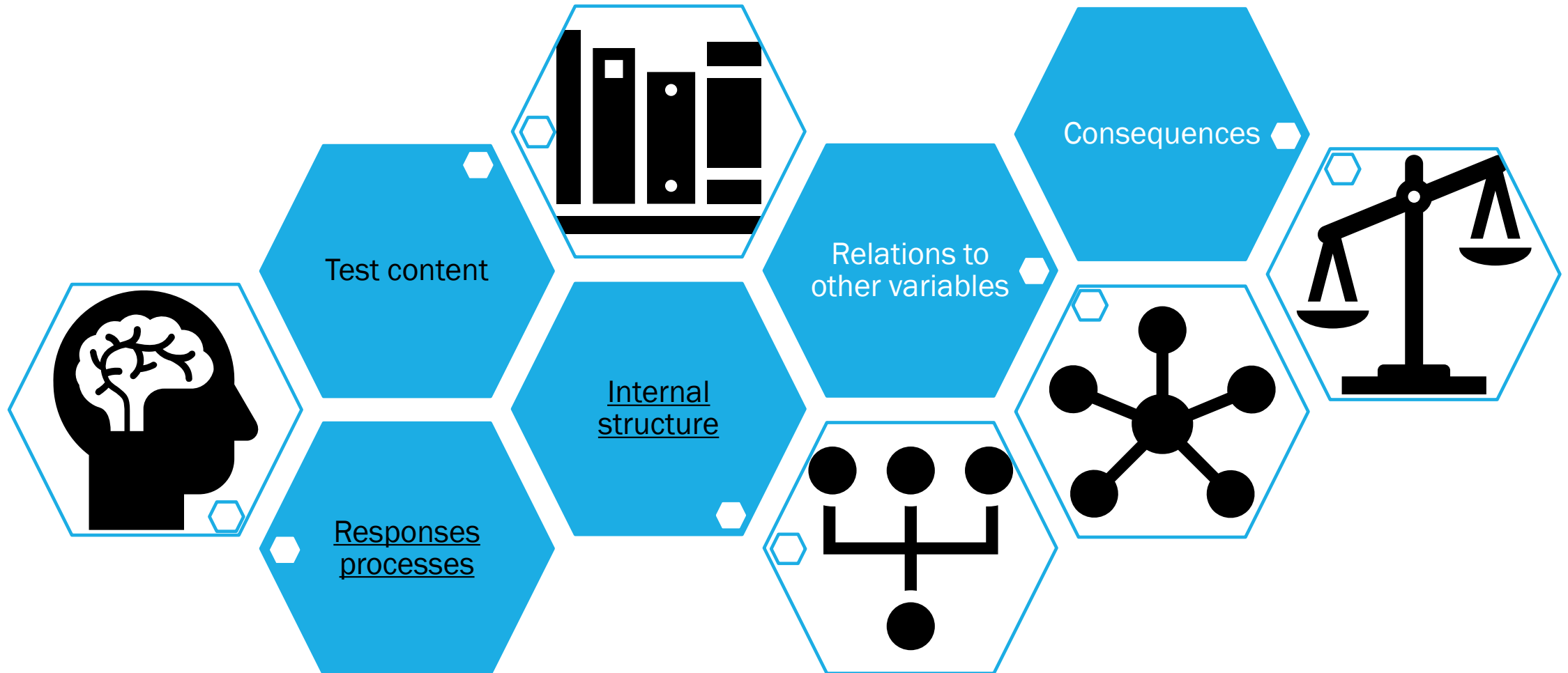
# WHERE HAVE WE BEEN AND WHERE ARE WE GOING?

- We've been following a scale development process to gather different types of validity evidence for the use of an instrument that measures a latent construct
- Our example is measuring pickle fanaticism, a construct a pickle company is interested to use to recruit pickle influencers so they can increase sales on the global market

# INSTRUMENT DEVELOPMENT PROCESS STATUS



# WHERE WE HAVE BEEN





## SUMMARIZING THE EVIDENCE FOR THE PF SCALE

- Take a few minutes to talk to your table about each type of evidence we've considered so far for PF
  - Content
  - Response Processes
  - Internal Structure
- What is the evidence against using this scale to identify pickle influencers and ultimately increase sales on the global market?
- What is the evidence in favor of the intended use and interpretation?



# TINY RECAP ON FACTOR ANALYSIS

- The point of doing a factor or components analysis is to test the hypothesis that you are measuring the number of somethings you intend
- We wrote items to measure three dimensions of pickle fanaticism
  - We can use factor analysis to test if the data exhibit properties of three somethings
- Factor analysis is based on covariance (i.e., correlation) with the assumption that if items measure the same thing, they will be highly correlated
- Factor and components analysis extracts from the data groups of highly correlated variables

# CHOOSING THE RIGHT EVIDENCE FOR YOUR PURPOSE

## CFA

You are measuring latent constructs that cause the item responses

You have the factors and items mapped out ahead of time and want to test if that measurement models fits the data (you've done this before)

## EFA

You are measuring latent constructs that cause the item responses

You want the maths to decide which items form which factors (this might be your first go)

## PCA

You aren't measuring a latent construct, but have a lot of variables and you'd like to reduce them

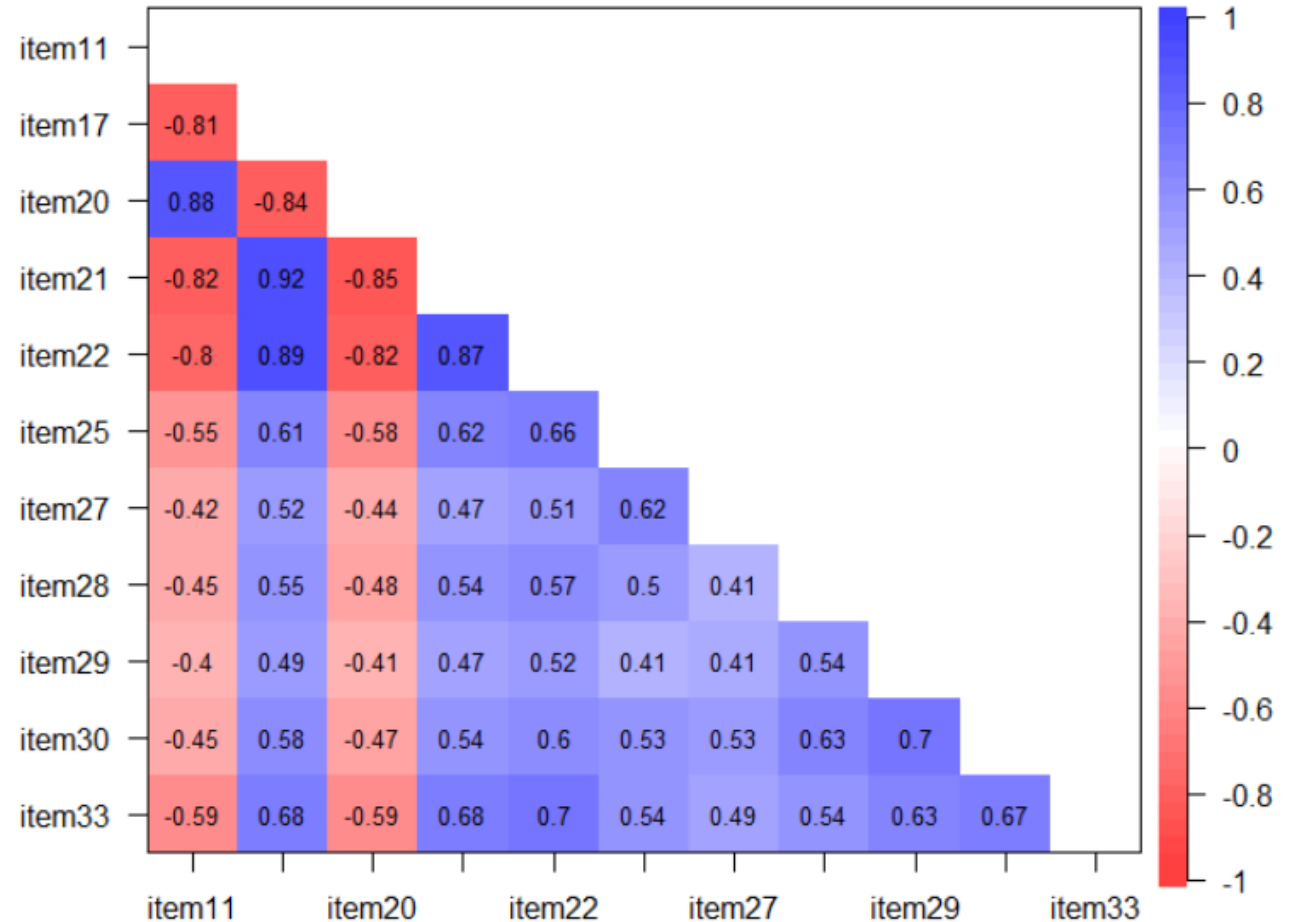
You want the maths to decide which items/variables form which components

# LET'S DIG INTO THE STEPS OF AN EFA

- Extract factors
  - Common sense,  $E < 1$ , Scree plot, parallel analysis
- Determine the number of factors to retain
  - Evaluate solution for interpretability
- Rotate the factors
  - Select a rotation method that will simplify the loading matrix, one that allows the factor to be correlated
- Interpret final factors
  - Think deeply
- Assumptions
  - Do this at the beginning
- Replicate and confirm the results

# COMMON SENSE, WHAT'S GOING ON WITH THE SMALLER SET OF ITEMS?

Pickle Liking	Pickle Evangelism
11. I avoid eating pickles	25. My family should eat pickles regularly
17. Pickles taste extremely good	27. Everyone should eat pickles
20. Pickles taste bad	28. Pickles are great gifts
21. Pickles are delicious	29. I like to tell people about my favorite pickles
22. I like pickles a lot	30. I want to share my love of pickles with the world
	33. I recommend pickles to people I know



# FIRST WE HAVE TO RUN A PARALLEL ANALYSIS TO DECIDE HOW MANY FACTORS TO EXTRACT

```
***** PARALLEL ANALYSIS COMMANDS
```

```
set mxloops=9000 printback=off width=80 seed = 1953125.  
matrix.
```

\* Enter the name/location of the data file for analyses after "FILE =";  
If you specify "FILE = \*", then the program will read the current,  
active SPSS data file; You can alternatively enter the name/location  
of a previously saved SPSS systemfile instead of "\*";  
you can use the "/ VAR =" subcommand after "/ missing=omit"  
subcommand to select variables for the analyses.

```
GET raw / FILE = "C:\Users\Jessica Flake\OneDrive - McGill University\Measurement Workshop\Data Demos\Clean Pickles.sav"  
/var item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33.
```

\* Enter the desired number of parallel data sets here.  
**compute ndatsets = 1000.**

\* Enter the desired percentile here.  
**compute percent = 95.**

\* Enter either  
1 for principal components analysis, or  
2 for principal axis/common factor analysis.  
**compute kind = 1 .**

\* Enter either  
1 for normally distributed random data generation parallel analysis, or  
2 for permutations of the raw data set (VERY time consuming).  
**compute randtype = 1.**

\* End of required user specifications.

There is not a procedure for this in SPSS, but there is a published macro. You paste it and fill in the blanks.

# R PACKAGES AND CODE

## 1. Load the required packages

```
library(ggplot2)
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --
## v tibble 3.1.0      v dplyr 1.0.5
## v tidyr 1.1.3      v stringr 1.4.0
## v readr 1.4.0     v forcats 0.5.1
## v purrr 0.3.3

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()

library(psych)

##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
## %+%, alpha

library(GPARotation)
```

The psych package in R can do nearly all the psychometric analyses we are talking about

# THE PARALLEL ANALYSIS IN R

```
summary(fa.parallel(finItems, (10, 10)))
```

There are a variety of other methods to determine the number of factors to extract. We will consider three common ones, and their PROs and CONS: Eigenvalues greater than 1 rule, a scree plot, and a parallel analysis.

**6. We can get this information by using the parallel analysis option in the psych package.**

```
PA <- fa.parallel(finItems)
```

# INTERPRET PARALLEL ANALYSIS OUTPUT

Run MATRIX procedure:

PARALLEL ANALYSIS:

Principal Components & Random Normal Data Generation

Specifications for this Run:

Ncases 563  
Nvars 11  
Ndatasets 1000  
Percent 95

Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues

Root	Raw Data	Means	Prcntyle
1.000000	7.047812	1.229806	1.288911
2.000000	1.212099	1.165778	1.206607
3.000000	.746199	1.115058	1.151352
4.000000	.505507	1.072044	1.103465
5.000000	.352545	1.032175	1.063049
6.000000	.318094	.994711	1.021822
7.000000	.273684	.958221	.987247
8.000000	.230975	.920124	.948673
9.000000	.133015	.882685	.912804
10.000000	.107250	.839964	.876326
11.000000	.072819	.789434	.832459

----- END MATRIX -----

What you get with your data

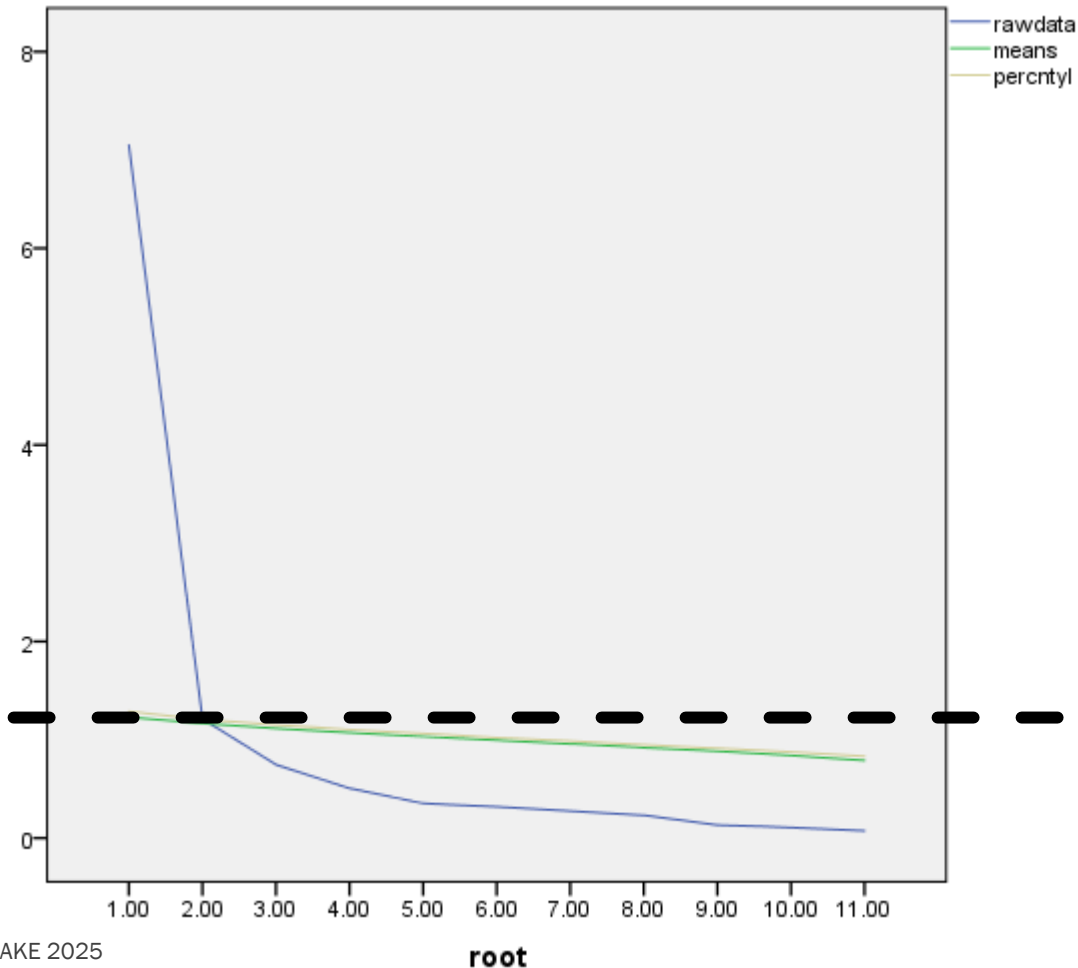
The mean of the distribution of 1000 fake datasets

The 95<sup>th</sup> percentile of the distribution of 1000 fake datasets

We are looking to see where our data produces larger numbers than the faked data – suggests a two factor solution



# INTERPRET PARALLEL ANALYSIS OUTPUT - SCREEPLOT



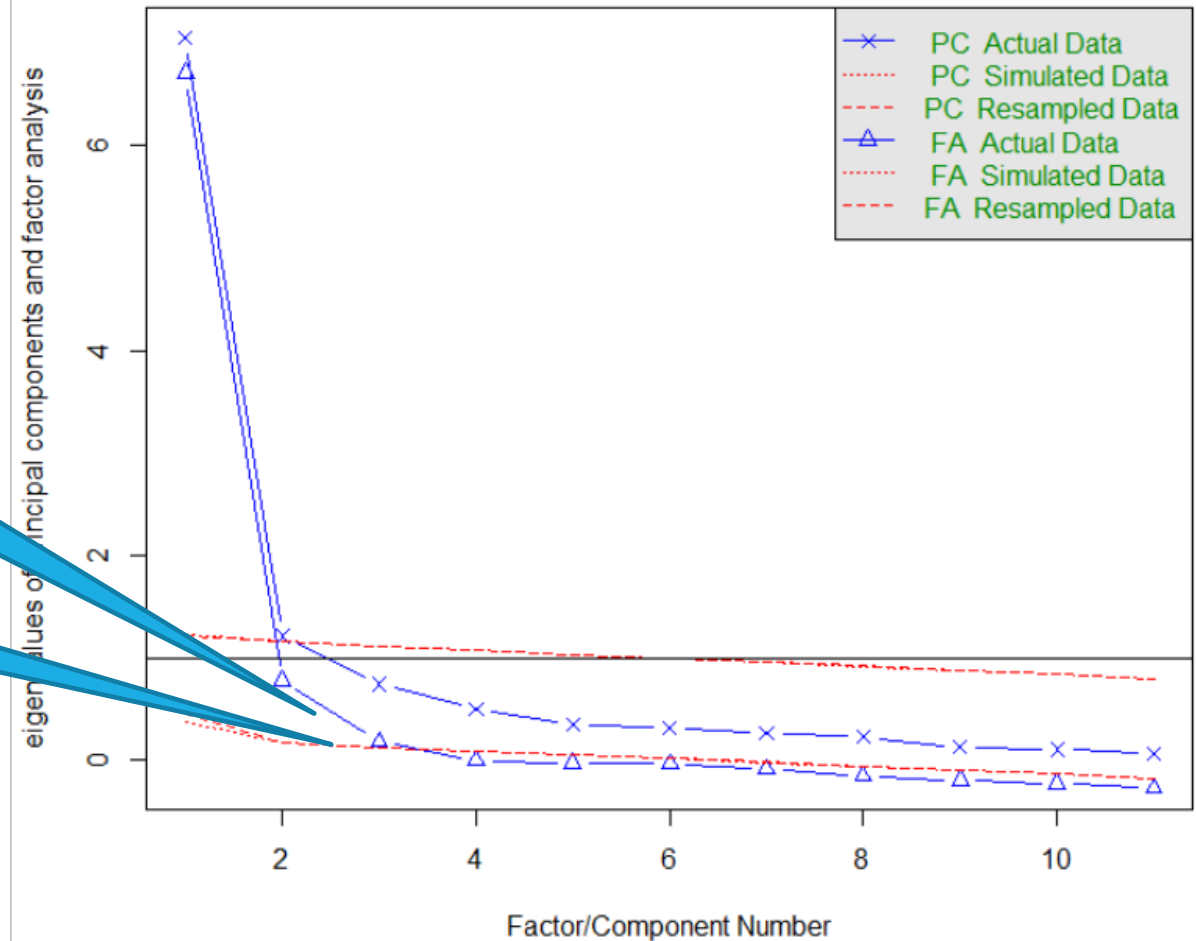
Screeplot interpretation is subjective, generally you want to retain factors before the elbow, excluding factors that form a flatline and are “the scree”  
Our data show an elbow at 2 factors, then a leveling off  
Our data also “beats” the random data above the elbow

# WHAT DOES THE OUTPUT LOOK

PF data using a FA

Random FA data

### Parallel Analysis Scree Plots



## PF DATA – WHAT ARE YOUR THOUGHTS?

- Methods can conflict
- Differences between software can produce slightly different numbers
  - Estimation differences between programs can produce different eigenvalues
    - In R the eigenvalue for the 3<sup>rd</sup> factor is .05 larger than the 3<sup>rd</sup> factor from the random
- Determine a range of plausible factors and then look to the interpretability of the solution to decide the final one
- For these data, looks like we should consider 2 and 3 factor solutions
  - PA suggests two
  - We do one extra to make sure we don't miss anything
  - Under extraction – not extracting enough factors is hard to detect, but over extraction, extracting too many, is easy to spot (I'll show you what I mean soon!)

# EFA; SPSS EXAMPLE

Clean Pickles.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform **Analyze** Graphs Utilities Extensions Window Help

Reports  
Descriptive Statistics  
Compare Means  
General Linear Model  
Generalized Linear Models  
Mixed Models  
Correlate  
Regression  
Loglinear  
Classify  
**Dimension Reduction** ▶ Factor...  
Scale  
Nonparametric Tests  
Forecasting  
Survival  
Multiple Response  
Simulation...  
Quality Control  
ROC Curve...  
Spatial and Temporal Modeling...

	V1	item5	item6
1	1		
2	2		
3	3		
4	4		
5	5		
6	6		
7	7		
8	8		
9	9		
10	10		
11	11		
12	12		
13	13		
14	14		
15	15		
16	16		
17	17		
18	18		
19	19		
20	20		
21	21		
22	22		

IBM SPSS Statistics Data Editor

Transform Analyze Graphs Utilities Extensions Window Help

	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10	item11	item12
1	1	2	1	1	5	1	3	1	2	1	2	5
2	2	4	1	2	3	2	4	4	5	4	1	5
3	1	1	1	1	5	1	1	1	1	1	5	5
4							4	2	3	1	1	2
5							5	3	4	2	1	5
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18	1	1	1	1	5	1						
19	1	1	1	1	5	1						
20	2	1	1	1	4	2						
21	1	1	5	5	5	1						
22	2	3	1	1	5	1						
23	2	4	1	1	4	2						
24	2	1	1	1	4	1						
25	1	5	1	1	5	1						

Factor Analysis

Variables: [V1]

Selection Variable:

Value:

OK Paste Reset Cancel Help

Factor Analysis: Extraction

Method: Principal components

Analyze: Principal components  
Unweighted least squares  
Generalized least squares  
Maximum likelihood  
Principal axis factoring

Extract:  Base  
Alpha factoring  
Image factoring

Eigenvalues greater than: 1

Fixed number of factors  
Factors to extract:

Maximum Iterations for Convergence: 25

Continue Cancel Help

# SPSS DEFAULTS TO KEEP AN EYE ON

- Extraction method default is PCA
  - For an EFA select 'principal axis factoring'
- Extract number of factors default is Eigenvalues > 1
  - Run a parallel analysis first, manually input the factors you want
- Rotation default method is None
  - This assumes your factors are not correlated
  - Oblimin with delta = 0
    - Negative delta makes the factors less correlated, positive delta makes the factors more correlated (max .80). 0 is suggested in methods papers, we trust them

```
DATASET ACTIVATE DataSet2.
```

```
FACTOR
```

```
/VARIABLES item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/MISSING LISTWISE  
/ANALYSIS item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/PRINT INITIAL KMO EXTRACTION ROTATION  
/PLOT EIGEN  
/CRITERIA MINEIGEN(1) ITERATE(25)  
/EXTRACTION PAF  
/CRITERIA ITERATE(25) DELTA(0)  
/ROTATION OBLIMIN  
/METHOD=CORRELATION.
```

\*this code shows how to change the number of factor option for manual specification, this specifies 2.

\*this produces the same output as above because 2 factors have an EV > 1

This is the default, Eigenvalues > 1 rule

```
DATASET ACTIVATE DataSet2.
```

```
FACTOR
```

```
/VARIABLES item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/MISSING LISTWISE  
/ANALYSIS item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/PRINT INITIAL KMO EXTRACTION ROTATION  
/PLOT EIGEN  
/CRITERIA FACTORS(2) ITERATE(25)  
/EXTRACTION PAF  
/CRITERIA ITERATE(25) DELTA(0)  
/ROTATION OBLIMIN  
/METHOD=CORRELATION.]
```

\*this code produces three factors for comparison.

This is manual specification of the factors, this specifies 2  
I also manually selected oblimin rotation and principal axis factoring

```
DATASET ACTIVATE DataSet2.
```

```
FACTOR
```

```
/VARIABLES item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/MISSING LISTWISE  
/ANALYSIS item11 item17 item20 item21 item22 item25 item27 item28 item29 item30 item33  
/PRINT INITIAL KMO EXTRACTION ROTATION  
/PLOT EIGEN  
/CRITERIA FACTORS(3) ITERATE(25)  
/EXTRACTION PAF  
/CRITERIA ITERATE(25) DELTA(0)  
/ROTATION OBLIMIN  
/METHOD=CORRELATION.
```

```
EFA3 <- fa(finItems,3, fm = "pa")  
EFA3
```

Using the Psych package in R you can say "3" factors and fm = PA for principal axis factoring

# EFA TWO FACTOR SOLUTION (ALSO IN YOUR HANDOUT)

## KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.929
Bartlett's Test of Sphericity	Approx. Chi-Square	5936.685
	df	55
	Sig.	.000

Assumptions check – that the data are amenable to EFA.

According to Kaiser, KMO > .90 is marvelous, in .80s meritorious, in .70s middling, in .60s mediocre, in .50s miserable, less than that unacceptable.

Bartlett's tests if the items are uncorrelated, the null states the matrix is an identity matrix, we want to reject the null.

The eigenvalues and the variance explained by the factors. By retaining two factors, we can explain 70% of the variance in the items. If we were to retain 11 factors, we would explain 100% of the variance.

## Communalities

	Initial	Extraction
11.I avoid eating pickles	.805	.808
17.Pickles taste extremely good	.889	.885
20.Pickles taste bad	.836	.856
21.Pickles are delicious	.881	.887
22.I like pickles a lot	.843	.861
25.My family should eat pickles regularly	.567	.511
27.Everyone should eat pickles	.456	.391
28.Pickles are great gifts	.474	.497
29.I like to tell people about my favorite pickles	.544	.582
30.I want to share my love of pickles with the world	.651	.792
33.I recommend pickles to people I know	.632	.654

The proportion of variance in the items that can be explained by all the other items (initial) and then by the factors you extracted (extracted). Items with low communality aren't correlated with the other items and should be considered for removal.

## Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings <sup>a</sup>
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	
1	7.048	64.071	64.071	6.790	61.730	61.730	6.200
2	1.212	11.019	75.090	.935	8.502	70.232	5.391
3	.746	6.784	81.874				
4	.506	4.596	86.469				
5	.353	3.205	89.674				
6	.318	2.892	92.566				
7	.274	2.488	95.054				
8	.231	2.100	97.154				
9	.133	1.209	98.363				
10	.107	.975	99.338				
11	.073	.662	100.000				

Extraction Method: Principal Axis Factoring.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

# EFA TWO FACTOR SOLUTIONS

**Factor Matrix<sup>a</sup>**

	Factor	
	1	2
11.I avoid eating pickles	-.829	.346
17.Pickles taste extremely good	.920	-.197
20.Pickles taste bad	-.857	.349
21.Pickles are delicious	.908	-.249
22.I like pickles a lot	.917	-.139
25.My family should eat pickles regularly	.712	.068
27.Everyone should eat pickles	.599	.180
28.Pickles are great gifts	.657	.257
29.I like to tell people about my favorite pickles	.633	.425
30.I want to share my love of pickles with the world	.738	.498
33.I recommend pickles to people I know	.783	.204

The table of loadings, as if we had not rotated the solution. This solution will assume the factors are not correlated and not do anything to simplify the structure.

**Pattern Matrix<sup>a</sup>**

	Factor	
	1	2
11.I avoid eating pickles	-.955	.087
17.Pickles taste extremely good	.840	.142
20.Pickles taste bad	-.977	.079
21.Pickles are delicious	.893	.071
22.I like pickles a lot	.769	.216
25.My family should eat pickles regularly	.390	.392
27.Everyone should eat pickles	.184	.487
28.Pickles are great gifts	.130	.611
29.I like to tell people about my favorite pickles	-.083	.816
30.I want to share my love of pickles with the world	-.100	.954
33.I recommend pickles to people I know	.277	.597

Extraction Method: Principal Axis Factoring.  
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 8 iterations.

The table of loadings, these are the relationships between the item and the factor, controlling for the other factors in the model. They are interpreted like regression coefficients.

**Structure Matrix**

	Factor	
	1	2
11.I avoid eating pickles	-.896	-.554
17.Pickles taste extremely good	.935	.706
20.Pickles taste bad	-.924	-.577
21.Pickles are delicious	.940	.671
22.I like pickles a lot	.914	.733
25.My family should eat pickles regularly	.653	.654
27.Everyone should eat pickles	.511	.611
28.Pickles are great gifts	.541	.699
29.I like to tell people about my favorite pickles	.466	.760
30.I want to share my love of pickles with the world	.541	.887
33.I recommend pickles to people I know	.678	.783

Extraction Method: Principal Axis Factoring.  
Rotation Method: Oblimin with Kaiser Normalization.

**Factor Correlation Matrix**

Factor	1	2
1	1.000	.672
2	.672	1.000

Extraction Method: Principal Axis Factoring.  
Rotation Method: Oblimin with Kaiser Normalization.

The correlation between the item and the factor, not conditional on the other factors. They are interpreted like zero order correlations. Something to note here is that all items are strongly correlated to both factors, this is consistent with the strong correlation between the two factors.



# THE INTERPRETABILITY OF THE SOLUTION

- We want a factor with at least 2 or 3 items that have strong loadings (>.32, 10% of the variance is explained by the factor)
- We want factors that are NOT mostly comprised of crossloaded items
  - Crossloaded items should be considered for removal because they are hard to score
- We do not want bloated specifics or methods factors – small factors that are only made up of wording or methods effects\*
  - If we over extracted, we would have little junk factors, this would suggest trying one less factors
  - If we under extract, we can't tell, this is why you should always take a look at a solution with one extra factor
- We want factors that make sense
- Take 5 (RP3: #13), what are some take-aways from this pattern matrix?

**Pattern Matrix<sup>a</sup>**

	Factor	
	1	2
11.I avoid eating pickles	-.955	.087
17.Pickles taste extremely good	.840	.142
20.Pickles taste bad	-.977	.079
21.Pickles are delicious	.893	.071
22.I like pickles a lot	.769	.216
25.My family should eat pickles regularly	.390	.392
27.Everyone should eat pickles	.184	.487
28.Pickles are great gifts	.130	.611
29.I like to tell people about my favorite pickles	-.083	.816
30.I want to share my love of pickles with the world	-.100	.954
33.I recommend pickles to people I know	.277	.597

Extraction Method: Principal Axis Factoring.  
 Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 8 iterations.

# TRY INTERPRETING THE 3 FACTOR ON YOUR OWN (TAKE 5, RP3 #15)

The eigenvalues and the variance explained by the factors. By retaining three factors, we can explain 74% of the variance in the items. 4% more than the previous solution. The question will now be: 'does the factor make enough sense to keep for explain 4% more of the variance?' The left side of this graph is unchanged.

**Total Variance Explained**

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings <sup>a</sup>
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	
1	7.048	64.071	64.071	6.822	62.014	62.014	6.159
2	1.212	11.019	75.090	.954	8.669	70.683	4.839
3	.746	6.784	81.874	.404	3.670	74.353	4.870
4	.506	4.596	86.469				
5	.353	3.205	89.674				
6	.318	2.892	92.566				
7	.274	2.488	95.054				
8	.231	2.100	97.154				
9	.133	1.209	98.363				
10	.107	.975	99.338				
11	.073	.662	100.000				

Extraction Method: Principal Axis Factoring.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

**Pattern Matrix<sup>a</sup>**

	Factor		
	1	2	3
11.I avoid eating pickles	-.955	.059	.028
17.Pickles taste extremely good	.833	.114	.047
20.Pickles taste bad	-.961	.071	-.008
21.Pickles are delicious	.894	.068	.011
22.I like pickles a lot	.742	.146	.116
25.My family should eat pickles regularly	.081	-.087	.882
27.Everyone should eat pickles	-.035	.134	.641
28.Pickles are great gifts	.140	.498	.140
29.I like to tell people about my favorite pickles	.000	.855	-.071
30.I want to share my love of pickles with the world	-.053	.809	.149
33.I recommend pickles to people I know	.317	.546	.045

Extraction Method: Principal Axis Factoring.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Take 5 to evaluate this solution, what are some of your take-aways?

- Number of strong items per factor?
- Presence of crossloadings?
- What does the 3<sup>rd</sup> factor represent?
- Under or over extraction?

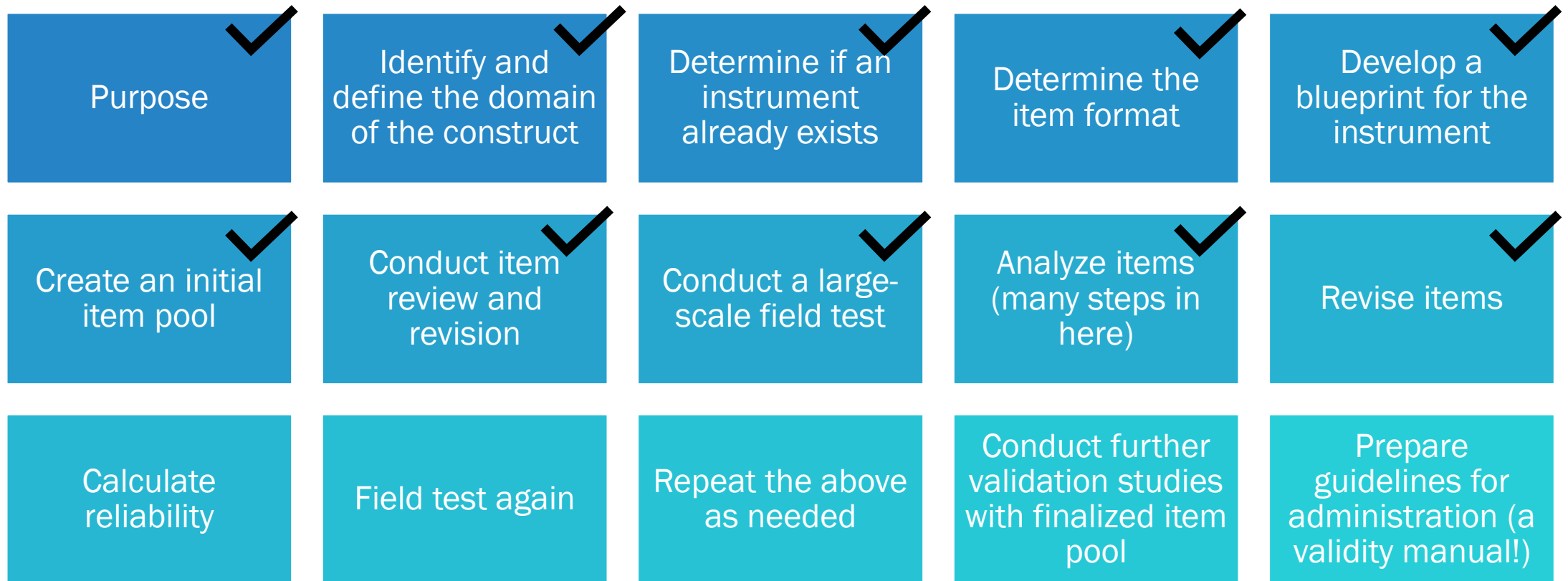
The first factor still looks like pickle liking and the second evangelism. The first only have two strong items, they both have the word "should" in them.

This could be wording effects or capturing a controlling pickle view (not something we wanted to measure anyway).

# EFA CONCLUSIONS

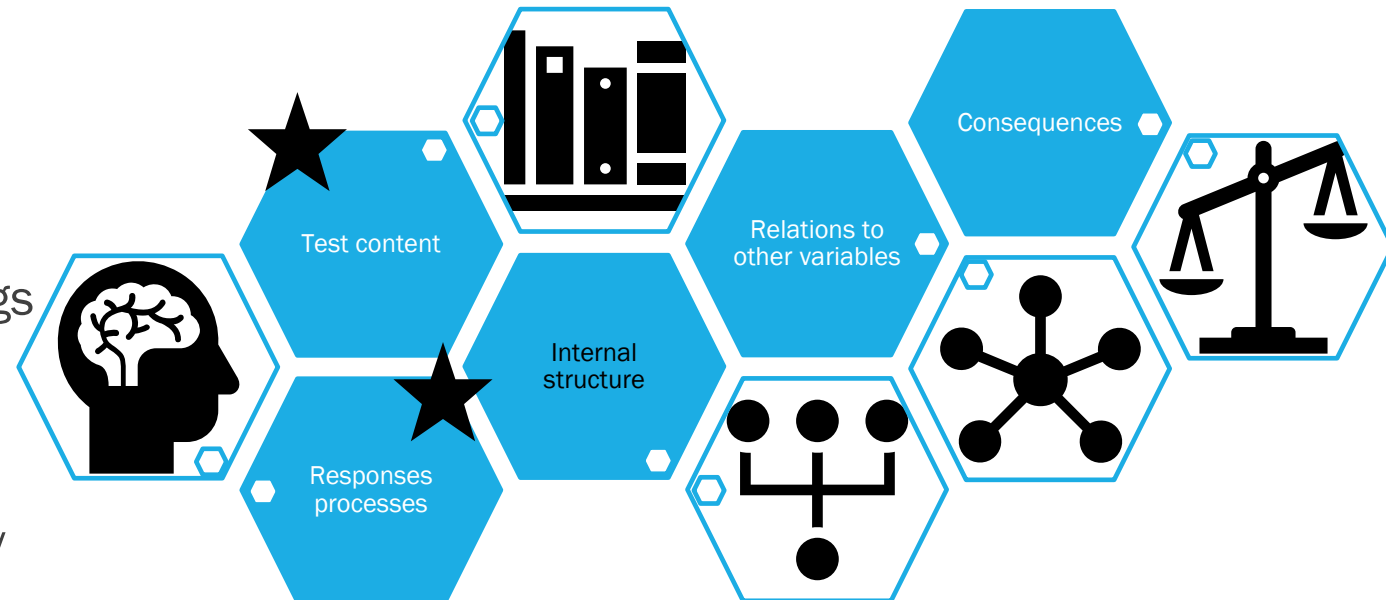
- Take a minute to take some notes about your conclusions from this analysis
- Discuss with your mates, what are the main take-aways from this analysis?
- Two factors seem to do it
  - Pickle liking, pickle evangelism
- Item 25 crossloads and forms a “should” factor with item 27
  - We didn’t intend to measure controlling pickle behavior, so we can nix these items
  - Well, maybe Big Pickle does want to measure this!

# INSTRUMENT DEVELOPMENT PROCESS STATUS FOR PF



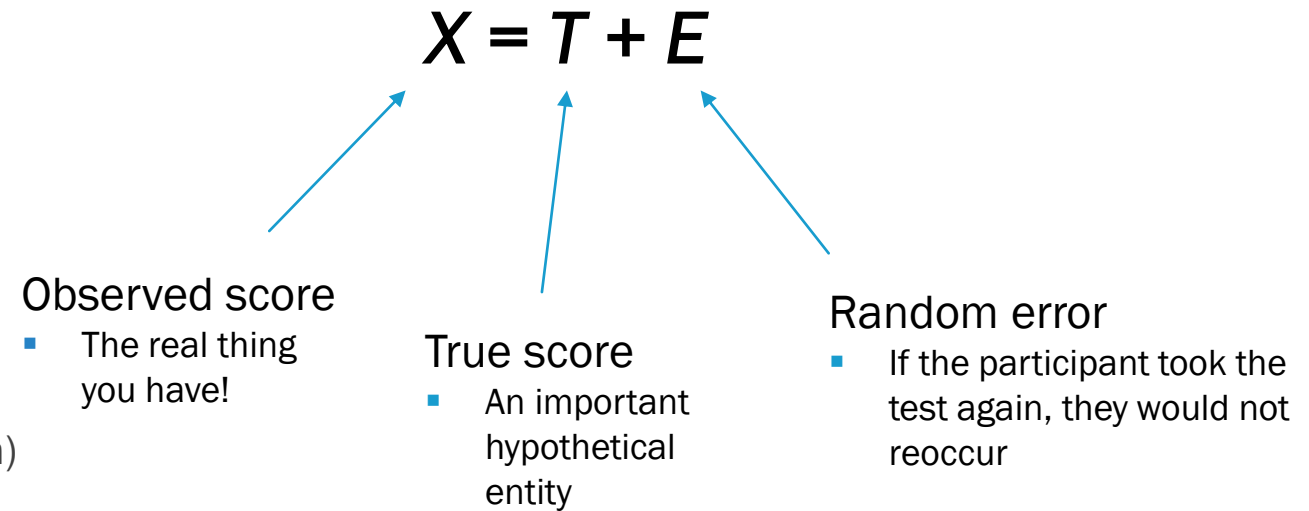
# BACK TO THE VALIDITY EVIDENCE

- Internal structure
  - Item analysis
  - Factor analysis
    - We've considered how many somethings we have
  - Reliability analysis
    - Now we need to know how reliable they are



# RELIABILITY

- Reliability is about consistency
  - If the thing we want to measure is stable (assumption)
  - And if we can do a “good” job of measuring it (assumption)
  - Do we get the same results?
- Reliability is based off of Classical Test Theory
  - The take-away is that test scores vary for two reasons: true differences in the thing being measured and completely random measurement error
  - Reliability is the variance attributable to true differences in the thing, it is the variance in the scores that is NOT measurement error



# METHODS OF ASSESSING RELIABILITY

Different reliability coefficients for different situations: what are you interested in?

- How consistent are responses within an instrument?
  - Internal consistency
- How consistent are scores on the same instrument at different times?
  - Test-retest reliability, or *coefficients of stability*
- How consistent are scores across different versions of the same instrument?
  - Alternate forms reliability, or *coefficients of equivalence*
- How consistent are raters when they use an instrument for the same person?
  - Inter-rater reliability

# INTERNAL CONSISTENCY

- Measures of internal consistency quantify how reliable the composite score from a set of items is
  - The idea is that the items on a single test can serve as a source of random measurement error
  - The less random error the better, more error means less accurate results and less statistical power
- Random measurement error
  - These errors happen differently across people from aspects of the items and environment
    - E.g., one person misreads a word, another person is in a hurry and doesn't answer thoughtfully, they are **random errors because they aren't caused by anything that is repeatable across people or items**
  - This is in contrast to the idea of systematic measurement error, which we will discuss later when we discuss bias
    - Systematic errors decrease validity (not reliability) because they come from measuring something other than the construct, they are repeatable across items and people
      - Which, in fact, can increase reliability



# MEASURES OF INTERNAL CONSISTENCY

- What measures of internal consistency have you heard of?
  - Alpha
  - Omega
  - Guttman
  - Split-half
- Coefficients have been developed based on the view of a test score as a composite of scores on individual items; items are treated as “mini-forms” of the test
- These measures differ in the assumptions they make about the “mini forms” (items)
  - The assumptions come down to how similar the items are in measuring the construct, if they measure it in the exact same way to totally not at all
  - This can be a bit technical without knowledge in confirmatory factor analysis

# PARALLEL, TAU-EQUIVALENT, AND CONGENERIC MEASURES

Parallel



A slightly broken Celsius thermometer – measures temperature with error

Parallel



Another slightly broken Celsius thermometer – measures temperature with the same amount of error

Tau-equivalent



Another slightly broken Celsius thermometer – measures temperature with a different amount of error

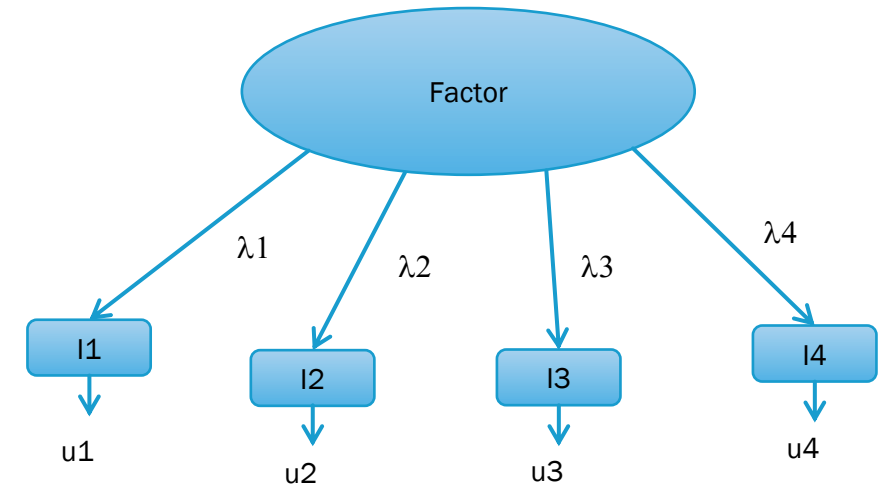
Congeneric



A Fahrenheit thermometer, measures on a different scale

# LEVELS OF EQUIVALENCE IN FACTOR ANALYTIC TERMS

- Parallel
  - Each item has the same amount of error, the  $u$ 's are equal
- Tau-equivalent
  - Each item has a different amount of error, but the loadings are equal
    - Alpha assumes this
    - Omega relaxes this assumptions and can accommodate different loadings
- Congeneric
  - The items measure the same thing, but in a totally different way



# ALPHA IS THE MOST COMMON RELIABILITY METRIC, AND THE MOST HATED

## Coefficient alpha and the internal structure of tests

LJ **Cronbach** - psychometrika, 1951 - Springer

... with those of Clark and **Cronbach** we have studies of ... **Cronbach**, permit him to make more comparable half-tests than would be obtained by random splitting. The data from **Cronbach's** ...

☆ Save [🔗](#) Cite Cited by 55926 Related articles All 18 versions

## [PDF] On the use, the **misuse**, and the very limited usefulness of **Cronbach's alpha**

K **Sijtsma** - Psychometrika, 2009 - Springer

... The goal of this paper is to illuminate the flaws and fallacies that surround both the “common” knowledge base and the practical use of **Cronbach's alpha**, and to provide alternatives. This paper is also meant to invite debate on topics that psychometricians often seem to ...

☆ [🔗](#) Cited by 2586 Related articles All 23 versions

## Uses and abuses of coefficient alpha.

N **Schmitt** - Psychological assessment, 1996 - psycnet.apa.org

The article addresses some concerns about how coefficient alpha is reported and used. It also shows that alpha is not a measure of homogeneity or unidimensionality. This fact and the finding that test length is related to reliability may cause significant misinterpretations of ...

☆ [🔗](#) Cited by 3229 Related articles All 14 versions

## Thanks coefficient alpha, we'll take it from here.

D **McNeish** - Psychological methods, 2018 - psycnet.apa.org

Empirical studies in psychology commonly report Cronbach's alpha as a measure of internal consistency reliability despite the fact that many methodological studies have shown that Cronbach's alpha is riddled with problems stemming from unrealistic assumptions. In many ...

☆ [🔗](#) Cited by 808 Related articles All 9 versions

## Thanks coefficient alpha, we still need you!

T **Raykov**, **GA Marcoulides** - Educational and psychological ..., 2019 - journals.sagepub.com

... (In the remainder of this note, when referring to “reliability” **we will** mean the **coefficient**  $\rho_Y$  defined in Equation 3, unless stated otherwise, and **will** assume that Inequality 2 holds.) Thereby, there is no assumption of continuity or normality of Y, which is needed for this reliability ...

☆ [🔗](#) Cited by 102 Related articles All 5 versions

## MY OPINION ON ALPHA

- Alpha is a good measure of internal consistency
- It is VERY inappropriately used
  - It does NOT measure validity
  - It does NOT provide evidence of dimensionality
  - It is NOT enough for it to be the *only* thing reported
- Omega is also a good measure, it requires running a factor analysis\* (in SPSS now?)
  - It does NOT measure validity
  - It does NOT provide evidence of dimensionality
    - However, you have to run a factor analysis to get it, so it encourages checking this
  - It is NOT enough for it to be the *only* thing reported
    - Encourages reporting on the factor analysis
- Either should be reported with a measure of precision to quantify sampling error
- Either should be reported with validity evidence and the mean inter-item correlation

Next step is to show you how to use alpha and conduct meaningful reliability analysis

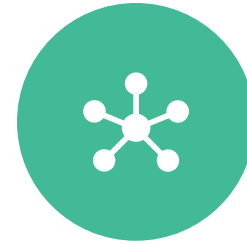
# WHEN YOU DO A RELIABILITY ANALYSIS



**SUBSTANTIVE PHASE** – BUILD THEORY ABOUT WHAT THE CONSTRUCT IS



**STRUCTURAL PHASE** – COLLECT EMPIRICAL EVIDENCE THAT SUPPORT THE ITEMS MEASURE THE CONSTRUCT (PSYCHOMETRICS, ITEM ANALYSIS, FACTOR ANALYSIS)



**EXTERNAL PHASE** – SEE IF CONSTRUCT CONNECTS TO OTHERS AS YOU EXPECT, TEST ASSOCIATIONS TO OUTCOMES AND THEORETICALLY RELEVANT CONSTRUCTS



Reliability quantifies how consistent people respond to items measuring a certain something. You need strong evidence you are measuring that certain something **FIRST**. Do a reliability analysis **AFTER** you've done substantive and factor analytic work on the instrument, not before or in place of it.