Creating Affective Scales: What Does the Research Say?

Deborah L. Bandalos
Professor and Director,
Assessment and Measurement PhD Program
James Madison University

York University 1/30/17

Overview

- Response models for affective items
  - Models of response process
  - Item responses as social encounters
  - Satisficing
- Implications of response models for scale development:
  - Negative item keying
  - Item order effects
  - Vague wording

Models of response processes

- Sudman, Bradburn, and Schwarz (1996) describe four components that underlie the response process (others have suggested very similar models):

1. Interpret the item
2. Generate a response
3. Format and report the response
4. Edit the response

- Let's use the following question to illustrate these steps:
- **Immigrants weaken our core values.**

Interpreting the item

- What is this item asking me?
- Some issues become apparent right away…
  - The item lack clarity – this is one of the major pitfalls in the interpretation process
  - Does the item refer to legal, illegal, or all immigrants?
  - What do they mean by "our core values"?

Interpreting the item

- Other issues in interpretation:
  - Use of politically or emotionally charged terminology….
    - "core values"
    - "socialized medicine"
    - "big business"
  - Use of vague terms such as "seldom" or "several"
  - Responses of 960+ students to the question "What % of time do these terms mean?"

Report

	a lot	as often as not	frequently	majority	many	most	nearly all	occasionally	often	rarely	sometimes
Mean	62.04	38.63	57.71	59.46	57.12	65.42	61.86	41.05	56.28	26.31	44.92
N	961	960	957	957	957	958	961	960	960	960	960
Std. Deviation	26.250	23.217	23.579	24.684	22.783	24.357	32.720	22.104	22.648	23.460	23.432

Interpreting the item

- Context effects can come from:
  - Item order
    - Suppose a question about support for seniors was preceded by questions about respondents' grandparents

Interpreting the item

- Scale options
  - Schwarz et al. (1985) reports a study in which respondents were asked how much TV they watched per week, on average.

Up to 1/2 hour	**½ to 1 hour**	**1 ½ to 2 hours**	**2 ½ to 3 hours**	**3 ½ to 4 hours**	**4 ½ or more hours**
Up to 2 hours	**2 - 3 hours**	**3 to 4 hours**	**4 to 5 hours**	**5 to 6 hours**	**6 or more hours**

  - %s of respondents reporting they watched more than two and a half hours were 16.2 % for those who received the first (low frequency) scale, but 37.5% for those receiving the high frequency scale.

Generating a response

- This requires respondents to retrieve the necessary information from memory
- If respondents have strong attitudes or are familiar with the topic, this may be easy
  - "I've thought about immigration a lot and I think immigrants actually strengthen American values."
- If they do not, they may draw on related information, experiences or attitudes of family or friends, stereotypes about the referent
  - "I don't know about core values, but my dad says immigrants take jobs away from people like him."

Generating a response

- The degree to which information is accessible also affects the consistency of responses.
- For those with strong attitudes, most of the retrieved information will be consistent.
- Those with weak attitudes may access a variety of information based on related attitudes, stereotypes, etc., and this information may be inconsistent.
  - "Everyone was an immigrant at some point, but they all became Americans", but "Some of these immigrants don't even learn English."
  - The response generated at any given timepoint will therefore depend on which pieces of information are drawn from memory.

Formatting and reporting the response

- In this step, respondents must map their response onto the response options provided.
- In some cases, those creating scales do not pay sufficient attention to the response scale, resulting in a mismatch between the items and response options.
- For example, what if the following response scale were used for our example item?

Never	Sometimes	Often	Always
1	2	3	4

Formatting and reporting the response

- Research suggests that respondents use the extremes of response scales to "anchor" or contextualize their responses.
- Schwarz et al. (1988)reports a study in which respondents were asked to report how many "irritating experiences" they had on an average day.
- When the response scale contained low frequencies, respondents assumed that major irritations were meant and reported fewer irritating experiences.
- When the response scale contained high frequencies respondents assumed this phrase referred to minor irritations and reported more events.

Editing the response

- In some cases, respondents edit their responses, either before or after reporting them.
- This can be due to the desire to respond in a socially desirable manner, a desire to appear consistent, or for many other reasons.

Item Responses as Social Encounters

Item responses as social encounters

- Another way to understand the response process is as a form of social encounter.
- Researchers have found that respondents apply many of the guidelines used in everyday conversations to answering self-report questionnaires (e.g., Strack and Schwartz, 1992).
- Theories in this area are based on Grice's (1975) *conversational maxims*:

Item responses as social encounters

- The maxim of quantity: Contributions to a conversation should contain as much information as necessary, but not more.
- The maxim of quality: Contributions to a conversation should be truthful and should be based on evidence.
- The maxim of relation: Contributions should be relevant to the conversation.
- The maxim of manner: Contributions should be clear and not overly wordy.

Item responses as social encounters

- How do these relate to the steps in the response process?
- Taking these one at a time:
- **Comprehension of the question**
  - the maxim of quantity suggests that respondents will assume that the question provides all the information they need to supply an answer
  - This is likely why respondents use information in the response options or in other items to inform their answers, especially if the item is unclear.
  - The maxim of relation also suggests that all information presented is relevant.

Item responses as social encounters

- **Generating a response**
  - The maxim of quantity – all information presented should be used in generating a response.
  - The maxim of relation suggests respondents should not include irrelevant or redundant information in their responses.
    - This may have implications for including very similar items on scales.
  - The maxim of quality implies that respondents should answer truthfully.

Item responses as social encounters

- **Formatting and reporting a response**
  - The maxim of relation suggests respondents will expect response options to be appropriate and to be sufficient for any response they may provide.
  - This maxim also suggests that respondents will assume all of the response options are meant to be used in some way, which may lead respondents to use these to anchor their responses

Item responses as social encounters

- **Editing the response**
  - Maxim of quantity suggests respondents should edit redundant information from their response
  - For example, if respondents are asked:
    - How many times in the past year have you gone to the beach?
    - And then asked:
    - How many times in the past year have you gone on vacation?
  - Most respondents will not include beach trips in their answer to the second question.

Satisficing

Satisficing

- Krosnick (1991) defines a continuum of responding that is anchored by two broad categories of respondents:
  - Optimizers carry out the steps in the response process to the best of their ability.
  - Satisficers skip some of the necessary processes entirely ("strong satisficing") or put forth less than their best effort ("weak satisficing").
- During the interpretation step of responding, satisficers may not attempt to fully understand the question.
  - This is especially problematic if the question is unclear or open to interpretation.

Satisficing

- During the response generation step, satisficers may base responses on the first piece of information that comes to mind, rather than fully searching memory for relevant info.
  - Satisficers may simply agree with all items, or skip them entirely (especially if the item is unclear or difficult to understand).
- During the response editing and reporting phase, satisficers may choose the first response option that seems reasonable, choose a neutral option, if one is available, or choose a labeled over an unlabeled response.
- Satisficers are unlikely to engage in any response editing, unless they are also susceptible to socially desirable responding.

Negative item keying

- A common recommendation in scale development is to include both positively and negatively keyed items:
- Postive keying:
  - Immigration is good for the U.S. economy.
- Negative keying:
  - Immigration is not good for the U.S. economy.
- Negative items can be either negatively ***worded*** as in the previous statement, which contains the word "not" or
- Negatively ***keyed***, as in the following item:
  - Immigration is bad for the U.S. economy.

Negative Item Keying

Negative Keying

- Inclusion of negatively keyed items is typical on affective scales, but is it a good idea?
- Pros
  - Balances the effects of *acquiescent* responding.
  - Allows researchers to detect those who are responding inconsistently.
- Cons
  - Some respondents have trouble answering negatively keyed items.
  - Inclusion of both negatively and positively keyed items results in lower reliability.
  - Inclusion of both negatively and positively keyed items often results in extra dimensionality (factors).

Negative Keying: Pro

- Balances the effects of *acquiescent* responding.
- The original idea was that the answers of acquiescent responders to positively keyed and negatively keyed items would cancel each other out (Cloud & Vaughn, 1970).
  - This would yield scale scores around the mean, so would not bias the overall sample mean.
- However, this assumes that responses to negatively keyed items, when reverse scored, are the same as responses to positively keyed items:
  - E.g., not being sad is the same as being happy
- There is evidence that this is not the case.

Negative Keying: Pro

- Allows researchers to detect those who are responding inconsistently.
- For example, if someone responds "strongly agree" to both positive and negative items, this is inconsistent and implies the person is:
  - Not really reading the items
  - Responding randomly
  - Otherwise satisficing

Negative Keying: Con

- Linguistic studies have shown that:
  - Affirmative statements are processed more quickly, and with fewer errors, than negative statements (Wason, 1959; 1961)
  - Negative statements result in more errors in remembering than positive statements (Peterson & Peterson, 1976)
  - Affirmative sentences are processed more quickly and accurately than negated sentences (especially those including the word "not") (Sherman, 1973; 1976).

Negative Keying: Con

- For example, assuming that Christian Bale if your favorite Batman actor, which of these statements is easier for you to answer, using the following scale?

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree	Strongly Agree

- Christian Bale is not my favorite Batman actor.
- Christian Bale is my favorite Batman actor.

Negative keying: Con

- Some respondents have trouble answering negatively keyed items, as evidenced by these statements from think-alouds we have carried out with such items:

  - "Like, definitely had to do a double-take for several of these…definitely #8, and then #6 [2 negatively keyed items]. Those two kinda just, threw me for a loop."
  - "Ah, these are making my head hurt!"
  - "Some of the double negatives, when you're trying to do it quickly on the spot, you're like, wait—am I…is it positive or negative? "

Negative keying: Con

- Inclusion of both positively and negatively keyed items results in lower reliability.
- This has been shown in most studies:
  - Barnette, 2000; Chang, 1995; Coleman, 2013; Finney, 2001; Schriesheim & Hill, 1981; Schriesheim et al., 1991
- This occurs because responses to negatively keyed and positively keyed items are not necessarily consistent.

Negative Keying: Con

- Inclusion of both negatively and positively keyed items often results in extra dimensionality (factors).
  - This can be a problem when studying scale dimensionality.
- Numerous studies have shown this (here are just a few):
  - Coleman, 2013; Corwyn, 2000; DiStefano, & Motl, 2003; Hazlett-Stevens, Ullman, & Craske, 2004; Magazine, Williams, & Williams, 1996; Marsh, 1996; Motl, Conroy, & Horan, 2000; Tomás & Oliver, 1999
- Although a few have not:
  - Bernstein & Garbin, 1985; Borgers, Hox, & Sikkel, 2004

Negative keying: Con

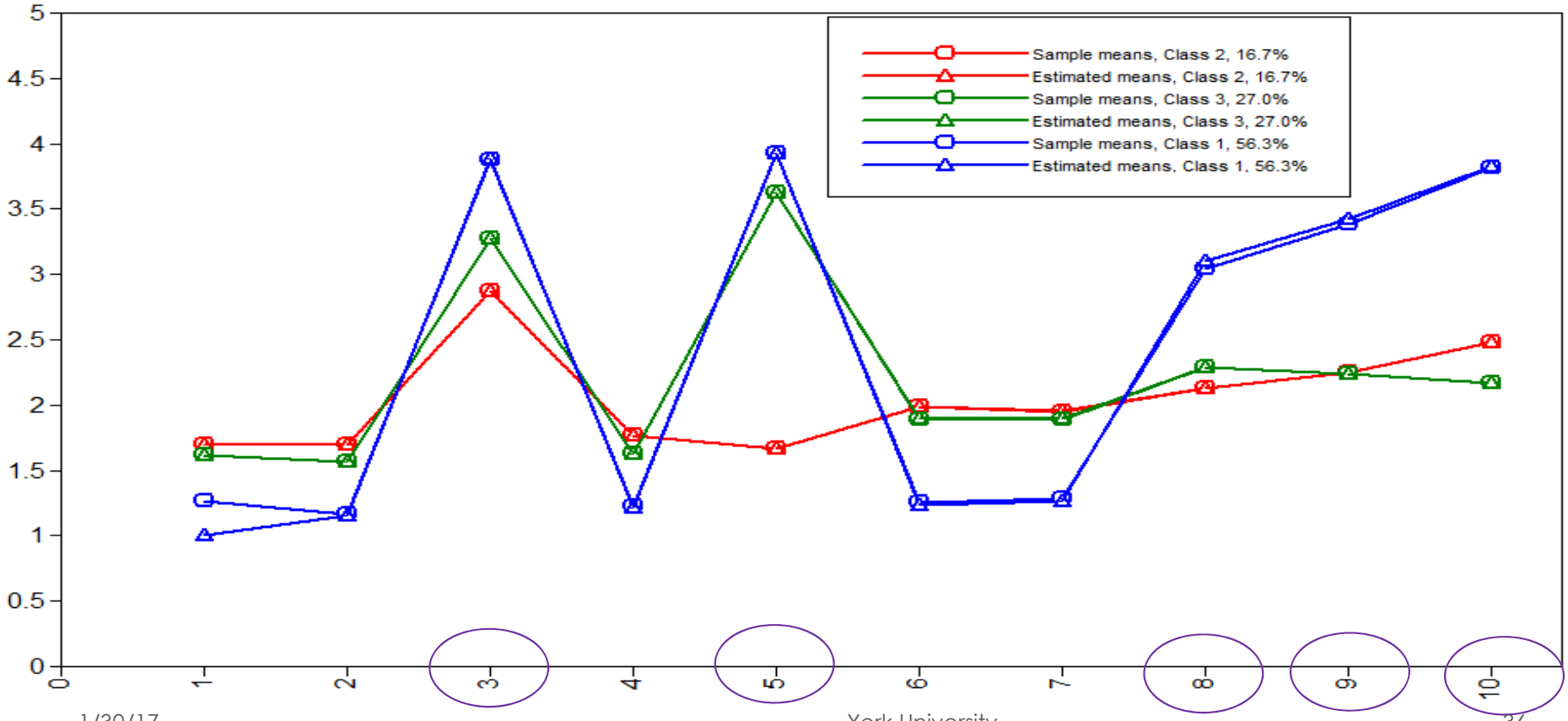
- Answering negatively keyed items appears to be especially difficult for:
  - Children (Benson & Hocevar, 1985; Marsh, 1986)
  - Those with lower verbal ability (Coleman, 2013; Corwyn, 2001; Dunbar, Ford, Hunt, & Der, 2000; Marsh, 1996)
  - Those with lower educational levels (Melnick & Gable, 1990)

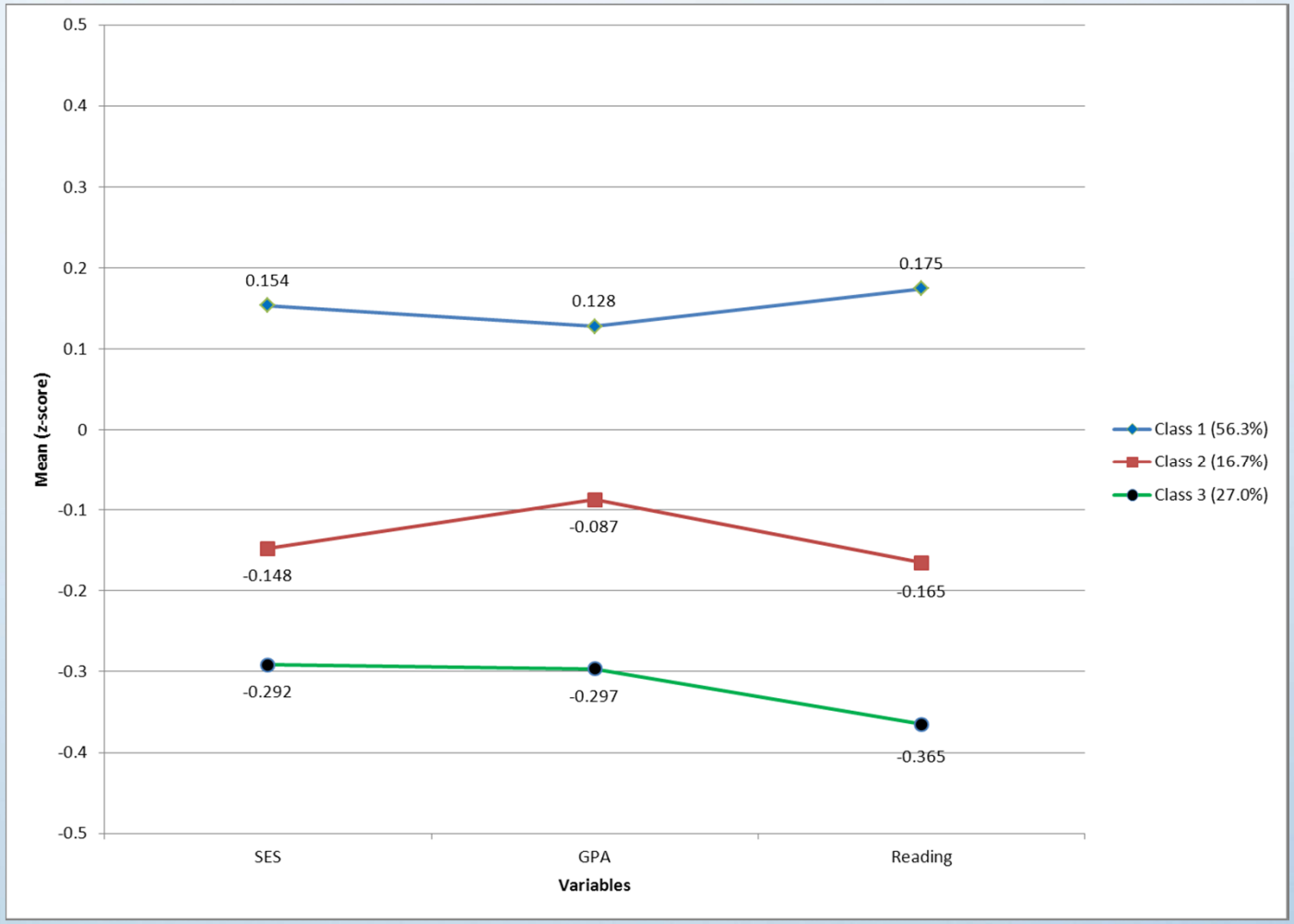
Negative keying: con

- Clearly, not everyone has trouble answering negatively keyed/worded items.
- In a recent study in my lab we investigated whether there were latent classes of respondents who responded inappropriately to negatively keyed items, and if so, what characterized these respondents.
- In Study 1 we used the Rosenberg Self-Esteem scale

Item	Wording
1. I feel that I'm a person of worth, at least on an equal plane with others.	Positive
2. I feel that I have a number of good qualities.	Positive
3. All in all, I am inclined to feel that I am a failure.	Negative
4. I am able to do things as well as most other people.	Positive
5. I feel I do not have much to be proud of.	Negative
6. I take a positive attitude toward myself.	Positive
7. On the whole, I am satisfied with myself.	Positive
8. I wish I could have more respect for myself.	Negative
9. I certainly feel useless at times.	Negative
10. At times I think I am no good at all.	Negative

We did find a class (in red, below) in which respondents appeared to answer in the same way to both positively and negatively keyed items

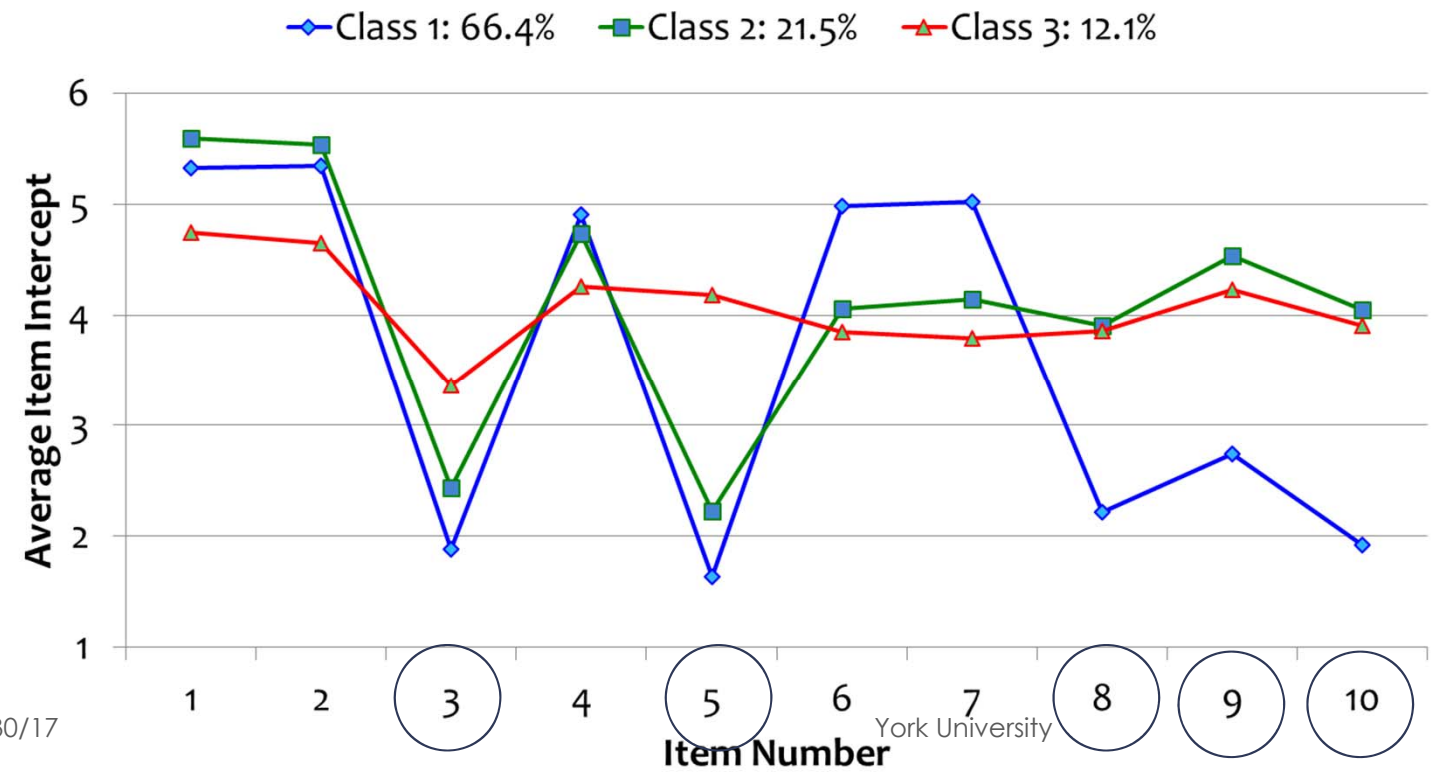




The classes differed significantly on reading ability, SES, and GPA.

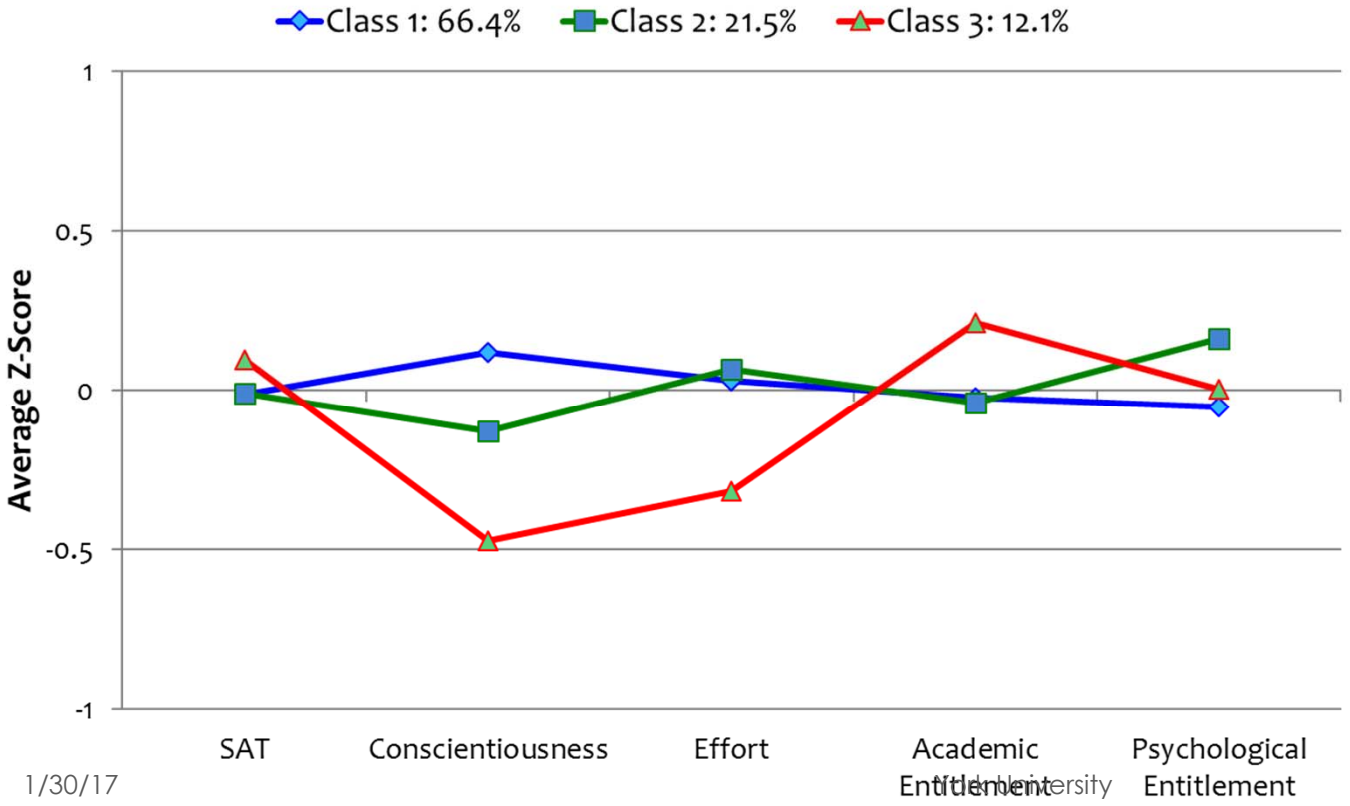
Negative Keying: Con

- We replicated these results on another sample using the RSE:



Negative Keying: Con

- Conscientiousness, effort, and psychological entitlement predicted the "red" class:



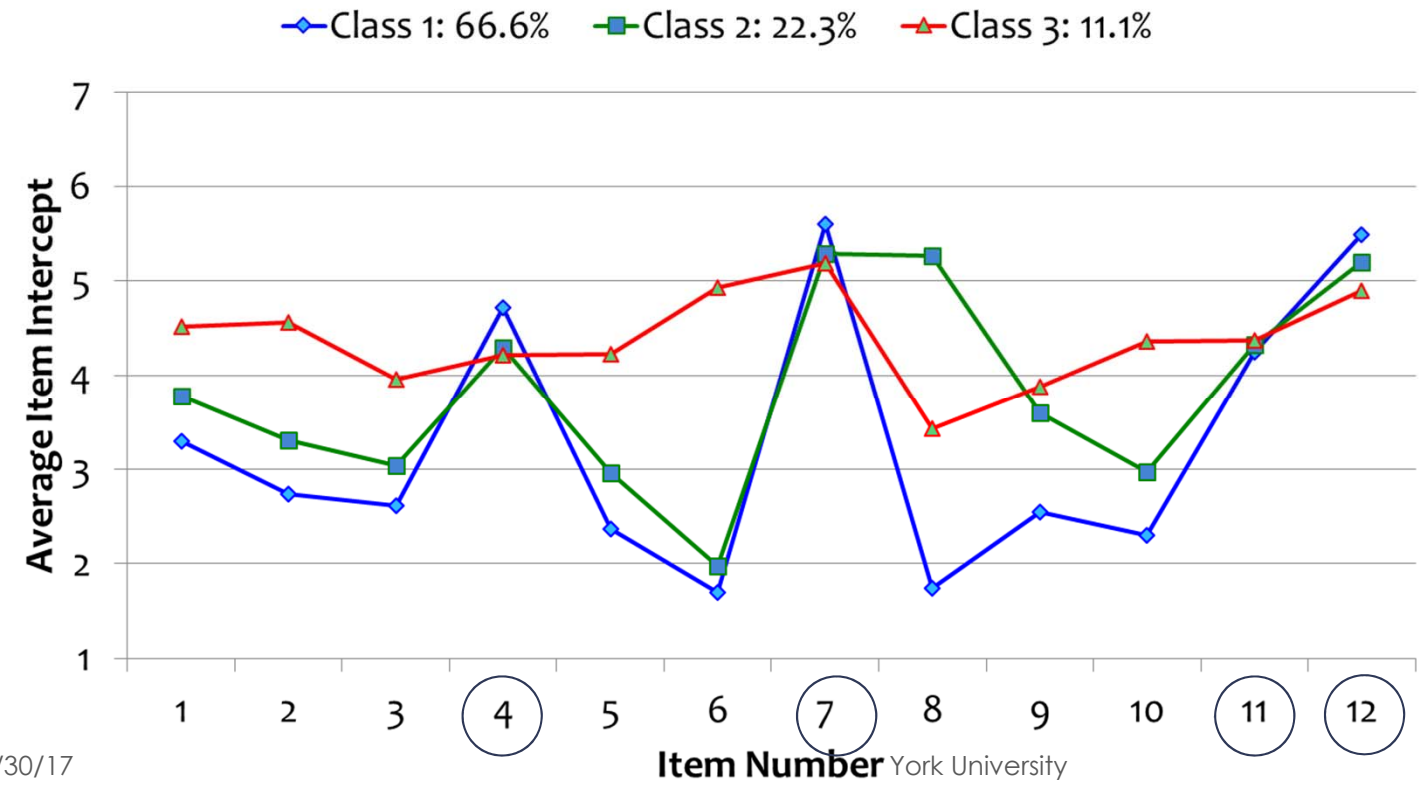
Negative Keying: Con

- In another replication, we found similar results using the Dickman Impulsivity Scale:

1. I will often say whatever comes into my head without thinking first.
2. Often, I don't spend enough time thinking over a situation before I act.
3. Many times the plans I make don't work out because I haven't gone over them carefully enough in advance.
4. **I enjoy working out problems slowly and carefully.***
5. I frequently make appointments without thinking about whether I will be able to keep them.
6. I often get into trouble because I don't think before I act.
7. **I am good at careful reasoning.***
8. I frequently buy things without thinking about whether or not I can really afford them.
9. I often make up my mind without taking the time to consider the situation from all angles.
10. I often say and do things without considering the consequences.
11. **I rarely get involved in projects without first considering the potential problems.***
12. **Before making any important decision, I carefully weigh the pros and cons.***

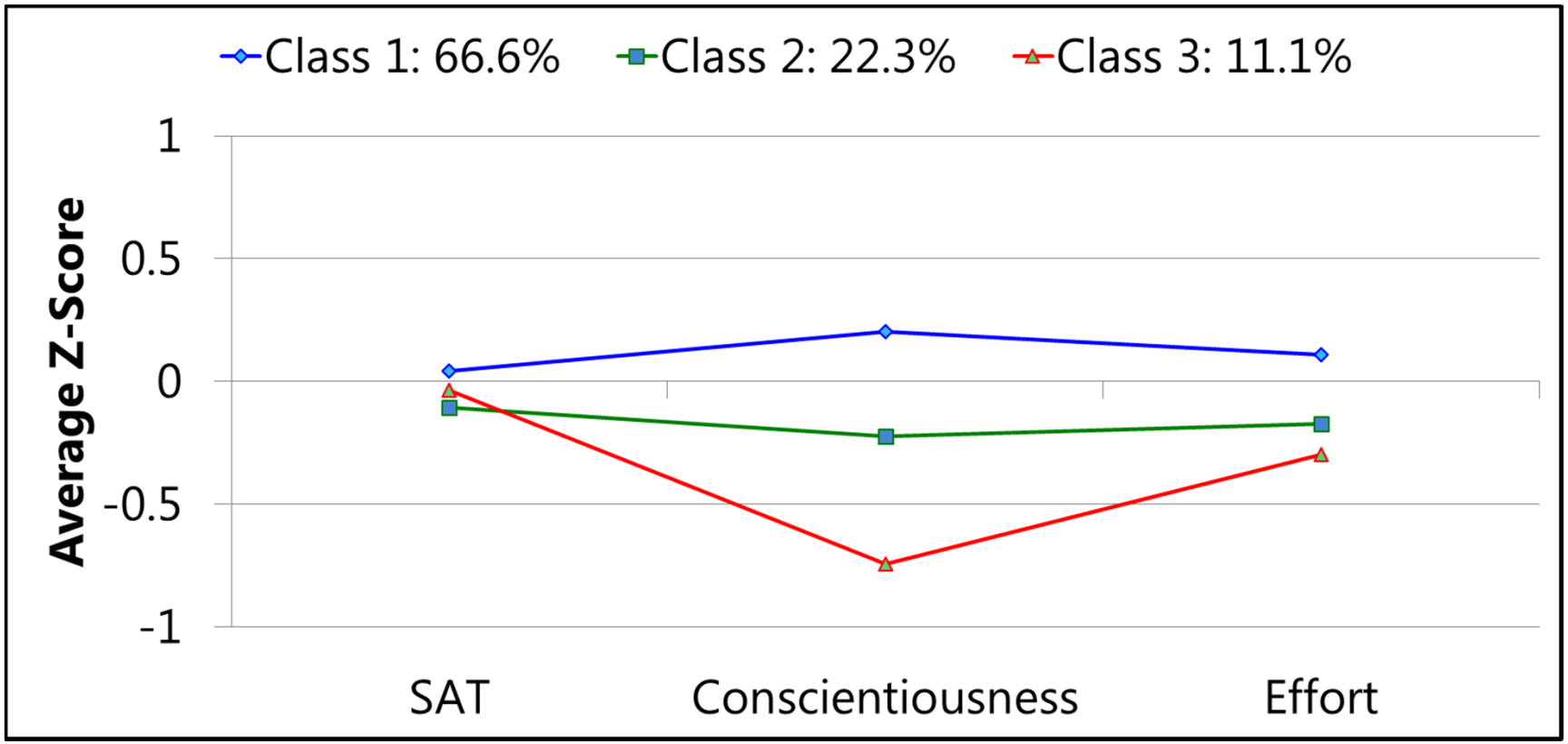
Negative Keying: Con

- In another replication, we found similar results using the Dickman Impulsivity Scale:



Negative Keying: Con

- Predictors of the "red" class were conscientiousness and effort



Take-Home Points

- I generally recommend against inclusion of negatively-keyed items.
  - Especially with children or those who may have limited English or cognitive ability.
- If you are going to use negatively-worded items, use items that don't contain words like "not", "none", "no", etc.
- Having said that, it may be possible to control for the negative effects of negatively keying through item ordering.

Item Order Effects

Item Order Effects

- For the two items below, does it matter whether item 1 or item 2 comes first in the questionnaire?
- In Lord of the Rings, Smeagol/Gollum's obsession with the Ring was beyond his control.
- Gollum's death in Return of the King was justly deserved.

- Some studies have found that item order matters, although this depends on the specific items used.

Item Order Effects

- There are two basic categories of item order effects: context effects and serial order effects.
- Context effects are those induced by exposure to material earlier in the questionnaire.
  - For example, suppose respondents were asked the question "Do you have fond memories of your grandparents?" and were then asked
  - "Do you support or oppose expansion of Medicare benefits for seniors?"
  - Respondents' answers to the first question will likely influence their responses to the second.
    - This is sometimes called a *priming effect.*
  - Context effects can also result as a cumulative effect of a series of previous items.

Order Effects

- In general, order effects occur because, even if a person's attitude is stable, the response process may not be.
- There may be instability in any of the steps in the response process:
  - Interpreting the question
  - Retrieving information
  - Integrating the information into an opinion
  - Selecting a response
- In the next few slides we will look at how order effects might occur for each of these.

Context Effects: Interpretation phase

- Context effects affect question interpretation in (at least) 2 ways (Torangeau & Rasinski, 1988):
  - By providing an interpretive framework (Knowles & Byers, 1996 refer to this as the construct-awareness hypothesis)
    - Most applicable to questions about unfamiliar or unclear topics.
    - For example, if questions about an unfamiliar government policy are preceded by questions about inflation, respondents will likely assume that the unfamiliar policy is related to inflation.
    - This is especially true if the questions are all blocked together.
    - Tourangeau & Rasinski (1988) refer to this as a ***carry-over effect.***

Context Effects: Interpretation phase

- Another way context affects question interpretation is by providing cues about what specific information is asked for – these effects are related to Grice's (1975) maxim of quantity.
  - This maxim states that information should not be provided if it is not necessary, or that redundant information should not be provided.
- For example, if respondents are asked several attitude items that are quite similar, they may infer different meanings to the questions, even though these are not intended.
  - Gollum's death in Return of the King was justly deserved.
  - In Return of the King, Gollum deserved to die.
  - I think Gollum should have died in Return of the King given all his misdeeds.

Context Effects: Retrieval phase

- The previous examples illustrate the effects of context on the interpretation phase of the response process.
- Context can affect the retrieval process by influencing the material that is retrieved from memory to form a judgment.
- The grandparent example is an example of the effect of context on retrieval, because the question about grandparents causes certain memories to be activated.
- Similar effects can occur if previous questions on the same scale have caused respondents to activate attitudes or beliefs that would not otherwise have become salient.
  - Effects on retrieval are strongest for unfamiliar topics, for those with mixed attitudes, and for those with little expertise on the topic.

Context Effects: Integration phase

- Context can also affect how people integrate the retrieved material into a single judgment.
- This is especially true when the standards for making a judgment or the definition of the attitude are unclear, or when the topic is complex.
- Shuman & Presser (1981) report a study in which people were asked whether "Communist reporters from other countries" should be allowed to come to the US and send back "the news as they see it."
  - In this example, people were more likely to agree if the question was preceded by one asking whether US reporters should be allowed the same freedom when reporting from other countries.

Context Effects: Response selection phase

- Scales consisting of many items are especially susceptible to context effects during the response selection phase of the response process.
- This is because answering a series of questions on the same topic makes the relations among them more obvious.
  - Knowles (1988) refers to this as the "construct-awareness hypothesis"
- Some respondents will try to make their responses consistent across the set of items.
- Answering a series of questions on the same topic can also affect respondents' internal anchors.
  - For example, if respondents answer a question with "strongly disagree" and encounter a later question with which they disagree even more, they may change their first answer to differentiate their levels of disagreement to the two questions.

Order Effects: Serial order

- Serial order effects refer to the phenomenon that, in a series of items, the order in which items are answered can affect item-level statistics.
- For example, several studies have found that item-total correlations are higher for items at the end of a set than for items at the beginning (Hamilton & Shuminsky, 1990; Knowles, 1988; Kraut, Wolson, & Rothenberg (1975); Steinberg, 1994)
- These order effects can lead to extra factors, or in CFA, to correlated residuals.
- This is because the serial order effects cause items to correlate more highly than they otherwise would.
  - This extra correlation results in additional factors or correlated residuals.

Order Effects: Serial order

- Coleman & Bandalos (2012) investigated the effect of serial order on correlated residuals.
- We manipulated the placement of positively and negatively worded items by creating 3 versions of the RSE scale:
  - Alternating (PNPNPN..)
  - Bunched (PPPPNNNN)
  - Random (via computer administration)
- We found that CFA model fit was best for the random and alternating versions and worst for the bunched version

Order Effects: Serial order

Rosenberg Self-Esteem Scale			
Version	**FF**	**RMSEA**	**CFI**
Randomized	**.56**	**.11**	.87
Alternating	.60	.13	**.88**
Bunched	1.17	.18	.81

- Also, correlated residuals were nearly always the result of contiguously-placed negatively worded items.

Order Effects: Serial order

- We investigated this effect in a series of further studies with different scales:

Conformity scale (Goldberg, et al. 2006)			
Version	**FF**	**RMSEA**	**CFI**
Randomized	**.41**	**.10**	**.75**
Alternating	.58	.13	.74
Bunched	.57	.12	.73

Order effects: Serial order

	FF	RMSEA	CFI
Conscientiousness Scale (John & Srivastava, 1999)			
Version	FF	RMSEA	CFI
Randomized	**.44**	**.12**	**.79**
Alternating	.59	.14	.75
Bunched	.58	.13	.73

Order Effects: Take-home message

- Item order can make a difference.
- Separate similarly worded items.
- Anticipate how context might affect responses…

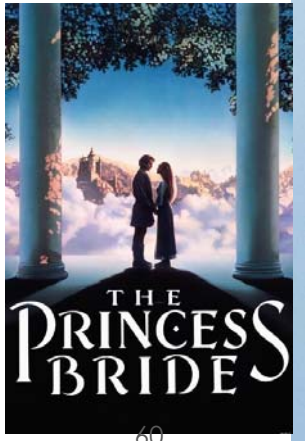
Vague Wording

Vague Wording

- I frequently re-watch The Princess Bride.

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

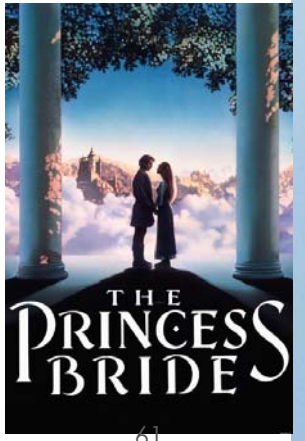
- How often do you think "frequently" is?



Vague Wording

- Would we obtain different answers if the question were worded:
- I re-watch The Princess Bride at least once a month.

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree



Vague Wording

- Pros
  - Vague wording is often less awkward than specific wording.
  - The meaning of responses can be easier to parse out when vague wording is used.
- Cons
  - Vague wording invites ambiguity.
  - Respondents will interpret vague words and phrases in different ways.

Vague Wording: Pros

- Vague wording is often less awkward than specific wording.
  - Consider this item from a personality scale: "I frequently forget to do things." How would you replace "frequently" with a more specific word or phrase?
    - "9 times out of 10, I forget to do things."
    - "80% of the time when I have to do something, I forget."
    - Imagine an entire scale with wording like that!
  - More ambiguous words like "frequently" may make items sound more natural and thus make them easier for respondents to answer.

Vague Wording: Pros

- The meaning of responses can be easier to parse out when vague wording is used.
  - 80% of my friends have seen The Princess Bride.

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

  - Why did this respondent disagree?
    - Because more than 80% of their friends have seen the movie? Or because fewer than 80% have?
    - If the item instead said "Most of my friends…" the meaning of a disagree response would be clearer – few or none of the respondent's friends have seen the movie.

Vague Wording: Cons

- Vague wording invites ambiguity.
  - By definition, vague wording is, well…vague!
  - This leads to ambiguity as respondents try to figure out what a question is asking.
  - Numerous studies have shown that people interpret various vague words differently.

	Mean	SD
a lot	62.04	26.250
as often as not	38.63	23.217
frequently	57.71	23.579
majority	59.46	24.684
many	57.12	22.783
most	65.42	24.357
nearly all	61.86	32.720
occasionally	41.05	22.104
often	56.28	22.648
rarely	26.31	23.460
sometimes	44.92	23.432

The Importance of Context

- Respondents' interpretations of vague phrases is also context dependent.
  - Earthquakes occurred frequently in California in 1980.
  - It frequently rained in Seattle in 1980.
- "Frequent" earthquakes – a relatively rare event – represent a much lower occurrence than "frequent" rainfall – a relatively common event (especially in Seattle!)
- The same is true for other questions we might ask:
  - "I study for classes frequently during the week."
  - "I use student affairs resources often."

Vague Wording

- We conducted two multiple-group CFA studies investigating the impact of items with vague vs. specific wording on the loadings and intercepts of items on two attitude scales.
- Intercepts are like "difficulty" – in this case, extremity of response.
- Loadings get at the ability of the item to discriminate among people with similar levels of the attitude.
- We hypothesized that specific wording would result in different intercepts (extremity of responses) and higher loadings (more discriminating).

Vague Wording

- **Side Note**
  - Why did we hypothesize that specific wording would result in different intercepts (extremity of responses)?
  - It was difficult to predict the direction of the effect because it would depend on whether the specific wording (e.g., 95%) was in line with how a respondent would interpret the vague wording (e.g., "Often").
  - Some respondents could interpret "95%" as more than "often" and thus provide a less extreme response, whereas some could interpret "95%" as less than "often" and provide a more extreme response.

Vague Wording

- In one study, we found that there were differences between vague and specific items in both the discriminative ability (loadings) and extremity of response (intercepts).
  - Which condition had higher discrimination was split – half the items had higher loadings in the vague condition and half had higher loadings in the specific condition.
  - But, loadings were higher in the specific wording condition for the two items with the greatest difference between versions.

Vague Wording

- The same was true for intercepts: 4 items had higher intercepts for the vague wording and five had higher intercepts for the specific wording.
- However, specifically worded items containing absolute words such as “everything” or “always” resulted in lower levels of agreement than their non-manipulated counterparts that did not contain these absolutes.
- Of the five items with the largest intercept differences between the vague and specific versions, four involved such absolutes in the specific version.

Vague Wording

- The greatest intercept difference by far between the two versions was found for item 4.
- Vague version: "Basically, I know what I need to know about the important things in life" – $\tau = 3.49$.
- Specific version: "Basically, I know everything I need to know about the important things in life" - $\tau = 4.96$.
- This item was recoded such that higher intercepts indicated *le*ss agreement.
- Inclusion of the more specific, but also more absolute word "everything" thus resulted in less agreement with the item.

Vague Wording

- So does vague wording make a difference?
  - In terms of loadings and intercepts, maybe – but the difference it makes is inconsistent.
  - Could be scale-specific.
- Anecdotally…
  - "Often is different than always, because often is every once in a while whereas always is always, like 'definitely has to happen'."
  - Often means "95%, like more times than not."
  - One interviewee chose a less extreme answer (slightly agree) because "95% seems pretty high."
  - I would have probably changed my answer [if there had been a vague phrase instead of a specific one] because I would have evaluated them myself…like, 'It's about 50/50.'"

Which to Use?

- Vague wording may make items sound more natural, and may eliminate some issues with interpretation of responses.
- However, specific wording eliminates the ambiguity associated with vague words, thus making interpretation easier.
- Interpret responses to items with vague wording carefully!
- If getting a precise idea of frequency is important, provide numerical responses (once a week, 1-2 times a week, etc.)
- If using specific wording is too awkward and exact frequency is unimportant, consider using vague wording – but interpret results accordingly.

References

References: Negative Wording

- Barnette, J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361-370.
- Benson, J. & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22,* 231-240.
- Bernstein, I., & Garbin, C. (1985). A comparison of alternative proposed subscale structures for MMPI Scale 2. *Multivariate Behavioral Research, 20,* 223-235.
- Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality & Quantity, 38,* 17-33.
- Chang, L. (1995a). Connotatively inconsistent test items. *Applied Measurement in Education, 8,* 199-209.
- Chang, L. (1995b). Connotatively consistent and reversed connotatively consistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement, 55,* 991-997.
- Cloud, J., & Vaughn, G. (1970). Using balanced scales to control acquiescence. *Sociometry, 33,* 193-202.
- Coleman, C. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study* (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.
- Cordery, J., & Sevastos, P. (1993). Responses to the original and revised Job Diagnostic Survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78,* 141-143.
- Corwyn, R. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34,* 357-379.

References: Negative Wording

- Dunbar, Ford, Hunt, & Der (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment, 16,* 13-19.
- Hazlett-Stevens, H., Ullman, J., & Craske, M. (2004). Factor structure of the Penn State Worry Questionnaire: Examination of a method factor. *Assessment, 11,* 361-370.
- Horan, P., DiStefano, C., & Motl, R. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10,* 435-455.
- Magazine, S., Williams, L., & Williams, M. (1996). A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's Affective and Continuance Commitment Scales. *Educational and Psychological Measurement, 56,* 241-250.
- Marsh, H. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22,* 37-49.
- Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810-819.
- Melnick, S. & Gable, R. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly, 14,* 31-36.
- Motl, R., Conroy, D., & Horan, P. (2000). The Social Physique Anxiety Scale: An example of the potential consequence of negatively worded items in factorial validity studies. *Journal of Applied Measurement, 1,* 327-345.
- Peterson, C. & Peterson, J. (1976). Linguistic determinants of the difficulty of true-false test items. *Educational and Psychological Measurement, 36,* 161-164.

References: Negative Wording

- Schriesheim, C., Eisenbach, R., & Hill, K. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, *51*, 67-78.
- Schriesheim, C., & Hill, K. (1981). Controlling acquiescence response bias by item Reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*, 1101-1114.
- Sherman, M. (1973). Bound to be easier? The negative prefix and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *12*, 73-84.
- Sherman, M. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15*, 143-157.
- Tomás, J., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84-98.
- Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *11*, 92-107.
- Wason, P. C. (1961). Response to the affirmative and negative binary statements. *British Journal of Psychology*, *52*, 133-142.

References: Vague Wording

- Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expression of frequency and amount. *Journal of Applied Psychology*, *59*(3), 313-320.
- Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, *15*(1), 83-94.
- Bradburn, N., & Miles, C. (1979). Vague quantifiers. *The Public Opinion Quarterly*, *43*(1), 92-101.
- Krosnick, J. A. (1989). A review: Question wording and reports of survey results: The case of Louis Harris and Associates and Aetna Life and Casualty. *The Public Opinion Quarterly*, *53*(1), 107-113.
- Pepper. S/. & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*, 95-101.
- Pohl, N. F. (1981). Scale considerations in using vague quantifiers. *The Journal of Experimental Education*, *49*(4), 235-240.
- Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *The Public Opinion Quarterly*, *55*(3), 395-423.
- Wänke, M. (2002). Conversational norms and interpretation of vague quantifiers. *Applied Cognitive Psychology*, *16*, 301-307.
- Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is "Quite a bit"? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, *8*, 479-496.

References: Order Effects

- Bandalos, D.L., & Coleman, C. (2012, October). The company you keep: The effects of serial order on correlated residuals. Paper presented at the annual meeting of the Society for Multivariate Behavioral Research, Vancouver, BC.
- Grice, H. (1975). Logic and conversation. In P. Cole & T. Morgan (Eds.), *Syntax and semantics: Vol.3: Speech Acts* (pp. 41-58). New York: Seminar Press.
- Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology, 59,* 1301–1307.
- Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.
- Knowles, E. S., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology, 70,* 1080–1090.
- Schuman, J., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording, and context.* New York: Academic Press.
- Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66,* 341–349.
- Tourangeau, R., & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.