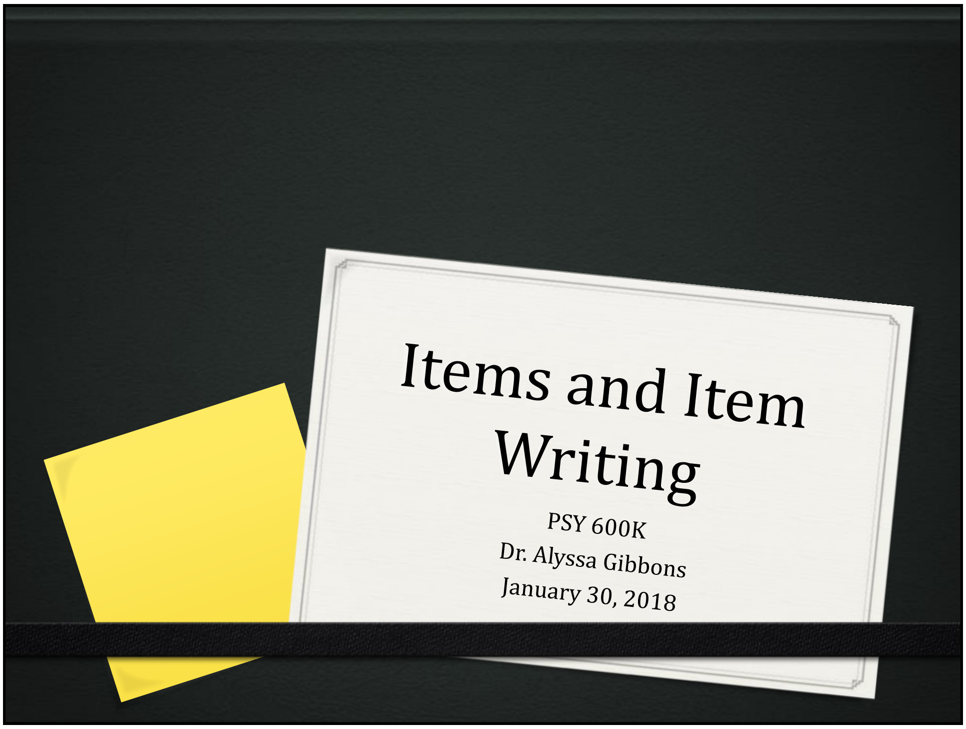
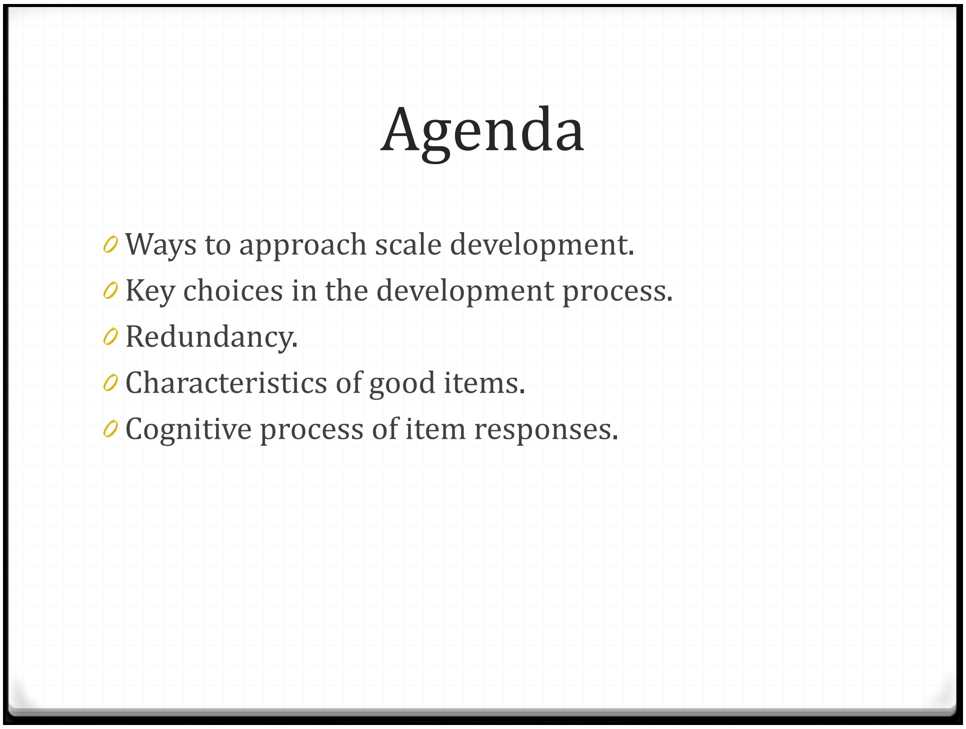


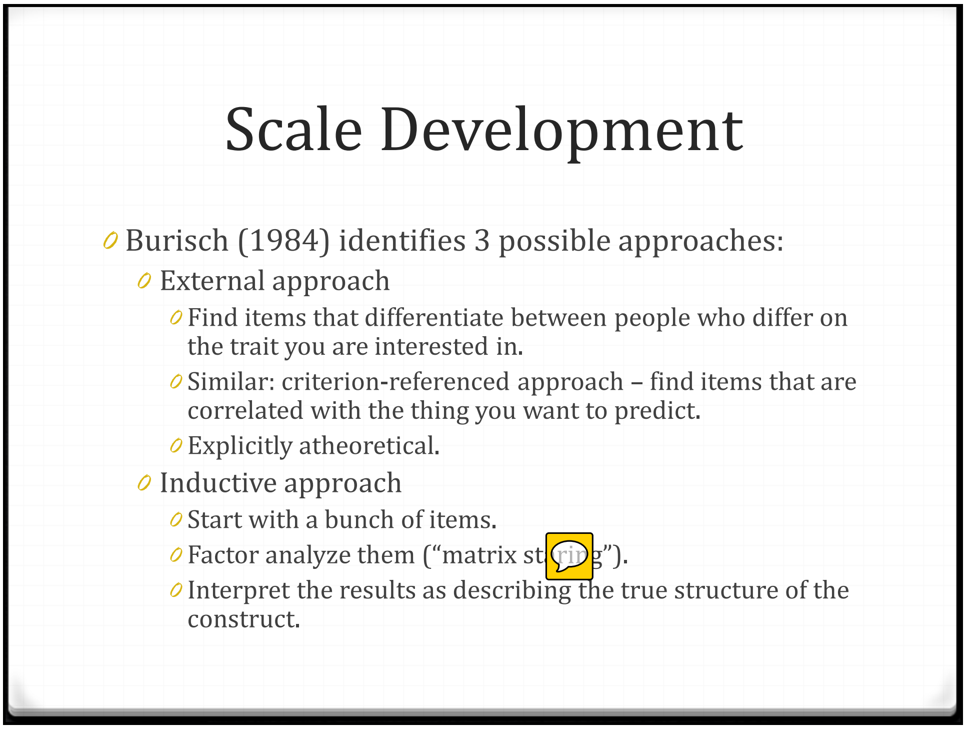
Items and Item Writing

PSY 600K
Dr. Alyssa Gibbons
January 30, 2018



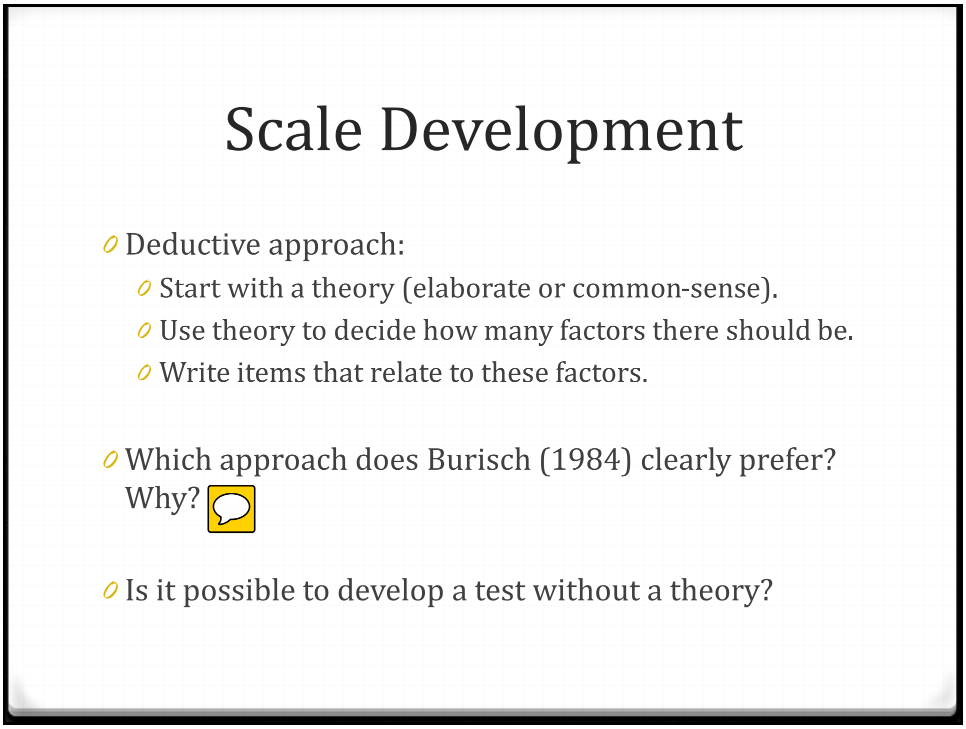
Agenda

- Ways to approach scale development.
- Key choices in the development process.
- Redundancy.
- Characteristics of good items.
- Cognitive process of item responses.



Scale Development

- Burisch (1984) identifies 3 possible approaches:
  - External approach
    - Find items that differentiate between people who differ on the trait you are interested in.
    - Similar: criterion-referenced approach – find items that are correlated with the thing you want to predict.
    - Explicitly atheoretical.
  - Inductive approach
    - Start with a bunch of items.
    - Factor analyze them ("matrix staring").
    - Interpret the results as describing the true structure of the construct.



Scale Development

- Deductive approach:
  - Start with a theory (elaborate or common-sense).
  - Use theory to decide how many factors there should be.
  - Write items that relate to these factors.

- Which approach does Burisch (1984) clearly prefer? Why?

- Is it possible to develop a test without a theory?

Be Intentional

- If you have no idea whatsoever about the probable structure of your measure, you're probably not ready to write it.
  - DeVellis: "The point is that scale developers should make this determination as an active decision and not merely generate a set of items and then see what they look like after the fact." (p. 75)
  - If there's not much theory, you need to start on one.
- Know the literature.
  - Those who do not remember the past are destined to republish it.
  - Learn from others' mistakes.

Key Choices

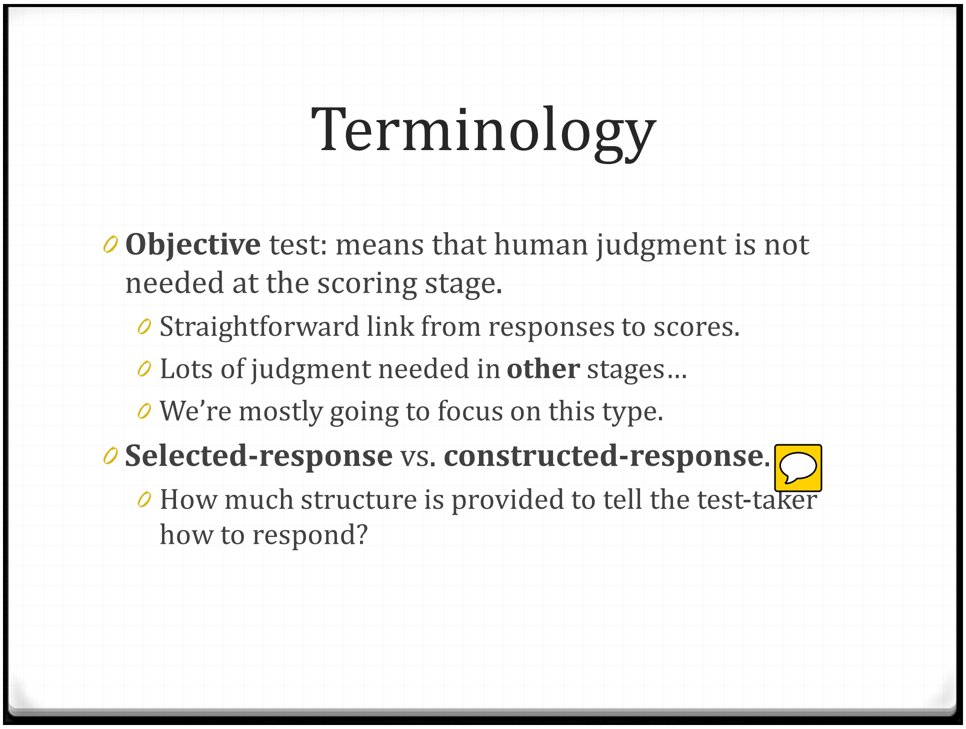
- Level of specificity:
  - Many constructs can be unidimensional or multidimensional… depending on your purpose.
  - Example: How many factors of intelligence are there?
    - Spearman: 2
    - Thurstone: 7
    - Guilford: 150!
Carroll: 9, in a hierarchical model
- Should we be trying to "discover" the "true" structure of our measures?

Key Choices

- What is and is not relevant?
  - Important to identify what does **not** belong in your scale.
  - DeVellis example: depression & physical health.
  - Other examples?
- Purpose
  - What do you want to be able to say about people after they complete your measure?
  - What part of the latent continuum are you most interested in?
- Population
  - Who will take the test? Under what circumstances?

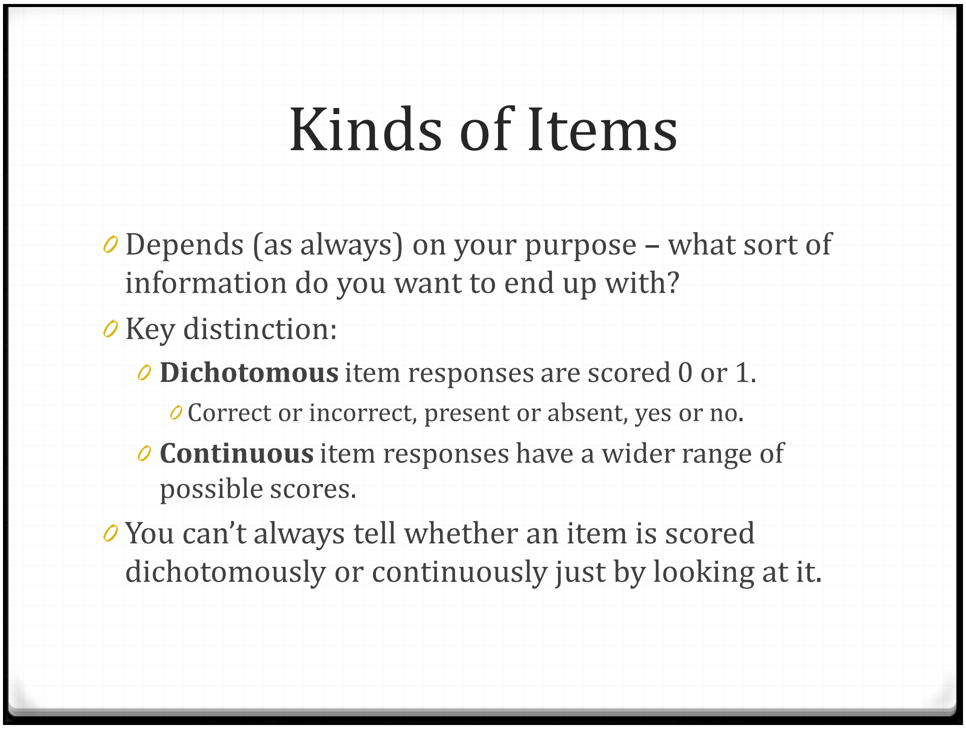
Every Test Needs:

- Item **stems:** the stimuli a test-taker responds to.
- Response options: what the test-taker can do in response to an item stem.
  - Minimum: 2
  - Maximum: infinity (or constrained by the scale developer).
- A method for assigning item scores (numbers) to those responses.
  - **Objective** test: means that human judgment is not needed at the *scoring* stage (there is plenty of judgment at other stages!)
  - Of course, not all tests are objective.
- A method for combining the item scores into test scores.



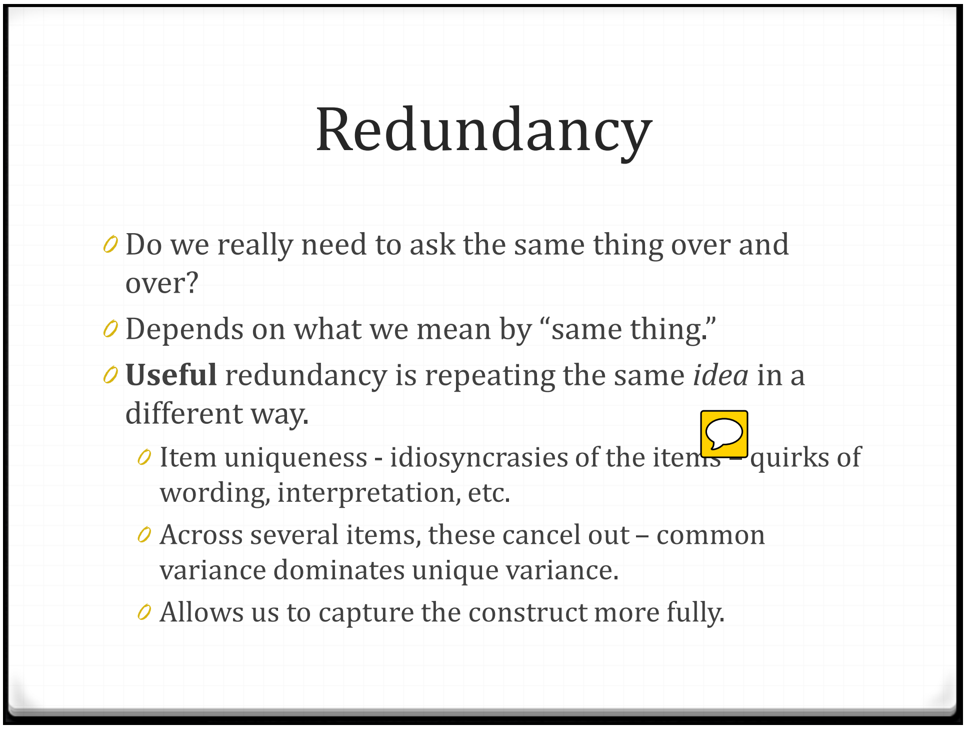
Terminology

- **Objective** test: means that human judgment is not needed at the scoring stage.
  - Straightforward link from responses to scores.
  - Lots of judgment needed in **other** stages…
  - We're mostly going to focus on this type.
- **Selected-response** vs. **constructed-response**.
  - How much structure is provided to tell the test-taker how to respond?



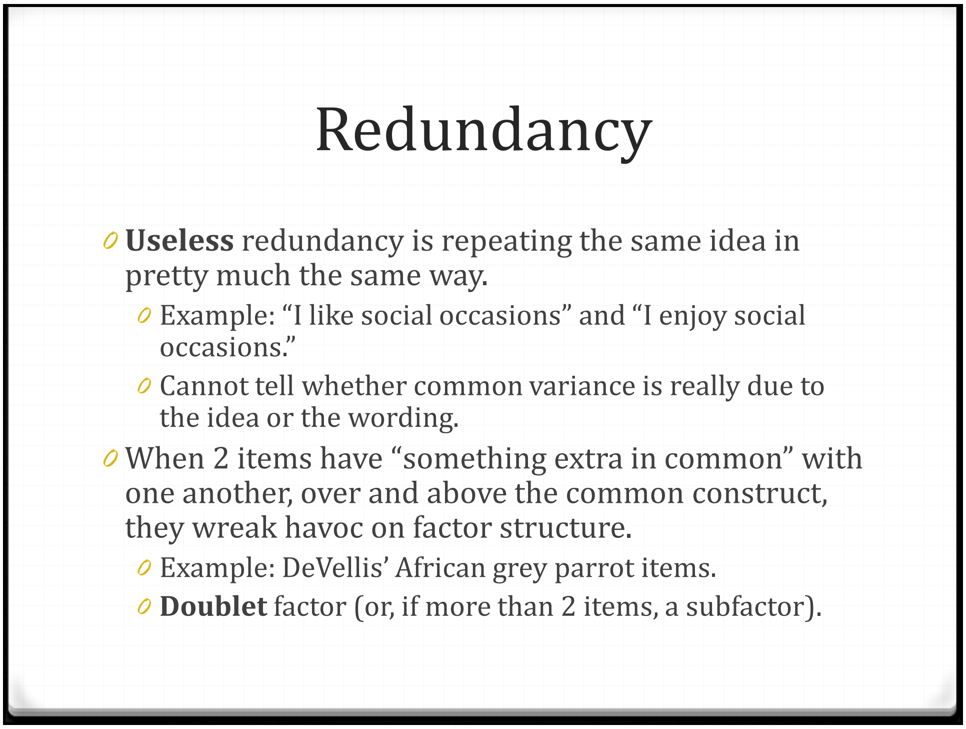
Kinds of Items

- Depends (as always) on your purpose – what sort of information do you want to end up with?
- Key distinction:
  - **Dichotomous** item responses are scored 0 or 1.
    - Correct or incorrect, present or absent, yes or no.
  - **Continuous** item responses have a wider range of possible scores.
- You can't always tell whether an item is scored dichotomously or continuously just by looking at it.



Redundancy

- Do we really need to ask the same thing over and over?
- Depends on what we mean by "same thing."
- **Useful** redundancy is repeating the same *idea* in a different way.
  - Item uniqueness - idiosyncrasies of the items – quirks of wording, interpretation, etc.
  - Across several items, these cancel out – common variance dominates unique variance.
  - Allows us to capture the construct more fully.



Redundancy

- **Useless** redundancy is repeating the same idea in pretty much the same way.
  - Example: "I like social occasions" and "I enjoy social occasions."
  - Cannot tell whether common variance is really due to the idea or the wording.
- When 2 items have "something extra in common" with one another, over and above the common construct, they wreak havoc on factor structure.
  - Example: DeVellis' African grey parrot items.
  - **Doublet** factor (or, if more than 2 items, a subfactor).

How Many Items?

- Absolute minimum: 3 per subscale
  - Mathematical necessity to have a stable factor structure.
- Having more items:
  - Increases reliability (not always for the right reasons).
  - Allows more thorough coverage of the construct.
- Having fewer items:
  - Is usually practical .
  - Prevents respondent fatigue.
- Your **initial** item pool should be 2x – 4x the number of items you hope to end up with.

Characteristics of Good Items

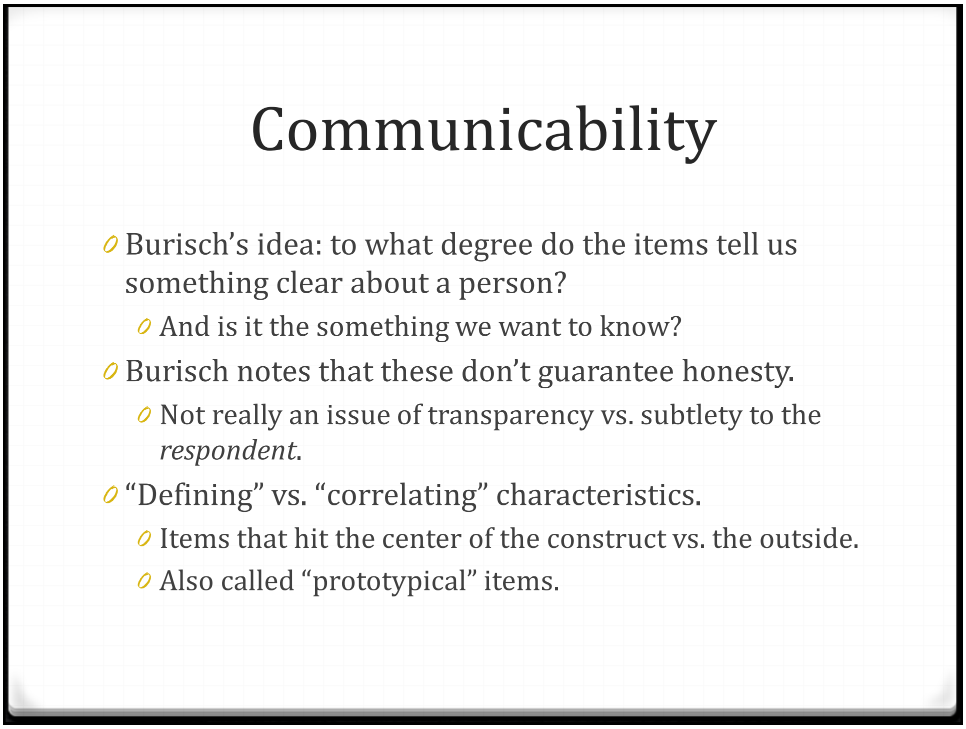
- #1. Short.
  - The more words in an item, the more opportunities there are for a respondent to misread or misinterpret one.
- #2. Readable.
  - Can evaluate reading level if you like.
  - Use simple, everyday language.
  - Avoid jargon.
    - Especially if you don't understand it yourself.

Good Items

- #3. Grammatically correct.
  - Avoid double negatives.
    - Or perhaps avoid negative words altogether.
  - Use adverbs & adjectives correctly.
  - Read for unintended alternative interpretations.
  - Complete sentences are often best.
- #4. Consistent.
  - Use the same referent ("I", "you", etc.) and general structure in all items.
  - Not:
    - "I am a warm and outgoing person."
    - "You like parties."
    - "Sociable."

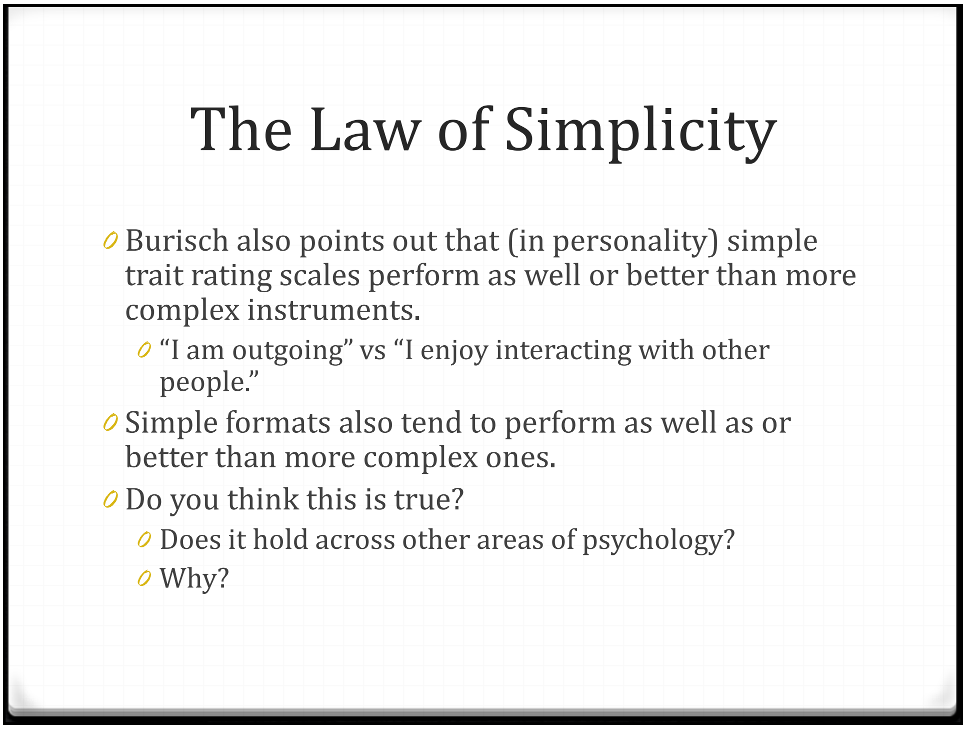
Good Items

- #5. Straightforward
  - Interpreting the question may be obvious to you, but is it obvious to the test-taker?
  - **Double-barreled** items: two questions in one.
    - "I believe graduate students are underpaid and overworked."
    - Does an "agree" response you agree with both parts of the statement? Or only one?
  - Clear relationship to the construct.
    - "I am confident that I am ready to be a parent."
    - Two possible response processes here…



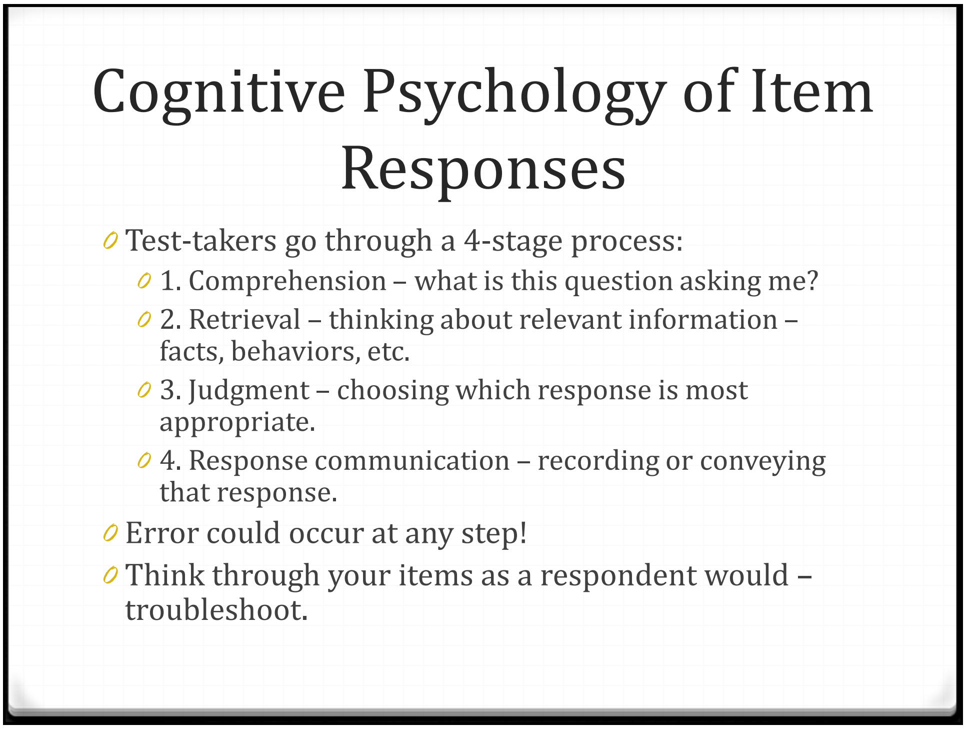
Communicability

- Burisch's idea: to what degree do the items tell us something clear about a person?
  - And is it the something we want to know?
- Burisch notes that these don't guarantee honesty.
  - Not really an issue of transparency vs. subtlety to the *respondent*.
- "Defining" vs. "correlating" characteristics.
  - Items that hit the center of the construct vs. the outside.
  - Also called "prototypical" items.



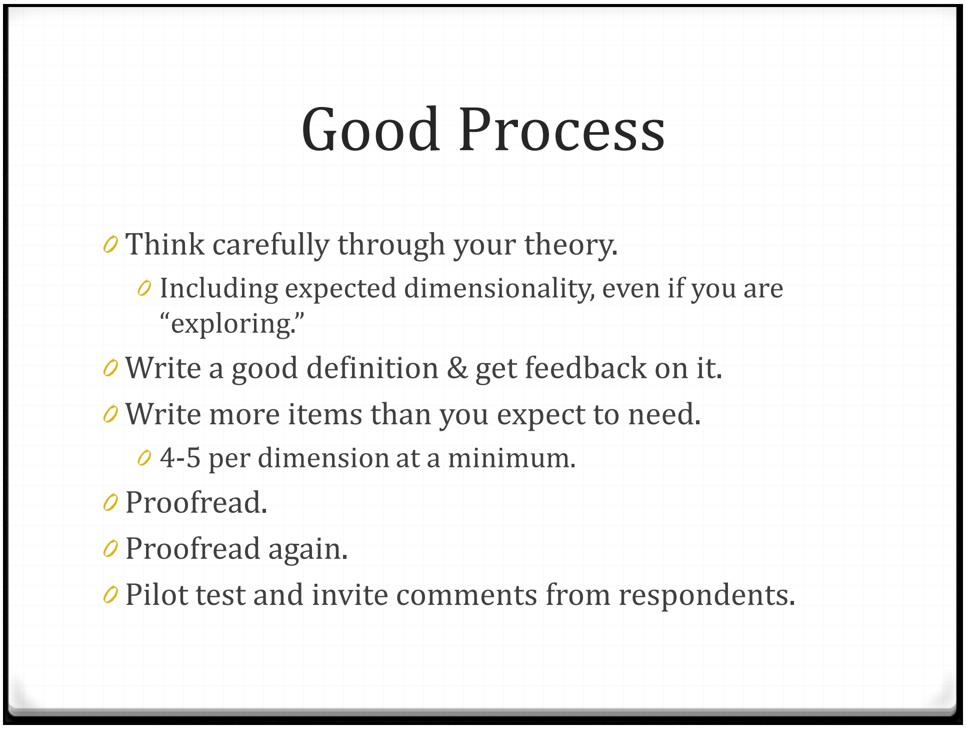
The Law of Simplicity

- Burisch also points out that (in personality) simple trait rating scales perform as well or better than more complex instruments.
  - "I am outgoing" vs "I enjoy interacting with other people."
- Simple formats also tend to perform as well as or better than more complex ones.
- Do you think this is true?
  - Does it hold across other areas of psychology?
  - Why?



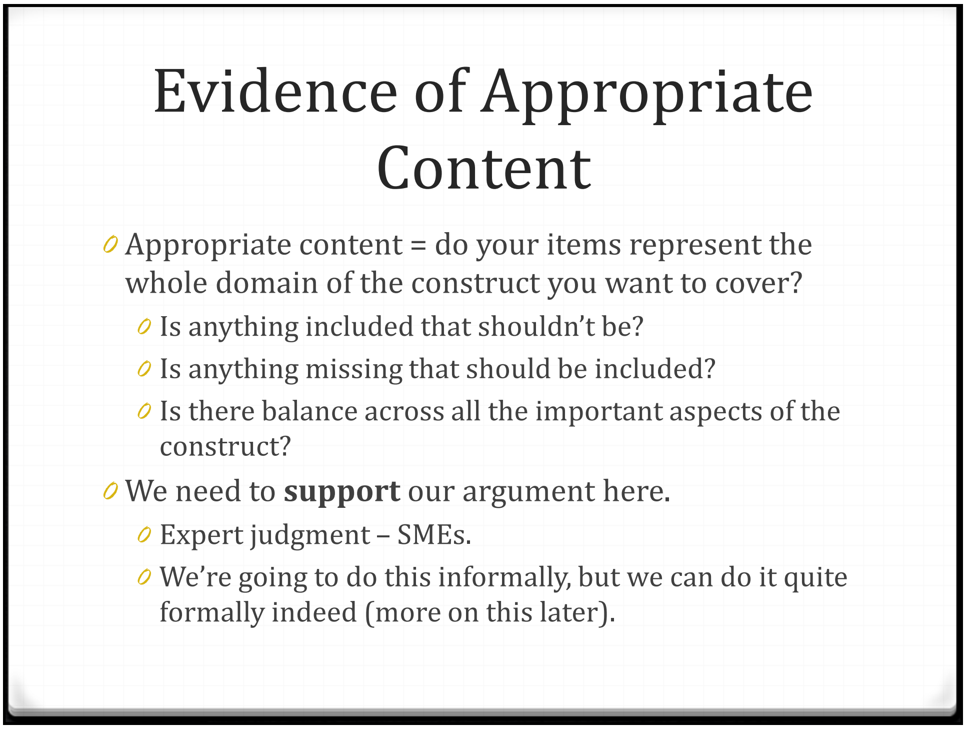
Cognitive Psychology of Item Responses

- Test-takers go through a 4-stage process:
  - 1. Comprehension – what is this question asking me?
  - 2. Retrieval – thinking about relevant information – facts, behaviors, etc.
  - 3. Judgment – choosing which response is most appropriate.
  - 4. Response communication – recording or conveying that response.
- Error could occur at any step!
- Think through your items as a respondent would – troubleshoot.



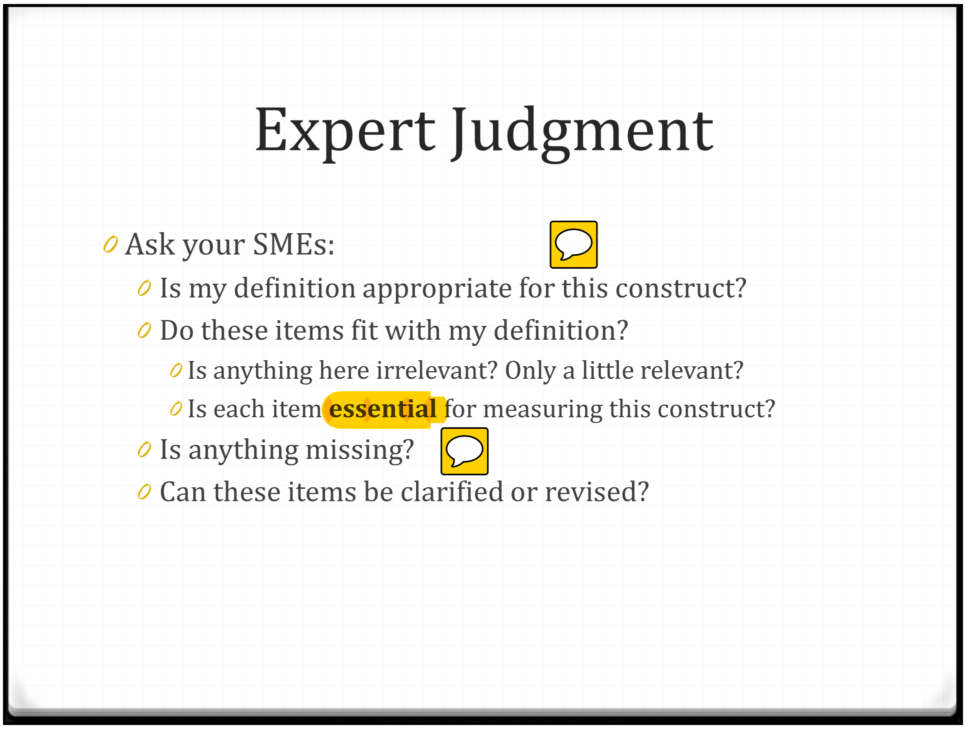
Good Process

- Think carefully through your theory.
  - Including expected dimensionality, even if you are “exploring.”
- Write a good definition & get feedback on it.
- Write more items than you expect to need.
  - 4-5 per dimension at a minimum.
- Proofread.
- Proofread again.
- Pilot test and invite comments from respondents.



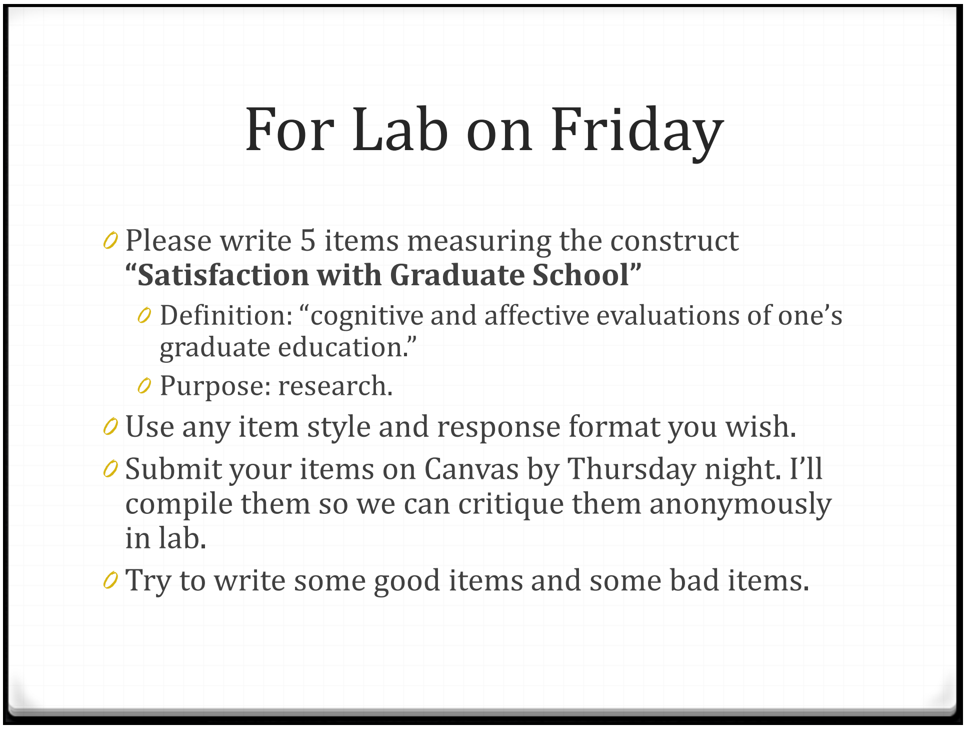
Evidence of Appropriate Content

- Appropriate content = do your items represent the whole domain of the construct you want to cover?
  - Is anything included that shouldn't be?
  - Is anything missing that should be included?
  - Is there balance across all the important aspects of the construct?
- We need to **support** our argument here.
  - Expert judgment – SMEs.
  - We're going to do this informally, but we can do it quite formally indeed (more on this later).



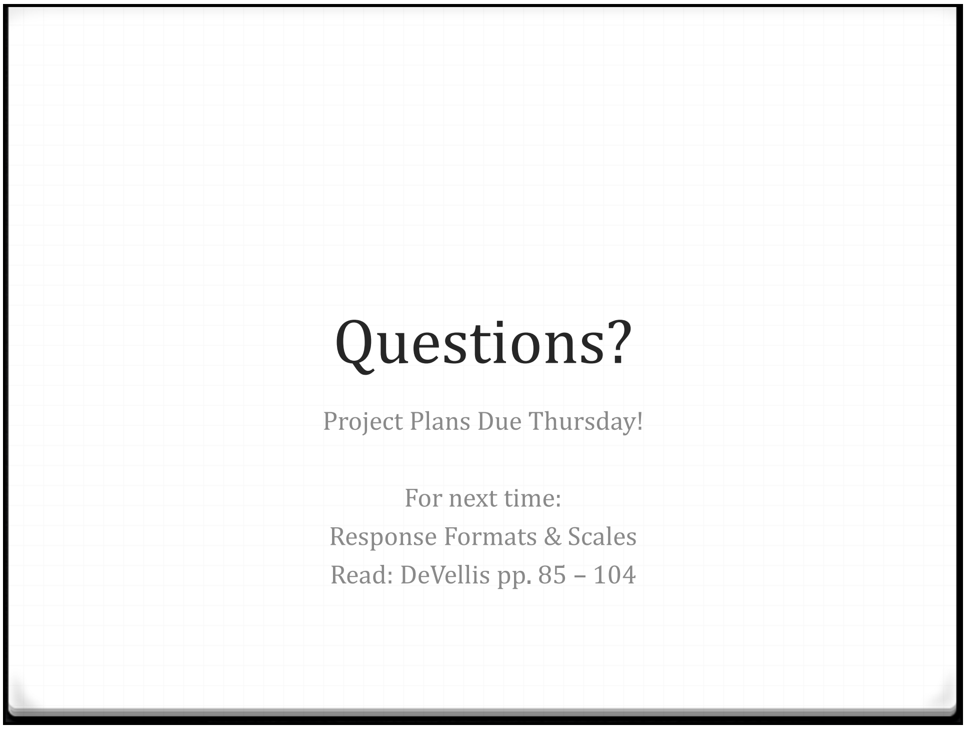
Expert Judgment

- Ask your SMEs:
  - Is my definition appropriate for this construct?
  - Do these items fit with my definition?
    - Is anything here irrelevant? Only a little relevant?
    - Is each item **essential** for measuring this construct?
  - Is anything missing?
  - Can these items be clarified or revised?



For Lab on Friday

- Please write 5 items measuring the construct **"Satisfaction with Graduate School"**
  - Definition: "cognitive and affective evaluations of one's graduate education."
  - Purpose: research.
- Use any item style and response format you wish.
- Submit your items on Canvas by Thursday night. I'll compile them so we can critique them anonymously in lab.
- Try to write some good items and some bad items.



Questions?

Project Plans Due Thursday!

For next time:
Response Formats & Scales
Read: DeVellis pp. 85 – 104