Summary of Fundamental of Data Collection I and II

Instructors: Frederick Conrad & Katharine Abraham

1.1 Comparing (designed) survey data and (organic) found data

Six dimensions that determine the fitness of use of a statistics.

- Relevance
- Accuracy
- Timeliness
- Coherence
- Accessibility
- Interpretability

	Survey Data	Found Data
Relevance/Accuracy	- Small scaled but designed to be representative	- Large scaled but not necessarily representative; typically convenience sample
Relevance	- Data units (households, individuals etc.) selected to meet statistical needs	- Data units meet other needs (e.g., Driving history, one record every 10 seconds, might not know the person behind these driving records)
Relevance	- Designed for statistical need, often include unique identifier to make linkage	- Designed for social, business needs, may not be linkable with other sources
Accuracy	- Quality control central to the survey organization, but various error occurs in the data collection process, measurement error in the survey responses	- Data elements relevant to the administrative and business needs are likely to be accurate (less measurement error)
Coherence	Comparability over time is controlled by the survey organization	Comparability over time may be disrupted by the changes of business or administrative requirement
Timeliness	Reasonably timely. Data release is a controlled process	Depends. Administrative data lag, but business and web interface data are quick

In a sentence, the major advantage of the survey data is that they are more controlled, so they can directly speak to our needs. But the drawback is that they might not be as timely and accurate as the large scaled found data.

2.1 Frame inadequacy results in coverage errors.

Examples of different frames:

	Area Frame	RDD	Address-based sampling frame	Business and organizational list frame
Suitable situations	- Mainly households. - Rarely for business establishment, since not all business have an obvious storefront presence. Exception could be for farms or private schools, both of which have physical presence	Households	Households	Business and organization
Steps	- Select subsets of areas - List housing units within these areas - Select subsets of housing units - Create rosters within the housing units	- Used to be landline only - Now incorporate cell phones	Commercial frame from US Postal Service Delivery Sequence (rather high coverage rate > 97%)	Typically based on administrative information compiled for some other purposes, for example, tax records
Challenges	- Creating the list of housing units is not trivial. Given a house, it is not clear how many housing units there are. - Creating rosters is not trivial. Consider the article read in TSL. In fluid households, who are residents is not clear - Under-coverage, duplication, clustering are all potential challenges	- Landline-only samples become increasingly problematic, given the large share of cell-phone samples - Cell phones are portable. Less accurate information about where the respondents live - Cell phone has to be dialed manually by law. More cost - Varying selection probabilities because people have multiple numbers (weighting) - Under-coverage of people who do not have phone access	- Better coverage in urban than in rural areas - Multiplicity: One household can have multiple addresses - Clustering: Multiple households can pick up from the same address - Foreign element: Vacant/ address can be business rather than residential	- Under-coverage: Tax records may not include the newly formed business due to lag in tax filing - Over-coverage: Failed to remove death - Duplication due to incorporating information from multiple sources - Difficulty in identifying linkage over time

2.2. Classifying response status

The response status of survey cases can be classified as

- Complete or partial complete
- Eligible but no response
- Illegible
- Unknown eligibility

Classification to one of the four categories depends on the method of the survey and the nature of the target population

	In-person household survey	RDD telephone survey	Mail survey
Complete or partial complete	Provide acceptable responses to a minimum set of questions	Same as in-person	Same as in-person
Eligible but no response	Refusal, non-contact	A substantive problem for telephone survey. Often difficult to determine the eligibility of the phone numbers. When no response, hard to tell whether it is eligible	A substantive problem for mail survey. Often difficult to determine the eligibility of the address, when no mail is returned.
Unknown eligibility	Interviewers unable to visit (e.g., dangerous areas, gated community)		
Illegible	Business, vacant, households with illegible residents	Business number; numbers associated with illegible respondents	Return mail stating that the respondents is not available (rare).

3.1 Representation of web panel survey (moved below)

3.2 Representation of non-survey data

Administrative data

Population covered by administrative data is generally well defined. E.g., tax records
Administrative data can be used on its own, or to be linked with and augment survey data

One problem with linking is that, typically, non all survey respondents can be linked with administrative data. This is analogous to non-response error. The inverse of linking propensity can be used to adjust the linking problem.

Transaction data

E.g., credit card transaction history.

The data are accurate. But such data typically cannot be generalized to the general population. For example, credit card transaction history excludes purchases by cash.

Statistical usage of commercial transaction data:
Even though transaction data suffer from coverage problem and does not give accurate estimates, they can be used to track trends and growth rates. Such data are often timely and sensitive to reflect changes.

Data on individual interactions with internets

Useful for providing qualitative insights

Track relative frequency of key terms → can reflect trends and growth rates.

4.1 Cognitive processes and suboptimal responses (measurement error)

Encoding	Information might not be encoded in the first place. Fail to retrieve might be a result of lack of encoding rather than forgetting, (Lee, 1999): parents never encode children's vaccinations
Comprehension	Different levels of comprehension: o Lexical level: words have different meaning to different people. For example, who are "children"? (Belson, 1981) Defining each term results in long and complex questions. → trade-off between clarity and complexity o Semantic level: survey concept maps differently to respondents' concept Schober & Conrad, 1997: conversational interview clarifies meaning of complex questions and improve response accuracy o Pragmatic level: cooperative principles in interactions guide people to infer intentions. For a scale that ranges from not at all successful to extremely successful, whether labelling the scale as -5 to 5 or 0 to 10 makes a difference.
Retrieval	- More events encoded during life periods of great changes: the reminiscence bump. People encode more events for their 20s. Two mechanisms that lead to retrieval failure: o Interference: The longer the time period, the more likely that the events blend with other similar events into a single generic memory o Decay: forgetting. Forgetting most rapid immediately after the event experienced. Forgetting continues after as many as 50 years. Improving recall: Generally, o Match between encoding and recall context affects memory performance In survey context: o Cues o Giving respondents more time o Decomposing recall task. E.g., recall smoking cigarettes in different situations at different times of days o Event history calendars Retrieval mistakes: o Telescoping and backward telescoping.
Estimation and judgement	o Recall and count: often underestimation

- Rate based estimation: often overestimation
  Typical for regularly occurring events, fail to take exceptions into account
- Impression based estimation
  Translate impression into numbers

Heuristics—informal rules to provide close enough estimates

- Availability heuristic:
  memory easily available means frequent or highly likely
  Schwarz 1991: recall 6 or 12 assertive situations → more likely to rate self as assertive in the 6-situation group because it was easy to recall 6 assertive situations than 12

Context effect:
The order of survey questions matters. Due to cooperation principles, people have expectations about the intentions of the questions.

- Assimilation
  Marriage satisfaction, then general satisfaction. Incorporate the former into account when evaluation the latter
- Contrast
  Explicitly separating the two aspects.

Report	Social desirability Recency effect: in telephone or in-person interview, when the options are read aloud, more likely to choose the recent options. Primary effect: in computer/ paper-pencil survey where the response options are visually presented., more likely to choose the earlier options

Optimal vs. acceptable responding

The above steps are the ideal response steps. Measurement error occurs despite good faith effort by the respondents.
However, in reality, we may not have respondents carefully engage in these steps at all:

Satisficing behaviors are common. Satisficing = satisfactory + suffice, but not optimal

- Weak satisficing: execute all response process, but superficially
- Skip entire steps in the response process
- The higher task difficult, the lower the ability and the lower the motivation, the higher the satisficing
- Two examples of satisficing are:
  - Speeding (young people speed more)
  - Straight lining (most likely for speeders)

5.1 Questionnaire design

Goal of question design process

1. Validity	Question development	Questions measuring the concepts that the researcher intend to measure Types of validity: o Face validity o Criterion validity: Indeed correlate with other variables that are theoretically associated o Construct validity: similar as above, correlate with proxy measures o Consistent with administrative records
2. Reliability	Measurement	Classical test theory: what proportion of the observed variance is variation in the true score In practice: reliability means agreement between repeated observable measures (refer to TSE)
3. Standard		Is everyone answering the same question. This is a mean to get to reliability. Standardized interview: standardize the questions and also the interactions with respondents. Behavior coding can be used to monitor the standardization. Open vs closed questions. Responses to closed questions are more standardized. From very open to very closed, it is a continuum. (multiple-word open → constrained open → field coding → fixed response categories) o Pros of open questions: Elicit answers that would not have been produced if fixed options presented o Pros of closed questions: Specific options can remind people of possible answers they would not think of with open form
4. Easy to administer		From interviewer end: o Interviewers have trouble delivering the question? From respondent end: o Respondents request clarifications? Have difficulty responding? Dependencies among questions: o Skip patterns (more difficult);

		○ bounding (more difficult); ○ tailoring of questions (e.g. plug in names from earlier questions). Easier for respondents, harder for the interviewers

Questions about facts and behaviors

Response dimensions:
- Occurrence
- Timing or dating
- Frequency:
  numerical
  relative (frequently/sometimes/rarely/never)← bad scale
- Interval or regularity
  every week?
- Duration

Guidelines:
- Include all reasonable options
- Provide recall cues
- Make the question specific
- Provide a clear time period
- Use bounded recall to reduce telescoping
- Encourage respondents to use records
- Avoid proxy reports

- For threatening questions: *self-administration*
- For threatening questions: random response technique. One threatening question and one unthreatening question. estimate the overall pattern of response without ever knowing any respondent's answer
- For threatening questions: use open questions so that respondents do not compare themselves to others
- For threatening questions: use longer questions; deliberately load the question
- For threatening questions: use informants
- For threatening questions: still use familiar terms

Questions about subjective states

Scales
- Bipolar vs. Unipolar term
  "Do you favor or oppose" is *bipolar*. Two opposing alternative, clear midpoint
  "I feel like I am up against the world. Not at all, sometimes, often" is unipolar. Varying levels with no clear midpoint.
  Bipolar scales are better.

- Balanced vs. unbalanced
  Are you in favor of xxx? "No" is ambiguous ← unbalanced
  You "favor", "oppose", "no opinion" ← balanced, better

- Number of scale point: *7* is good
- Fully labeled scales are generally good
- Ranking vs. rating
  Ranking forces an order. It is a more difficult task.

Guidelines (examples)

- Attitude objects should be clearly specified
- Measure the strength of the attitudes. "Moderately favor" or "strongly favor"
- Avoid double-barreled items. Bipolar scale rather than unipolar statement + agree/disagree
- DK category can make a big difference

6.1 Cognitive interview

There are a lot of details about cognitive interviewing.

- Two basic methods are think aloud and direct probing.
- I only highlight one point that I did not realize before.
  Whether to revise the question also depends on magnitude of the problems: how many will be impacted by the problem? How large is the impact?

7.1 Business survey

Editing and imputation plays especially important roles in the production of business survey data sets. Often a lot of missing in business survey.

Cognitive process of responses	Cognitive model of business survey response	A response process model
○ Encoding ○ Comprehension ○ Retrieval ○ Judgement ○ Report	○ Encoding/recording (business records are constructed to serve the business goal. They may not have records on what the survey asks) ○ Comprehension ○ Source decision: from informant's memory or record? ○ Retrieval/ record look up ○ Judgement and estimation (Respondent may decide whether to report based on memory or records. In the case that the organization does not have exact information that the survey is asking for, the respondent needs to determine whether to put in extra effort.) ○ Communication	In the context of business survey, reporting is a separately story. These are the new elements added: ○ Respondent selection (respondent is chosen by the organization rather than by the survey. Informants can distribute across multiple levels.) ○ Assessment of priorities (motivation Investor are of highest priority. Mandatory government request may get some attention. Other surveys might be ignored) ○ Release of the information

Editing and imputation in business surveys (fixing measurement errors)
Significant effort in business survey to correct questionable items and fill in missing items

- Recontact key respondents
- Use administrative or other data sources
- Impute

(Can refer to TSE for more discussion on editing). The basic idea is using logic to flag suspicious data and try to correct it
Imputation for business surveys (refer to TSE and applied sampling)

- One point: historical ratio imputation. Leveraging historical data of the company: $\frac{X_{lastyear}}{Y_{lastyear}} = \frac{X_{thisyear}}{?}$

7.2 Found data (Social media)

Big data are probably good to obtain qualitative insights and monitor trends, but not to obtain accurate and quantitative estimates

Elaboration on 1.1 Comparing survey data and social media data

	Survey data	Social media data
Cost	High, especially with interviewers	Secondary data; relative cheap
# of observations	No larger than necessarily	Very large
# of variables	As needed	Small (usually only one)
Information on the individuals	Frame info Can ask for plenty of information from respondents	Rare. Profile, geotags can provide some info
Support population estimate?	Designed for this purpose	Usually no
Timeliness	Lag for weeks/months	Yes. Instant
Data quality	Quality control by the survey organization	Depends. Face validity
Burden on providers	Yes. Depends on # of questions	No burden

Challenges of social media data

Representation.

- coverage error(internet coverage + social platform coverage)+
- pre-selected respondents (users post on topics of their choices)+
- little user level info (case rich but variable poor)
- textual analysis not nuanced
- spurious correlations

Measurement.

- In traditional survey, extensive effort are put in standardizing stimuli that are provided to respondents in order to reduce measurement error
- In the context of social media, stimuli that generate the content are unknown and uncontrollable. The concept of "true value" loses its relevance. Errors function at a completely level.

Triggers (if no questions as stimuli, then what about imagined audience?)
For social media content, to one effort to understand what stimuli trigger the response is to query the imaged audience of the content. Whom and what purpose are these content generated for?

- Nobody
- Friends
- Impression management and self-censor for unknown audience

→ How to understand accuracy of the content? The motivations and actions are diverse

How to use social media content:

- extract content from site
- extract meaning from text

Success stories of social media data. Can be divided into three categories:

1. *Using social media data by themselves.*
   Extracting emotions from tweets and profile how people's emotion changes across the week (Golder and Macy, 2009)
2. *Social media as supplement to survey data.*
   Survey data builds a projection model for candidates' senate races. Adding the candidates' Wikipedia page view significantly improve the model
3. *Social media as alternative to survey data.*
   Can social media provide population as survey data?
   O'connor et al. 2010. Found that sentiment ratio (positive/negative words) of tweets containing "jobs" correlated highly with Michigan's index of consumer confidence. The smoothing constant is arbitrary defined though. The correlation did not stay robust.

8.1 Mode comparisons

	Mail	Telephone		Face-to-face	
		CATI	IVR, touchtone data entry	CAPI	CASI
Agent	Self-administered	Interviewer-administered	Self-administered	Interviewer-administered	Self-administered
Characteristic		- Primarily centralized - Direct supervision - Telephone survey center maintain equipment		- Decentralized - No direct supervision - Interviewers maintain equipment - No control over environmental conditions	Have variations: - ACASI (audio) - Video-CASI - Recruit-and-switch/outbound IVR
Frame consideration	- Addressed needed - Addresses based sampling (ABS) typically use commercial list frame from USPS delivery sequence files - When address frame not available, in-person listing will be conducted (expensive)	RDD generates telephone number, which is very inefficient Landline coverage decreases. Now changing to RDD for landline + purchased cell phone		Addressed needed (same as mail survey)	
Non-response consideration [modes vary in types and sources of nonresponse]	Non-response confounded with coverage problem. No response could be	Non-response confounded with coverage problem. No response could	Computerization increases cognitive burden.	Highest response rate	Self- administration increases cognitive burden. Not all respondents are able and willing

	Mail	Telephone		Face-to-face	
		CATI	IVR, touchtone data entry	CAPI	CASI
	because address illegible. Research on how different practices increase mail return rate. But not clear what combination works the best.	be because number illegible.	Break-off during outbound IVR is common.		to do CASI: education and age are associated with reluctance
Data quality consideration ~ interviewer vs. self	- Pros: self-administration generates more socially undesirable responses. Good for sensitive questions.	- Pros: Interviewer increases response rate and produce higher quality data. They can for example motivate, probe and clarify. - Cons: But they can elicit socially desirable response. And they are expensive.		- Pros: Interviewer increases response rate and produce higher quality data. They can for example motivate, probe and clarify. - Cons: But they can elicit socially desirable responses. And they are expensive.	- Pros: self-administration generates more socially undesirable responses. Good for sensitive questions. - computerization has extra benefit on top of self-administration.
Mode comparison [To establish legitimate comparisons, need to make sure the	In 1991, similar response rate as telephone (>70%)… Has the benefits of self-administration	- Pros: Faster than face-to-face - Cons: lower response rate	Has the benefits of self-administration	- Cons: Face-to-face survey takes much longer than telephone	Self-administration uniformly results less desirable and more honest responses.

	Mail	Telephone		Face-to-face	
		CATI	IVR, touchtone data entry	CAPI	CASI
two modes are implemented on comparable samples: - Panel study alternating modes across waves - Use frame with information for both modes]		than personal visit - Few substantive differences from face-to-face		- Pros: higher response rate than telephone; Preferred; - Few substantive differences from telephone	(ACASI> CASI< CAPI)

9.1 TSE perspective on web survey +
3.1 Representation of web panel survey

Web survey is on the rise, in response to decreasing survey response rate+ increasing costs of the traditional method+ growing internet coverage

Representation

There is no good general purpose frame to sample from. **Web panel recruited in various ways:**

- ***Probability***
  For specific purposes:
  1. Intercept based method
     Pop-up invitation every nth user. Systematic sampling.
     Useful for specific task such as evaluation of websites
  2. List-based samples
     Invitation with link to survey via email/letter.
     - Address based: send letters with a survey link
     - Within a bounded circle: E.g., sampling from a list of university emails, sending emails with survey link

  For general-population web surveys:
  - Start with RDD or ABS
  - Provide devices and internet access if the sampled elements do not have internet access: Netherlands LISS panel
- ***Non-probability***
  For specific purposes:
  1. Entertainment pool
     - "question-of-the-day" type of survey
  2. Unrestricted self-selected survey
     - Open invitation through banner or link
     - Problem:
       - vague definition of population
       - no control over multiple reply
       - no probability sampling

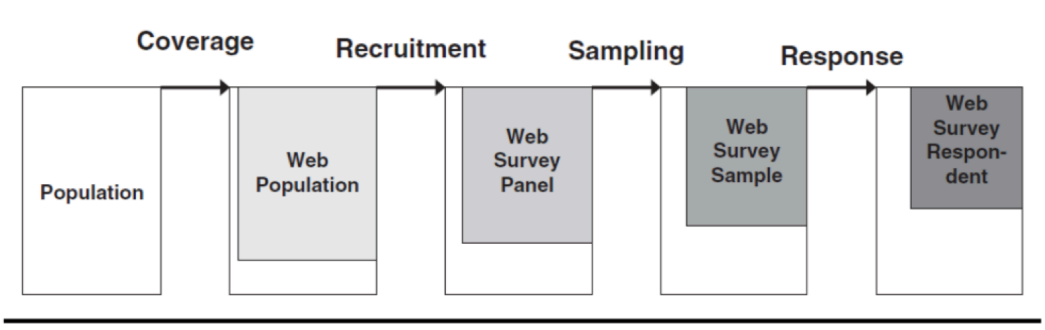
  For general-population web surveys:
  3. Web panel:
     - Volunteer panel
       Express interests and supply email address
       Confirm interest and complete the survey
     - Web panel used as a frame for specific surveys
       Specific survey looking for respondents with characteristics xxx
     - Web panelists are likely to be different from the general public.
     - Web panel has considerable information about its members
       Reweight the non-probability panel to match certain characteristics of the target population

Non-response error in web survey
Non-response error is a function of non-response rate and the difference between respondents and non-respondents.
Lacking a general frame, little is known about non-respondents. Hence, we typically focus on response rate.

- Web survey has lower response rate than other modes

- Response rates of opt-in web panel are conditional response rates. The initial response rate for joining opt-in panels are usually unknown and unknowable.
  The non-representativeness accumulate and typically how the representation dwindles in the earlier stages is hard to known.

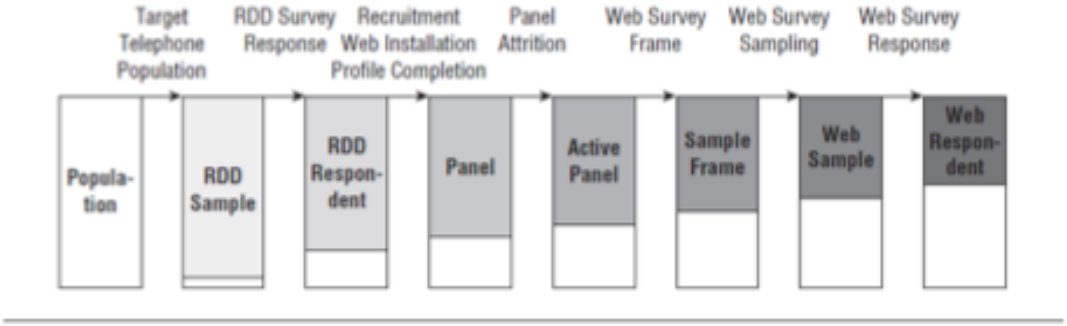
Volunteer Panel Web Survey Protocol



Source: Lee (2004).

- Response rates are more calculatable for probability based web panel that starts with RDD or ABS.

Protocol of Prerecruited Probability Web Panel Surveys



Note: Figure not drawn to scale.

In response to the non-probability nature of the web surveys,

- can reweight non-probability web panel by combining it with probability samples and building model to estimate pseudo-sampling probability. (refer to the Elliott notes Module 12)
- Population-based adjustment. Match the survey distribution to population distribution
- If MAR, can make population inference. But this assumption typically does not hold

Measurement

Web surveys are self-administered

Pros:

- general advantage of self-administration. Less social desirability.

Cons:

- No interviewer presented to motivate, probe, clarify tec.
- Less guarantee that the respondents will complete the questionnaires. Breakoff is a problem.
- Has no control over the environment

Web surveys are computerized

Pros:

- Program in complex routing, skips, edits etc.
- Easier for experiments: can randomize order, content etc.
- Scrolling web survey
- Paging web survey (comparing to scrolling web survey)—one or a few item per screen; data transmitted after each screened

Cons:

- Automatic instrument generally cannot accommodate unexpected answers

Web survey can be interactive

Pros:

- Can provide feedback (e.g., on-site edit check, instructions mimicking interviewer's behaviors)

Cons:

- Implementing these features might require plug-in, reducing accessibilities for some sample members
- Reinforced social presence, may risk reducing the benefit of self-administration

Web survey is a rich visual medium

Pros:

- Easy to add graphical enhancement

Cons:

- Use of images can convey unintended meaning. Images can provide a context for the respondents to contrast, prime the respondents etc.
- May change measurement properties of questions

10.1 Mobile web

Web survey on mobile phone.

Mobile web respondents started as "unintentional mobile respondents" are were excluded from analysis because the questionnaires were designed for PC web survey and were not optimized for mobile phone

Mobile web different from conventional web:

- Technology:
  Different display
  Different operating systems

- User characteristics:
  demands on motor skills
  motivation & interest

- Context of use
  mobile use subject to more distractions, presence of others, multi-tasking, interstitial activities

Does mobile web count as a separate mode or a variant of PC web survey? How does its error properties differ?

Coverage

- Smartphone ownership is growing fast. Higher coverage than desktop/laptop. Growing population that has smart phone only
  In the US, it's the lower-educated population that is becoming phone only; whereas in Europe, it's the higher-education population that is becoming phone only.
- Non coverage bias does contribute to total survey error (Antoun, 2018)

Sampling

- Same as the conventional web surveys: list-based
- RDD, specific to mobile phones, but low hit rate. Then CATI interviewers invite respondents to participate in mobile web survey

Nonresponse

- Comparing to PC, lower start and completion rates and higher breakoff rates in mobile web, even when the surveys optimized for mobile devices
  Breakoff because of cognitive burden and less satisfying survey experiences
- Burden: several studies show that responding on mobile devices takes longer than on PC and thus require more effort from the respondents

Measurement error

- Small screen
- Multitasking/distractions
- Reduced privacy because of others' presence
- Little supporting evidence. Antoun et al., 2017. Almost no difference in mobile and PC

10.2 Text message survey

Texting can be used to invite sample elements to participate in surveys or as the vehicle of the surveys directly

Comparing to face-to-face and telephone survey, text message surveys are:

- Less synchrony—no obligation to reply immediately
- Medium—visual rather than auditory
- Persistence of entire conversation—the texting history is preserved
- Privacy—others unlikely to see conversation on screen
- Connection to the network—can be intermittent. (no breakoff due to internet connection problem)

Coverage

- No differences on most demographic variables
- Not all smartphone users are texters (texters tend to be younger and higher-educated)
- Depending on literacy level, manual and visual ability etc.

Nonresponse

- Higher response rate than voice invitations
- But higher breakoff rate too

Measurement

- There is empirical evidence that text are associated with less satisficing and more disclosure (Schober, 2015) than voice interviews

Texting can also be differentiated between self-administered (automated text) and interviewer-administered (human text)

11.1 Mixed mode

Mixed mode for pre-notification
E.g., SMS prenotification + email invitation to web survey seems to be most effective

Mixed mode for data collection

	Concurrent mixed-mode design	Sequential mixed model	Switch modes within questionnaire	Longitudinal mixed mode	Parallel mixed mode
	Different modes for different subgroups of sample e.g., web for those with internet, mail for those without internet	Sample members recruited in increasingly effective and expensive modes if previous contact attempts are not successful Mail→ telephone → face-to-face	Different modes for different parts of the questionnaire. e.g., face-to-face + ACASI	The mode changes or alternates in different waves. Change: e.g., face-to-face to recruit members into the panel, then change to telephone in subsequent waves Alternate: e.g., in each wave, half of the sample face-to-face, the other half telephone	Different sample, different modes.
Pros	Good for reducing coverage and non-response	Reduce non-response; Reduce costs for non-response follow up	Switch to self-administration for the sensitive questions.	Save costs Tease apart the mode effect	Mainly for comparative studies to accommodate different regional survey traditions or practical constraints.
Cons	Mode effects confound with subgroups	Cons of non-response follow up in general: Follow up can increase response rate, but it does little to resolve demographic			

		bias due to nonresponse.			

Concurrent mixed model ~ offering mode choices

What's the effect of having mode choices on response rate?

Seems to reduce response rate but increase data quality for those who participate

- Reduced response rate:
  Seem to because of switching costs. Offering mode choices means breaking in the response process. This increases the probability of dropping out
- Increased data quality:
  less satisficing, less breakoff, greater satisfaction with interview

Switching mode within questionnaire

Increase reported sensitive information

Audio (ACASI) has an extra benefit on top of CASI? Mixed finding. The audio function is not frequently used.

What's inhibiting honest report in face-to-face?

It's not just the presence of a human interviewer, presence of human-like features such as a face can also inhibit disclosure.

Computerization has extra benefit on top of self-administration

IVR

Has the benefit of self-administration, but breakoff is high (breakoff is generally a concern when no interviewer is present)

Speech IVR

Speech recognition system. Accurate but could be frustrating. response rate is lower, satisfaction is lower

An interesting recruitment method—IVR and Redirected Inbound Call Sampling (RICS)

Millions of misdialed numbers every month

These calls can be redirected for other purpose including survey research "This number does not exist. Do you want to participate in a survey?"

The rationale is that since people initiate the call, the survey does not interrupt their daily routine.

11.1 Interviewer's roles

Interviewers can affect multiple sources of error

- coverage error: interviewers list, roster, and determine eligibility.
- non-response error: interviewer contact and recruit.
- measurement error: interviewers implement questionnaires, probe, clarify
- largest cost. Using interviewers means that cost in other aspects need to be reduced.

Effect of interviewer speech

- E.g., a little disfluent is the most efficient in recruitment. Perfectly fluent interviewers are the least efficient. (Conrad et al., 2013)
- Possible to train interviewers to better recruit?
  Yes. Interviewer's responsiveness matters

Interviewers select respondents often from households

- Kish method with tables
  Most precise selection method, each household member has equal probability of selection
  can be automated
  But, intrusive and time consuming
- Birthday method
  easy to implement, less intrusive
  But, exact chance of selection is unknown. Difficult to verify.

- Proxy respondents
  Pros:
  Less costly to use proxy reports
  Higher response rate
  Cons:
  Generally, self-report is more accurate than proxy-report. Proxy relies more on typicality and traits rather than actual events.
  Though:
  Exactly how data quality is affected by proxy responding depend on item content

11.2 Techniques

Standardization debates

- Standardized interview proponents: All respondents should get the exact stimuli. Interviewers should not result in variability. They should be neutral "tool" for delivering the questionnaires and they read the questions verbatim.
- Criticism of standardized wording:
  Standardized wording prevents conversational grounding. It leaves room for respondents to have various interpretation and deviate from the researchers' intention.
- Better to standardize meaning. Interviewers clarify the questions until the respondents understand the intention and give valid responses → conversational interview

- Interviewer clarification to standardize meaning is important when the questions are complicated.

- But conversational interviews take longer than standardized interviews.

Formal vs. Personal style? Rapport?

More business-liked or interpersonally-engaged?
Is personal style motivating or ingratiating?
Similarly, what about rapport?

- Personal style and rapport encourage more conscientious and more honest responses.

How is rapport established? Typically through nonverbal behaviors: nodding, smiling etc.

- Using a set of technique to motivate respondents increase data quality

To sum, interviewing technique can have a major effect on response quality. Interviewers must use discretion and judgement. More improvisation is needed in the recruitment stage, but less in the data collection stage. All interview techniques have pros and cons.

11.3 Interviewer effect

- Intra-interviewer correlation: rho= between-interviewer variance/total variance
  → design effect= 1+ (m-1)rho→ shrinks effective sample size
- Interviewer effect inflate variance about as much as geographical clustering
- Attitudinal questions and open-ended questions have more room for interviewer effect
- Between-interviewer variance becomes large when probes are involved

- Does conversational interview increase interview variance?
  For the most part, conversational interviews do not increase interviewer variance
  Conversational interviews definitely increase accuracy and thus reduce bias
  Thus, all in all, conversational interviews are good for reducing MSE

Interviewer characteristics: e.g., race and gender

- Affect questions that are related to those characteristics
- Stronger effect in face-to-face surveys. But the perceived characteristics on phone make a difference too

II.1.1 Paradata

Paradata of a survey are data about the process by which the survey data were collected

Different models allow different kinds of paradata to be collected

1. Mail survey
   Very limited paradata
2. CATI or CAPI
   interviewer characteristics
   history of contact attempts
   audio recording
   with computerized surveys, can record response behaviors
   interviewer observations (including observations about non-respondents)
3. Web survey
   respondent's devices
   response behaviors: navigation through the questionnaire

Uses for paradata:

1. Improving sample frames
   e.g., is the lister walking or driving → quality of the sample frame
   - Think about the study that I am working on. Using paradata to evaluate whether extra frames are useful
2. Improving survey response
   e.g., history of contact attempts → guide effective contact
   - Also think about Felicitas Mittereder's dissertation. Intervention to prevent breakoff using respondents' behaviors
3. Improving nonresponse adjustment
   e.g., interviewer observation on the frame can be used to make non-response adjustment → these are the information on the sample frame → building response propensity model
   - recall the example of interviewer judging whether the sampled individuals are sexually active.
   - other examples: interviewers can observe neighborhood income quartile
   Paradata need to correlate both response propensity and the variables of interest to be helpful for non-response adjustment. Correlating with the variables of interest is often hard the establish.
4. Reducing measurement error
   e.g., navigation behaviors can help to identify problematic items
5. Reducing processing error
   e.g., Frequency of coders consulting the codebook may help to identify potential coding problems

Caveats and cautions of using paradata:

- Interviewer-provided information may be incomplete or erroneous
- Computerized paradata seems objective, but respondents' multitasking and interruption may distort response latencies
- Privacy concern

Different types of paradata and what they mean

Response time
Short response time associates with acquiescent response style, lack of motivation and inaccurate answers.

Long response time indicate uncertainty, ambivalence and inaccurate answers.

Respondent's speech
Paralinguistic utterance (um and uh) can indicate how well the questions are understood and the amount of difficulty in answering
E.g., response on the health question is uncertain when the respondent's health history is inconsistent and disfluent
Respondents may not directly ask for help when they don't understand the question, but may exhibit cues of confusion or uncertainty
In fact, for complicated situations, the more evidence of difficult and uncertainty displayed, the *greater* the response accuracy

Visual behaviors
Nonverbal paradata can be collected in face-to-face interviews, such as facial expression, head movement, and gaze aversion.
Respondents more likely to change initial answer if averted gaze while answering. Presumably gaze aversion reflects comprehension difficulty

Respondent's inactivity
Online surveys do not offer the clarification that interviewers can offer.
Plausible indication of confusion/difficulty is inactivity; automatically provide clarification when respondent is inactive improves response accuracy?
Providing clarification when the respondents seem to need it increases response accuracy

Mouse movement
Average number of mouse movement increases as the level of question difficulty increases

Multitasking
Multitasking is widespread for web survey.
Multitasking can be indicated by behaviors such as focus-out events (leave browse and return) or time-out event (inactive for a longer than a page-specific threshold)
Multitasking does significantly predict non-response, but the effect is not too large and the data quality does not seem to be significantly compromised

Possible to prompt to increase satisficing?
Prompting does slow the respondents down, but it does not always increase accuracy

II.2.1 Record linkage application

1. Sample frame development
   - Records linkage necessarily to determine which units are in both frame, which can be subsequently removed→ useful for multi-frame surveys

2. Estimation of under-coverage in a census count
   - Record linkage identifies which units are in both lists.
     Example application:
     Capture-recapture estimates of Syrian civil war death.

		Sample B	
		Present	Absent
Sample A	Present	X_{11}	X_{12}
	Absent	X_{21}	X_{22}

     Assuming the two lists are independent (capture-recapture), then x_{22} can be calculated → estimate total death
   - Similarly, enumeration + re-enumeration allow estimating the potential under-coverage of the frame.

3. Augment information contained in a survey data file
   - Example application:
     Improving survey data on income and self-employment status. There are significant discrepancies in income and self-employment captured by survey and administrative data.

II.2.2 Record linkage methods

1. Deterministic linkage
   - Link occurs only if identifiers match exactly; any disagreement results in mismatch.
   - Potential match keys are SSN or a combination of first name, last name, age, address
   - Deterministic linkage works well only when the files are accurate and comparable

2. Probabilistic linkage
   Most common approach uses multiple soft identifiers that may contain errors, calculates the relative likelihood between match and non-match. Link occurs if the likelihood pass a certain threshold.

   Example: Consider two data sets A and B, each has 8 records (name, age, gender…) → $n_A * n_B$ pairs, each pair is either a match M or a mismatch U.
   Consider one pair $\gamma = \{1, 1, 1, 1, 0, 1, 0, 1\}$ that matches on 6 records but mismatch on 2 records → Calculate $\frac{P(\gamma|M)}{P(\gamma|U)}$ → if passes a threshold → consider this pair a match

   But this method is computational extensive. Given 8 records, need to compute 2^8 probabilities to compute the probability of one specific pattern $P(\gamma|M)$.
   The computation can be greatly simplified if we are willing to accept independence between the records: $P\{1, 1, 1, 1, 0, 1, 0, 1|M\} = P(1|M)P(1|M)P(1|M)P(1|M)P(0|M)$ …

   Why do we need to consider this probabilistic linkage?
   Because the records are not perfectly accurate.
   Suppose that the population is half male and half female, but the sex is coded incorrectly on 10% of

the records.
Select a pair from the two list:
???

- Pre-processing and standardizing the records are essential for matching
- Rather than directly recording the records as match or mismatch $\{1, 1, 1, 1, 0, 1, 0, 1\}$, can record the extend of which they mismatch (e.g., distance between one string and another)
- Not feasible to compare all possible pairs to determine matching, solution is the block the records—divide them into groups based on one or more variables and only make comparisons within groups. E.g., only try to link people within the same area. The implicit assumption is that the blocking variables do not contain errors. If the addresses are entered incorrectly or if households moved, then we would miss the true matches.

- Accuracy is not an informative index—Sometimes, matching is the exception. Keep classifying the pairs as non-match can result in a decent accuracy rate. Recall and precision are better indicators.

3. Statistical matching
   Using information from two non-overlapping data sets.
   Refer to Elliott notes for more information. The main assumption is conditional independence. The method is basically imputation. Imputing information to one dataset based on the assumed model on another dataset.

II.4.1 Consent and Privacy Concern

Bias in record linkage
Bias depends on the percentage of linkage and also on the difference between the linked and the unlinked.

- Non-consent bias is generally small comparing to other error (non-response and measurement error)
- Even if non-consent does not cause bias, fewer cases increase variance

Agreement to record linkage
Demographic variables often correlated with probability of consent

- Pattern by age and sex is mixed; education generally positively associate with consent

Research on optimizing consent rates

- Placement of consent request
  Better either at the beginning of the survey or in context

- Framing of consent request
  Framing the question as avoiding loss (would be much less valuable if you do not allow us to link) is more useful than framing the question as producing gain (a lot more useful if you allow us to link)

- Opt in versus opt out
  matters a lot

Plausible practices to maintain data security

- Employee only access

- Access via secure system
- Password+ multi-factor authentication to access
- Only making linkage on the parts of the data that are related to the specific interest and only for the time being; do not store large linked data files → expensive
- Multi-party computing (MPC): computation using data from multiple sources without the source data ever having to be shared → method not well developed for all calculations, costly
- Micro-data are identifiable with soft identifiers such as gender, date of birth and zip codes. Tabular data is one way to avoid the danger of releasing micro data.
  However, even though individual table is save, allowing multiple queries against a dataset may create disclosure risk. If sufficient tables are published, may be possible to recreate the underlying microdata.
  (Refer to TSL for techniques on protecting privacy)
- Differential privacy a mathematical definition of privacy
  Adding noise to the data so that any person's information cannot change the overall statistics "too much". Amount of noise needed depends on the number of people in the dataset and the value distribution. Also having a privacy budget and only release a certain amount of information.
  Generally differential privacy works well on large dataset and will result in inaccurate results for small datasets.
  Pros: can tell data user the form and magnitude of the noise, making it possible for them to draw statistically valid conclusion from the published output
  Challenge: trade-off between accuracy and privacy. What's the right amount of privacy budget?

II.5.1 Situated and passive data collection

Situated data collection

Survey self-report

- Lag between occurrence of the event and survey report → recall error
- Questions are often out of context, but respondents' internal state and social context at the time of measurement can affect reports
- In panel studies, resolution is typically low, repeated measurement implemented monthly or even annually

→ ***Situated self-report***: repeatedly ask about "now" separated by brief intervals (usually multiple times a day). This is a direct descendant of diary studies with prompt reporting.
Two traditions, basically same method:

- *Experience sampling method*
  in psychology
  widely used to study time usage and mood
  e.g., mood is more situational than dispositional. Increased negative affect immediately after the company of others

- *Ecological momentary assessment*
  Event-based: subjects determine when event has occurred and initiate assessment (e.g., pain)
  Time-based: resolution depends on frequency and duration of the target behaviors
  e.g., real-time assessment of smoking cessation.

Caveats

- Not clear whether real-time assessment is necessarily more accurate and predictive
  Even though forgetting occur in retrospective recall, situated report can omit information because the events are underway and experiences of the events are still developing
  In some case, recall is better predictor of subsequent behaviors than situated-report. E.g., pain recall rather than the real-time pain experiences is a better predictor of return for procedures.
- Reactivity
  Changes in behaviors and experiences as a result of its being measured?
  No evidence, so far
- Compliance
  Success of situated report depends on timely responses. Missing data, if systematic, can bias findings.

Passive data collection

- Data collection
  Collected via sensors, increasingly on smart phone
  Measurement occurs without any action by participant
  Self-reports are used to verify and interpret the sensory data

- Error of representation
  Smartphone coverage is a problem (coverage rate is high, but those who are not covered are likely to be very different)
  Participation rate is low. Often rely on convenience samples.

- Measurement error
  Sensors can be inaccurate. Different devices can produce different measures.
  Target behaviors are inferred from sensory data. E.g., darkness → sleep; movement → walk
  Compliance: accidentally or deliberately turn off the devices → bias (though compliance seems high so far)

- Cost
  Does not impose extra burden on participants
  But app development, IT infrastructure for big dataset, data preparation etc can lead to far more expensive projects

- Frequency of passive measurement
  Situational: questions delivered when certain conditions are met
  High frequency discrete: More often than self-report but data collection not always on
  Continuous: fine-grained behaviors such as moving

- Implication for privacy
  High frequency measurement, especially when participants do not control, are likely to be seen as intrusive.
  Practices to reduce the concern:
  Explain why passively collecting data is necessarily
  Let participant to turn the tracker off
  Limit sensor's abilities
  - Generally, incentive is the most important factor motivating people to participate in passive data collection.
  - Higher willingness among women, smartphone active users, individuals who trust tech companies and institutions.

- Empirical example.
  At this point, empirical studies demonstrate the feasibility of using sensory data, rather than actually using the sensory data
  E.g., Using blue tooth, smart phone can detect the presence of other phones → + surveys to verify activities → train models to predict friends based on proximity/ infer social interactions
  E.g., Investigate relationships between workload, stress, sleep, activity, mood, sociability, academic performance over 10 week academic term

II.7.1 Coding open responses

Open responses:

- Numerical responses
- Verbal/linguistic responses. Can be as text or as speech
- Graphics or photos occasionally

Why open response:

- Complex coding categories, impossible to present as closed options
- Collecting insights to develop closed questions
- Give respondents opportunities to explain. More likely to provide truthful responses when respondents know that they will not be misunderstood

Human coding and measurement error

- Coding is largely conducted by human codes.
  Coding according to a coding system. Individual coders cannot extend the coding the system but the entire system can be updated.

- It is hard to discuss validity and true value in coding. Validity is usually measured as agreement with experts. (i.e., expert's judgement is the true value)

- Inter-coder reliability is typically used to evaluate coding quality.
  Coders may jointly create informal rules to code difficult responses that are not covered by the instructions. But if these rules are not documented and followed through and the next batch of coders develop other rules, then the results would be both biased and varying.
  This kind of practice would increase reliability but not necessarily validity.
  Also, coders have the incentives to agree with the coding whenever possible.
  → It is better to have coders code independently and then compared, than having one confirming the work of another.

- Correlated coder error: 1. Intra-coder correlation; 2. If informal coding rules are developed, a group of coders can have correlated coder error as well.
  One coder often code a large number of responses, even if the intra-coder correlation is small, the large workload can result in a large inflation in sampling variance.

- Empirical findings:
  For easy responses, shorter → more agreement
  For hard responses, longer → more agreement

Semi-automated coding

- Hybrid approach: easy answers classified automatically and hard answers classified manually
- Set accuracy threshold to determine how many cases are classified automatically

Automatic coding

- Rule-based:
  Rules based on combination of words or other features. Can be accurate. But building and maintaining these rules is expensive

- Machine learning:
  E.g., Naïve Bayes—pros: fast, robust to irrelevant features, good in domains with many equally important features (dominate decision trees for this sort of problems)
  cons: suboptimal if assumption of independence does not hold

II.9.1 Longitudinal surveys

Different panel designs

1. Repeated cross sectional survey

	Time 1	Time 2	Time 3	Time 4
Sample 1	X			
Sample 2		X		
Sample 3			X	

   - *Definition*: independent samples at multiple points in time. Possible to examine time series of estimates
   - *Changes* in estimates result from population changes + differences in the independent samples
2. Panel survey, no rotation

	Time 1	Time 2	Time 3	Time 4
Sample 1	X	X	X	X

   - *Definition1~cohort study*: Follow a particular group/cohort of persons over time
     Definition2~repeated panels: Surveys that consist of a sequence of short panels
   - *Changes* in estimates result from population changes + difference in samples due to attrition
3. Rotating panel survey

	Time 1	Time 2	Time 3	Time 4
Sample 1	X	X		
Sample 2		X	X	
Sample 3			X	X

   - *Definition*: Equally sized sets of sample units are brought in and out of the sample in some specified pattern.
   - *Example*: CPS—Sample units in for 4 months, out for 8 months, in for another four months. (→ a four-month overlap in calendar month)
   - *Changes* in estimates result from population changes + sample changes due to the rotation.
   - *Advantages* of the rotating design is that it is possible to tease apart these two elements of changes. In case of unexpected events, at any time point, there is always a before-after sample.
4. Split panel survey

	Time 1	Time 2	Time 3	Time 4
Sample 1a	X			
Sample 1b	X	X		
Sample 2a		X		
Sample 2b		X	X	
Sample 3a			X	
Sample 3b			X	X

   - *Definition*:
   - *Changes* in estimates result from population changes+ sample changes.
     Similar to rotating panel survey in a way. If only the orange samples are included, then it is exactly rotating panel survey
   - *Advantages* of the split panel is its flexibility of the cross-sectional change sizes, yet it still can continually update information.
     Can be used to identify panel conditioning.

Extra design considerations

- Lifetime
  longer lifetime → richer analyses
- Number of waves
  More waves, shorter reference periods reduce recall error
  But increases respondent burden and the risk of panel conditioning
- Mode
  Different modes can be used in different waves
  But would increase measurement inconsistencies across waves
  Good to alternate modes
- Dependent interviewing
  Feeing back answers from previous waves
  E.g., Wave 1: "How long have you been unemployed?" "5 months"
  Wave 2: "Are you employed?" "No"
  → 5+3= 8 month unemployed
  May makes answers more consistent, but acquiescence bias is a risk
- Incentives
  Recruit initially but also encourage continued participation
- Respondent rules
  Allowing proxy?
- Sample design
  Cluster design's advantages decrease over time. People move and no longer cluster together.
  Increased traveling cost
- Updating the sample
  Bring in new members to maintain the representation?
- Tracking and tracing
  Need to keep up with participants' whereabout. Effort to retain respondents even in the gap years.

Representation issues

- Coverage
  Even with perfect sample retention, initial panel becomes unrepresentative over time because the *population changes*
  Need panel refreshment to augment the coverage.
  ` Recent immigrants had no chance to be included in the sample → top-up sample of immigrants
  ` New birth (whether it's people or business)

- Sampling
  Panel refreshment complicates the calculation of sampling weights
  ` Can either top up a sample including the people who had no change of being included in the previous waves (e.g., new immigrants). This require intensive screening
  ` Or can top up a sample of general public. This is complicated from a weighting perspective. Some respondents have two chances of being included, others only one.

- Nonresponse and attrition
  Wave nonresponse and panel attrition lead to the responses becoming less representative over time
  Even modest rate of attrition can accumulate over waves to become significant attrition

Tracking and tracing:
` Failure to locate respondents after a move
` Failure to obtain cooperation
Respondents have different move propensity both due to personal level characteristics and societal level characteristics. Can pay special attention to the respondents who have a high tendency to move.
Some strategies:
Obtain contact info of families and friends
Contact between waves
Maintain full records of panel members

Panel studies are often reluctant to experiment techniques that may reduce panel attrition because they don't want to risk losing members.

Evaluating representation of panel continuers:

- Use information from initial waves to access whether attrition is systematic. If so, determine how to reweight to account for differential attrition

  → Interestingly, differential attritions do not necessarily have a large impact on the estimates and reweighting may have very little impact.

- Compare estimates from later waves to estimate from high-response rate cross-sectional surveys

Measurement issues

- Change of mode over waves
  Expensive mode (face-to-face) to recruit respondents, then less expensive modes in the subsequent waves
  Concern:
  mode effects confound with population changes
  difference in mode affects retention and attrition
- Panel conditioning
  Participation in previous surveys affect people's actual behaviors and attitudes
  ` Affect willingness to report socially undesirable behaviors (though it could also be the other way around)
  ` Learn to manipulate survey instruments to minimize survey burden: interleafed design (Yes immediately lead to follow up) vs. grouped design (first all yes-no questions, then followed up questions grouped together)
  ` May learn to provide more accurate and complete response (assemble records in preparation for the survey)
- Recall
  Forgetting—harder for respondent to recall more distant events. Example:

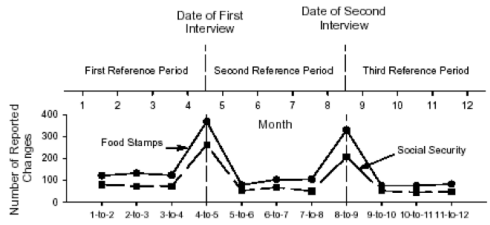
	Survey Year			
Response Year	1984	1986	1988	1990
1979	1.228			
1980	1.703			
1981	2.331	1.593		
1982	3.188	2.144		
1983	3.390	2.003	1.535	
1984		2.196	1.680	
1985		3.383	2.186	1.453
1986			2.414	1.787
1987			3.065	2.001
1988				1.913
1989				2.925

  Telescoping—when an event is temporally misreported as occurring in a time period different from when it actually occurred. Forward telescoping is more common than backward telescoping.

` Can use bounded interviews to correct. Example:

	Jobs (unbounded)		Expenditures	
Job Size	Unadjusted	Adjusted (bounded)	Unadjusted	Adjusted
Under $10	29.3*	9.0	34.9*	9.7
$10-19	33.0*	4.7	33.5*	-1.4
$20-49	52.3*	19.6*	56.5*	17.2
$50-99	89.6*	53.1*	86.6*	51.0*
$100-499	61.1*	61.1*	76.2*	76.2*
$500 and over	31.4	31.4	35.6	35.6
Total	39.7*	15.4*	55.2*	39.1*

- Seam effect
  Spurious transitions: different rates of period-to-period change when periods span two interviews than when the periods are within the reference period for a single interview



  Possible cause:
  ` satisficing behaviors: constant wave responding. Using the heuristic of anchor and adjustment → under adjust → seam effect
  counter strategy 1: blocking the questions by week rather than by topic reduce the amount of constant wave responding. [Blocking the question by topic: how much you spent last week/two weeks ago/three weeks ago? → more likely to anchor and adjust]

  ` Memory issues: less likely to remember events that occurred early in the reference period
  counter strategy 2: calendar-aided interviewing provides better cues and anchors

- Dependent interviewing
  Makes use of information provided in previous waves in administration of the current wave
  - Proactive dependent interview
    Remind the previous responses, then ask the questions/seek confirmation/or ask for changes
  - Reactive dependent interview
    Use information from the previous waves only when having item non-response or inconsistency

  Advantage is bounding interviews, but could risk acquiescence
  dependent interview significantly reduced seam effect

Longitudinal administrative data

- Administrative records allow researchers to look at changes over a long period of time. The sample size is very large and the information is generally accurate. Few panel surveys can cover very long period of time.
- But administrative records contain limited information and may miss segments of the population.
- Empirical example:
  - using tax data to study economic mobility
- Business administrative data files. Data file constructed by linking business units over time have yielded important insights regarding economic activity.
  But linkage needs to made carefully. Establishment are bought and sold and thus moving from one firm to another. Failing in making linkages accurately results in underestimating stability.
  One way is to identify workers. If two establishment have very similar workers, it is likely that it is the same establishment.

II.11.1 Evaluation methods—Design

Causal inference
Potential pitfalls in causal inference

- Internal validity
  Confounders? Can the observed difference be attributed to the treatment effect?
- External validity
  Can experiments performed on specific population under controlled settings be generalized to the general population
- Construct validity
  Experimental variables capture the constructs of interest?
- Statistical validity
  Are the observed effects statistically significant?

Observational study
Central challenge of observational research is "holding all else constant"?
Are sufficient control variables included? Can we rule out alternative explanation because of unobserved variables?
E.g., if we do not observe a relationship between incentives and response rate, can we observe that the incentives are useless?

Experimental studies

- Post-test only design
  Randomly assignment the sample to treatment/control group → compare the two groups' outcome

R	X1 (Treatment)	O_1
R	X2 (Control)	O_2

- Pre-test post-test design
  On top of post-test only design, also collect pre-treatment outcomes.
  Pros: Better for accounting for the randomness in treatment assignment.
  Cons: Panel conditioning

R	O_1	X1 (Treatment)	O_2
R	O_3	X2 (Control)	O_4

- Factorial design
  E.g., simple two way factorial design. Receive one version of treatment A and one version of treatment B.
  Pros: less experiment for testing main effects of treatment + can also capture interactions

R	XA1B1	O_1
R	XA1B2	O_2
R	XA2B1	O_3
R	XA2B2	O_4

- Within subject design
  Administering all treatments in sequence to the same set of subjects. Fully control for differences in individual characteristics
  Cons: conditioning effect is a potential concern

- Crossover design

R	X1 (Treatment A)	O_1	X1 (Treatment B)	O_2
R	X2 (Treatment B)	O_3	X2 (Treatment A)	O_4

Address the problem of the within subject design

II.11.2 Evaluation methods—Measurement

Direct measure of accuracy/bias

- Individual level, comparing survey responses with administrative records
  Example: to measure non-response bias and measurement bias, the actual values of the respondents and non-respondents are needed. Administrative record is one way to go. Recall West and Olson's study on the Wisconsin divorce study
  Challenges:
  Unlikely records will be available
  Timeliness of administrative data
  Consent bias
  If the records are self-reported, then get back to the same problem
- Individual level, comparing survey responses with sensor data
  No forgetting or lapse in attention. No over- or under-estimation
  Challenge:
  Sensory data not always perfectly accurate
  Target behaviors are inferred from sensory data

- Scenarios
  Asking respondents to evaluate based on fictional situations for which the true value is known.
  Can be used to evaluate the accuracy of responses.
  Example: comparing the accuracy of standardized interviews and conversational interviews.
  Ecological validity is a problem:
  Response process differs, so results may not be generalizable.

Indirect measures

- Response changes/Reliability
  Higher consistency is an indicator of higher quality.
  But response changes can also mean a better understanding and thus improvement.
- Aggregate level: compare survey estimates with administrative records or a trusted survey (i.e., benchmark)
- Disclosure: More disclosure of undesirable behaviors is regarded as more accurate
- When there is no gold standard, can use concurrent and predictive validity. Do answers to a question correlate with answers to other questions that are known to be theoretically related?
  Concurrent → two variables in the same wave
  Predictive → two variables in different waves