EXPERT COURSE 3

RIGHT QUESTIONS, VALID ANSWERS IN DEVELOPING QUESTIONNAIRES

PD Dr. Isabelle Arnet

Dr. Christiane Eickhoff

WHAT TO EXPECT TODAY

- Typical „rules“ for questionnaire design
  - Simple, short questions
  - No hypothetical questions
  - No suggestive or leading questions
  - Unambiguous terms
  - Clear reference period
  - Use exhaustive response categories
  - …
- Instead we will discuss:
  - Your examples and issues
  - Share our experiences
  - Include evidence-based considerations

PROPOSED SCHEDULE

1:00 - 1:20 p.m.	20’	Introducing the moderators and participants. Clarifying expectations.
1:20 - 2:00 p.m.	40’	Case discussion: What do participants use in their work / research? What problems do they encounter? What can we learn from them?
2:00 - 2:45 p.m.	45’	Refresh on questionnaire development, question and response options, pitfalls (connected to participants’ cases)
2:45 - 3:00 p.m.	15’	Break
3:00 - 4:30 p.m.	90’	Option to either: - cont.: refresh on questionnaire or - directly start working in groups according to interest: develop questions and answer options on individual topics
4:30 – 5:00 p.m.	30‘	Presentation of the groups, discussion, wrap-up, conclusions

REFRESHER - DEVELOPING A QUESTIONNAIRE

- Question answer process
  - What happens on the respondent‘s side?
  - Satisficing
- Questions and scales
  - Type of questions
  - Question wording
  - Response options and scales
- Common pitfalls
  - Response errors
  - Asking sensitive questions
- From questions to questionnaire
  - Survey modes
  - How to group questions
  - Question order effects
  - Pre-testing methods

THE BASIS FOR A QUESTIONNAIRE: THEORY AND CONCEPT

- Step 1: define the research question (theoretical concept)
- Step 2: „translation“ into suitable indicators and variables (operationalization)
- Aim is construct validity:
  - All aspects are covered by the instrument (quantitative aspect)
  - The indicators and variables are suitable (qualitative aspect)

Question answer process

QUESTION ANSWER PROCESS

- Before 1980: it was believed that responses are always true
- Today we know many (*actually all…* ☹) aspects have an influence:
  - Type of questions
  - Answering options
  - Wording
  - Context
  - Scales
  - Survey mode
  - Visual aspects
  - …

AN EXAMPLE

„Tennis and soccer are the two sports that draw the largest audience. When you compare both sports with regard to how entertaining they are for a TV audience, would you say that tennis is more exciting to watch than soccer, or that tennis is less exciting to watch than soccer?“

- Tennis is much more exciting than soccer
- Tennis is more exciting that soccer
- Tennis in somewhat more exciting than soccer
- Tennis and soccer are equally exciting
- Tennis is somewhat less exciting than soccer
- Tennis is less exciting that soccer
- Tennis is much less exciting than soccer

- 85% considered tennis to be more exciting
- BUT: if you change the direction of comparison: only 35% consider tennis to be more exciting

Waenke, Schwarz (1992) https://www.ssoar.info/ssoar/handle/document/6978

QUESTION ANSWER PROCESS

- Meaning of words and context
- Respondents assume that all has a purpose *(wording, response options, question order, graphical information, scales, size of boxes….)*

- Fitting the response in the response options

- Cognitive process
- Searching relevant information, decision whether to include it or not and generating a subjective response

- Also depends on social desirability

Cannell, Miller & Oksenberg (1981). Research on interviewing techniques. In: Social Methodology. San Francisco: Jossey-Bass Publishers.
Tourangeau, Rips, & Rasinski (2000). The Psychology of Survey Response. Cambridge: Cambridge University Press.

SATISFICING

- How respondents answer depends on: motivation, capability and satisficing
- Satisficing (composed of “satisfy” and “suffice”)
  - Decision-making strategy aiming at a satisfactory result rather than the best (pragmatic approach)
- Weak satisficing: superficial question answer process
  - E.g.: choosing the first acceptable response, non-differentiation, estimating, guessing
  - *Very widely found => should be assumed* (no big problem)
- Strong satisficing: certain stages of the question answer process are skipped
  - E.g.: keyword search, random answers, neutral answers or “don’t know”

Krosnick (1999). Maximizing questionnaire quality. In: Measures of Political Attitudes (pp. 37-57). San Diego: Academic Press.
Krosnick (1999). Survey research. Annual Review of Psychology, 50, 537-567.

SATISFICING

- Factors leading to satisficing:
  - Too difficult
  - Low education level (if you pilot -> include lower educated people)
  - Low motivation
  - Long questionnaires, boring issues, …

Type of questions and question wording

QUESTION TYPES

- Closed questions
  - Limited choice vs. multiple response question
  - Matrix question
  - Forced choice questions / check all that apply
  - Ranking questions
  - …
- Open-ended questions
  - Narrative questions
  - Frequency questions
  - …

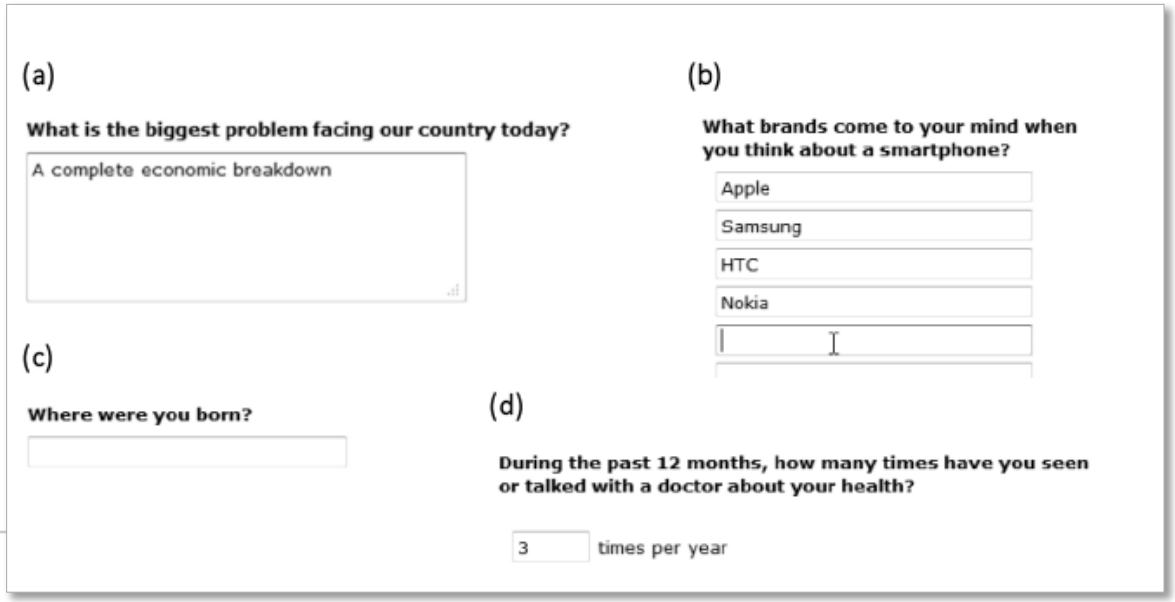
CLOSED-ENDED QUESTIONS

Closed-ended question (vs. open-ended question)

- Easier to answer and to analyse
- BUT: response options define the range of potential answers
  - Respondents will not think “outside the box”
  - Extensive prior knowledge is needed
    => *if you don’t know the proper categories, ask open questions*
- Partly closed questions: not favored
  - Only 2% of respondents will fill in “others” if offered. You might miss a lot!

OPEN-ENDED QUESTIONS

- Advantage: Respondents are not influenced by response categories
- Disadvantages:
  - Takes time and effort to answer them => burdensome
  - Depends on the skills (oral and writing) of the respondent
  - In self-administed surveys: often low data quality
  - Huge effort needed to analyse the data
- Types:
  - (a) narrative
  - (b) list-style
  - (c) short
  - (d) frequency



OPEN-ENDED QUESTIONS – VISUAL DESIGN

- Larger text areas (in web and paper based surveys)[1]:
  - Response length increases
  - Number of reported topics increases
- Increasing the number of text boxes[2,3]:
  - Number of items reported increases
  - Number on non-responses increases

[1] Christian et al. (2004). Public Opinion Quarterly 68(1): 57-80.
[2] Smyth et al. (2007). 62nd Annual Conference of the American Association for Public Opinion Research (AAPOR).
[3] Fuchs (2013). Annual Conference of American Association for Public Opinion Research, Boston.

OPEN-ENDED QUESTIONS - VERBAL DESIGN

- Effect of verbal design in narrative open-ended questions (e.g. in web surveys)

Version (1)

In your own words, how would you describe your advisor(s)?

Version (2)

This question is very important to our research. In your own words, how would you describe your advisor(s)?

Version (3)

Please take your time answering this question. In your own words, how would you describe your advisor(s)?

Version (4)

This question is very important to our research. Please take your time answering this question. In your own words, how would you describe your advisor(s)?

- Clarifying and motivating instructions
  - Increase response length
  - Increase number of reported topics

	Mean number of words		Mean number of themes		Percent elaborated		Mean response time (seconds)	
	Early	Late	Early	Late	Early	Late	Early	Late
Survey 3								
Q6: No intro. (1)	28.1	23.0	2.6	2.2	46.2	38.0	58.7	52.0
Q6: "Important" (2)	37.3	37.4	2.8	2.8	53.1	56.7	77.7	77.4
Q6: "Take time" (3)	37.8	31.2	3.0	2.3	60.1	61.4	80.2	71.4
Q6: Both intros. (4)	42.5	37.5	3.0	2.7	63.6	58.9	85.6	85.5

Smyth et al. (2009). Public Opinion Quarterly, 73(2): 325-337.

BEHAVIORAL FREQUENCY QUESTIONS

» Many options…

When you get together with your friends, would you say you discuss political matters frequently, occasionally or never?
Frequently
Occasionally
Never

(EVS, 2008)

On an average weekday, how much time, in total, do you spend watching television?
No time at all
Less than ½ hour
½ hour to 1 hour
More than 1 hour, up to1½ hours
More than 1½ hours, up to 2 hours
More than 2 hours, up to 2½ hours
More than 2½ hours, up to 3 hours
More than 3 hours

(ESS, 2014)

When you participate in sports, hobbies, or other activities that can cause eye injury, on average, do you wear eye protection always, most of the time, some of the time, or none of the time?
Always
Most of the time
Some of the time
None of the time

(NHIS, 2016)

About how many minutes or hours per week do you spend sending and answering electronic mail or e-mail?
Hours:
Minutes:

(GSS, 2016)

BEHAVIORAL FREQUENCY QUESTIONS

- Task for the participants (question answer process):

- Strategies to respond:
  - Recall and count => *for low numbers and irregular events*
  - Estimation => *for high numbers and regular events*
    (done by guessing or rate based estimation; takes less time and effort)
- Factors to consider:
  - Frequency and regularity of the behavior
  - Reference period (length, distance)
  - Adjustments (for special period like holidays, seasons etc.)
    => Ask for a typical week? Or just live with it?
  - Format (open or closed)

BEHAVIORAL FREQUENCY QUESTIONS

Which scale to use?

Version A	Version B
How many hours a day do you watch television? • up to ½ hour • ½ to 1 hour • 1 to 1½ hour • 1½ to 2 hours • 2 to 2½ hours • more than 2½ hours	How many hours a day do you watch television? • up to 2½ hours • 2½ to 3 hour • 3 to 3½ hour • 3½ to 4 hours • 4 to 4½ hours • more than 4½ hours
>2 ½ hours: 16.2%	>2 ½ hours: 37.5%

» When estimating:

› The response scale is used as an orientation

› Respondents assume that the middle is the average => scale effect

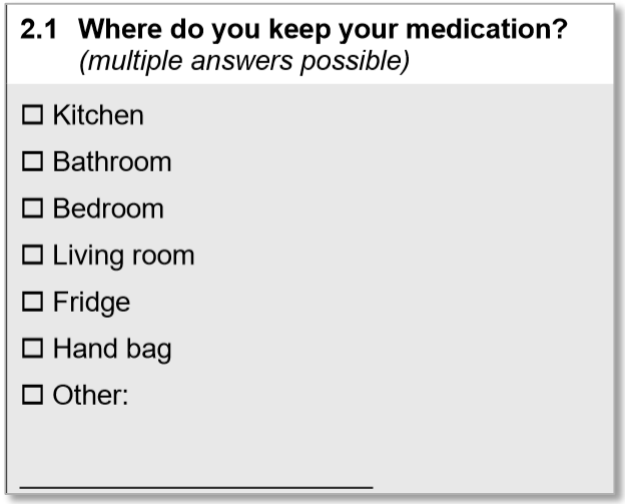
Fuchs 2005

BEHAVIORAL FREQUENCY QUESTIONS

- Open or closed frequency questions?
  - If you don't know the distribution: *use open frequency questions*
  - If you know the distribution: *6 categories are optimal*
- Things to consider:
  - Labelling boxes produces better quality (date e.g. DD/MM/YY)
  - Open questions suggest that the response is larger than zero => zero filter
  - When estimating: respondents prefer rounded numbers

MULTIPLE RESPONSE QUESTIONS

- Several options can be selected:



- Primacy effect => dominant in visual presentations
  - Selection of the first acceptable answer
  - Fewer cognitive effort is spend on items at the bottom of lists
- Recency effect => dominant in oral presentations
  - Limited memory (most processing time is spend on final items)
- *Expected especially in lower educated, children and older people*

MULTIPLE RESPONSE QUESTIONS

What to do?

- To reduce the primacy effect:
  - Splitting the items into two columns => works only for the first items
  - Blocks with subheadings
  - Randomizing the order => bias goes to random error
- Forced choice format (yes/no)
  - Longer response time
  - Increases number of “yes” answers (Satisficing? Deeper processing?)
- Limit the number of items (max. 7-8)
- Ask respondents to click only the 3 (e.g.) most important items

=> *Discussion if these kind of questions should be used at all*

MATRIX QUESTIONS

- Consist of several related items measuring one latent variable
- Respondents rate on a ordinal scale, e.g. agree to disagree

How do you see yourself and your situation? Please assess the following statements:				
	strongly agree	*agree*	*disagree*	*strongly disagree*
I know for each of my medications when and how to take it.	☐	☐	☐	☐
I know for each of my medications why I take it.	☐	☐	☐	☐
My drug therapie is much too complicated for me (e.g. how to take it and how often).	☐	☐	☐	☐

MATRIX QUESTIONS

- Advantages:
  - Space saving
  - Relation between items is emphasized

  => *Easy and efficient to complete*
- Disadvantages:
  - Much information in rows and columns
  - Potentially less attention to individual items and response options
  - People don't like them (break-offs, item non-response/omission)

  => *Risk of superficial processing and reduced data quality*

MATRIX QUESTIONS

strongly agree	agree	partly agree, partly disagree	disagree	strongly disagree
❍	❍	❍	❍	❍

What to consider:

- Symmetry: equal number of positive and negative options with equal distance
- Midpoint: mixed results concerning reliability and validity:
  - if offered: invite less motivated respondents to satisfice
  - it not: forced choice may result in “mental coin flipping”
- No of item: 4 to 10 per matrix question
- No. of scale points: 7±2 (more detailed if you want to rank items)
- Verbal labeling: full labeling is more meaningfull (increases data quality)

WORDING OF THE QUESTION TEXT

- Issue labeling: choosing "accepted" wordings or adding a positive verb/phrase increases support, e.g.[1]
  - Using "assistance to the poor" instead of "welfare": 62.8. vs. 23.1%
  - Using "protecting social security" instead of "social security": 68.2 vs. 53.2%
- Including an alternative may modify responses, e.g.[2]
  - "Are you in favor of giving special priority to buses in the rush hour?" vs. "Are you … *or should cars have just as much priority as buses?*": agreement 69 vs. 55%

[1] Schuman & Presser (1981). *Questions and Answers in Attitudes Surveys (reprint 1996 by Sage ed.). San Diego, California: Academic Press*

[2] Rasinski (1989)

WORDING OF THE QUESTION TEXT

» Balancing the question wording gives equal credit to both viewpoints (otherwise a bias in a particular direction is introduced)

Unbalanced form		Formally balanced form	
1. Fuel shortage (SRC-74 Fall)			
If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes?		*If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law?*	
Yes, should be a law	38.3%	Should be a law	29.4%
No, not a law	61.7	Oppose such a law	70.6
	100		100
	(507)		(494)
$\chi^2 = 8.90$, $df = 1$, $p < .01$			

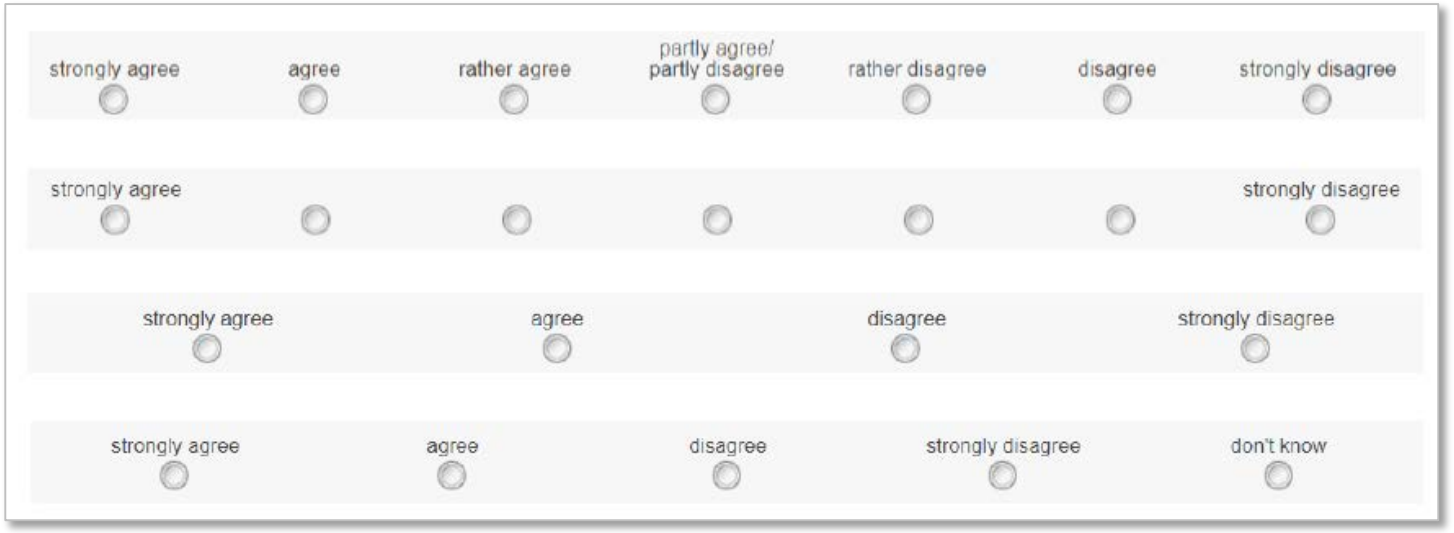
=> Also minimal balancing is effective (Should there be… or not…)

Schuman & Presser (1981). *Questions and Answers in Attitudes Surveys (reprint 1996 by Sage ed.). San Diego, California: Academic Press*

Response options and scales

RESPONSE CATEGORIES

- Number of response options
- Polarity of response options
- Labeling of response options
- Using a middle category
- Providing „don‘t know“



NUMBER OF RESPONSE CATEGORIES

- Should include neutral, moderate and extreme responses
- Not too many response options
  - In telephone less than with visual options
  - Respondents may use only part of the scale
    - Depends on how detailed people can differentiate
    - Respondent is not willing to use the full range
- Thumb rule:
  - *Bipolar scales: 5 or 7 response options*
  - *In unipolar scales: 4 to 7 response options*

BIPOLAR AND UNIPOLAR SCALES

» Bipolar:

strongly agree	agree	partly agree/ partly disagree	disagree	strongly disagree

› Both endpoints should have the same intensity (but of opposite value)

› Conceptual midpoint, e.g. partly agree/disagree, neutral, neither/nor

› *People are likely to choose the middle => try to avoid it*

» Unipolar:

very important	important	moderately important	rather not important	not at all important

› Scale differs in intensity

› Midpoint is a moderate position

LABELING RESPONSE CATEGORIES

- Verbal vs. numeric labels:

- In general: Labeling helps:
  - Interpreting response scales
  - Clarifying meaning of the question
- Verbal labels:
  - Limit the respondent's burden (formatting process!)
  - *Get rid of numbers*: you never know what they mean to a respondent

Krosnick & Fabrigar (1997). Designing rating scales for effective measurement in surveys. In: Survey Measurement and Process Quality. New York: Wiley.

LABELING RESPONSE CATEGORIES

- Numbers may change the meaning of scales
- Famous experiment:

13. What would you say: Overall, how successful have you been in life?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
−5 −4 −3 −2 −1 0 +1 +2 +3 +4 +5
not at all successful — very successful

Lower 5 categories selected by 13%
Mean: 7.3

13. What would you say: Overall, how successful have you been in life?

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
0 1 2 3 4 5 6 7 8 9 10
not at all successful — very successful

Lower 5 categories selected by 34%
Mean: 6.4

- Respondents think of “0” as no success and negative numbers as failure
  - The scale with “0-10” is represented by “0 to +5” in the left scale
  - Mismatch: success and failure

Schwarz (1991); Fuchs (2005)

RATING VERSUS RANKING

- Rating:
  - Items are rated on a pre-defined scale
  - Method: matrix question
  - Items can then be put in order by the researcher
  - Challenge: possible that some items get the same values

- Ranking:
  - All items are put in an order
  - Challenges: time consuming, satisficing
  - Ranking looks better (but in fact is not)

Blasius (2012)

Common pitfalls

RESPONSE ERRORS

- Systematical selection of a specific response option, irrespective of the content
- Most common types:
  - “Don’t know” responses
  - Acquiescent responding
  - Extreme vs. midpoint responding
  - Non-differentiation

Relevant factors:
- survey mode
- respondent (satisfycing!)
- the questionnaire itself

SHOULD WE OFFER A „DON‘T KNOW“ CATEGORY?

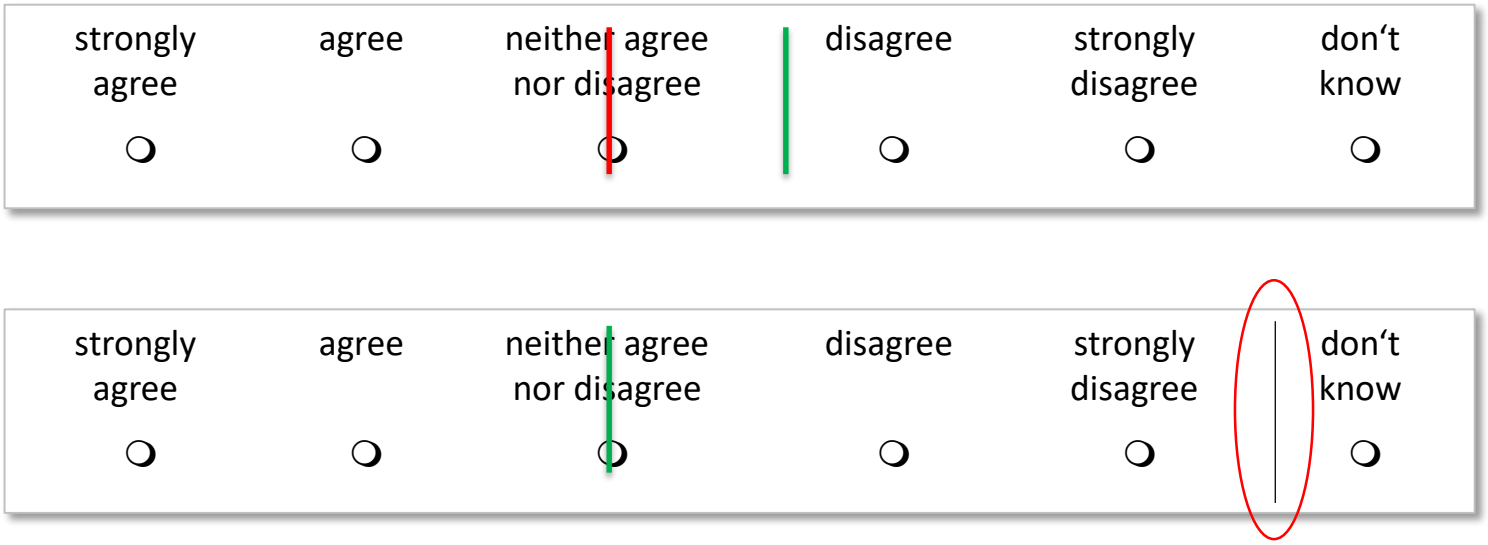
- In interviews: not needed
- Self-administered questionnaires:
  - If provided: more respondents report don‘t know“ => an easy way out
  - If not provided: respondents without an option might express moderate option or select midpoint of scales[2]
- *Don‘t provide it* when you can assume that respondents have an optinion (e.g. attitude questions)

SHOULD WE OFFER A MIDDLE CATEGORY?

- Only matters in bipolar scales
- Unclear (mixed results in terms of validity and reliability):
  - If not provided: forced choice my reult in „mental coin flipping“
  - If provided: less motivated respondents might satisfice
  - Reasons unclear: maybe indifference (only moderate opinion) or ambivalence (unstable opinion)

SHOULD WE OFFER A MIDDLE CATEGORY?

- The middle is often used as an anchor (considered average or typical)
- If „don‘t know“ or „no opinion‘“ are added, the conceptual and visal midpoint is not aligned any more => significant shift of the mean



ACQUIESCENCE

- Tendency to agree rather than disagree irrespective of the content
  - Weak satisficing: premature termination of cognitive processing
  - Strong satisficing: reliance on the social convention to be polite
- More likely with
  - Items that are ambiguous or vague in content
  - Items of little or no personal relevance for the respondent
- Consequence: observed means may be inflated (or deflated)

NON-DIFFERENTIATION

- Tendency to use (nearly) the same response options to answer a matrix question
- Often associated with speeding
- What can be done? => Change the „direction“:

	strongly agree	agree	neither agree, nor disagree	disagree	strongly disagree
On the whole, I am satisfied with myself.	○	✕	○	○	○
At times I think I am no good at all.	○	✕	○	○	○
I feel that I have a number of good qualities.	✕	○	○	○	○
I am able to do things as well as most other people.	○	✕	○	○	○
I feel I do not have much to be proud of.	○	✕	○	○	○
I certainly feel useless at times.	○	○	○	○	○

sdfgsdgsdsdgsdgsdfgsdf

ASKING SENSITIVE ISSUES

- Typical sensitive issues: income, sex, drug addiction, voting, discrimination, donating
- Behind this is: intrusiveness, thread of disclosure, social desirability
- Consequences:
  - Block or item non-response
  - Misreporting (systematic editing)
- *What to do?*
  - Assurance of confidentiality
  - Use self-administered modes
  - Use of alternative question strategies

ASKING SENSITIVE ISSUES

- Alternative question strategies:
  - “Forgiving” wordings: “It is very common that… “ (Adherence!)
  - Wordings presupposing the attitude or behavior in question (trap effect): “How many cigarettes do you smoke each day?”
  - Ask for income ranges instead of exact amounts (unfolding bracket technique)
  - Encouraging question wordings: “Have you had time …?”
  - Offer an extra explanation in a box => but careful: cannot be used too often because it is an additional burden (length goes up, drop off rate…)

From questions to questionnaire

SURVEY MODES

Technology	Type of administration		
	Interviewer administration		Self-administration
	via telephone	face-to-face	
Computer-assisted data collection	CATI	CAPI	Web survey
Paper and pencil	PATI	PAPI	Paper & pencil, mail

CAPI: computer-assisted personal interviewing, CATI: computer-assisted telephone interviewing, PAPI: paper-and-pencil interviewing, PATI: paper-and-pencil telephone interviewing.

» The survey mode has a huge influence on the results

SURVEY MODES

- Example[1]:
  - Same sample, field period, question
  - Different modes
- Results depend on the mode
- Causes[2]:
  - Mode characteristics
    - presence or absence of interviewer
    - channel of communication
  - Different question wording and response options
  - Measurement error

Figure 6.3 Differences between percent of self-administered (mail) and interview (telephone or face-to-face) respondents selecting most positive response categories in three surveys.

	Average Percent "Not a Problem"		Range of Differences Interview–Self-Administered
Dillman and Mason, 1984	Interview	Mail	
Nine community issues from two-state general public survey, telephone vs. mail	47.9%	31.9%	6.7–18.8
Nine community issues from same two-state general public survey, face-to-face vs. mail	43.8%	31.9%	4.2–18.7
Tarnai and Dillman, 1992			
Five community issues from student survey; self-administered vs. telephone (aural only)	37.8%	13.0%	12.6–38.1
Five community issues from student survey; mail vs. telephone while viewing questionnaire (aural + visual)	31.3%	13.0%	4.3–20.0

Question: *"For each issue, please tell me whether you believe it is Not a Problem, a Small Problem, a Medium Problem, a Serious Problem, or if you Don't Know."*

[1] Dillman et al. (2008). Internet, mail, and mixed-mode surveys: The tailored design method. New Yoik: Wiley.
[2] de Leeuw & Hox (2015). Survey mode and mode effects. In: Improving Survey Methods (pp. 22-34). New York. Routledge.

SURVEYS ON MOBILE DEVICES

- Web surveys are often designed for big screens
  - More break off if surveys are not optimized to mobile devices
  - Responsive design: only a technical solution => visual design differs
    - Size of the screen
    - Orientation (landscape vs. portrait) and
    - Input method (no keyboard)
    - Privacy
    - Different users
- Interesting to collect para data (e.g. how long took the answer, device used, interruptions)

Kosnick & Fabricar (1997); [2] Rasinski (,1989); 3

HOW TO START AND GROUP QUESTIONS

- Initial, general instructions
- First question as an icebreaker
  => ideally simple, interesting and applicable to all respondents
- Related questions are grouped and put in logical order
  (like a conversation)
- Arranging the modules:
  - Easy block first, then increasing the difficulty
  - Block length should ideally decrease

POSITION OF SOCIO-DEMOGRAPHIC QUESTIONS

- Beginning of questionnaire:
  - Higher break-off because not interesting and sometimes personal
- End of questionnaire => *often recommended*
  - Build up trust first, then ask sensitive issues
  - But: in case of break-off, responses are of little value

QUESTION ORDER EFFECTS

- Preceeding questions are used to interpret the meaning of following questions
- Most relevant between consecutive questions
- Effects:
  - Assimilation effect (including information)
    - Most common and happens automatically
  - Contrasting effects (exclusion)

QUESTION ORDER EFFECTS

Should be anticipated => cannot be avoided

- Adapt your questionnaire to everyday thinking
  - Most common responses should come first (otherwise: risk of anchoring)
- To avoid (minimize) a context effect => add a buffer question
  (e.g. a knowledge question from the same context)
- Specific or general questions first (part-whole questions)? => both possible
  - Start with general questions first, then go into details
  - If you want a summary judgement: general question at the end
- Cognitive pretest!

PRE-TEST METHODS

"If you don't have the resources to pilot test your questionnaire, don't do the study."

"Even after years of experience, no expert can write a perfect questionnaire"

Sudman & Bradburn 1983

- Methods before constructing the questionnaire (focus groups, expert discussions)
- Observational pre-tests
- Cognitive pre-tests

STANDARD PRE-TEST

- Aim: to see if the questionnaire is functioning
- The questionnaire is tested under realistic conditions
  - Interviewer observes and reports problems (response latency etc.)
  - Respondent usually does not know that it is a pilot
  - Conclusions are drawn from the respondents answers and reactions
- 20-50 interviews (quota or random)
- Advantages:
  - Easy, quick, cheap, you gain information on: technical issues, handling, realistic Estimation on the average duration of the interview
- Disadvantages:
  - Only superficial information, rather rough

COGNITIVE PRE-TESTS

- Aim: to identify problems of comprehension
- 10 interviews
- Techniques:
  - Thinking aloud: *most popular*!! => Special competencies required (e.g. depending on the educational level)
  - Probing: Additional questions asked after each question/section, e.g. "Why did you pick this answer?", "How do you define the term 'xyz'?"
  - Paraphrasing: e.g. "Please repeat the question in your own words"
  - Confidence rating: e.g. "How accurately could you rate the number of visits at the doctor in the past 12 month?"

Universität Basel

ABDA

And many more issues to discuss...