Carnegie Mellon University

Effective User Survey Design and Data Analysis

Hanan Hibshi and Travis Breaux

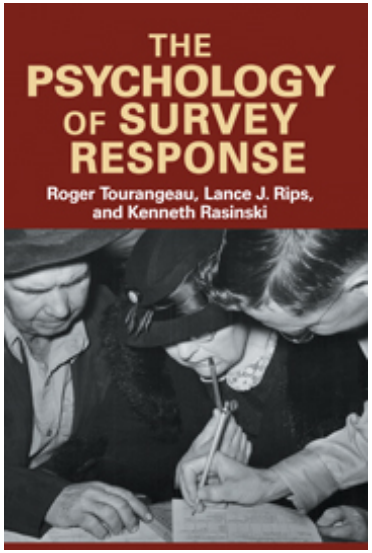
Monday, September 12, 2016

Full Day Tutorial

23rd IEEE International Requirements Engineering Conference

isr institute for SOFTWARE RESEARCH

Recommended Reading



Tourangeau, Rips, Rasinski. *The Psychology of Survey Response*, Cambridge University Press, 2000

Portions of this tutorial were based on research cited by Tourangeau et al.

A Model of Survey Response

Comprehension	Attend to questions and instructions Represent logical form of questions Identify question focus (information sought) Link key terms to relevant concepts
Retrieval	Generate retrieval strategy and cues Retrieve specific, generic memories Fill in missing details
Judgement	Assess completeness and relevance of memories Draw inferences based on accessibility Integrate material retrieved Make estimate based on partial retrieval
Response	Map judgement on to response category Edit response

Tourangeau, Rips, and Rasinski. *The Psychology of Survey Response*, Cambridge University Press, 2000.

COMPREHENSION

Cognitive load and working memory

- Sweller theorized that people have limited amounts of working memory, which they use to solve problems
- Cognitive load describes the amount of working memory that is consumed during a given problem exercise
- Survey questions can increase cognitive load and exhaust working memory, which can reduce answer quality over time

Sweller, J. (1988). “Cognitive load during problem solving: Effects on learning.” *Cognitive Science*. 12(2): 257-285.

Constructing the question representation

Cognitive operations for deriving the question representation:

- Representing the question in some format
- Picking out the question's focus
- Linking the nouns and pronouns to the relevant concepts in memory
- Assigning meanings to the predicates in the underlying representation

Example:

- How many people do you typically need to deliver a project on time?

Constructing the representation-of the question

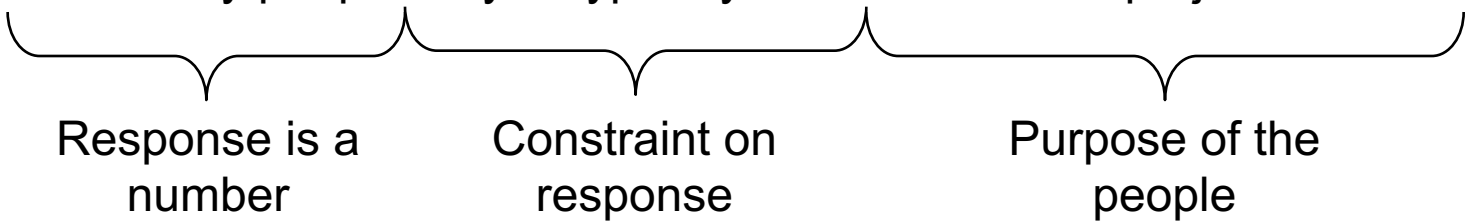
Cognitive operations for deriving the question representation:

- Representing the question in some format
- Picking out the question's focus
- Linking the nouns and pronouns to the relevant concepts in memory
- Assigning meanings to the predicates in the underlying representation

Format may be lexico-syntactic, or semantic, frame-based (e.g., Jackendoff, 1991 or Schank, 1975)

Example:

- How many people do you typically need to deliver a project on time?



Construction the representation-of the question

Cognitive operations for deriving the question representation:

- Representing the question in some format
- Picking out the question's focus
- Linking the nouns and pronouns to the relevant concepts in memory
- Assigning meanings to the predicates in the underlying representation

What are the roles of project-related people?

Example:

- How many people do you typically need to deliver a project on time?

What is my role?

Which projects am I reporting on?

Representation-of versus representation-about

- The representation-of is *required* to respond to the question
- The representation-about is *optional* and includes the respondent's interpretation of question elements

Do I include legal, marketing, users with developers?

Example:

- How many people do you typically need to deliver a project on time?

As a project lead, team member or witness of another project?

For projects within the last 6, 12 or 36 months?

Representation-of versus representation-about

- The representation-of is *required* to respond to the question
- The representation-about is *optional* and includes the respondent's interpretation of question elements

Example:

- How many people do you typically need to deliver a project on time?

- Graesser, Singer, Trabasso, 1994, identified 13 types of inferences: only two concern representations-of
- Respondent inferences depend upon several factors:
  - Amount of time allotted to think about question
  - Understanding of the survey purpose
  - Amount of information known about the topic

Interrogative Forms

- How many people do you typically need to deliver a project on time?

- For wh- questions, an implicit marker or trace determines the focus of the sentence (Radford, 1997)
- Monitoring for the trace position consumes working memory, making the question more difficult to parse (Just & Carpenter, 1992)
- *Who*, *which* and *what* questions are easier to parse than *where*, *when*, *why* or *how*
  - For example: What is a typical project size delivered on time?

Overloading working memory

- Some people feel that their company should adopt agile development, which reduces documentation by increasing communication, to reduce development times by a lot. Others feel that their companies should place more emphasis on architecture and incremental development, without reducing documentation and team sizes. Where would you place yourself on this [seven-point] scale?

- Some people feel that their company should adopt agile development, which reduces documentation by increasing communication, to reduce development times by a lot. What do you think? Do you agree strongly, agree…?

Overloading working memory

Just and Carpenter, 1992, identify several forms of overloading:

- Degree of embeddedness
  - When did you complete the last project in the healthcare domain where four or more people participated, one of whom was your current supervisor?
- Syntactic ambiguity
  - How many scenarios and use cases, which are testable, did you create?
- Garden path sentences
  - Which window frames individual and complex components and layouts?
- Working memory capacity of individuals

Consequences of overloading: (1) items may drop out, and / or (2) cognitive processing may slow down

Grice's cooperative principle

- Paul Grice (1989) proposed that speakers and listeners aspire to follow at least three maxims:
  - Quantity – make contributions minimally and maximally informative
  - Quality – contribute only truth for which you have sufficient evidence
  - Relation – make contributions relevant

Example:

Speaker tells Fred that he is out of gas. Fred responds by telling speaker, "there is a gas station around the corner." For Fred to be cooperative, he must believe the gas station is not closed.

- Examples where speakers flout the maxims by withholding information, such as in a letter of recommendation

Grice's maxims in survey design

- Respondents assume survey designer expects questions to be relevant (Grice's maxim of relation)
- If two questions appear similar, respondents assume they are different and may subtract answers from latter questions or inflate semantic differences in question comprehension (Strack, Schwartz, Wanke, 1991)

- Factors affecting implicatures:
  - Overlaps in concept meaning among items
  - Item sequencing
  - Range and labeling of rating scales

Semantic Effects

"The survey designer's job is to ask a question in such a way as to convey the intended space of interpretation, and the respondent's job is to reconstruct that space and say where the correct answer lies"

— Tourangeau, Rips and Rasinski

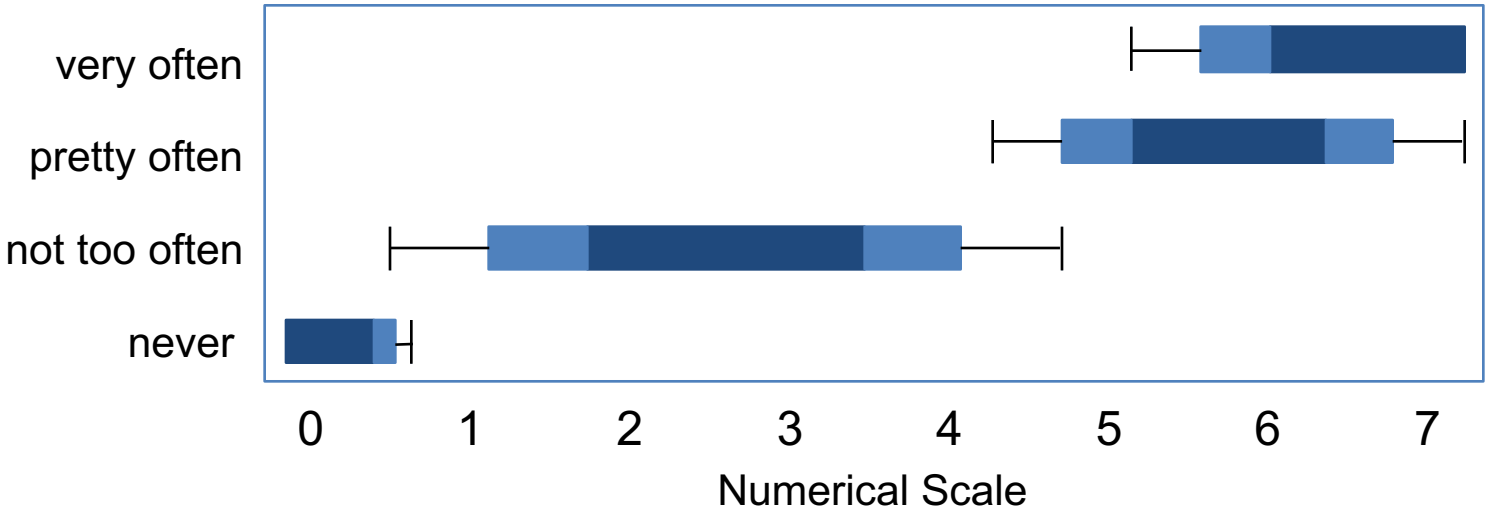
Semantic Effects: Presupposition

1. At what time do you usually participate in stand-up meetings?
   - Question 1 assumes you participate in stand-up meetings, and that these meetings occur at regular times during the day (also called double-barreling)
2. At what time in the morning do you usually participate in stand-up meetings?
   - Question 2 further assumes that the meetings only occur in the morning, and not the afternoon

- Leading questions may cause respondents to misremember events as though the presuppositions were true (Loftus, 1979)
- Filter questions to route participants around assumptions
  - Do you participate in stand-up meetings? If yes, during what times [select from a list of time intervals]?
- Including a “I don’t know” or “No opinion” response option

Semantic Effects: Vagueness

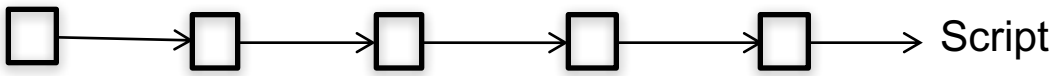
- Vague concepts
  - Do stakeholders typically expect to participate in designing acceptance tests?
- Vague quantifiers
  - At what time do you usually participate in stand-up meetings?
  - Response options: never, not too often, pretty often, very often



RETRIEVAL AND JUDGEMENT

Autobiographical reflection

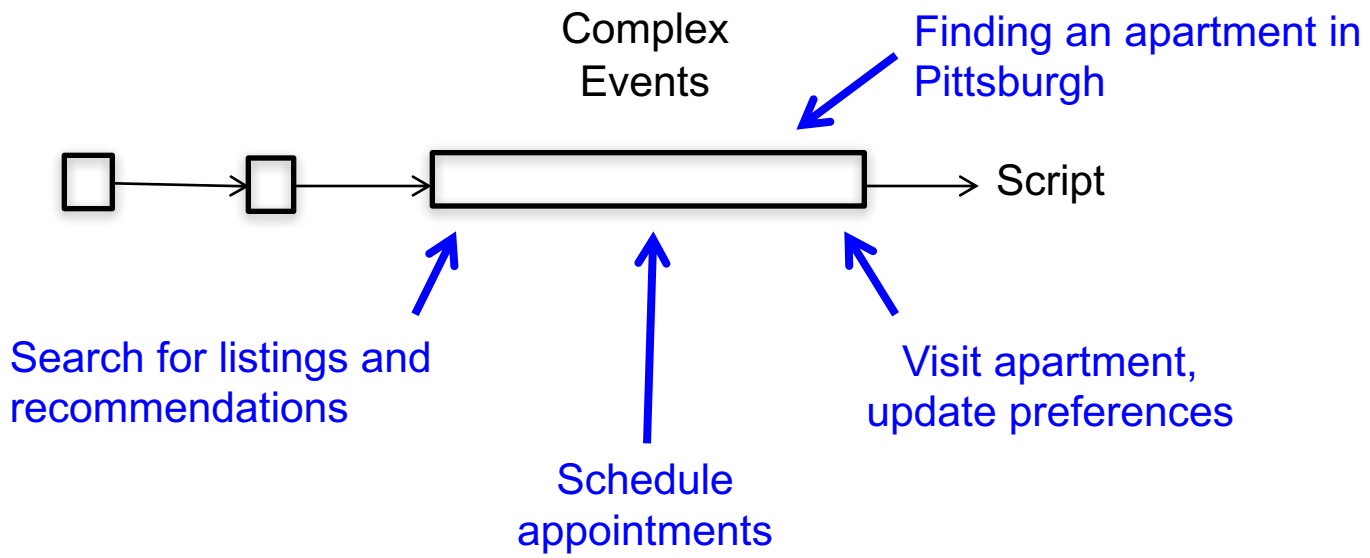
Schank and Abelson's theory of **scripts** (1977): a casual sequence of events, such as the steps one takes when one shops at a grocery store.



Respondents often report general patterns of facts, even when asked about specific time periods (A.F. Smith, 1991; Smith, Jobe and Mingay, 1991).

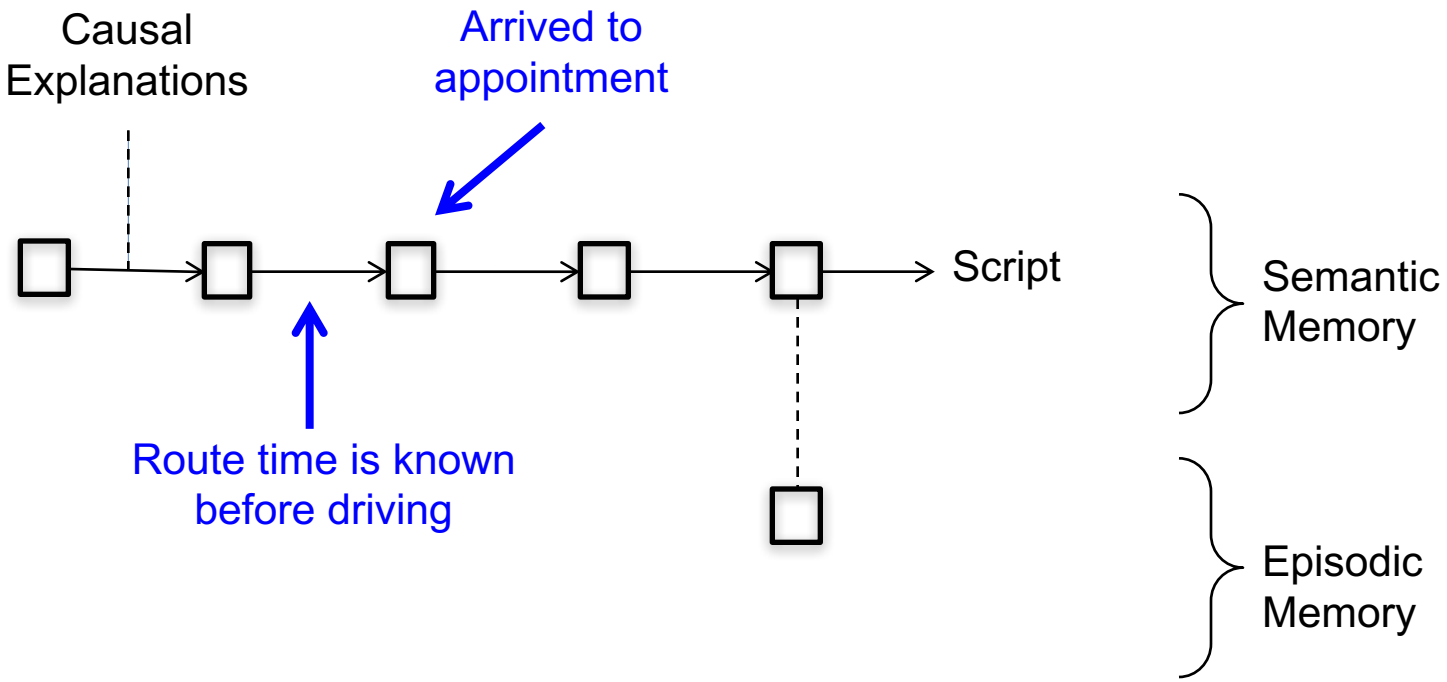
What can you do to overcome this limitation?

Recalling complex events



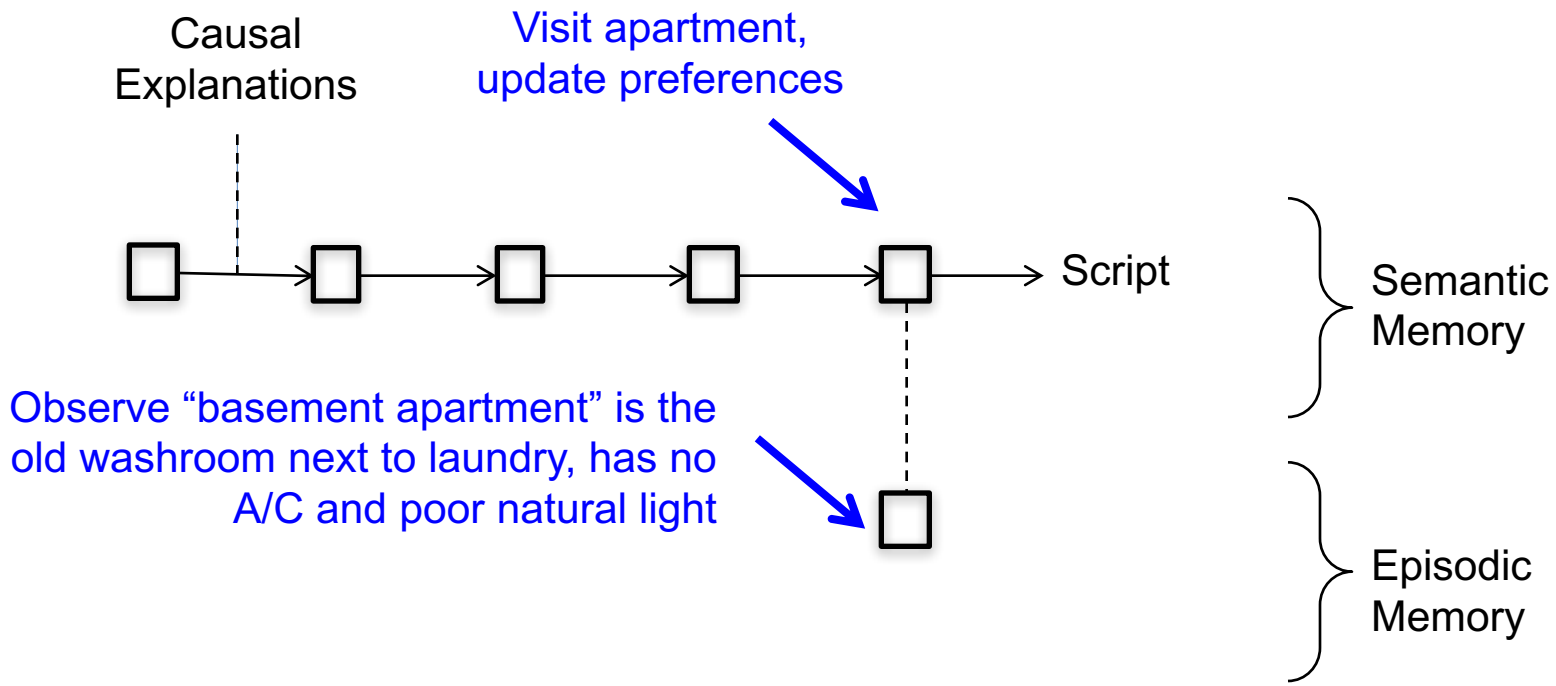
People recall complex events as a single event (Barsalou,1988)

Semantic v. episodic memory



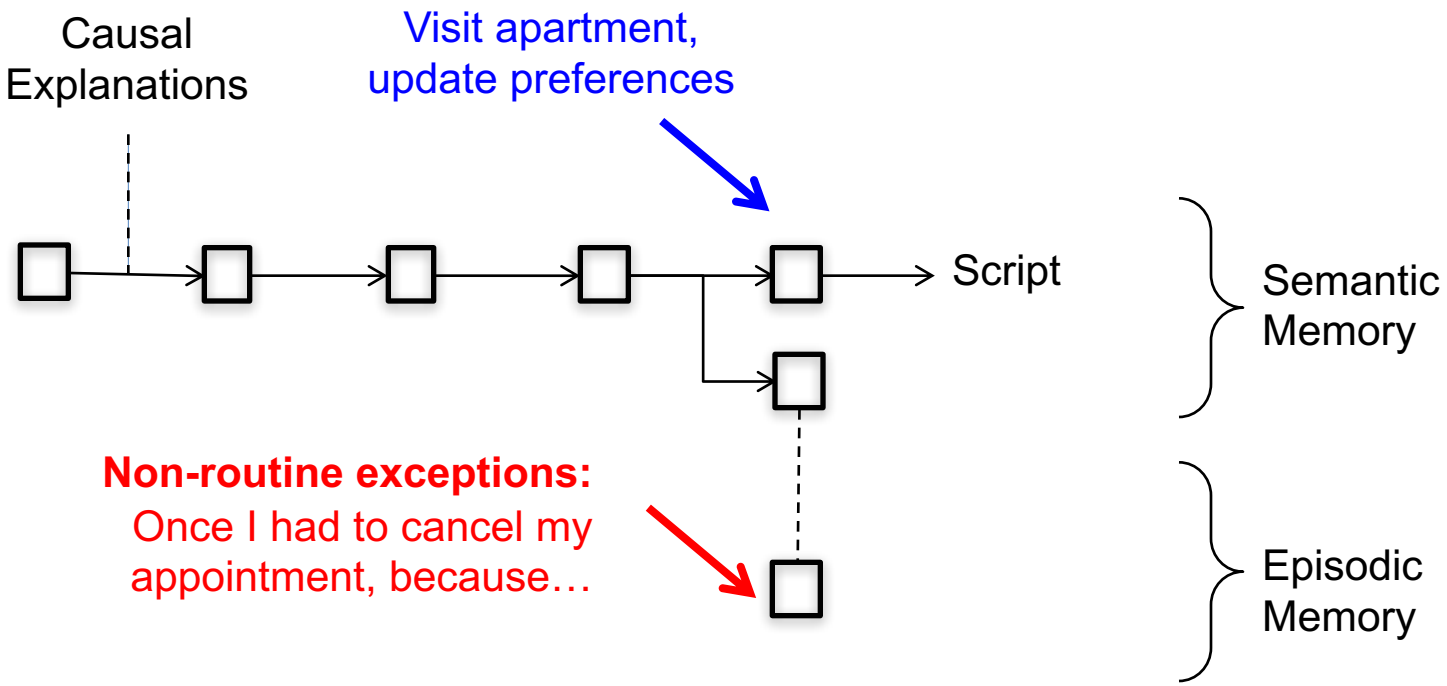
Tulving (1983) early minimalism model distinguishes semantic memory – an abstract network of concepts that affords rich inferential ability – from episodic memory –specific events from which inference is difficult

Semantic v. episodic memory



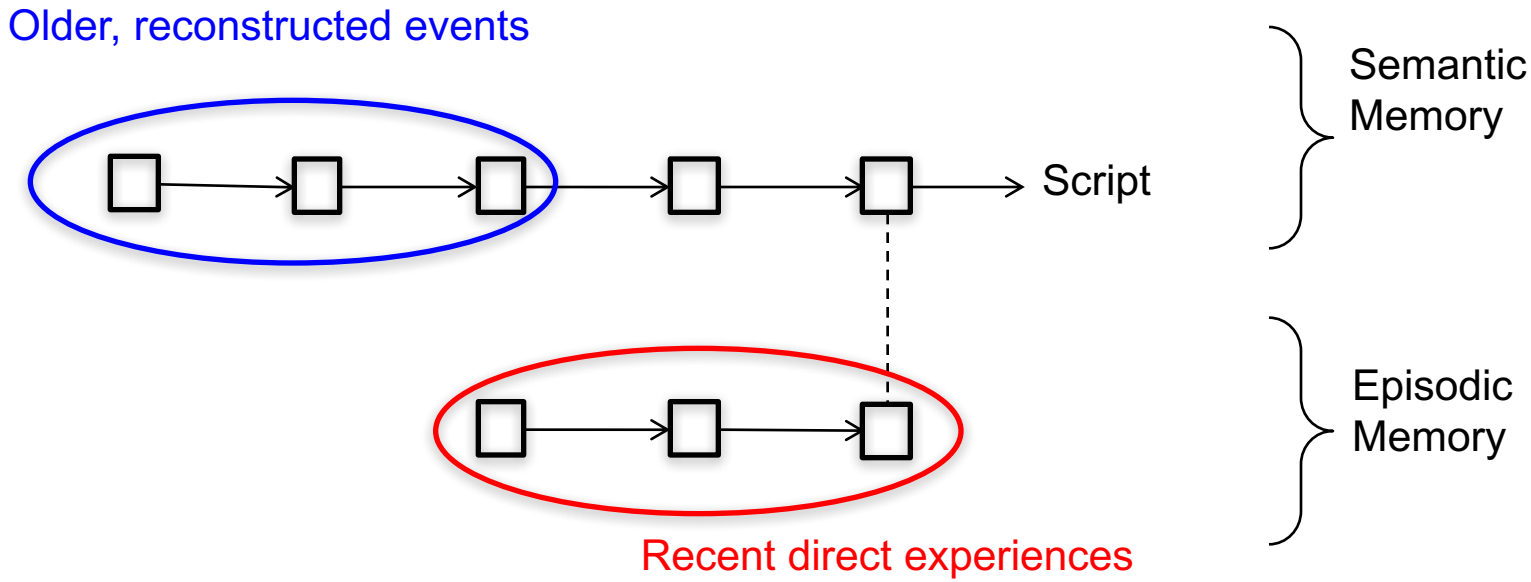
Tulving (1983) early minimalism model distinguishes semantic memory – an abstract network of concepts that affords rich inferential ability – from episodic memory –specific events from which inference is difficult

Semantic v. episodic memory



Tulving (1983) early minimalism model distinguishes semantic memory – an abstract network of concepts that affords rich inferential ability – from episodic memory –specific events from which inference is difficult

Estimation from scripts



Bradburn, Rips and Shevell, 1987 review how people use inference to reconstruct experiences as opposed to using direct experience.

As the retention interval gets longer, respondents rely more on inference and estimation than on recall to report events.

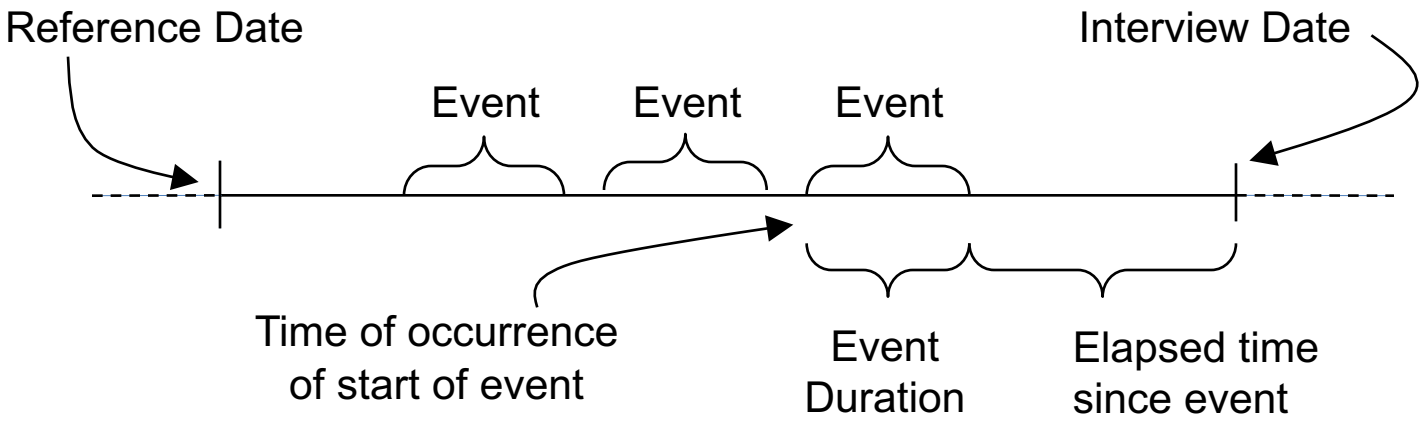
How to apply memory to surveys and interviews

- Distinguish semantic recall from episodic, and probe for specific examples, counter examples
- Prioritize people with recent domain experiences, who presently fill the role of interest
- Ask to engage the person in the domain activity, allow them to describe events in their own words

Remembering dates and durations

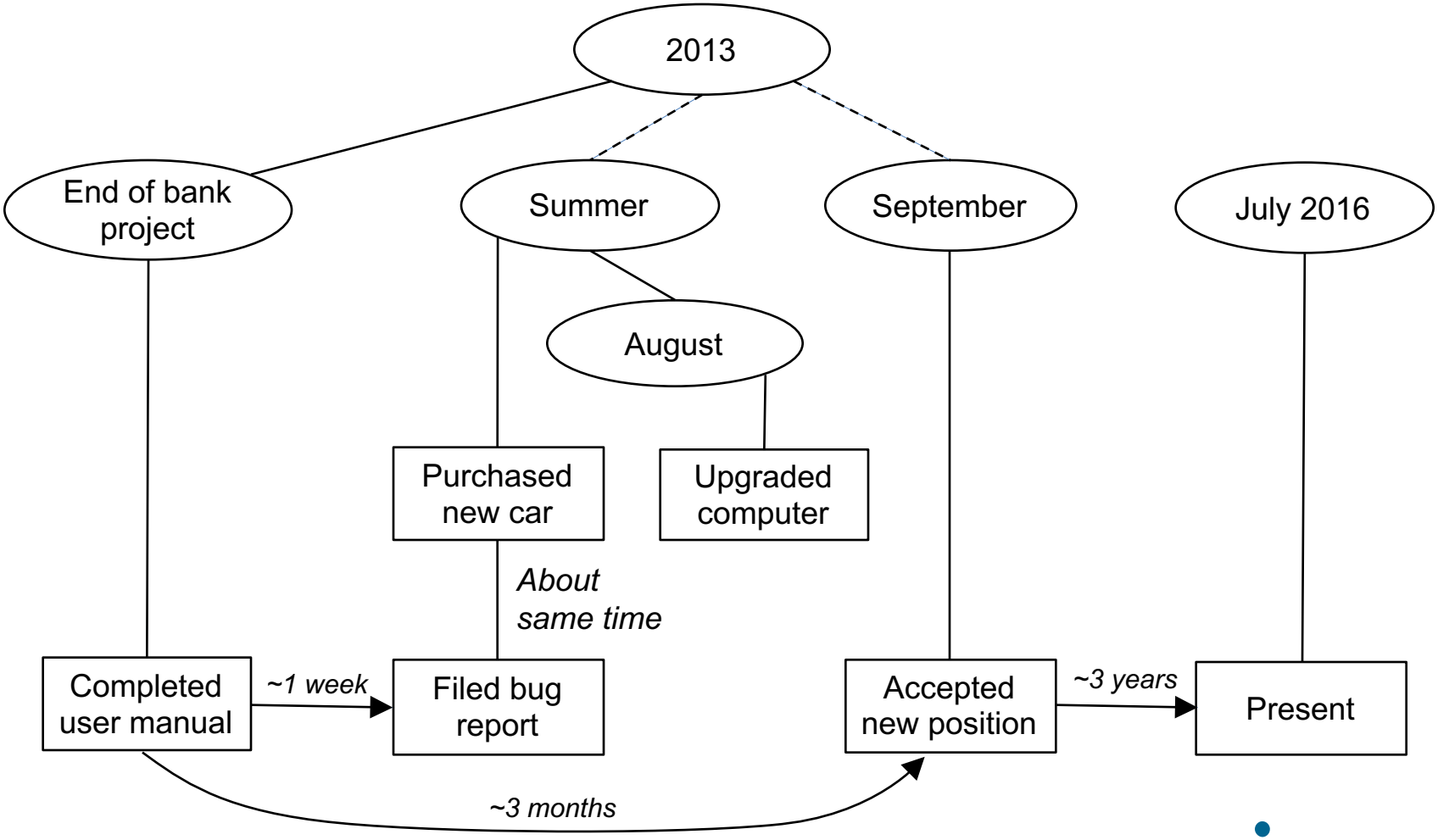
- Landmark dates: People may not remember exact dates, but remembered dates correlate highly with exact dates (Shum, 1997)
- Elapsed time: Estimates of length of temporal interval usually increase linearly with actual duration (Waterworth, 1985)
- Temporal compression: People may mistakenly report events before the a reporting boundary, or omit events from within the boundary (forward and backward telescoping, Neter and Waksberg, 1964)

Typology of Temporal Questions



- Time-of-occurrence questions: on what date did the event happen?
- Duration questions: How long did the event take?
- Elapsed-time questions: How long since the event?
- Temporal-frequency questions: In how many [shorter time units] per [longer time units] did the event occur?

When did you last file a bug report?



Numerical estimation

- When a project succeeds, what is the probability that project success was due to using an agile process?

What kinds of inferences are needed to answer this question?

- Number of projects
- Number of successful projects
- Number of agile projects
- Ratio of successful agile projects to all successful projects
- Ratio of successful to unsuccessful agile projects

To improve accuracy in numerical estimation, decompose complex judgements into simpler component questions

Strategies for answering frequency estimates

- Recall-and-count (episodic enumeration): recall each event and count the events to get the total
- Recall-and-count-by-domain (additive decomposition): recall and count events separately by domain
- Recall-and-extrapolate (rate estimation): recall a few events to estimate the rate, then project rate across the reference period
- Tally: recall current tally of events
- Retrieve rate: retrieve existing information about the rate
- Retrieve recommended rate: retrieve existing information about the recommended rate and then adjust upward or downward
- Guess: rough approximation, direct estimation
- Context-influenced-estimate: use value given by middle response category and then use as anchor and adjust based on impression

Attitudes and beliefs

- Attitudes refer to an individual's opinion or feeling about a topic
- Attitudes have reliability, but not accuracy or validity
- Examples questions:
  - What is your level of discomfort with [event]?
  - How secure or usable is the [system]?

Three sources of answers to attitude questions:

- Impressions or stereotypes (Sanbonmatsu & Fazio, 1990)
- General values and predispositions (Rokeach, Ball-Rokeach, 1989)
- Specific beliefs about the target (Zaller and Fledman, 1992)

Attitude judgements are often created in response to the question

Belief-sampling model (Tourangeau et al.)

- Disparate attitude responses likely based on:
  - Small, non-randomly selected sample of considerations
  - Large, potentially diverse set of considerations that attitudes may encompass
- Homogenous views lead to stable answers over time, whereas mixed views lead to disparate answers depending on considerations

Example:

- Will agile development practices improve your company's ability deliver high quality products?
  - Which specific practices?
  - What are acceptable measures of improved ability?
  - What are preferred measures of product quality?

Context Effects

- Correlational effects refer to the relation among answers to context and target questions
  - If context questions suggest considerations for target questions, and cause divergent views on target questions, this is a correlation effect
- Directional effects refer to changes in the overall direction of answers to target questions
- Context effects include: ordering effects and priming

Context and question comprehension

- Assimilation effects – use of higher-level structures, narratives and survey purpose, to complete missing information
- Contrast effects – specific questions precede more general questions, causing respondents to subtract from responses to second question

Assimilation		**Contrast**
Moderate	Target	Extreme
Hypothetical	Target	Actual / Real
Distracted	Task	Uninterrupted
Low	Motivation	High
Present	Events	Past

Tversky and Kahneman, 1974

- Representativeness – similarity of A and B dominates estimation, ignoring prior probabilities, sample sizes (e.g., B is an outcome of A, if A and B seem related, ignoring evidence and missing information)
- Availability – probabilities increase, if instances or occurrences are more easily brought to mind
- Adjustment and Anchoring – estimates begin from an initial value that is adjusted to fit the situation (different starting points yield different estimates)

Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, New Series, v. 185, n. 4157, pp. 1124-1131, 1974.

RESPONSE OPTIONS

Open-ended v. closed-ended questions

- Open-ended questions allow responses in participants own words
  - Good for developing constructs
  - Responses are typically coded by investigators
- Closed-ended questions limit responses to select choices, for example:
  - Dichotomous questions (e.g., yes/no)
  - Multiple choice
  - Likert and semantic scales
  - Card sorting

Closed-ended Questions

Multiple choice, and ordinal and interval scales provide participants with a limited set of response options (e.g., What is your age range?)

Good practices –

- Solicit response options in a pilot study
- Randomize order, if concerned about order effects
- Avoid bias from unequal response options
- Check all that apply vs. forced-choice

Choosing labels for multiple choice

1. How fast is your Internet connection?
   - Responses: Economy, Business, Gamer
2. How fast is your Internet connection?
   - Responses: fast, slow, crawl
3. How fast is your Internet connection?
   - Responses: 32Kbps, 56Kbps, 384Kbps
4. How fast is your Internet connection?
   - Responses: T1, DSL, Cable, …

Regarding response labels, choose technically accurate terminology for your audience

Eliciting response labels from participants

In a pilot study, ask participants to describe the construct of interest

Example:

- Please list all the words that you could use to describe the speed of an Internet connection.

Constructing an ad hoc scale

Ad hoc scales refer to ordinal and interval scales that relate a construct to numerical measures

Example:

- The following words have been used to describe the speed of an Internet connection. Please rank-order the words from lowest to highest speeds.

- Please assign a numeric value to each of the words above beginning with 10 for the lowest speed. For the next highest speed, assign a value relative to the lowest: assign 20, if the next speed is twice as fast, and 30 if the next speed is three times as fast, and so on

Unequal response options

- How likely are you to share your location to meet friends after work?

  - ❑ Absolutely never
  - ❑ Sometimes
  - ❑ Occasionally
  - ❑ Once or more a week
  - ❑ Everyday

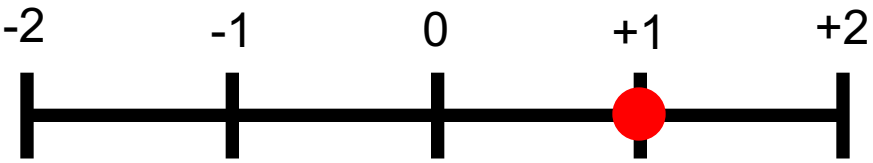
Is it easy or difficult to distinguish between these three categories?

If difficult, why?

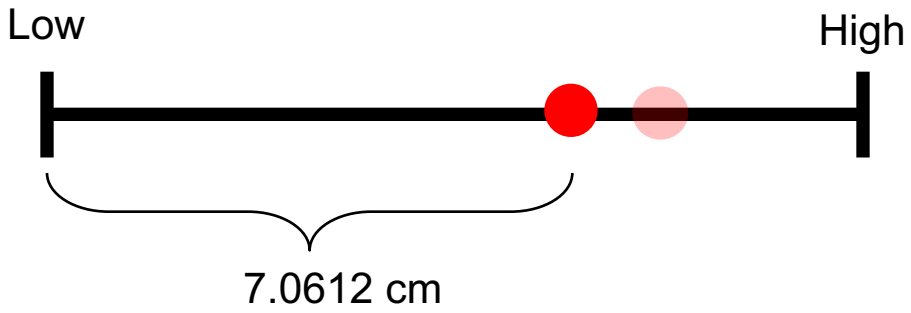
Ordinal scales

- Ordinal or interval scales ask interviewees to choose a "level" of the variable of interest

Numbered Scale:
(choose your number)



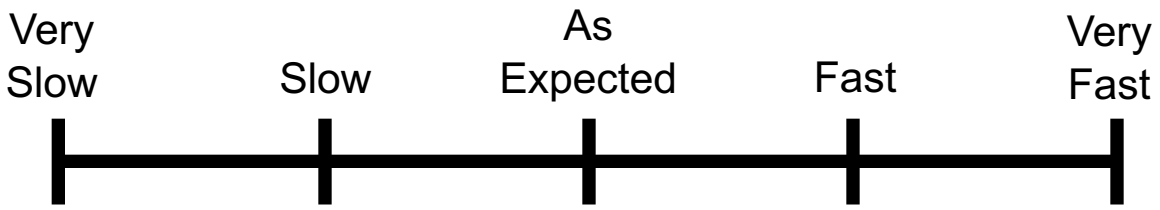
Visual Analogue Scale:
(mark your level)



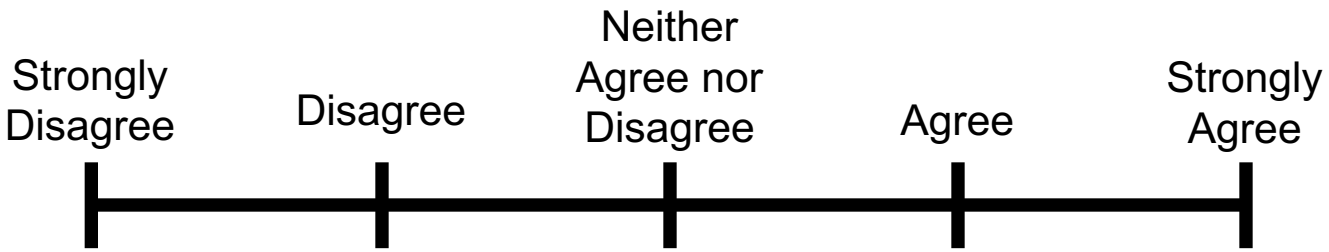
Cowley, Youngblood. "Subjective response differences between visual analogue, ordinal and hybrid response scales," *Human Factors and Ergonomics Society Annual Meeting*, 53(25): 1883-1887, 2009.

Construct-specific v. value-laden

What is the speed of your Internet connection?



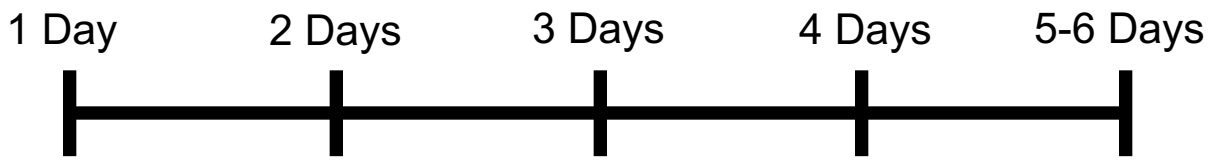
Is your Internet connection too slow?



Distance between anchors

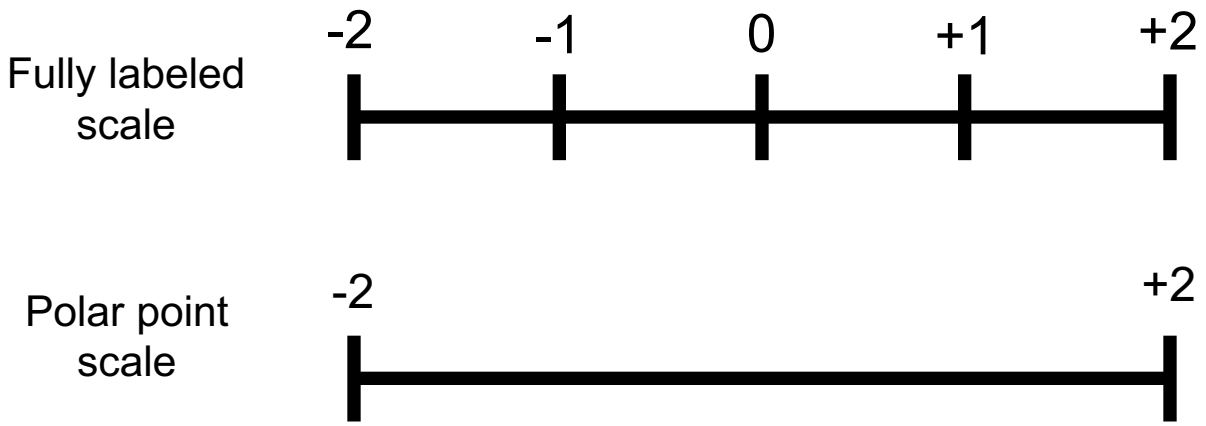
- Frequency in comparable terms (days, weeks, months v. regularly, rarely, occasionally)
- Scales should approximate the actual distribution in the population; balance anchor labels to be equidistant; order should be linear in one row or column, not random

How often do you attend classes each week?

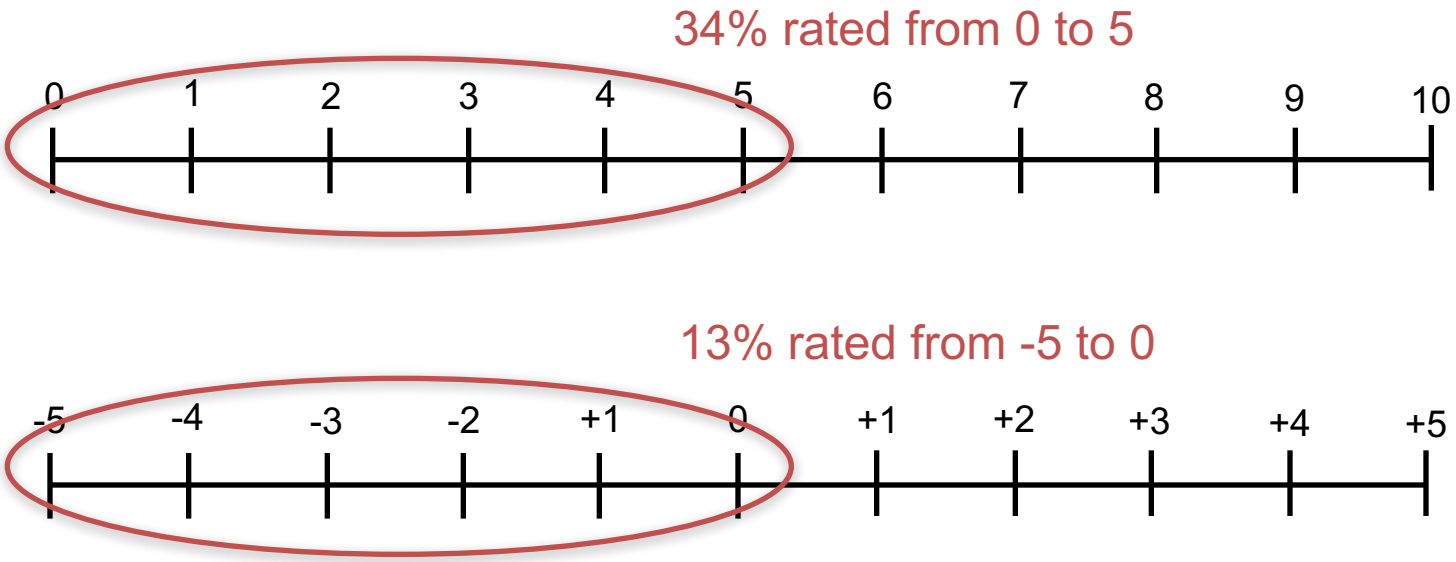


Fully-labeled vs. polar-point

- Fully labeled scales elicit more positive responses than polar-point scales (Dillman & Christian, 2005); this makes comparison across studies difficult
- Fully labeled scales show higher reliability and validity (Krosnick & Fabringer, 1997)



Numbering scale anchors



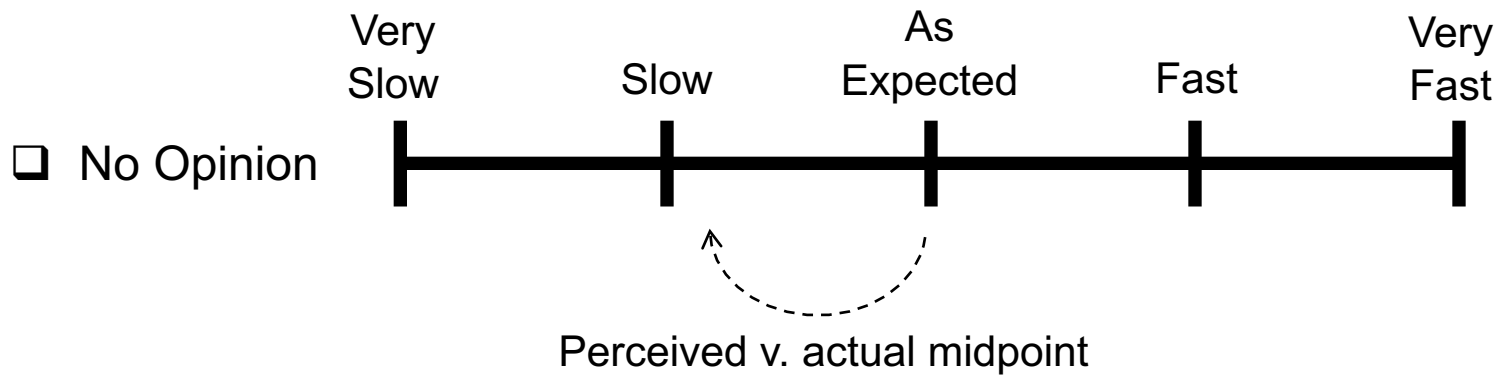
- The 0 was perceived as “no success at all”, whereas -5 was perceived as “explicit failure”
- Verbal labeling can reduce but not remove this effect

Schwarz et al., 1991

Non-substantive options

- Note and place non-substantive options at the end and apart from the scale (“don’t know” or “unsure” or “no opinion”)
- Tourangeau et al. (2004) found that non-apart shifted the visual midpoint and responses toward the negative (lower part of a column)

What is the speed of your Internet connection?



Survey Organization & Execution

- Begin with salient questions that respondents can easily answer
- Keep questions simple to reduce inference and cognitive load
- Group questions by topic
- Keep in mind context effects and biases
  - *Assimilation and contrast effects*
  - *Acquiescence*: the tendency to agree
  - *Social desirability*: the need to present oneself in a desirable light
- Limit surveys to 30-45 minutes
- Focus group the survey on friends and colleagues